# 12
# Classical Scaling

Because the first practical method available for MDS was a technique due to Torgerson (1952, 1958) and Gower (1966), *classical scaling* is also known under the names *Torgerson scaling* and *Torgerson–Gower scaling.* It is based on theorems by Eckart and Young (1936) and by Young and Householder (1938). The basic idea of classical scaling is to assume that the dissimilarities are distances and then find coordinates that explain them. In (7.5) a simple matrix expression is given between the matrix of squared distances $\mathbf{D}^{(2)}(\mathbf{X})$ (we also write $\mathbf{D}^{(2)}$ for short) and the coordinate matrix $\mathbf{X}$, which shows how to get squared Euclidean distances from a given matrix of coordinates and then scalar products from these distances. In Section 7.9, the reverse was discussed, that is, how to find the coordinate matrix given a matrix of scalar products $\mathbf{B} = \mathbf{X}\mathbf{X}'$. Classical scaling uses the same procedure but operates on squared dissimilarities $\boldsymbol{\Delta}^{(2)}$ instead of $\mathbf{D}^{(2)}$, because the latter is unknown. This method is popular because it gives an analytical solution, requiring no iterations.

## 12.1 Finding Coordinates in Classical Scaling

We now explain some fundamental issues in classical scaling. How do we arrive at a scalar product matrix $\mathbf{B}$, given a matrix of squared distances $\mathbf{D}^{(2)}$? Because distances do not change under translations, we assume that $\mathbf{X}$ has column means equal to 0. Remember from (7.5) that the squared distances are computed from $\mathbf{X}$ by

$$\mathbf{D}^{(2)} = \mathbf{c1}' + \mathbf{1c}' - 2\mathbf{XX}' = \mathbf{c1}' + \mathbf{1c}' - 2\mathbf{B}, \tag{12.1}$$

where $\mathbf{c}$ is the vector with the diagonal elements of $\mathbf{XX}'$. Multiplying the left and the right sides by the centering matrix $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{11}'$ and by the factor $-\frac{1}{2}$ gives

$$\begin{aligned} -\tfrac{1}{2}\mathbf{JD}^{(2)}\mathbf{J} &= -\tfrac{1}{2}\mathbf{J}(\mathbf{c1}' + \mathbf{1c}' - 2\mathbf{XX}')\mathbf{J} \\ &= -\tfrac{1}{2}\mathbf{Jc1}'\mathbf{J} - \tfrac{1}{2}\mathbf{J1c}'\mathbf{J} + \tfrac{1}{2}\mathbf{J}(2\mathbf{B})\mathbf{J} \\ &= -\tfrac{1}{2}\mathbf{Jc0}' - \tfrac{1}{2}\mathbf{0c}'\mathbf{J} + \mathbf{JBJ} = \mathbf{B}. \end{aligned} \tag{12.2}$$

The first two terms are zero, because centering a vector of ones yields a vector of zeros ($\mathbf{1}'\mathbf{J} = \mathbf{0}$). The centering around $\mathbf{B}$ can be removed because $\mathbf{X}$ is column centered, and hence so is $\mathbf{B}$. The operation in (12.2) is called *double centering.* To find the MDS coordinates from $\mathbf{B}$, we factor $\mathbf{B}$ by eigendecomposition, $\mathbf{Q\Lambda Q}' = (\mathbf{Q\Lambda}^{1/2})(\mathbf{Q\Lambda}^{1/2})' = \mathbf{XX}'$. The method of classical scaling only differs from this procedure in that the matrix of squared distances $\mathbf{D}^{(2)}$ is replaced by the squared dissimilarities $\mathbf{\Delta}^{(2)}$.

The procedure for classical scaling is summarized in the following steps.

1. Compute the matrix of squared dissimilarities $\mathbf{\Delta}^{(2)}$.

2. Apply double centering to this matrix:

$$\mathbf{B}_{\mathbf{\Delta}} = -\tfrac{1}{2}\mathbf{J\Delta}^{(2)}\mathbf{J}. \tag{12.3}$$

3. Compute the eigendecomposition of $\mathbf{B}_{\mathbf{\Delta}} = \mathbf{Q\Lambda Q}'$.

4. Let the matrix of the first $m$ eigenvalues *greater than zero* be $\mathbf{\Lambda}_+$ and $\mathbf{Q}_+$ the first $m$ columns of $\mathbf{Q}$. Then, the coordinate matrix of classical scaling is given by $\mathbf{X} = \mathbf{Q}_+\mathbf{\Lambda}_+^{1/2}$.

