PRINCIPAL COORDINATES ANALYSIS

Principal Coordinates Analysis (PCO), also known as "Metric" or "Classical" Multidimensional Scaling (MDS), is analogous to Principal Components Analysis (PCA) in providing simple low-dimension summary graphs of multivariate data considered as points of objects in high-dimension "hyperspaces" with as many dimensions as variables. In both techniques, the object is to provide heurestic graphs for interpretation of the data rather than formal statistical tests. In PCA, data consisting of objects (rows) measured by multiple variables (columns) are converted into square matrices representing either variance-covariance (S) or correlation (R) between the variables. Eigenvectors and eigenvalues of either S or R provide directions and magnitudes respectively of maximum variance or correlation. PCO takes a similar approach using a defined distance matrix (D) between the objects. Distance matrix D is a square matrix with zeros (no distance) along the main diagnonal and mirror-image positive values on each side. Distance can be defined in terms of Euclidean distance, Manhattan, Distance, or many other ways. Many kinds of data naturally occur as matrices of similarity or difference, so can be analyzed by PCO in a similar way. Data here is from RA Johnson & DW Wichern *Applied Multivariate Statistical Analysis 4th Edition* 1998.

$ORIGIN \equiv 1$

The 1970 Census provided tract information on 5 socioeconomic variables for the Madison, Wisc. area. The data for 14 housing tracts (objects), from JW Table 8.5, p. 470.

Read in Data:

M := READPRN("c:\DATA\Multivariate\T8-5.DAT")

M =

	1	2	3	4	5
1	5.935	14.2	2.265	2.27	2.91
2	1.523	13.1	0.597	0.75	2.62
3	2.599	12.7	1.237	1.11	1.72
4	4.009	15.2	1.649	0.81	3.02
5	4.687	14.7	2.312	2.5	2.22
6	8.044	15.6	3.641	4.51	2.36
7	2.766	13.3	1.244	1.03	1.97
8	6.538	17	2.618	2.39	1.85
9	6.451	12.9	3.147	5.52	2.01
10	3.314	12.2	1.606	2.18	1.82
11	3.777	13	2.119	2.83	1.8
12	1.53	13.8	0.798	0.84	4.25
13	2.768	13.6	1.336	1.75	2.64
14	6.585	14.9	2.763	1.91	3.17

Sizing data array:

$X := M^T$

$n := cols(X) \qquad n = 14$

$p := rows(X) \qquad p = 5$

$i := 1 .. n \qquad j := 1 .. p$

$1_i := 1$

Mean vector (X_{bar}):

$$X_{bar} := \frac{1}{n}\cdot X\cdot 1 \qquad X_{bar} = \begin{pmatrix} 4.323 \\ 14.014 \\ 1.952 \\ 2.171 \\ 2.454 \end{pmatrix}$$

< means for each variable

Variance/covariance (S):

$I := identity(n)$

$$S := \frac{1}{n-1}\cdot X\cdot\left(I - \frac{1}{n}\cdot 1\cdot 1^T\right)\cdot X^T$$

$$S = \begin{pmatrix} 4.308 & 1.684 & 1.803 & 2.155 & -0.253 \\ 1.684 & 1.767 & 0.588 & 0.178 & 0.176 \\ 1.803 & 0.588 & 0.801 & 1.065 & -0.158 \\ 2.155 & 0.178 & 1.065 & 1.969 & -0.357 \\ -0.253 & 0.176 & -0.158 & -0.357 & 0.504 \end{pmatrix}$$

Calculating Squared Euclidean Distances between the objects in M:

On many occasions, data is collected in distance form - as for instance, base differences in alligned nucleotide sequences. As long as "distances" utilized obey the triangle inequality, PCO allows ordination of the data and plotting i a way that is analogous to PCA.

$i := 1 .. \text{cols}(X) \quad k := 1 .. \text{cols}(X)$

$$D_{i,k} := \left|\left(X^{\langle i\rangle} - X^{\langle k\rangle}\right)^{T} \cdot \left(X^{\langle i\rangle} - X^{\langle k\rangle}\right)\right|$$

