Multidimensional Scaling

statgraphics 18®
centurion

Revised: 10/11/2017

Summary

The *Multidimensional Scaling* procedure is designed to display multivariate data in a low-dimensional space. Given an *n* by *n* matrix of distances between each pair of *n* multivariate observations, the procedure searches for a low-dimensional representation of those observations that preserves the distances between them as well as possible. The primary output is a map of the points in that low-dimensional space (usually 2 or 3 dimensions).

Input to the procedure may be either:

1. An *n* by *n* matrix of distances or "dissimilarities".
2. An *n* by *p* matrix of observations for *p* variables, from which a distance matrix may be constructed.

The calculations are performed by R using the "cmdscale" and "isoMDS" functions. To run the procedure, R must be installed on your computer together with the *MASS* package. For information on downloading and installing R, refer to the document titled "R – Installation and Configuration".

Sample StatFolios: *mds1.sgp and mds2.sgp*

Sample Data

The fist sample data set *city distances.sgd* contains a 15 by 15 matrix with the flying distance between each pair of 15 U.S. cities. The matrix is shown below:

City	ATL	BOS	DCA	DEN	HOU	JFK	LAX	MIA	MKE	MSY	ORD	PHL	PHX	SEA	SFO
ATL	0	946	547	1198	689	760	1944	597	671	426	607	666	1585	2184	2184
BOS		0	399	1751	1597	187	2608	1261	860	1368	865	280	2297	2490	2700
DCA			0	1474	1207	213	2308	923	635	970	611	119	1976	2326	2439
DEN				0	863	1624	862	1710	893	1062	887	1555	602	1029	966
HOU					0	1417	1378	964	985	304	926	1324	1008	1881	1635
JFK						0	2472	1093	746	1183	739	94	2151	2417	2583
LAX							0	2341	1753	1669	1743	2399	370	967	338
MIA								0	1262	675	1200	1017	1971	2729	2584
MKE									0	905	67	690	1458	1691	1842
MSY										0	839	1089	1299	2091	1910
ORD											0	677	1439	1719	1844
PHL												0	2073	2375	2518
PNX													0	1118	651
SEA														0	691
SFO															0

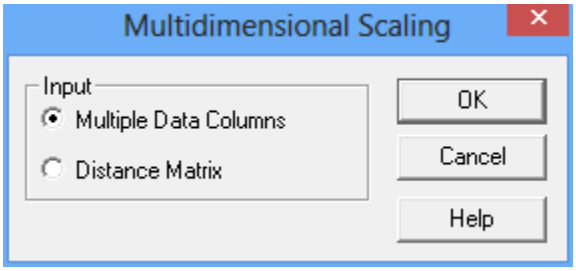
When entering a distance matrix, you may supply the entire matrix or just one of the diagonals (upper or lower).

The second sample data set *oecd2007.sgd* contains statistics for 22 countries, from which dissimilarities between each pair of countries will be calculated:

Country	Year	Population	Central government debt	Inflation	Long-term interest rates	Unemployment rate	Infant mortality	Female life expectancy	Male life expectancy
			% of GDP	% change in consumer prices	percent per annum		deaths per 1,000 live births	at birth	at birth
Australia	2007	21015040	5.181	2.332361	5.994521	4.4	4.2	83.7	79
Austria	2007	8300954	57.829	2.168556	4.2975	4.4	3.7	83.1	77.4
Belgium	2007	10625700	85.295	1.823149	4.333333	7.5	3.9	82.6	77.1
Canada	2007	32929730	25.183	2.138384	4.269834	6	5.1	83	78.3
Denmark	2007	5457415	27.765	1.714031	4.286608	3.8	4	80.6	76.2
France	2007	61965050	52.118	1.488074	4.304167	8	3.8	84.4	77.4
Germany	2007	82266370	39.55	2.26378	4.216667	8.7	3.9	82.7	77.4
Iceland	2007	311396	23.237	5.051564	9.419559	2.3	2	82.9	79.4
Ireland	2007	4339000	19.834	4.91622	4.328333	4.6	3.1	82.1	77.4
Italy	2007	59375290	95.627	1.829738	4.487258	5.9	3.5	84.2	78.7
Japan	2007	127771000	164.546	0.05795182	1.6655	3.9	2.6	86	79.2
Korea	2007	48456370	20.861	2.535017	5.350833	3.3	3.6	82.7	76.1
Mexico	2007	105790700	20.861	3.96685	7.595	3.7	15.7	77.4	72.6
Netherlands	2007	16381690	37.552	1.61418	4.288167	3.2	4.1	82.3	78
Norway	2007	4709156	11.681	0.7289971	4.774167	2.5	3.1	82.9	78.3
Portugal	2007	10608330	66.622	2.451603	4.424141	8	3.4	82.2	75.9
Spain	2007	44873570	30.019	2.786319	4.306417	8.3	3.5	84.3	77.8
Sweden	2007	9148093	36.406	2.212169	4.1675	6	2.5	83	78.9
Switzerland	2007	7551117	23.216	0.7323495	2.92675	3.4	3.9	84.4	79.5
Turkey	2007	70256000	39.551	8.756181	8.8	10.3	15.9	75.6	71.1
United Kingdom	2007	60124000	42.744	2.341667	5.011667	5.3	4.8	81.8	77.6
United States	2007	301393600	35.703	2.852673	4.629167	4.6	6.8	80.4	75.4