If $\mathbf{\Delta}$ happens to be a Euclidean distance matrix, then classical scaling finds the coordinates up to a rotation. Note that the solution $\mathbf{Q}_+\mathbf{\Lambda}_+^{1/2} = \mathbf{X}$ is a principal axes solution (see Section 7.10). In step 4, negative eigenvalues can occur but not if $\mathbf{\Delta}$ is a Euclidean distance matrix (see Chapter 19). In classical scaling, the negative eigenvalues (and its eigenvectors) are simply ignored as error.

Classical scaling minimizes the loss function

$$\begin{aligned} L(\mathbf{X}) &= ||-\tfrac{1}{2}\mathbf{J}[\mathbf{D}^{(2)}(\mathbf{X}) - \mathbf{\Delta}^{(2)}]\mathbf{J}||^2 \\ &= ||\mathbf{XX}' + \tfrac{1}{2}\mathbf{J\Delta}^{(2)}\mathbf{J}||^2 \\ &= ||\mathbf{XX}' - \mathbf{B}_{\mathbf{\Delta}}||^2, \end{aligned} \tag{12.4}$$

sometimes called *Strain* (see Carroll & Chang, 1972). Gower (1966) proved that choosing the classical scaling solution solves (12.4).[1]

A nice property of classical scaling is that the dimensions are nested. This means that, for example, the first two dimensions of a 3D classical scaling solution are the same as the two dimensions of a 2D classical scaling solution. Note that MDS by minimizing Stress does not give nested solutions.

It remains to be seen what dimensionality one should choose. Sibson (1979) suggests that the sum of the eigenvalues in $\mathbf{\Lambda}_+$ should approximate the sum of all eigenvalues in $\mathbf{\Lambda}$, so that small negative eigenvalues cancel out small positive eigenvalues. For a rationale of this proposal, see Chapter 19.

## 12.2 A Numerical Example for Classical Scaling

As an example, we use the faces data from Table 4.4. Here, we consider the first four items only; that is,

$$\mathbf{\Delta} = \begin{bmatrix} 0 & 4.05 & 8.25 & 5.57 \\ 4.05 & 0 & 2.54 & 2.69 \\ 8.25 & 2.54 & 0 & 2.11 \\ 5.57 & 2.69 & 2.11 & 0 \end{bmatrix}, \text{ so that } \mathbf{\Delta}^{(2)} = \begin{bmatrix} .00 & 16.40 & 68.06 & 31.02 \\ 16.40 & .00 & 6.45 & 7.24 \\ 68.06 & 6.45 & .00 & 4.45 \\ 31.02 & 7.24 & 4.45 & .00 \end{bmatrix}.$$

The second step in classical scaling is to compute

$$\begin{aligned} \mathbf{B}_{\mathbf{\Delta}} &= -\tfrac{1}{2}\mathbf{J}\mathbf{\Delta}^{(2)}\mathbf{J} \\ &= -\frac{1}{2} \begin{bmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{bmatrix} \begin{bmatrix} .00 & 16.40 & 68.06 & 31.02 \\ 16.40 & .00 & 6.45 & 7.24 \\ 68.06 & 6.45 & .00 & 4.45 \\ 31.02 & 7.24 & 4.45 & .00 \end{bmatrix} \begin{bmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{bmatrix} \\ &= \begin{bmatrix} 20.52 & 1.64 & -18.08 & -4.09 \\ 1.64 & -.83 & 2.05 & -2.87 \\ -18.08 & 2.05 & 11.39 & 4.63 \\ -4.09 & -2.87 & 4.63 & 2.33 \end{bmatrix}. \end{aligned}$$

