Multidimensional scaling

In this section, we introduce mapping methods on visual and geometric presentation of multivariate data, with the focus of maintaining the concepts of similarity and ranking among data points. These methods can be viewed as a type of unsupervised learning. Traditionally they belong to the ordination methods.

1 The concepts of Multidimensional Scaling

Associations among subjects or interdependence between variables are often quantified by pairwise relations, such as distance, similarity, and correlations. **Multidimensional scaling** (MDS) is a set of visualization techniques using pairwise information to map high dimensional data to a lower, usually two dimensional space, such that more "similar" objects are closer in the low dimensional configuration. The geometric representation provides insight into the relationships among objects of interest. The MDS process can be described as

Pairwise similarity or dissimilarity in high dimensions $\Longrightarrow$ A configuration, a map, in low dimensions

MDS Data

The data are $n(n-1)/2$ pairs of pairwise dissimilarities (or distances) of n items.

Presumably the n items were from a space of high dimension $p \geq 3$.

The $n(n-1)/2$ dissimilarity measures, denoted as d_{ji}, are often displayed as the lower triangular part of an $n \times n$ matrix, in which the (i,j)th entry is the dissimilarity or distance between item i and item j.

$$\begin{bmatrix} 0 & & & & & \\ d_{21} & 0 & & & & \\ \vdots & \vdots & & & & \\ d_{i1} & d_{i2} & \cdots & 0 & & \\ \vdots & \vdots & \vdots & \vdots & & \\ d_{n1} & d_{n2} & \cdots & d_{nj} & \cdots & 0 \end{bmatrix}$$

Quite often, the pairwise meausres are given in the following misleading format:

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 \\ d_{21} & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{i1} & d_{i2} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nj} & \cdots & 0 \end{bmatrix}$$

If we assume the dissimilarity measure is symmetric with $d_{ij} = d_{ji}$, as dissimilarity measures most often are, it is suitable to convert the given $n(n-1)$ values to a symmetric $n \times n$ **dissimilarity matrix**, which has diagonal elements $= 0$.

$$\begin{bmatrix} 0 & d_{12} & \cdots & d_{1j} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2j} & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{i1} & d_{i2} & \cdots & d_{ij} & \cdots & d_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nj} & \cdots & 0 \end{bmatrix}$$

It is not uncommon that the data are $n(n-1)/2$ measurements of pairwise similarities of n items, instead of dissimilarities. A **similarity matrix** should have the largest value on the diagonal, often $= 1$, indicating that an item is the most similar to itself.

$$\begin{bmatrix} 1 & s_{12} & \cdots & s_{1j} & \cdots & s_{1n} \\ s_{21} & 1 & \cdots & s_{2j} & \cdots & s_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{i1} & s_{i2} & \cdots & s_{ij} & \cdots & s_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nj} & \cdots & 1 \end{bmatrix}$$

A dissimilarity measure can be created based on the given similarity measure, and vice versa.

MDS objective

A $q < p$ dimensional map of data maintaining as much as possible the original dissimilarity relations.
Usually $q \leq 3$, preferably $q \leq 2$.

Remarks on MDS

- The data format used by MDS is different from the datasets used by many other multivariate methods, which typically apply to a dataset with n observations of p-variate vectors.

  MDS uses so called proximity data, which consists of similarity (or dissimilarity) information for pairs of objects. If the dataset is of n objects (n observations) of p-variate vectors, a similarity or dissimilarity matrix of dimensions $n \times n$ needs to be created. This matrix is the input data of MDS.

- Similar to PCA and FA, MDS achieves dimension reduction. However PCA and FA are linear methods with respect to the original data, while MDS is a non-linear procedure. The low dimensional configuration achieve by MDS aims to maintain similarity, at the loss of linearity if necessary.