Data Input

When the *Multidimensional Scaling* procedure is selected from the Statgraphics menu, the first dialog box displayed requests the type of format in which the data are stored:

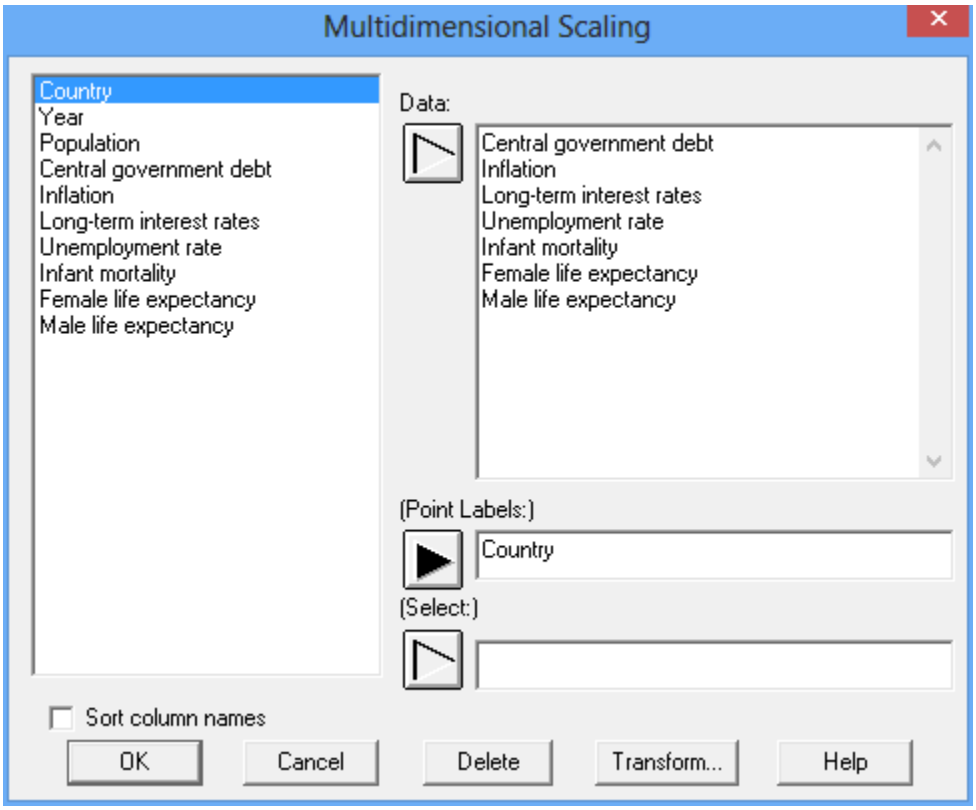


- **Multiple Data Columns:** each column contains the values for a separate data variable. Given *n* rows of data, an *n* by *n* distance matrix will be constructed.

- **Distance Matrix:** the data contain an *n* by *n* matrix of distances between *n* pairs of points.

Multiple Data Columns

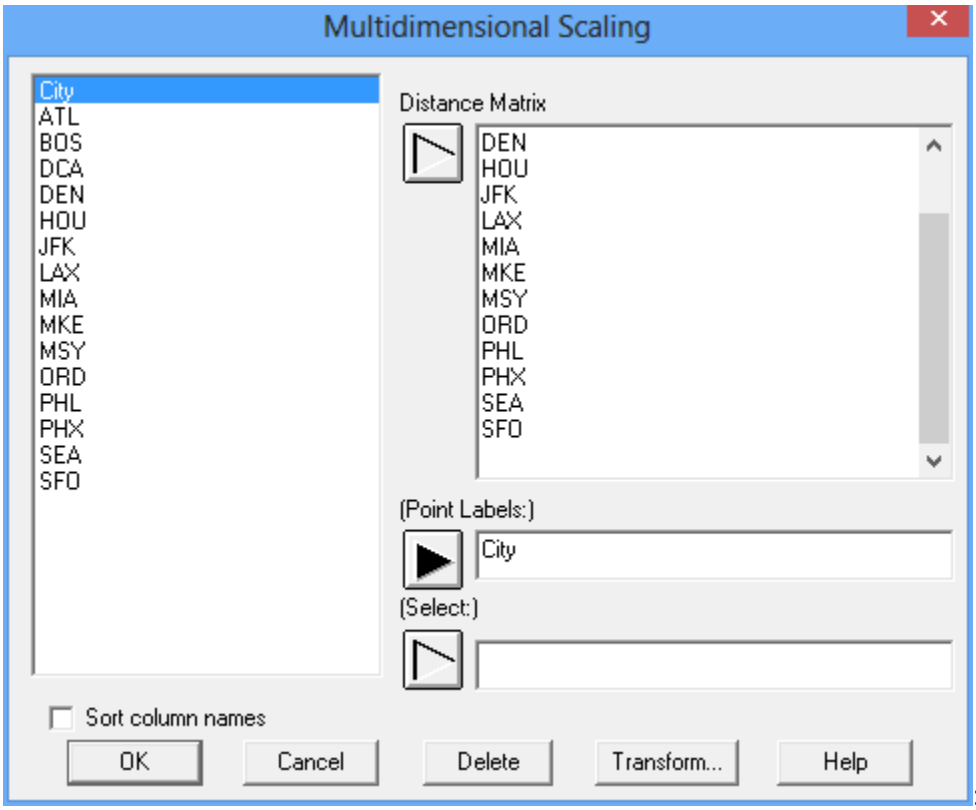
If the data consist of multiple data columns, the second dialog box requests the names of the columns containing the values from which the distance or dissimilarity matrix will be constructed:



- **Data:** numeric columns containing the *n* values to be used to calculate the distance matrix.
- **Point Labels:** *n* optional labels used to identify each point.
- **Select:** subset selection.

Distance Matrix

If the data consist of an already calculated distance matrix, the second dialog box requests the names of the columns containing that matrix



:

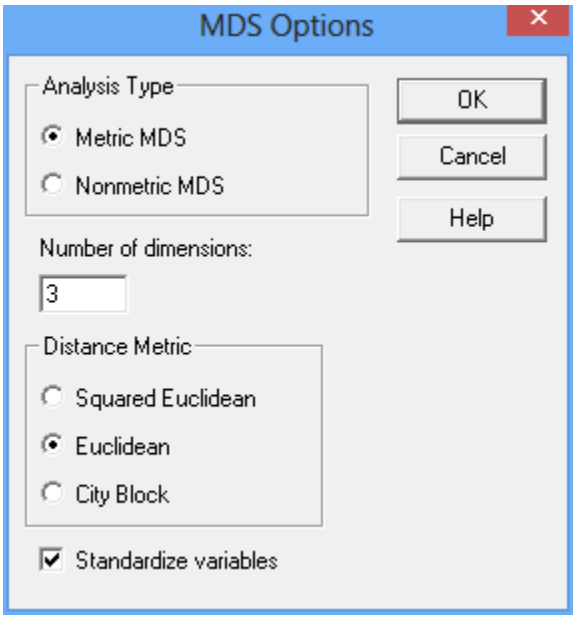
- **Distance Matrix:** numeric columns containing each column of the distance matrix. Since the matrix is symmetric, only the upper diagonal or lower diagonal part of the matrix is required.

- **Point Labels:** *n* optional labels used to identify each point. If not supplied, the column names will be used to create the labels.

- **Select:** subset selection.

Analysis Options

The *Analysis Options* dialog box sets various options for the procedure:



- **Analysis type**: specifies the type of multidimensional scaling to be performed. Nonmetric MDS may be used on data which are ordinal rather than continuous. *Metric MDS* uses the *cmdscale* function in R, while *Nonmetric MDS* uses the function *iosMDS*. See the links in the *References* section for more details.

- **Number of dimensions:** the number of dimensions for which coordinates are obtained.

- **Distance metric:** the metric used to measure distance between cases. 3 options are available for measuring the distance between observation *x* and observation *y*:

    1. Squared Euclidian distance: $d(x,y) = \sum_{i=1}^{p} (x_i - y_i)^2$ (1)

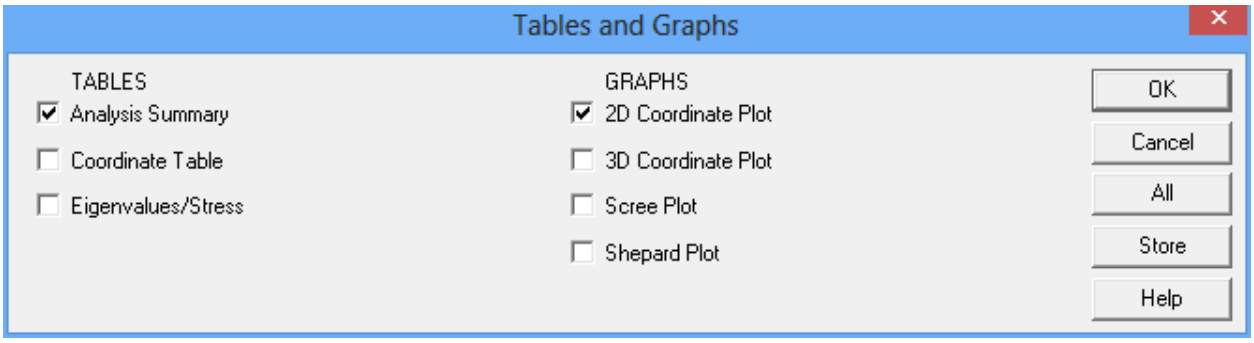
    2. Euclidian distance: $d(x,y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$ (2)

    3. City Block distance: $d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$ (3)

- **Standardize variables**: If checked, the variables will be standardized before calculating the distance matrix (*Multiple Data Columns* input format only). If selected, each variable is standardized by subtracting its sample mean and then dividing by its sample standard deviation.