In the third step, we compute the eigendecomposition of $\mathbf{B}_{\mathbf{\Delta}}$; that is, $\mathbf{B}_{\mathbf{\Delta}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$ with

$$\mathbf{Q} = \begin{bmatrix} .77 & .04 & .50 & -.39 \\ .01 & -.61 & .50 & .61 \\ -.61 & -.19 & .50 & -.59 \\ -.18 & .76 & .50 & .37 \end{bmatrix} \text{ and } \mathbf{\Lambda} = \begin{bmatrix} 35.71 & .00 & .00 & .00 \\ .00 & 3.27 & .00 & .00 \\ .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & -5.57 \end{bmatrix}.$$

There are two positive eigenvalues, one zero eigenvalue due to the double centering and one negative eigenvalue.[2] For this example, we can construct

[1] Note that Kloek and Theil (1965) also derived the classical scaling solution, but more in the sense of a how-to-do construction scheme than in terms of algebra.

[2] Double centering introduces a linear dependency, because if the columns of a matrix add up to the zero vector, then any column can be expressed as a linear combination

at most two dimensions in Euclidean space. Step 4 tells us that the configuration $\mathbf{X}$ is found by

$$\begin{aligned}\mathbf{X} &= \mathbf{Q}_+\mathbf{\Lambda}_+^{1/2}\\ &= \begin{bmatrix} .77 & .04\\ .01 & \text{-}.61\\ \text{-}.61 & \text{-}.19\\ \text{-}.18 & .76\end{bmatrix}\begin{bmatrix} 5.98 & .00\\ .00 & 1.81\end{bmatrix} = \begin{bmatrix} 4.62 & .07\\ .09 & \text{-}1.11\\ \text{-}3.63 & \text{-}.34\\ \text{-}1.08 & 1.38\end{bmatrix}.\end{aligned}$$

## 12.3 Choosing a Different Origin

Usually, $\mathbf{X}$ is constructed so that its columns sum to zero. This means that the origin of configuration $\mathbf{X}$ coincides with the center of gravity of its points (centroid). Choosing this origin is, however, not necessarily the best choice. In psychological research, for example, some objects may be less familiar to the respondents, and thus lead to less reliable distance estimates than others. In such a case, it is wiser to pick as an origin a point that is based more on the points associated with less error. How could this be accomplished?

For a general solution, consider picking some arbitrary point $s$ as the new origin, with the restriction that $s$ lies in the space of the other points. That is, in terms of algebra, point $s$ should lie in the row space of $\mathbf{X}$; that is, the coordinate vector of $s$ is a weighted sum of the rows of $\mathbf{X}$, $\mathbf{s}' = \mathbf{w}'\mathbf{X}$, where $\mathbf{w}'$ is an $m$-element row vector of weights. With $s$ as the new origin, the point coordinates become

$$\mathbf{X}_s = \mathbf{X} - \mathbf{1}\mathbf{s}' = \mathbf{X} - \mathbf{1}\mathbf{w}'\mathbf{X} = (\mathbf{I} - \mathbf{1}\mathbf{w}')\mathbf{X} = \mathbf{P}_w\mathbf{X}. \tag{12.5}$$

If the weight vector $\mathbf{w}$ is chosen such that $\mathbf{w}'\mathbf{1} = 1$, then $\mathbf{P}_w$ is a *projector*[3]. If $\mathbf{B} = \mathbf{X}\mathbf{X}'$, one obtains $\mathbf{B}_s = \mathbf{X}_s\mathbf{X}_s'$ after projecting $\mathbf{X}$ to a new origin $s$. In terms of the old origin, $\mathbf{B}_s = \mathbf{P}_w\mathbf{B}\mathbf{P}_w'$.

If one chooses a particular object $i$ as the origin, then $\mathbf{w}' = [0, \ldots, 1, \ldots, 0]$, where the 1 is in the $i$th position. If one picks the centroid as the origin, then $\mathbf{w}' = [1/n, \ldots, 1/n]$. Another choice is to pick the weights in $\mathbf{w}$ so that they reflect the reliability of the objects. In this case, unreliable elements should have a weight close to zero and reliable elements a high value. In this way, the origin will be attracted more towards the reliable points.

of the other columns. Hence, a doubly centered matrix does not have full rank, and, therefore, it has at least one zero eigenvalue (see Chapter 7). The negative eigenvalue shows that $\mathbf{\Delta}$ is not a matrix of Euclidean distances (see Section 19.1).

