Metric and non-metric scaling

Multi-dimensional scaling (MDS) provides various alternatives to dendrograms for visualizing distances between cases, so facilitating the recognition of potential groupings in a space of lower dimension than the number of variables. Given a n by m data set, the idea is to generate a set of n points in a Euclidean sub-space of dimension $1 < k << n - 1$ that have a distance matrix as close as possible to the distance matrix for the original data, so that distances can be visualized in the subspace for say $k = 2$, or $k = 3$. There are two cases.

1. **Classical metric scaling**
   This technique is used when the original data are in the form of observed quantities measured in terms of coordinates where distance is meaningful.

2. **Non-metric (ordinal) scaling**
   This technique is resorted to when the original data are of categorical or similar type that have been observed on a scale where only ranking is important and not actual differences.

SIMFIT can perform classical metric and/or non-metric (ordinal) scaling using a distance matrix calculated interactively or by supplying a pre-calculated distance matrix.

From the main SIMFIT menu choose [Statistics], [Multivariate], then [Scaling] using a distance matrix, read in the test file `g03faf.tf1`, and analyze using both metric and non-metric techniques to obtain the results as follows.

Eigenvalues from MDS (divided by the trace of the E matrix)

0.787130
0.280850
0.159630
0.077476
0.031624
0.020654
0.000000
-0.012186
-0.013685
-0.030479
-0.045469
-0.056206
-0.079207
-0.117400

[Sum 1 to 2]/[sum 1 to 13] = 0.9558 (95.58%) (actual values)
[Sum 1 to 2]/[sum 1 to 13] = 0.6709 (67.09%) (absolute values)
STRESS = 0.12557 (start = Metric 0%)
S-STRESS = 0.14962 (start = Metric 0%)

This table first lists the eigenvalues from classical metric scaling, where each eigenvalue has been normalized by dividing by the sum of all the eigenvalues, then the $STRESS$ and $SSTRESS$ values are listed.

Note that the type of starting estimates used, together with the percentages of the metric values used in any random starts, are output by SIMFIT and it will be seen that, with this distance matrix, there are small but negative eigenvalues, and hence the proportion of the distances captured by the lower dimensional subspace

would be inflated, and in addition two-dimensional plotting could be misleading. However it is usual to consider such small negative eigenvalues as being effectively zero, so that metric scaling in two dimensions is probably justified in this case as most of the proportion is in the first two eigenvalues.

The indication (actual values) is for the case where the sum of eigenvalues is used in the denominator when calculating the proportion P, i.e.

$$P = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{n-1} \lambda_i,$$

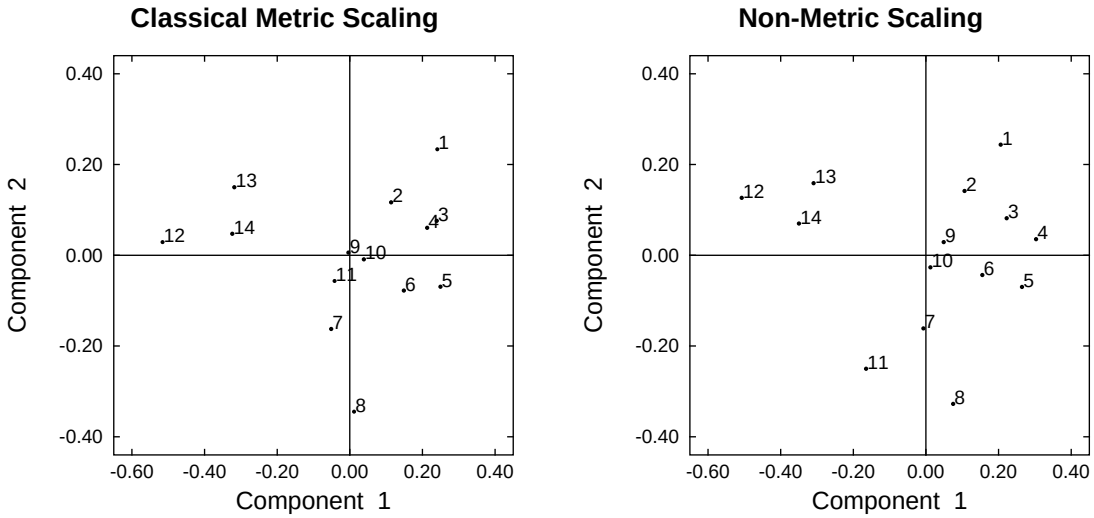
while the indication (absolute values) is for the case where the sum of the absolute values is used as the denominator, i.e.

$$P = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{n-1} |\lambda_i|,$$

as discussed later.

In an ideal case all the eigenvalues would be positive and these two values would be the same. For this reason a warning is issued when negative eigenvalues are encountered to alert users that caution is required when regarding the subspace plot as a valid representation of the distance between cases.

The next figures confirm the validity of using metric scaling with these data by showing considerable agreement between the two dimensional plots from metric scaling, and also non-metric scaling involving the $STRESS$ calculation. Note that the default labels in such plots may be integers corresponding to the case numbers, and not case labels, but such plot labels can be edited interactively, or overwritten from a labels file if required.



Format for data input

The data required for scaling must either be in the form of a multivariate matrix from which a distance matrix is calculated interactively, then either used directly or saved to a file for retrospective analysis. Of course a distance matrix $D = d_{ij}$ from a n by m data matrix is a symmetric n by n matrix but, because the diagonals are zero and generally $d_{ij} = d_{ji}$, then only $n(n-1)/2$ differences need to be available. For that reason, distance matrices are stored and analyzed by SIMFIT in strict lower triangular format as now described.

