11

MDS Fit Measures, Their Relations, and Some Algorithms

A problem in MDS is how to evaluate the Stress value. Once a solution is found, how good is it? In Chapter 3, several statistical simulation studies were reported. Here we give an interpretation of normalized Stress in terms of the proportion of the explained sum-of-squares of the disparities. We also show that normalized Stress is equal to Stress-1 at a minimum and that the configurations only differ by a scale factor. Then, other common measures of fit for MDS are discussed. For these fit measures, we refer to some recent algorithmic work. Finally, it is discussed how weights in MDS can be used to emphasize different aspects of the data, to approach other MDS loss functions, or to take the reliability of the data into account.

Throughout this chapter, we refer to the data as being dissimilarities δ_{ij} for notational simplicity. However, all definitions of Stress measures and their relations remain valid when the dissimilarities are replaced by $\widehat{d}_{ij}$ obtained by optimal transformation (see the approach taken in Section 9.1).

11.1 Normalized Stress and Raw Stress

In Section 3.2, we saw that σ_r depends on the "size" of $\mathbf{X}$. Changing the scale of the coordinates of $\mathbf{X}$ changes σ_r accordingly. To avoid this scale dependency, one can use the implicit normalization used in Kruskal's Stress-1. Here, we elaborate on a different measure, which we call *normalized Stress*. This coefficient shows (after convergence) the proportion of the sum-

of-squares of the δ_{ij}s that is *not* accounted for by the distances. We define normalized Stress $\sigma_n(\mathbf{X})$ as

$$\sigma_n(\mathbf{X}) = \frac{\sigma_r(\mathbf{X})}{\eta_\delta^2} = \frac{\sum_{i<j} w_{ij}(\delta_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{i<j} w_{ij}\delta_{ij}^2}. \tag{11.1}$$

Clearly, if $\sum_{i<j} w_{ij}\delta_{ij}^2 = 1$, then $\sigma_n(\mathbf{X}) = \sigma_r(\mathbf{X})$.

De Leeuw (1977) (and, among others, Commandeur, 1993) show how $\sigma_n(\mathbf{X})$ is related to the square of Tucker's coefficient of congruence.[1] This relation can be explained as follows. Suppose that $\mathbf{X}^*$ is a local minimum of $\sigma_r(\mathbf{X})$. This implies that $b\mathbf{Y}^* = \mathbf{X}^*$ (with $b > 0$) also must be a local minimum. Note that $\mathbf{Y}^*$ has coordinates that are proportional to $\mathbf{X}^*$. We show that for optimal b normalized Stress is equal to one minus the square of Tucker's coefficient of congruence.

To find an optimal b, we use the property that the Euclidean distance is a positively homogeneous function in $\mathbf{X}$; that is, $d_{ij}(b\mathbf{Y}^*) = bd_{ij}(\mathbf{Y}^*)$ for $b \geq 0$. Then $\sigma_r(b\mathbf{Y}^*)$ can be written as

$$\begin{aligned}
\sigma_r(b\mathbf{Y}^*) &= \sum_{i<j} w_{ij}(\delta_{ij} - d_{ij}(b\mathbf{Y}^*))^2 \\
&= \sum_{i<j} w_{ij}\delta_{ij}^2 + b^2 \sum_{i<j} w_{ij} d_{ij}^2(\mathbf{Y}^*) - 2b \sum_{i<j} w_{ij}\delta_{ij} d_{ij}(\mathbf{Y}^*) \\
&= \eta_\delta^2 + b^2\eta^2(\mathbf{Y}^*) - 2b\rho(\mathbf{Y}^*).
\end{aligned} \tag{11.2}$$

The minimum of (11.2) over b is obtained by setting the first derivative of $\sigma_r(b\mathbf{Y}^*)$ with respect to b equal to zero, $2b\eta^2(\mathbf{Y}^*) - 2\rho(\mathbf{Y}^*) = 0$. Thus, the optimal b is $b^* = \rho(\mathbf{Y}^*)/\eta^2(\mathbf{Y}^*)$ (see, e.g., Mathar & Groenen, 1991). Inserting b^* in $\sigma_r(b\mathbf{Y}^*)$ gives

$$\sigma_r(b^*\mathbf{Y}^*) = \eta_\delta^2 - \left(\frac{\rho(\mathbf{Y}^*)}{\eta(\mathbf{Y}^*)}\right)^2. \tag{11.3}$$