Illustrative examples

- Example 1

  Four observations A, B, C, D are in a high dimensional space $\mathbb{R}^p, p > 1$. Pairwise similarities of the points form a similarity matrix:

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 1 & & & \\ 0.980 & 1 & & \\ 0.995 & 0.955 & 1 & \\ 0.875 & 0.955 & 0.82 & 1 \end{bmatrix} \end{array}$$

  The similarity matrix gives an ordering of the pairwise similarities:

$$s_{AC} \geq s_{AB} \geq s_{BC} \geq s_{BD} \geq s_{AD} \geq s_{CD} \tag{1}$$

  There are many ways to convert a similarity measure s_{ik} between objects i and k to a distance measure d_{ik}, we may assign d to be $1/s$, $1/(s+1)$, $\sqrt{1-s^2}$, etc. Here we choose (actually with theoretical motivation)

$$d = 10\sqrt{2(1-s)}$$

  Then the (rounded) distance or dissimilarity matrix of the four points is

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 2 & 0 & & \\ 1 & 3 & 0 & \\ 5 & 3 & 6 & 0 \end{bmatrix} \end{array}$$

From the distance matrix we have

$$d_{AC} \leq d_{AB} \leq d_{BC} \leq d_{BD} \leq d_{AD} \leq d_{CD} \tag{2}$$

which, by construction, matches the ordering of the original similarity measures in the opposite direction perfectly.

A question in MDS is: Can we assign coordinates $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ to the points in a low dimensional space such that the corresponding distances maintain the ordering in (2)? For this example, the answer is yes. In fact, we can even find a $q = 1$ dimensional representation by assigning one-dimensional coordinates as

$$\hat{A} = 1, \quad \hat{B} = 3, \quad \hat{C} = 0, \quad \hat{D} = 6.$$

The distances of the one-dimensional representation of the points satisfy

$$d_{\hat{A}\hat{C}} \leq d_{\hat{A}\hat{B}} \leq d_{\hat{B}\hat{C}} \leq d_{\hat{B}\hat{D}} \leq d_{\hat{A}\hat{D}} \leq d_{\hat{C}\hat{D}}$$

thus maintaining the distance ordering of the points in the original high dimensional space.

- Example 2

Three observations A, B, C are in a high dimensional space $\mathbb{R}^p, p > 2$, with dissimilarity matrix

$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 0 & & \\ 4 & 0 & \\ 5 & 3 & 0 \end{bmatrix} \end{array}$$

From the distance matrix we have

$$d_{BC} \leq d_{AB} \leq d_{AC}$$

Can we assign coordinates $\hat{A}, \hat{B}, \hat{C}$ in a low dimensional space such that the corresponding distances maintain the ordering above? Again the answer is yes. In fact, we can even find a $q = 1$ dimensional representation by assigning one-dimensional coordinates as

$$\hat{A} = 1, \quad 3.5 < \hat{B} < 6, \quad \hat{C} = 6.$$

The distances of the one-dimensional representation of the points satisfy $d_{\hat{B}\hat{C}} \leq d_{\hat{A}\hat{B}} \leq d_{\hat{A}\hat{C}}$. thus maintaining the distance ordering of the points in the original high dimensional space.

However the pairwise similarity values are not perfectly maintained. A "loss" of information has occurred in the reduced dimension representation.

A $q = 2$ dimensional representation of

$$\hat{A} = (1, 1), \quad \hat{B} = (1, 4), \quad \hat{C} = (4, 6).$$

not only satisfying $d_{\hat{B}\hat{C}} \leq d_{\hat{A}\hat{B}} \leq d_{\hat{A}\hat{C}}$, but also maintaining the original distance perfectly, an ideal case of lower dimension representation.

- Example 3

Four observations A, B, C, D are in a high dimensional space $\mathbb{R}^p, p > 3$, with dissimilarity matrix

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

A tetrahedron in dimension $q = 3$ gives a geometric representation with perfect match of the dissimilarity. No geometric representation with perfect match of the original dissimilarity exists in lower dimension spaces $\mathbb{R}^2$ and $\mathbb{R}$.