[3] For every projector matrix $\mathbf{P}$ it holds that $\mathbf{PP} = \mathbf{P}$ (*idempotency*). In the given case, it is also true that $\mathbf{P}_s\mathbf{1} = \mathbf{0}$ (Schönemann, 1970).

Instead of using $\mathbf{J}$ in the double-centering formula, we can also use the projector $\mathbf{P}_w$. Then, step 2 in classical scaling becomes $\mathbf{B}_\Delta = -\frac{1}{2}\mathbf{P}_w\mathbf{\Delta}^{(2)}\mathbf{P}'_w$. The zero eigenvalue of $\mathbf{B}_\Delta$ has eigenvector $\mathbf{w}$, so that the weighted average (using weights $\mathbf{w}$) of the classical scaling coordinate matrix $\mathbf{X}$ is equal to zero.

## 12.4 Advanced Topics

A solution for classical scaling with linear constraints was discussed by Carroll, Green, and Carmone (1976), De Leeuw and Heiser (1982), and Ter Braak (1992). The linear constraints imposed on $\mathbf{X}$ require $\mathbf{X} = \mathbf{YC}$, where $\mathbf{Y}$ is an $n \times r$ matrix of $r$ external variables, and $\mathbf{C}$ are weights to be optimized by classical scaling. (This type of constraint was also discussed in Section 10.3 for constrained MDS with Stress.)

How can the weights in $\mathbf{C}$ be computed? Let $\mathbf{Y} = \mathbf{P\Phi Q}'$ be the singular value decomposition. Then $\mathbf{X} = \mathbf{YC} = \mathbf{P\Phi Q}'\mathbf{C} = \mathbf{PC}_*$, where $\mathbf{P}$ is orthonormal ($\mathbf{P}'\mathbf{P} = \mathbf{I}$). The Strain loss function (12.4) used by classical scaling can be written as

$$\begin{aligned} L(\mathbf{C}) &= ||\mathbf{B}_\Delta - \mathbf{YCC}'\mathbf{Y}'||^2 = ||\mathbf{B}_\Delta - \mathbf{PC}_*\mathbf{C}'_*\mathbf{P}'||^2 \\ &= ||\mathbf{B}_\Delta||^2 - ||\mathbf{P}'\mathbf{B}_\Delta\mathbf{P}||^2 + ||\mathbf{P}'\mathbf{B}_\Delta\mathbf{P} - \mathbf{C}_*\mathbf{C}'_*||^2, \qquad (12.6) \end{aligned}$$

which can be verified by writing out all of the terms in the equation. Only the last term of (12.6) is dependent on $\mathbf{C}_*$. $L(\mathbf{C})$ is solved for $\mathbf{C}$ by the eigendecomposition of $\mathbf{P}'\mathbf{B}_\Delta\mathbf{P} = \mathbf{Q\Lambda Q}'$ and choosing $\mathbf{C}_* = \mathbf{Q}_+\mathbf{\Lambda}_+^{1/2}$ (as in Step 4 in Section 12.1), so that $\mathbf{C} = \mathbf{Q\Phi}^{-1}\mathbf{C}_*$.

To illustrate constrained classical scaling, we reanalyze the constrained MDS of the facial expression data in Section 10.3. The external constraint matrix $\mathbf{Y}$ is defined as in Table 10.2. The total loss of the constrained 2D classical scaling solution is 8366.3, which explains 75% of the sum-of-squares of $\mathbf{B}_\Delta$, against 92% for the unconstrained classical scaling solution (with loss 2739.9). The corresponding solution is shown in Figure 12.1, where the external variables are represented by lines. The optimal weight matrix $\mathbf{C}$ obtained by constrained classical scaling is

$$\mathbf{C} = \begin{bmatrix} 1.283 & .482 \\ -.219 & .300 \\ -.445 & .782 \end{bmatrix}.$$

To get the coordinates $\mathbf{X}$ in Figure 12.1, we compute $\mathbf{X} = \mathbf{YC}$. This solution does not differ much from the constrained MDS solution in Figure 10.4. The main difference lies in the location of point 8.