Dividing both sides by η_δ^2 yields

$$\sigma_n(b^*\mathbf{Y}^*) = \frac{\sigma_r(b^*\mathbf{Y}^*)}{\eta_\delta^2} = 1 - \left(\frac{\rho(\mathbf{Y}^*)}{\eta_\delta\eta(\mathbf{Y}^*)}\right)^2, \tag{11.4}$$

where the last term is equal to the square of Tucker's congruence coefficient with distances and dissimilarities. The congruence coefficient is always between -1 and 1, due to the Cauchy–Schwarz inequality. Moreover,

[1] The congruence coefficient of two variables X and Y, c, is the correlation of these variables about their origin or "zero", not about their means (as in Pearson's correlation coefficient). The coefficient c was first used by Tucker (see, e.g., Tucker, 1951) to assess the similarity of corresponding factors resulting from factor analyses of different samples. It is defined as $c = (\sum_i (x_i y_i)/[(\sum_i x_i^2)(\sum_i y_i^2)]^{1/2}$.

negative congruence coefficients are impossible because distances and dissimilarities are nonnegative. Hence, at a stationary point $\mathbf{X}^*$, it holds that $0 \leq \sigma_n(\mathbf{X}^*) \leq 1$. The value of $\sigma_n(\mathbf{X}^*)$ is the proportion of variation of the dissimilarities not accounted for by the distances, and $1 - \sigma_n(\mathbf{X}^*)$ is the fitted proportion, a *coefficient of determination*. Because distances and dissimilarities both are positive, congruence coefficients tend to be close to 1 in practice. Therefore, values of $\sigma_n(\mathbf{X}^*) < .10$ are usually not difficult to obtain.

Using the normalized Stress (as defined in this section) gives a clear interpretation that does not depend on the scale of the dissimilarities.

Relation Between Normalized Stress and Stress-1

Fortunately, there exists a simple relation between the normalized Stress σ_n and Stress-1 σ_1. In fact, we show here that $\sigma_1^2 = \sigma_n$ at a local minimum if we allow for a rescaling of the solution. Note that Raw Stress σ_r and normalized Stress σ_n differ from most other Stress measures in that no square root is taken.

Let $\mathbf{X}^*$ be a local minimum obtained by minimizing σ_n. De Leeuw and Heiser (1980) and De Leeuw (1988) proved that for $\mathbf{X}^*$ it holds that $\eta^2(\mathbf{X}^*) = \rho(\mathbf{X}^*)$. This result implies that

$$\sigma_n(\mathbf{X}^*) = 1 - \frac{\eta^2(\mathbf{X}^*)}{\eta_\delta^2}. \tag{11.5}$$

Now, for the same configuration, Stress-1 can be expressed as

$$\begin{aligned} \sigma_1^2(\mathbf{X}^*) &= \frac{\eta_\delta^2 + \eta^2(\mathbf{X}^*) - 2\rho(\mathbf{X}^*)}{\eta^2(\mathbf{X}^*)} = \frac{\eta_\delta^2 - \eta^2(\mathbf{X}^*)}{\eta^2(\mathbf{X}^*)} \\ &= \frac{\eta_\delta^2}{\eta^2(\mathbf{X}^*)} - 1. \end{aligned}$$

From (11.5) we have $\eta_\delta^2/\eta^2(\mathbf{X}^*) = 1/(1 - \sigma_n(\mathbf{X}^*))$, so that

$$\sigma_1^2(\mathbf{X}^*) = \frac{\sigma_n(\mathbf{X}^*)}{1 - \sigma_n(\mathbf{X}^*)}.$$