Types of multidimensional scaling

Generally, N items form $N(N-1)/2$ pairwise measures of dissimilarities or distance. Multidimensional Scaling aims to find a map, a graphical representation of the items in lower dimensions such that the inter-item proximities match the original similarities as closely as possible.

If only the rank orders of the $N(N-1)/2$ original similarities are used, the process is called **non-numeric** scaling. If the actual magnitudes of the original similarities are used, the process is called **metric** scaling. The original metric MDS with Euclidean distance is **classical** scaling, also known as **principal coordinates** or principal coordinates analysis.

Multidimensional scaling method

Denote the similarity between items i and j by s_{ij}. Assume there are no ties in similarities. Arrange the similarities in a strictly ascending order as

$$s_{i_1k_1} < s_{i_2k_2} < \cdots < s_{i_Mk_M}, \quad M = N(N-1)/2 \tag{3}$$

If instead the measure is in terms of dissimilarity or distance d_{ik}, then strictly decreasing order is used. Multidimensional scaling attempts to find a lower q-dimensional configuration of the N items such that the distance $d_{ik}^{(q)}$ between the pairs of items match the order in (3). A perfect match occurs if

$$d_{i_1k_1}^{(q)} > d_{i_2k_2}^{(q)} > \cdots > d_{i_Mk_M}^{(q)} \tag{4}$$

Perfect match can be achieved if $q =$ the dimension of the original data space, using the same proximity metric. However the interest is on finding (4) or its approximation in $q = 2$ or 3 dimensional space.

2 Measure of goodness of fit in MDS

For a given dimension q, there may not exist a configuration of the points whose pairwise distances match the ordering of the original pairwise dissimilarities. **Stress** (Kruskal, 1964) is a measure of closeness between the original similarities and the fitted values of similarities.

$$Stress(q) = \left\{ \frac{\sum\sum_{i<k} \left(d_{ik}^{(q)} - \hat{d}_{ik}^{(q)}\right)^2}{\sum\sum_{i<k} \left(d_{ik}^{(q)}\right)^2} \right\}^{1/2} \in [0, 1]$$

where $\hat{d}_{ik}^{(q)}$'s are the fitted distances in a q dimensional configuration, $d_{ik}^{(q)}$'s are ideal q-dimensional pairwise numbers (with respect to a reference dissimilarity matrix) that correspond to the perfect match with the original similarities s_{ik} or their dissimilarity counterpart.

An alternative measure of closeness is **SStress** (Takane, 1977). Dropping the index q in $d_{ik}^{(q)}$,

$$SStress(q) = \left\{ \frac{\sum\sum_{i<k}(d_{ik}^2 - \hat{d}_{ik}^2)^2}{\sum\sum_{i<k} d_{ik}^4} \right\}^{1/2} \quad \in [0,1]$$

Remarks

- Typically the value of *Stress* or *SStress* less than 0.1 is considered a good representation of the objects by the points in the given low dimension configuration. In practice, the goodness of fit of Stress is based on the range of the values:

	Perfect	Excellent	Good	Fair	Poor
Stress value	0%	2.5%	5%	10%	20%

- In the formula of *Stress* or *SStress*, the ideal distance $d_{ik}^{(q)}$ between the ith and kth objects often uses the corresponding (i,k)th entry (denoted as d_{ik} or D_{ik}) in the input $n \times n$ dissimilarity matrix D, presumably the distance of the two objects in a higher dimensional space.

  In other word, in practice, when only one distance matrix D is given, the (i,k) entry d_{ik} in the given distance matrix D is used in the place of $d_{ik}^{(q)}$ in the calculation of *Stress* and *SSress*.

- The two stress measures often give comparable results. The newer one, $SStress$, is preferred sometimes. In our example demo, $SStress$ consistently yields smaller values than $Stress$.