Tables and Graphs

The following tables and graphs may be created:



Analysis Summary

The *Analysis Summary* begins with a list of the R commands that were executed.

Multidimensional Scaling

```
d<-
read.csv("C:\\\\Users\\\\Neil\\\\AppData\\\\Local\\\\Temp\\\\distancemat.csv"
,dec=".",sep=",")
setwd("c:\\temp")
fit <- cmdscale(as.matrix(d),eig=TRUE,k=3)
str(fit)

## List of 5
##  $ points: num [1:22, 1:3] -0.402 -0.704 -0.515 -0.429 0.103 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : NULL
##  $ eig   : num [1:22] 86.3 28.13 14.88 10.88 3.43 ...
##  $ x     : NULL
##  $ ac    : num 0
##  $ GOF   : num [1:2] 0.88 0.88

write(fit$points,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\coordinates.csv
",sep=",",ncolumns=3)
write(fit$eig,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\eigenvalues.csv",s
ep=",",ncolumns=1)
```

For metric scaling, the *cmdscale* function in R is used to perform the scaling. Nonmetric scaling uses the *isoMDS* function from the *MASS* package and produces similar output:

Multidimensional Scaling

```
d<-
read.csv("C:\\\\Users\\\\Neil\\\\AppData\\\\Local\\\\Temp\\\\distancemat.csv"
,dec=".",sep=",")
setwd("c:\\temp")
library("MASS")
fit <- isoMDS(as.matrix(d),k=3)

## initial  value 6.014079
## iter   5 value 3.010912
## iter  10 value 2.842646
## final  value 2.816370
## converged

str(fit)

## List of 2
##  $ points: num [1:22, 1:3] -0.582 -0.662 -0.598 -0.568 0.167 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : NULL
##  $ stress: num 2.82

write(fit$points,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\coordinates.csv
",sep=",",ncolumns=3)
write(fit$stress,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\eigenvalues.csv
",sep=",",ncolumns=1)
```

Coordinate Table

The MDS procedures produce a new set of coordinates for each observation or sample point. If d dimensions are requested, then we can represent the coefficients by

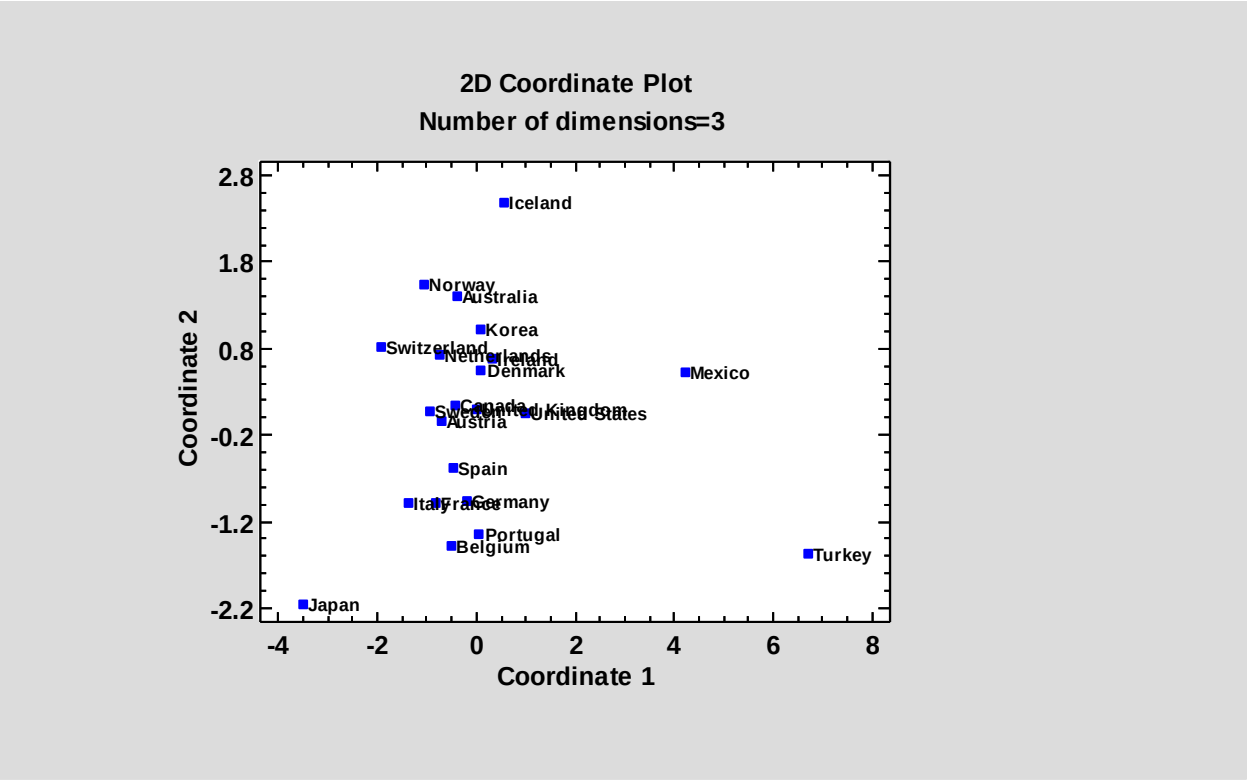
$c_{i,j}$ = location of sample i along dimension j, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, d$