For instance, the data contained in test file `cluster.tf1` is the following 12 by 8 matrix

1.0	4.0	2.0	11.0	6.0	4.0	3.0	9.0
8.0	5.0	1.0	14.0	19.0	7.0	13.0	21.0
3.0	1.0	3.0	1.0	3.0	6.0	23.0	37.0
9.0	0.0	7.0	7.0	1.0	2.0	21.0	2.0
7.0	12.0	9.0	5.0	14.0	9.0	12.0	14.0
2.0	13.0	15.0	2.0	23.0	6.0	34.0	8.0
11.0	7.0	2.0	1.0	4.0	17.0	11.0	4.0
6.0	3.0	7.0	12.0	11.0	8.0	8.0	0.0
8.0	21.0	1.0	10.0	31.0	9.0	3.0	18.0
19.0	14.0	12.0	9.0	16.0	10.0	0.0	27.0
17.0	18.0	10.0	6.0	19.0	14.0	1.0	24.0
15.0	21.0	8.0	7.0	17.0	12.0	4.0	22.0

leading to the strict lower triangle of the 12 by 12 distance matrix below.

22.0										
36.2	28.8									
22.9	29.7	36.6								
1.95	16.6	31.1	24.5							
39.8	32.7	40.6	31.8	26.1						
21.7	28.3	38.2	21.3	19.3	36.2					
14.1	24.1	42.6	18.8	18.9	34.2	18.5				
32.7	23.0	45.4	44.9	23.6	38.7	36.6	33.4			
31.6	23.9	37.2	41.0	22.2	43.9	33.5	33.9	24.7		
32.2	24.4	39.1	41.8	20.2	41.4	31.3	33.4	19.9	8.25	
29.9	22.7	37.7	39.0	17.2	38.4	29.2	31.4	18.1	11.4	6.24

However this lower triangle would be stored packed by rows as follows

22.0
36.2
28.8
22.9
29.7
36.6
...
6.24

and distance matrices supplied for analysis by SIMFIT must be formatted in this way.

Theory for metric scaling

For instance, once a distance matrix $D = (d_{ij})$ has been calculated for n cases with m variables, as described for dendrograms, it may be possible to calculate principal coordinates. This involves constructing a matrix E defined by

$$e_{ij} = -\tfrac{1}{2}(d^2_{ij} - d^2_{i.} - d^2_{.j} + d^2_{..}),$$

where $d^2_{i.}$ is the average of d^2_{ij} over the suffix j, etc., in the usual way. The idea is to choose an integer k, where $1 < k << n - 1$, so that the data can be represented approximately in a space of dimension less than the number of cases, but in such a way that the distance between the points in that space correspond to the distances represented by the d_{ij} of the distance matrix as far as possible. If E is positive semi-definite, then the ordered eigenvalues $\lambda_i > 0$ of E will be nonnegative and the proportionality expression

$$P = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{n-1} \lambda_i$$

will show how well the cases of dimension n are represented in this subspace of dimension k. The most useful case is when $k = 2$, or $k = 3$, and the d_{ij} satisfy

$$d_{ij} \leq d_{ik} + d_{jk},$$

so that a two or three dimensional plot will display distances corresponding to the d_{ij}.

If this analysis is carried out but some relatively large negative eigenvalues result, then the proportion P may not adequately represent the success in capturing the values in distance matrix in a subspace of lower dimension that can be plotted meaningfully.

It should be pointed out that the principal coordinates will actually be the same as the principal components scores when the distance matrix is based on Euclidean norms. Further, where metrical scaling succeeds, the distances between points plotted in say two or three dimensions will obey the triangle inequality and so correspond reasonably closely to the distances in the dissimilarity matrix, but if it fails it could be useful to proceed to non-metrical scaling, which is discussed next.

Theory for non-metric (ordinal) scaling

Often a distance matrix is calculated where some or all of the variables are ordinal, so that only the relative order is important, not the actual distance measure. Non-metric (i.e. ordinal) scaling is similar to the metric scaling previously discussed, except that the representation in a space of dimension $1 < k << n - 1$ is sought in such a way as to attempt to preserve the relative orders, but not the actual distances. The closeness of a fitted distance matrix to the observed distance matrix can be estimated as either $STRESS$, or $SSTRESS$, given by

$$STRESS = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}(\hat{d}_{ij} - \tilde{d}_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\hat{d}_{ij}^2}}$$

$$SSTRESS = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}(\hat{d}_{ij}^2 - \tilde{d}_{ij}^2)^2}{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\hat{d}_{ij}^4}},$$

where $\hat{d}_{ij}$ is the Euclidean squared distance between points i and j, and $\tilde{d}_{ij}$ is the fitted distance when the $\hat{d}_{ij}$ are monotonically regressed on the d_{ij}. This means that $\tilde{d}_{ij}$ is monotonic relative to d_{ij} and is obtained from $\hat{d}_{ij}$ with the smallest number of changes.

It should be noted that this is a nonlinear optimization problem which may depend critically on starting estimates, and so can only be relied upon to locate a local, not a global solution. For this reason, starting estimates can be obtained in SimFIT by a preliminary metric scaling, or alternatively the values from such a scaling can be randomly perturbed before the optimization, in order to explore possible alternative solution points.

As mentioned previously, SimFIT can save distance matrices to files, so that dendrogram creation, classical metric, and non-metric scaling can be carried out retrospectively, without the need to generate distance matrices repeatedly from multivariate data matrices. Such distance matrices will be stored as vectors, corresponding to the strict lower triangle of the distance matrix packed by rows, (i.e. the strict upper triangle packed by columns).