- If Euclidean distance in the lower dimension configuration space $\mathbb{R}^q$ is used for dissimilarity, then $\hat{d}_{ik}^{(q)}$ in *Stress* has the simple form $\hat{d}_{ik}^{(q)} = \|x_i - x_k\|$ where $x_i = (x_{i1}, \cdots, x_{iq})$ represents the coordinates of the ith object under the q-dimensional configuration, $\|\cdot\|$ is the usual ℓ_2 Euclidean norm.

- The Stress measure can be viewed as a function on $\mathbb{R}^{qn}$. An optimization algorithm can be carried out by numerical methods such as gradient descent.

Algorithm outline

1. Obtain the N(N-1)/2 similarities between distance pairs of items. Order the similarities as in (3).

   Typically, dissimilarities or distances are used instead of similarities.

2. Using a trial configuration in q dimensions, determine the inter-item distances $d_{ik}^{(q)}$ and estimate $\hat{d}_{ik}^{(q)}$ to minimize the Stress or SStress measure.

3. Using $\hat{d}_{ik}^{(q)}$, moving the points around to obtain an improved configuration.

4. Choose the dimension q by evaluating the Stress or SStress measure.

3 General measure of closeness

Similarity, dissimilarity, Proximity, distance, distance metric

- Similarity measure reflects how close two objects are. The "closer" the two objects are to each other, the larger is their similarity value.
- Dissimilarity measure indicates how different two objects are. The farther the two objects are to each other, the larger is their dissimilarity value.
- Proximity can be either similarity or dissimilarity.
- Distance sometime loosely refers to a dissimilarity measure.

  On the other hand, distance as a metric is a more strictly defined mathematical concept.
- Definition of metric

  A function $d(x,y)$ is a (distance) metric or a distance function if the following holds.

  1. $d(x,y) = d(y,x)$ for any x, y. (Symmetry)
  2. $d(x,y) \geq 0$
  3. $d(x,y) = 0$ if and only if $x = y$.
  4. $d(x,y) \leq d(x,w) + d(w,y)$ for all x, y, w. (Triangle inequality)
- All sensible dissimilarity measures fulfill conditions 1 and 2. Most but not all dissimilarities satisfy condition 3. However it is not so uncommon that a dissimilarity measure fails to satisfy condition 4, which is usually the harder condition to verify.

Example

If correlation $\rho(\boldsymbol{x}, \boldsymbol{y})$ is used for similarity measure, then $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{2(1-\rho)}$ satisfies 1,2, 4 but not the "only if" part of 3. Such measure is called a pseudometric.

Proof.

1. By the definition of correlation, $\rho(\boldsymbol{x}, \boldsymbol{y}) = \rho(\boldsymbol{y}, \boldsymbol{x})$, hence $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) \geq 0$, property 1 holds.
2. By the definition of correlation, $\rho \leq 1$, $2(1-\rho) \geq 0$, thus $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$, property 2 holds.
3. By the definition of correlation, if $\boldsymbol{y} = c\boldsymbol{x}$ $(c \neq 0)$ is a non-zero constant multiple of $\boldsymbol{x}$, then $\rho(\boldsymbol{x}, \boldsymbol{y}) = 1$ thus $d(\boldsymbol{x}, \boldsymbol{y}) = 0$. Consequently, $d(\boldsymbol{x}, \boldsymbol{x}) = 0$ is true, thus the "if" part in property 3 holds. However $d(\boldsymbol{x}, 2\boldsymbol{x}) = 0$ also, which means the "only if" part in property 3 is not true.
4. To use the triangular property of Euclidean norm, let's consider centered, normed vectors.
   Let $\boldsymbol{x}_c = \boldsymbol{x} - \bar{\boldsymbol{x}}, \boldsymbol{y}_c = \boldsymbol{y} - \bar{\boldsymbol{y}} \in \mathbb{R}^n$, where $\bar{\boldsymbol{x}}$ represents an n-vector of constant component value $\bar{x}$. Since correlation does not chance when the variables subtract constants,

   $$\rho(\boldsymbol{x}, \boldsymbol{y}) = \rho(\boldsymbol{x} - \bar{\boldsymbol{x}}, y) = \rho(\boldsymbol{x} - \bar{\boldsymbol{x}}, y - \bar{\boldsymbol{y}}) = \rho(\boldsymbol{x}_c, \boldsymbol{y}_c)$$