$$D = \begin{pmatrix}
0 & 25.852 & 17.197 & 7.233 & 2.339 & 13.621 & 14.316 & 9.466 & 14.107 & 12.5 & 7.664 & 25.557 & 11.596 & 1.358 \\
25.852 & 0 & 2.667 & 11.861 & 18.735 & 72.245 & 2.505 & 47.727 & 53.953 & 7.721 & 12.406 & 3.195 & 3.347 & 35.204 \\
17.197 & 2.667 & 0 & 10.188 & 11.697 & 55.807 & 0.457 & 37.568 & 38.058 & 2.052 & 5.22 & 9.019 & 2.104 & 25.799 \\
7.233 & 11.861 & 10.188 & 0 & 4.645 & 34.535 & 6.47 & 14.44 & 36.702 & 12.802 & 10.684 & 10.343 & 5.226 & 9.199 \\
2.339 & 18.735 & 11.697 & 4.645 & 0 & 17.905 & 9.014 & 8.959 & 16.213 & 8.896 & 4.041 & 19.945 & 6.584 & 5.096 \\
13.621 & 72.245 & 55.807 & 34.535 & 17.905 & 0 & 51.155 & 10.029 & 11.214 & 43.795 & 30.42 & 70.796 & 44.845 & 10.806 \\
14.316 & 2.505 & 0.457 & 6.47 & 9.014 & 51.155 & 0 & 31.67 & 37.522 & 2.986 & 5.147 & 7.211 & 1.066 & 21.667 \\
9.466 & 47.727 & 37.568 & 14.44 & 8.959 & 10.029 & 31.67 & 0 & 26.92 & 34.503 & 24.068 & 46.795 & 28.45 & 6.406 \\
14.107 & 53.953 & 38.058 & 36.702 & 16.213 & 11.214 & 37.522 & 26.92 & 0 & 23.897 & 15.497 & 57.464 & 31.944 & 18.543 \\
12.5 & 7.721 & 2.052 & 12.802 & 8.896 & 43.795 & 2.986 & 34.503 & 23.897 & 0 & 1.54 & 14.096 & 3.188 & 21.223 \\
7.664 & 12.406 & 5.22 & 10.684 & 4.041 & 30.42 & 5.147 & 24.068 & 15.497 & 1.54 & 0 & 17.397 & 3.863 & 14.633 \\
25.557 & 3.195 & 9.019 & 10.343 & 19.945 & 70.796 & 7.211 & 46.795 & 57.464 & 14.096 & 17.397 & 0 & 5.282 & 32.936 \\
11.596 & 3.347 & 2.104 & 5.226 & 6.584 & 44.845 & 1.066 & 28.45 & 31.944 & 3.188 & 3.863 & 5.282 & 0 & 18.602 \\
1.358 & 35.204 & 25.799 & 9.199 & 5.096 & 10.806 & 21.667 & 6.406 & 18.543 & 21.223 & 14.633 & 32.936 & 18.602 & 0
\end{pmatrix}$$

Scaling distances to produce matrix Δ:

$A := -0.5D$

$$A_{bar} := \frac{1}{n} \cdot A \cdot 1$$

$$\Delta_{i,k} := A_{i,k} - A_{bar_i} - A_{bar_k} + \text{mean}(A)$$

^ centered & scaled distance matrix Δ

$$A_{bar} = \begin{pmatrix} -5.815 \\ -10.622 \\ -7.78 \\ -6.226 \\ -4.788 \\ -16.685 \\ -6.828 \\ -11.679 \\ -13.644 \\ -6.757 \\ -5.449 \\ -11.43 \\ -5.932 \\ -7.91 \end{pmatrix}$$

< Column means of distance matrix D scaled by -0.5

$$\Delta = \begin{pmatrix} 2.95 & -5.17 & -3.69 & -0.26 & 0.75 & 7.01 & -3.2 & 4.08 & 3.72 & -2.36 & -1.25 & -4.22 & -2.73 & 4.36 \\ -5.17 & 12.56 & 8.39 & 2.24 & -2.64 & -17.5 & 7.52 & -10.24 & -11.39 & 4.84 & 1.19 & 11.77 & 6.2 & -7.75 \\ -3.69 & 8.39 & 6.88 & 0.23 & -1.96 & -12.12 & 5.7 & -8.01 & -6.29 & 4.83 & 1.94 & 6.02 & 3.98 & -5.89 \\ -0.26 & 2.24 & 0.23 & 3.77 & 0.01 & -3.04 & 1.14 & 2 & -7.16 & -2.1 & -2.35 & 3.8 & 0.86 & 0.85 \\ 0.75 & -2.64 & -1.96 & 0.01 & 0.89 & 3.84 & -1.57 & 3.31 & 1.64 & -1.58 & -0.46 & -2.44 & -1.25 & 1.47 \\ 7.01 & -17.5 & -12.12 & -3.04 & 3.84 & 24.69 & -10.75 & 14.67 & 16.04 & -7.14 & -1.76 & -15.97 & -8.49 & 10.51 \\ -3.2 & 7.52 & 5.7 & 1.14 & -1.57 & -10.75 & 4.97 & -6.01 & -6.97 & 3.41 & 1.02 & 5.97 & 3.55 & -4.78 \\ 4.08 & -10.24 & -8.01 & 2 & 3.31 & 14.67 & -6.01 & 14.68 & 3.18 & -7.5 & -3.59 & -8.97 & -5.3 & 7.7 \\ 3.72 & -11.39 & -6.29 & -7.16 & 1.64 & 16.04 & -6.97 & 3.18 & 18.61 & -0.23 & 2.66 & -12.34 & -5.08 & 3.6 \\ -2.36 & 4.84 & 4.83 & -2.1 & -1.58 & -7.14 & 3.41 & -7.5 & -0.23 & 4.83 & 2.75 & 2.46 & 2.41 & -4.63 \\ -1.25 & 1.19 & 1.94 & -2.35 & -0.46 & -1.76 & 1.02 & -3.59 & 2.66 & 2.75 & 2.22 & -0.5 & 0.77 & -2.64 \\ -4.22 & 11.77 & 6.02 & 3.8 & -2.44 & -15.97 & 5.97 & -8.97 & -12.34 & 2.46 & -0.5 & 14.18 & 6.04 & -5.81 \\ -2.73 & 6.2 & 3.98 & 0.86 & -1.25 & -8.49 & 3.55 & -5.3 & -5.08 & 2.41 & 0.77 & 6.04 & 3.18 & -4.14 \\ 4.36 & -7.75 & -5.89 & 0.85 & 1.47 & 10.51 & -4.78 & 7.7 & 3.6 & -4.63 & -2.64 & -5.81 & -4.14 & 7.14 \end{pmatrix}$$