The table below shows the coefficients calculated for the OECD data, where $n = 22$ and $d = 3$:

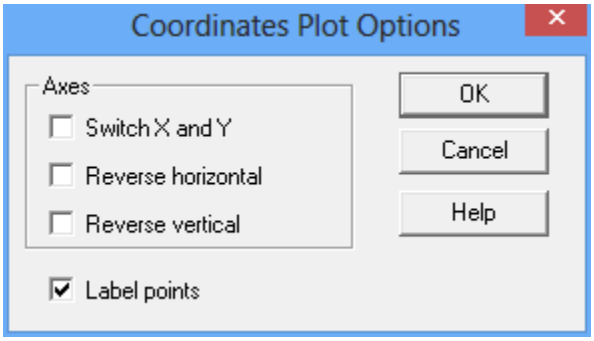
	Coordinate 1	Coordinate 2	Coordinate 3
Australia	-0.401865	1.40166	-0.644157
Austria	-0.703926	-0.0396514	0.355375
Belgium	-0.514776	-1.47011	-0.108095
Canada	-0.429374	0.140136	-0.390335
Denmark	0.103423	0.539862	0.798559
France	-0.827797	-0.984143	-0.756043
Germany	-0.187946	-0.957675	-1.02354
Iceland	0.575781	2.48838	-1.01818
Ireland	0.303604	0.674355	-0.553902
Italy	-1.35992	-0.986957	0.0022706
Japan	-3.50367	-2.153	1.60838
Korea	0.0824342	1.02136	0.323492
Mexico	4.2104	0.528406	1.9275
Netherlands	-0.729427	0.718513	0.663687
Norway	-1.04515	1.54019	0.515723
Portugal	0.0653589	-1.35462	-0.409345
Spain	-0.479637	-0.588843	-1.38193
Sweden	-0.931508	0.0795587	-0.629003
Switzerland	-1.91865	0.819206	0.3577
Turkey	6.68997	-1.57144	-0.477375
United Kingdom	-0.00382102	0.0999217	0.0453969
United States	1.00649	0.054891	0.79383

2D Coordinate Plot

This plot shows the location of each sample with respect to the 2 first dimensions:



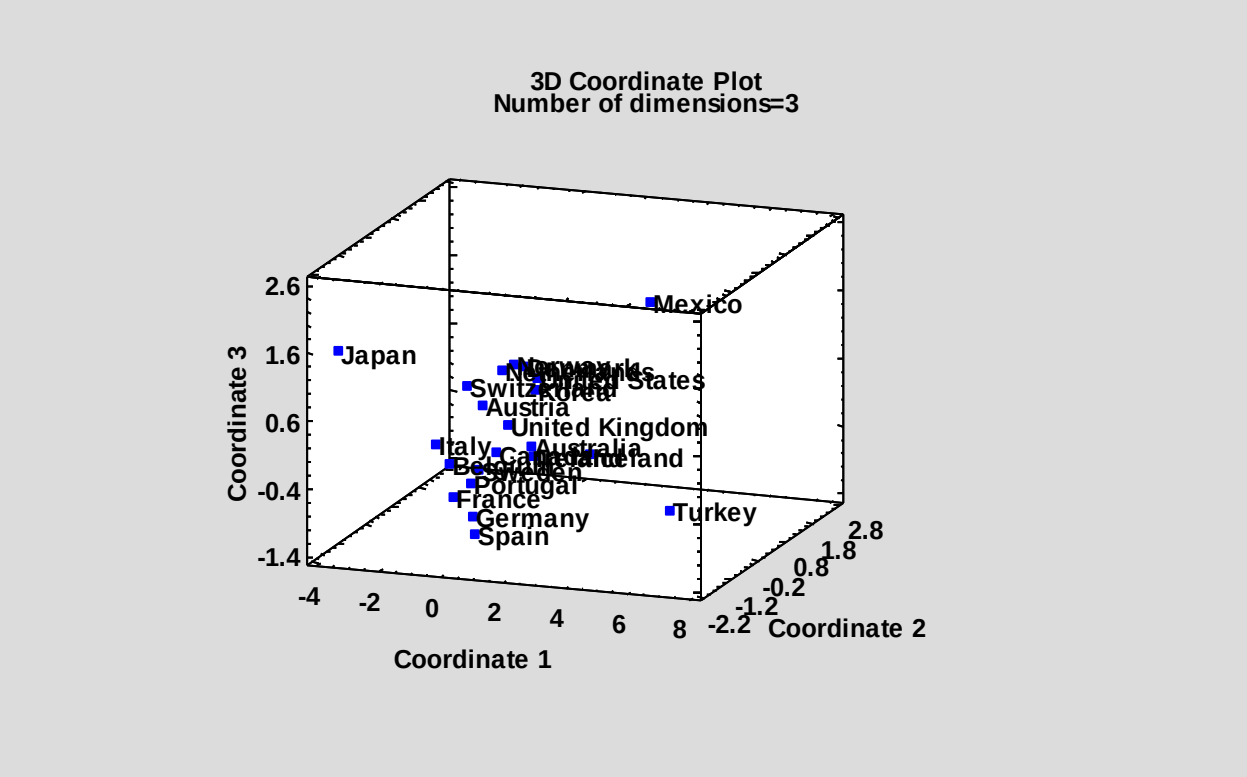
Pane Options



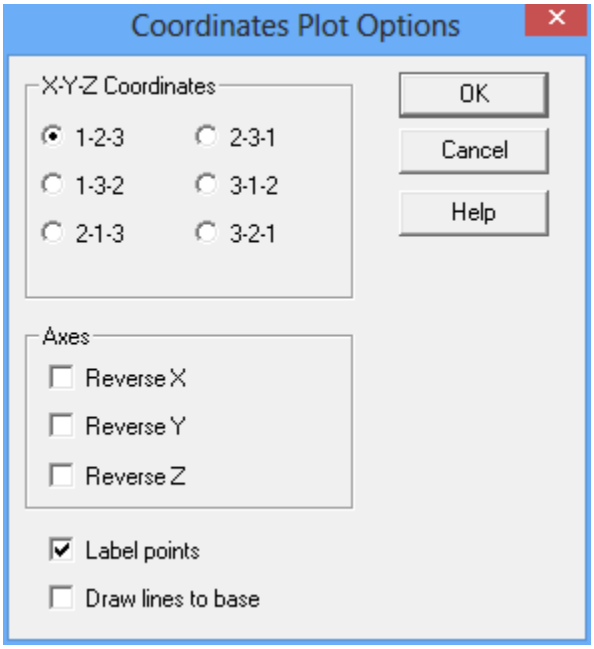
- **Switch X and Y:** if checked, corrdinate 2 will be plotted along the horizontal axis.
- **Reverse horizontal:** if checked, the location of points along the horizontal axis will be reversed by multiplying the coordinate on that axis by -1.
- **Reverse vertical:** if checked, the location of points along the vertical axis will be reversed by multiplying the coordinate on that axis by -1.
- **Label points:** if checked, each point will be labeled.

3D Coordinate Plot

This plot shows the location of each sample with respect to the 3 first dimensions:



Pane Options



- **X-Y-Z coordinates:** selects the coordinates to be plotted on the X, Y and Z axes, respectively.
- **Reverse X:** if checked, the location of points along the X axis will be reversed by multiplying the coordinate on that axis by -1.
- **Reverse Y:** if checked, the location of points along the Y axis will be reversed by multiplying the coordinate on that axis by -1.
- **Reverse Z:** if checked, the location of points along the Z axis will be reversed by multiplying the coordinate on that axis by -1.
- **Label points:** if checked, each point will be labeled.
- **Draw lines to base:** if checked, vertical lines will be drawn from each point to the bottom of the graph.

Eigenvalues/Stress

The metric MDS procedure generates eigenvalues for $d = 1$ through $n - 1$. The smaller the eigenvalue, the better the representation of the original distance matrix. The eigenvalues for the OECD data are shown below:

Eigenvalues/Stress

	Eigenvalue
1	86.297
2	28.1329
3	14.8799
4	10.8785
5	3.42961
6	2.28269
7	1.09927
8	0.0000842907
9	0.0000357323
10	0.000027731
11	0.0000246961
12	0.000010291
13	0.00000543245
14	0.00000120979
15	1.86234E-15
16	-0.00000271756
17	-0.00000950384
18	-0.0000149066
19	-0.000016884
20	-0.0000220292
21	-0.0000363245