   The Euclidean distance between $\boldsymbol{x}_c$ and $\boldsymbol{y}_c$ is

   $$\|\boldsymbol{x}_c - \boldsymbol{y}_c\|^2 = (\boldsymbol{x}_c - \boldsymbol{y}_c)'(\boldsymbol{x}_c - \boldsymbol{y}_c) = \boldsymbol{x}_c'\boldsymbol{x}_c + \boldsymbol{y}_c'\boldsymbol{y}_c - 2\boldsymbol{x}_c'\boldsymbol{y}_c = \|\boldsymbol{x}_c\|^2 + \|\boldsymbol{y}_c\|^2 - 2\|\boldsymbol{x}_c\|\|\boldsymbol{y}_c\|\rho(\boldsymbol{x}, \boldsymbol{y})$$

   Consider the normed vectors

   $$\boldsymbol{x}^* = \frac{\boldsymbol{x}_c}{\|\boldsymbol{x}_c\|}, \quad \boldsymbol{y}^* = \frac{\boldsymbol{y}_c}{\|\boldsymbol{y}_c\|}$$

Since the normed vectors are constant multiple of the original vectors, we have $\rho(\boldsymbol{x}, \boldsymbol{y}) = \rho(\boldsymbol{x}_c, \boldsymbol{y}_c) = \rho(\boldsymbol{x}^*, \boldsymbol{y}^*)$.

$$\|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 = \|\boldsymbol{x}^*\|^2 + \|\boldsymbol{y}^*\|^2 - 2\|\boldsymbol{x}^*\|\|\boldsymbol{y}^*\|\rho(\boldsymbol{x}^*, \boldsymbol{y}^*) = 2(1 - \rho(\boldsymbol{x}^*, \boldsymbol{y}^*))$$

Therefore

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{2(1 - \rho(\boldsymbol{x}, \boldsymbol{y}))} = \|\boldsymbol{x}^* - \boldsymbol{y}^*\|.$$

That is, the distance $d(\boldsymbol{x}, \boldsymbol{y})$ is the Euclidean distance of the centered, normed vectors $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$, and Euclidean norm has triangular property. We have

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x}^* - \boldsymbol{y}^*\| \leq \|\boldsymbol{x}^* - \boldsymbol{w}^*\| + \|\boldsymbol{w}^* - \boldsymbol{y}^*\| = d(\boldsymbol{x}, \boldsymbol{w}) + d(\boldsymbol{w}, \boldsymbol{y}).$$

Therefore the distance measure $d(\boldsymbol{x}, \boldsymbol{y})$ satisfies the triangular inequality, thus property 4 holds.

□

4 Classical metric scaling

How do we find a low dimension approximation of a given distance matrix?

Given coordinations of points, we can find their pairwise distances. Metric multidimensional scaling (including classical multidimensional scaling) deals with the inverse: from pairwise distance, can we find a set of coordinates for the points?

Metric scaling is an algebraic reconstruction from dissimilarity to a configuration of points.

Construction of coordinates from pairwise distance

If we knew the coordinates of the rth observation point to be $(x_{r1}, \cdots, x_{rp})$, we can obtain the Euclidean distance (or any other distance) between the rth and sth observations as d_{rs},

$$d_{rs}^2 = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2$$

Classical scaling works in the other direction: it aims to construct a set of coordinates

$$X = [x_{rs}]_{n \times p}$$

based on

$$D = [d_{rs}^2]_{n \times n}$$

that is, recover the coordinates from the knowledge of pairwise distances only.