Eigenvalues & eigenvectors of matrix Δ:

$$\Lambda := \mathrm{reverse}(\mathrm{sort}(\mathrm{eigenvals}(\Delta)))$$

$$E^{\langle j\rangle} := \mathrm{eigenvec}(\Delta, \Lambda_j)$$

Remember:

Eigenvectors are automatically scaled to unit length!

$$\left|E^{\langle 1\rangle}\right| = 1$$

$$\Lambda = \begin{pmatrix} 90.104 \\ 23.207 \\ 5.065 \\ 2.984 \\ 0.184 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -0 \\ -0 \\ -0 \\ -0 \end{pmatrix}$$

$$E = \begin{pmatrix} -0.15145 & -0.06074 & -0.19572 & 0.43333 & -0.46903 \\ 0.37239 & -0.01715 & 0.01617 & -0.10156 & -0.38977 \\ 0.25286 & 0.13377 & 0.33077 & 0.22166 & 0.06355 \\ 0.06271 & -0.38318 & -0.02063 & 0.03658 & 0.11175 \\ -0.08077 & -0.05561 & 0.11427 & -0.21951 & 0.36973 \\ -0.52226 & 0.00931 & -0.09195 & -0.14727 & 0.09172 \\ 0.22457 & 0.01306 & 0.2746 & 0.12188 & -0.00529 \\ -0.31513 & -0.43536 & 0.4435 & -0.32588 & -0.2918 \\ -0.33415 & 0.59293 & -0.22959 & -0.19214 & -0.23294 \\ 0.14967 & 0.33215 & 0.14369 & 0.22348 & -0.0465 \\ 0.03844 & 0.28659 & 0.14195 & -0.1084 & 0.46906 \\ 0.34749 & -0.20274 & -0.64019 & -0.29896 & 0.08661 \\ 0.18303 & 0.01716 & -0.06128 & -0.2146 & -0.06271 \\ -0.22739 & -0.23019 & -0.22557 & 0.57137 & 0.30561 \end{pmatrix}$$

Scaling the Eigenvectors of Δ:

$$EE_{i,j} := E_{i,j}\cdot\sqrt{\Lambda_j} \qquad \left|EE^{\langle 1\rangle}\right| = 9.492$$

^ Each eigenvector is scaled by the square root of non-zero eigenvalues in Λ:

Scaled eigenvectors (matrix EE) represent the object coordinates in PCO space.
Each row represents an object. As in PCA, choose only the first few PCO coordinates to represent the data. Here, the first two coordinates would seem to be enough:

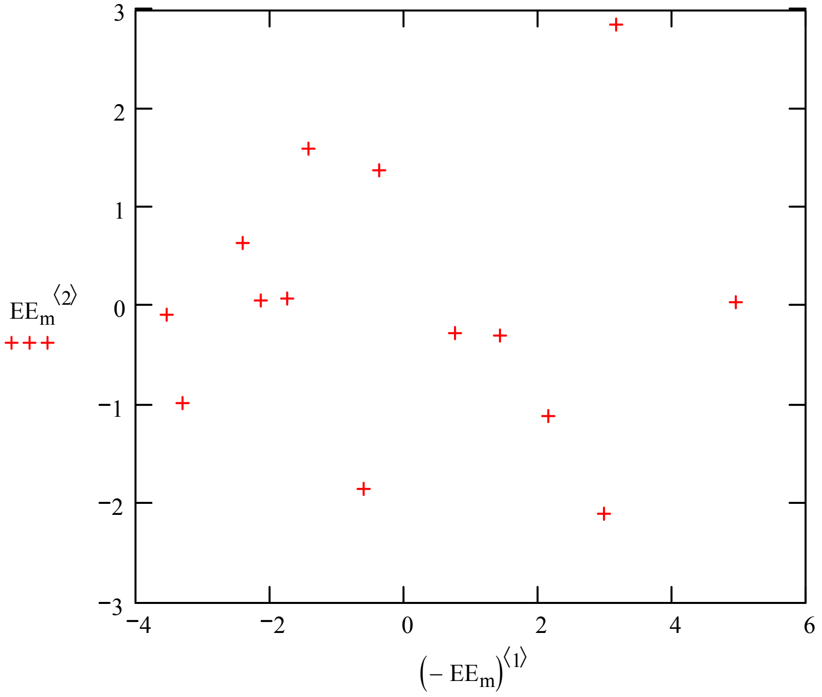
$$EE = \begin{pmatrix} -1.438 & -0.293 & -0.441 & 0.749 & -0.201 \\ 3.535 & -0.083 & 0.036 & -0.175 & -0.167 \\ 2.4 & 0.644 & 0.744 & 0.383 & 0.027 \\ 0.595 & -1.846 & -0.046 & 0.063 & 0.048 \\ -0.767 & -0.268 & 0.257 & -0.379 & 0.159 \\ -4.957 & 0.045 & -0.207 & -0.254 & 0.039 \\ 2.132 & 0.063 & 0.618 & 0.211 & -0.002 \\ -2.991 & -2.097 & 0.998 & -0.563 & -0.125 \\ -3.172 & 2.856 & -0.517 & -0.332 & -0.1 \\ 1.421 & 1.6 & 0.323 & 0.386 & -0.02 \\ 0.365 & 1.381 & 0.319 & -0.187 & 0.201 \\ 3.298 & -0.977 & -1.441 & -0.516 & 0.037 \\ 1.737 & 0.083 & -0.138 & -0.371 & -0.027 \\ -2.158 & -1.109 & -0.508 & 0.987 & 0.131 \end{pmatrix}$$

PCO Plot:

$m := 1 .. 2$

$EEm_{i,m} := EE_{i,m}$

Note that PCA/PCO directions & therefore handedness of the coordinate system for the plots may differ, but are unimportant.



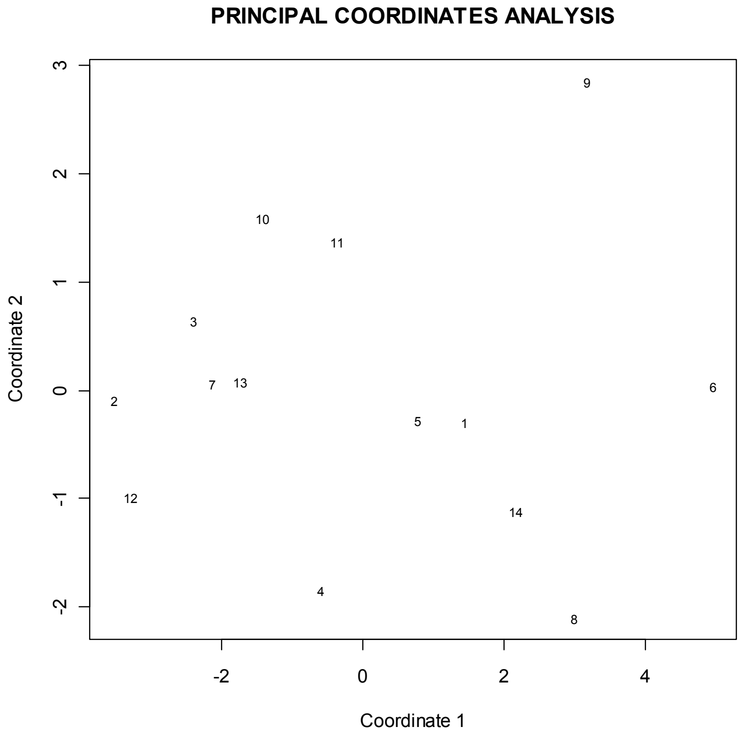
Prototype in R:

```
# PRINCIPAL COORDINATES ANALYSIS:
M=read.table("c:/MultivariateDATA/T8-5.DAT")
M
D=dist(M,method="euclidean")
D

# When importing a distanc matrix directly see:
?as.matrix

PCO=cmdscale(D,k=2,eig=TRUE,x.ret=TRUE)
PCO

# PLOT:
x <- PCO$points[,1]
y <- PCO$points[,2]
plot(x, y, xlab="Coordinate 1", ylab="Coordinate 2",
  main="PRINCIPAL COORDINATES ANALYSIS", type="n")
text(x, y, labels = row.names(M), cex=.7)
```


PRINCIPAL COORDINATES ANALYSIS
Coordinate 2
Coordinate 1