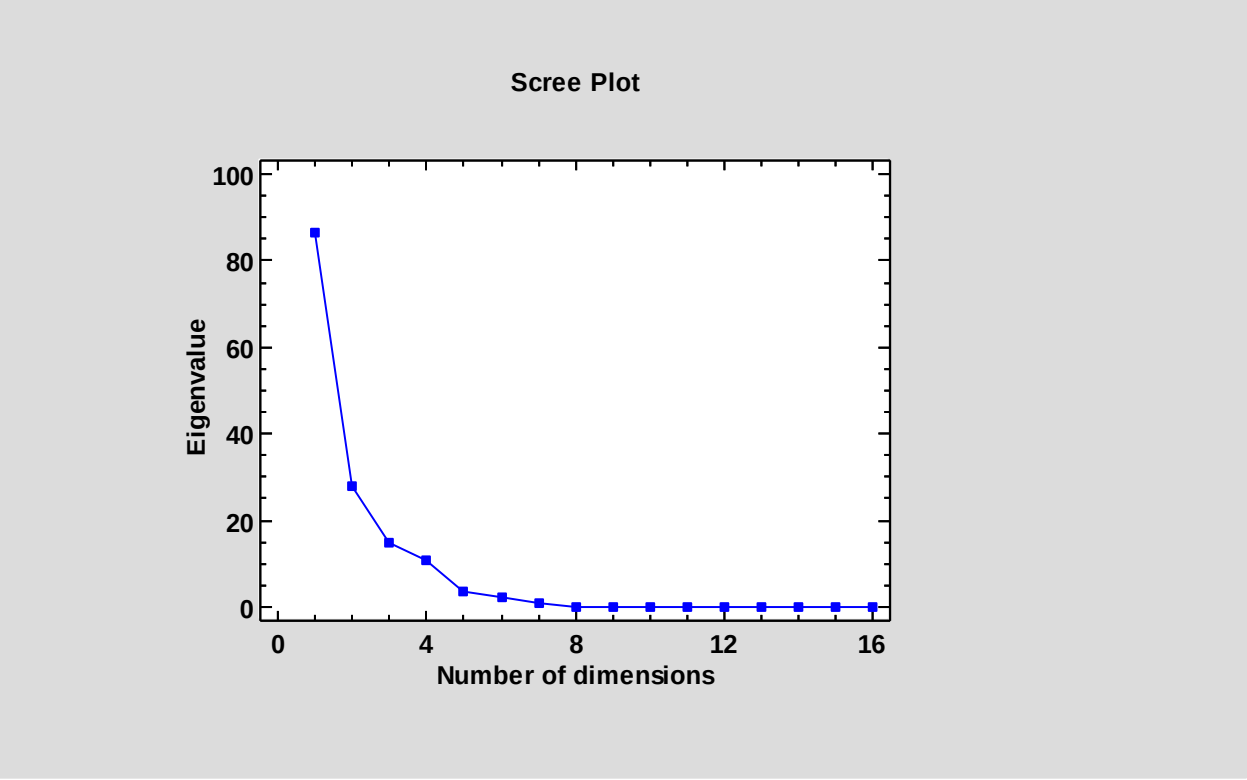
The nonmetric MDS procedure generates a quantity called “stress”. As with the eigenvalues, smaller stress values correspond to better representations of the original distance matrix. The table below displays eigenvalues for dimensions d = 1 through the largest dimension that generated a positive value of stress:

Eigenvalues/Stress

	Stress
1	14.0212
2	6.26252
3	2.83315
4	1.57293
5	0.61866
6	0.264071
7	0.13029

Scree Plot

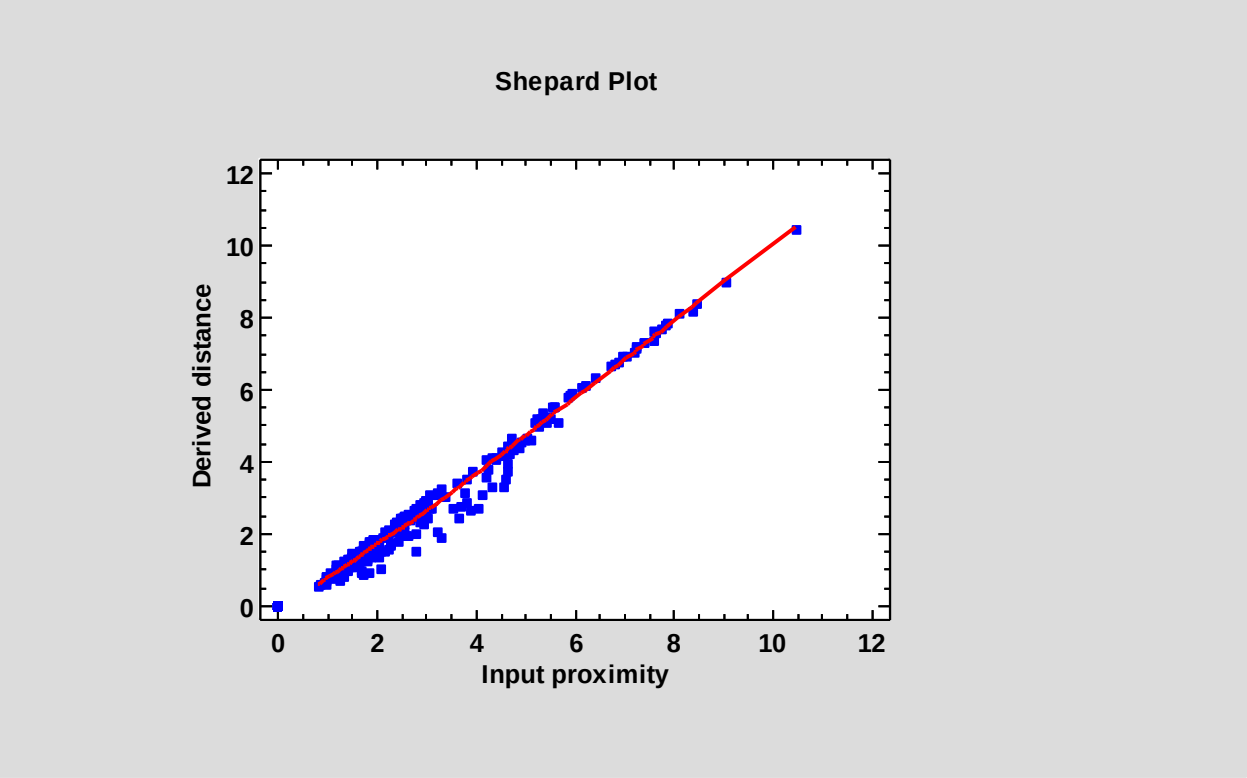
The scree plot displays the eigenvalues or stress values calculated by the MDS procedures:



To determine the number of dimensions necessary to represent adequately a set of data, it is common to look for a “knee” in the plot. The knee is the number of dimensions at which the plot begins to level out. Although no well-defined knee is present in the plot above, there is a noticeable change in the slope after *d* = 3.

Shepard Plot

The *Shepard Plot* is used to help judge the goodness of fit after the MDS procedure is run:

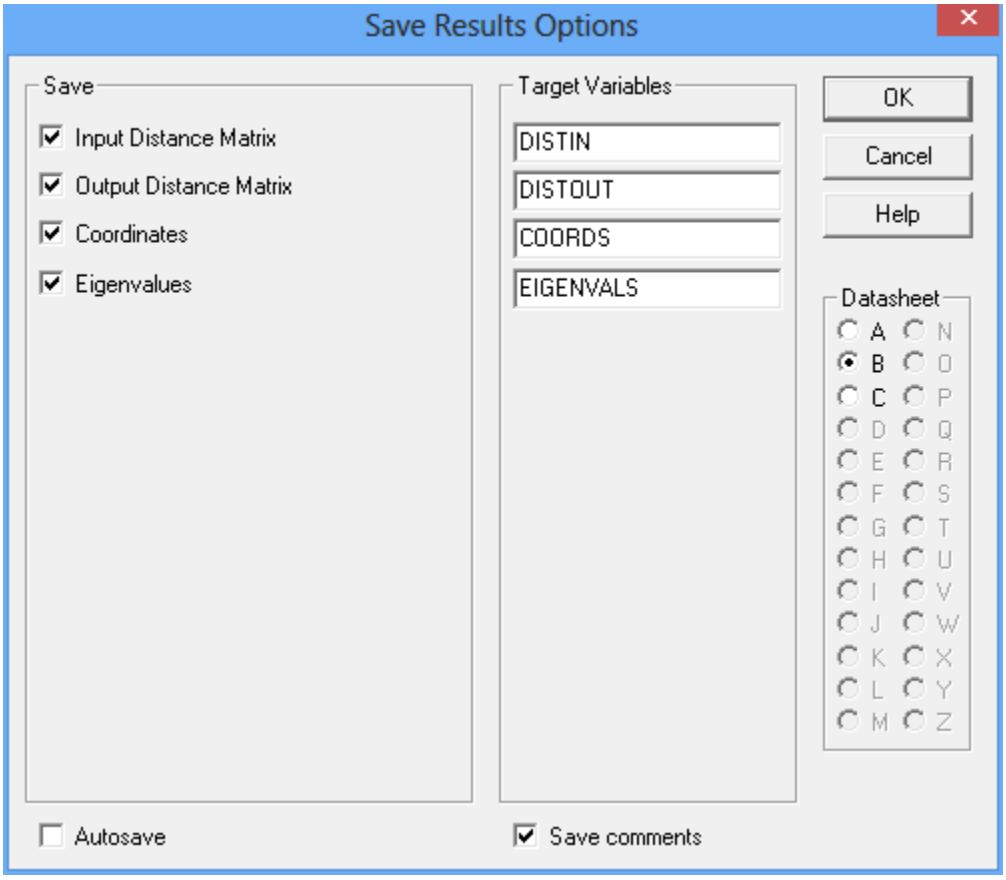


There are $n(n\text{-}1)$ points on the plot, one for each pair of samples. The horizontal axis shows the input distance or "proximity" of the pair, while the vertical axis shows the distance between the samples based on the derived coordinates. If MDS fits the data well, the points should lie close to a diagonal line.

To help judge the goodness-of-fit, a LOWESS smooth has been added to the plot using the *Smooth* button on the analysis toolbar.

Save Results

Various results can be saved to a datasheet by pressing the *Save Results* button on the analysis toolbar. The following dialog box will be presented:



Select:

- **Save:** select the items to be saved.
- **Target Variables:** enter names for the columns to be created.
- **Datasheet:** the datasheet into which the frequencies will be saved.
- **Autosave:** if checked, the results will be saved automatically each time a saved StatFolio is loaded.
- **Save comments:** if checked, comments for each column will be saved in the second line of the datasheet header.

References

R function “cmdscale”
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cmdscale.html

R Package “MASS” (2016)
https://cran.r-project.org/web/packages/MASS/MASS.pdf

Forrest M. Young, “Multidimensional Scaling”.
http://forrest.psych.unc.edu/teaching/p208a/mds/mds.html

Brian S. Everitt and Torsten Hothorn “A Handbook of Statistical Analyses using R”
https://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_multidimensional_scaling.pdf