The construction considers an intermediate matrix

$$B = [b_{rs}]_{n \times n} = XX'$$

where X is the coordinate matrix that we aim to recover. The (r, s) entry of B can be expressed as

$$b_{rs} = \sum_{j=1}^{p} x_{rj} x_{sj}$$

Then

$$d_{rs}^2 = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2 = \sum_{j=1}^{p} x_{rj}^2 + \sum_{j=1}^{p} x_{sj}^2 - 2\sum_{j=1}^{p} s_{rj} x_{sj}$$

So the pairwise distance squares in $D = [d_{rs}^2]$ can be expressed in terms of entries of matrix B,

$$d_{rs}^2 = b_{rr} + b_{ss} - 2b_{rs} \tag{5}$$

Define

$$d_{r\cdot}^2 = \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2, \qquad d_{\cdot s}^2 = \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2, \qquad d_{\cdot\cdot}^2 = \frac{1}{n^2}\sum_{s=1}^{n}\sum_{r=1}^{n} d_{rs}^2,$$

which are the row, column and overall averages of the square-distance matrix $D = [d_{rs}^2]$ respectively.

Note that rotations, reflections, and translations do not change the pairwise distance of the n points. Therefore, to recover or to solve for a set of coordinates X in $\mathbb{R}^p$, the solutions are not unique, not even when the MDS dimension is as large as $q = p$, the dimension of the original data space.

We now put in the constraint that X has column means zero, that is,

$$\sum_{k=1}^{n} x_{kj} = 0, \qquad j = 1, \cdots, p.$$

This is equivalent to have each component variables centered at 0. Then

$$\sum_{r=1}^{n} b_{rs} = \sum_{r=1}^{n} \left(\sum_{j=1}^{p} x_{rj} x_{sj} \right) = \sum_{j=1}^{p} \left(\sum_{r=1}^{n} x_{rj} \right) x_{sj} = 0$$

Given pairwise distances d_{ij}, sum over (5) with respect to r,

$$\sum_{r=1}^{n} d_{rs}^2 = \sum_{r=1}^{n} b_{rr} + \sum_{r=1}^{n} b_{ss} - 2\sum_{r=1}^{n} b_{rs} = trace(B) + nb_{ss}$$

since

$$\sum_{r=1}^{n} b_{rs} = 0$$

We may state the result of sum (5) over r as

$$nd_{\cdot s}^2 = T + nb_{ss}$$

where

$$T = trace(B) = \sum_{j=1}^{n} b_{jj}$$

Now, sum over (5) with respect to s, we have

$$nd_{r\cdot}^2 = T + nb_{rr}$$

Sum over the above with respect to r,

$$n\sum_{r=1}^{n} d_{r\cdot}^2 = \sum_{r=1}^{n} T + n\sum_{r=1}^{n} b_{rr} = nT + nT$$

We have

$$n^2 d_{\cdot\cdot}^2 = 2nT$$

The three summation results are

$$nd_{\cdot s}^2 = T + nb_{ss}, \quad nd_{r\cdot}^2 = T + nb_{rr}, \qquad T = \frac{1}{2} nd_{\cdot\cdot}^2$$

Then b_{ss} can be expressed in terms of D matrix entries as

$$b_{ss} = d_{\cdot s}^2 - \frac{1}{n} T = d_{\cdot s}^2 - \frac{1}{2} d_{\cdot\cdot}^2$$

Likewise,

$$b_{rr} = d_{\cdot r}^2 - \frac{1}{n} T = d_{\cdot r}^2 - \frac{1}{2} d_{\cdot\cdot}^2$$

From (5)

$$d_{rs}^2 = b_{rr} + b_{ss} - 2b_{rs} \qquad \Rightarrow \qquad b_{rs} = -\frac{1}{2}\left(d_{rs}^2 - b_{rr} - b_{ss}\right)$$