Even for loss functions other than $L(\mathbf{X})$, classical scaling is optimal. Let $\mathbf{E} = \mathbf{XX}' - \mathbf{B}_\Delta$, so that $L(\mathbf{X}) = ||\mathbf{E}||^2$. The loss can also be expressed as the

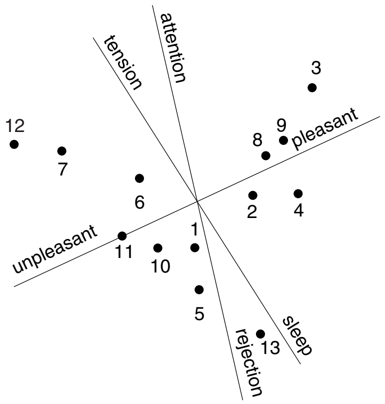

FIGURE 12.1. Constrained classical scaling of the facial expression data of Abelson and Sermat (1962).

sum of the squared eigenvalues of $\mathbf{E}$; that is, if $\mathbf{E} = \mathbf{K\Phi K}'$, then $||\mathbf{E}||^2 = \sum_i \phi_i^2$. This loss function is an example of an orthonormal invariant norm, because the value of the loss function remains invariant under pre- and postmultiplication of the orthonormal matrix $\mathbf{K}$. Mathar and Meyer (1993) prove that the classical scaling solution is also optimal for the minimization of any orthonormal invariant norm on $\mathbf{E}$. For example, the classical scaling solution is optimal if the loss is defined as $L(\mathbf{X}) = \sum_i |\phi_i|$.

In contrast to the MDS method discussed in Chapter 9, it is difficult to incorporate transformations of the proximities in classical scaling. An algorithm was proposed by Trosset (1993) that optimally transforms the proximities for Strain.

Classical scaling can even be used to study or discover the "intrinsic geometry" of highly nonlinear structures contained in high-dimensional spaces. For example, a point configuration that forms a helix in 3D space is intrinsically one-dimensional in the sense that if you move back and forth on this helix, the distances along the helix are additive. As long as you stay on the helix, Euclidean distances $d_{ij}$ are only approximately correct measures for the length of the path from point $i$ to point $j$ if $i$ and $j$ are close (or, in the case of highly nonlinear structures, "very close") to each other. For points that are far apart, Euclidean distances can grossly underestimate the intrinsic distance of $i$ and $j$. To study such geometries and to unroll them into low-dimensional Euclidean geometries, Tenenbaum, De Silva, and Langford (2000) first define the radius of a small neighborhood, $\epsilon$, and then set $\delta_{ij} = d_{ij}$ for all $ij$ where $d_{ij} < \epsilon$, and $\delta_{ij} = \infty$ otherwise. Then, in a second cycle, these values are replaced by computing distances over the network of point triples as follows: $\delta_{ij} = \min_k(\delta_{ij}, \delta_{ik} + \delta_{jk})$, for all $k$. If there are many points that are well spread out, this generates graph distances that approximate the lengths of the paths within the curved structure. Applying classical scaling to dissimilarities generated in such a way from nonlinear structures allowed Tenenbaum et al. (2000) to unroll these structures successfully.

## 12.5 Exercises

*Exercise 12.1* Use classical scaling on the data in Table 4.1, p. 65. (Note: You first have to transform the similarity data into reasonable dissimilarities.) Compare the solution to the one obtained by ordinal MDS (Figure 4.1).

*Exercise 12.2* Use matrix $\mathbf{X}$ computed in Section 12.2, p. 264, to reconstruct both $\mathbf{B}_{\boldsymbol{\Delta}}$ and $\boldsymbol{\Delta}$. Assess how well this $\mathbf{X}$ "explains" $\boldsymbol{\Delta}$.

*Exercise 12.3* Take matrix $\boldsymbol{\Delta}$ from Section 12.2. Instead of centering this matrix, choose one of its entries as the element serves as the origin of the MDS space.

(a) Compute $\mathbf{B}_{\boldsymbol{\Delta}}$ relative to this particular origin.

(b) Find the classical scaling representation for this $\mathbf{B}_{\boldsymbol{\Delta}}$.

(c) Compare this solution to the solution $\mathbf{X}$ found in Section 12.2.