Combining the three equations above, we can express the entries of B in terms of the entries of D,

$$b_{rs} = -\frac{1}{2}\left(d_{rs}^2 - d_{r\cdot}^2 - d_{\cdot s}^2 + d_{\cdot\cdot}^2\right)$$

We have gone halfway toward our goal, that is, we can express elements in B in terms of the given elements in D,

$$d_{rs}^2 \in D \qquad \Longrightarrow \qquad b_{rs} \in B$$

Next we need to express the coordinates x_{rs} in terms of elements in B,

$$b_{rs} \in B \qquad \Longrightarrow \qquad x_{rs} \in X$$

By the symmetry of matrix B, the Spectral Theorem gives an eigenvalue-eigenvector decomposition of B,

$$B = V\Lambda V' = (V\Lambda^{1/2})(V\Lambda^{1/2})',$$

where

$$V = [e_1 \ \cdots \ e_n]_{n\times n}$$

is the orthonormal eigenvector matrix of B, the matrix

$$\Lambda = diag\{\lambda_1, \cdots, \lambda_n\}$$

has B's eigenvalues on the diagonal, and the relation

$$Be_i = \lambda_i e_i, \ i = 1, \cdots, n,$$

By the positive semi-definiteness of symmetric matrix $B = XX'$, for $k = rank(B) \le n$, the non-zero eigenvalues of B can be ordered as

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0,$$

with

$$\lambda_1 \ge \cdots \ge \lambda_k > 0.$$

So we may denote

$$\Lambda^{1/2} = diag\{\lambda_1^{1/2}, \cdots, \lambda_n^{1/2}\}$$

We construct X by defining

$$X = V\Lambda^{1/2} = \left[\sqrt{\lambda_1} e_1 \ \cdots \ \sqrt{\lambda_\ell} e_\ell \ \cdots \ \sqrt{\lambda_p} e_p\right]_{n\times p}$$

The $n \times p$ matrix provides a set of coordinates for the n items in $\mathbb{R}^p$, with pairwise Euclidean distance matrix D.

The above shows the process of finding a set of coordinates from pairwise distance:

$$d_{rs}^2 \in D \qquad \Longrightarrow \qquad b_{rs} \in B \qquad \Longrightarrow \qquad x_{rs} \in X$$

A "best" set of coordinates $\{(x_{r1}, \cdots, x_{rq}), r = 1, \cdots, n\} \in \mathbb{R}^q$ as a q-dimensional representation ($q \le p$) of the n data points can be defined by the first q columns,

$$X^{(q)} = \left[\sqrt{\lambda_1} e_1 \ \cdots \ \sqrt{\lambda_q} e_q\right]_{n\times q}$$

Pairwise distance, usually Euclidean distance, can be obtained and compared with the original distance, presumably from a space with (unknown) higher dimensions $p > q$.

In particular, a "best" two-dimensional map of the n points are given the coordinates

$$X^{(2)} = \left[\sqrt{\lambda_1}\, e_1 \ \ \sqrt{\lambda_2}\, e_2\right]_{n\times 2}$$

which are used in MDS $q = 2$ dimensional plots.

Remarks

- The reconstruction of the coordinates assumes or treated the given pairwise distances as Euclidean distance in the $p < n$ dimensional space. Therefore the method is particularly appropriate when the dissimilarities are actually or at least approximately Euclidean distances.
- In practice, the distance matrix D, sometimes converted from a similarity matrix, may not be an Euclidean distance matrix in any $p < n$ dimensions. Then the classical metric scaling cannot recover a set of coordinates that match the original pairwise distances.
- The classical scaling method is by singular value decomposition, or equivalently, using principal components, therefore sometimes classical scaling is also called "principal coordinates analysis".

Note: Relevant section in Johnson and Wichern: section 12.6.