However, the scale of $\mathbf{X}^*$ is not optimal for Stress-1. By allowing for a scaling factor b, Stress-1 becomes

$$\sigma_1^2(b\mathbf{X}^*) = \frac{\eta_\delta^2 + b^2\eta^2(\mathbf{X}^*) - 2b\rho(\mathbf{X}^*)}{b^2\eta^2(\mathbf{X}^*)} = \frac{\eta_\delta^2 + (b^2 - 2b)\eta^2(\mathbf{X}^*)}{b^2\eta^2(\mathbf{X}^*)}.$$

An optimal b can be found by differentiating $\sigma_1^2(b\mathbf{X}^*)$ with respect to b; that is,

$$\frac{\partial \sigma_1^2(b\mathbf{X}^*)}{\partial b} = \frac{2b^2[b-1]\eta^4(\mathbf{X}^*) - 2b\eta^2(\mathbf{X}^*)[\eta_\delta^2 + (b^2 - 2b)\eta^2(\mathbf{X}^*)]}{b^4\eta^2(\mathbf{X}^*)}$$

$$= \frac{2b\eta^2(\mathbf{X}^*) - 2\eta_\delta^2}{b^3},$$

which is equal to zero for $b^* = \eta_\delta^2/\eta^2(\mathbf{X}^*)$. Inserting b^* into $\sigma_1^2(b\mathbf{X}^*)$ yields

$$\begin{aligned}
\sigma_1^2(b^*\mathbf{X}^*) &= \frac{\eta_\delta^2 + \frac{\eta_\delta^4}{\eta^4(\mathbf{X}^*)}\eta^2(\mathbf{X}^*) - 2\frac{\eta_\delta^2}{\eta^2(\mathbf{X}^*)}\eta^2(\mathbf{X}^*)}{\frac{\eta_\delta^4}{\eta^4(\mathbf{X}^*)}\eta^2(\mathbf{X}^*)} \\
&= \frac{\eta_\delta^2/\eta^2(\mathbf{X}^*) - 1}{\eta_\delta^2/\eta^2(\mathbf{X}^*)} \\
&= 1 - \frac{\eta^2(\mathbf{X}^*)}{\eta_\delta^2} = \sigma_n(\mathbf{X}^*).
\end{aligned}$$

This proves that Stress-1 is equal to normalized Stress at a local minimum if the scale is calibrated properly.

11.2 Other Fit Measures and Recent Algorithms

A whole variety of MDS loss functions have been proposed in the literature. In this section, we describe some of them. A summary of different fit measures and their relations is given by Heiser (1988a). Here, we restrict ourselves to the most commonly used MDS loss functions. One of the reasons for our emphasis on using Stress in MDS is that the majorization algorithm is a simple procedure for which nice theoretical convergence results have been derived (De Leeuw, 1988). In this section, we assume that the weights $w_{ij} = 1$, for all i, j. We start with a brief overview of other algorithms for minimizing Stress.

Algorithms for Minimizing Raw Stress

Let us first turn to raw Stress. Apart from majorization, several other approaches for minimizing raw Stress have been reported in the literature. Some of these approaches are equivalent to the majorization algorithm discussed in Section 8.6. For example, De Leeuw (1993) reparameterized the raw Stress function, where the coordinates are restricted to be a sum of some other fixed coordinate matrices. The algorithm is also based on majorization. A convex analysis approach for minimizing raw Stress (De Leeuw, 1977; Mathar, 1989; Mathar & Groenen, 1991; Meyer, 1993) leads to the same algorithm as the majorization approach. A relation between the convex analysis approach and the majorization approach (for the more general case of Minkowski distances) was discussed by Mathar (1994). A genetic algorithm to minimize raw Stress was proposed by Mathar and Žilinskas (1993), who found this a promising approach for small MDS

problems. Glunt, Hayden, and Raydan (1993) proposed a spectral gradient algorithm, which was, in one example, 10 times faster than the majorizing algorithm.

Implicitly Normalized Stress

In Section 3.2, it was indicated that raw Stress $\sigma_r(\mathbf{X})$ can be misleading, because it is dependent on the normalization of the dissimilarities. To circumvent this inconvenience, normalized Stress $\sigma_n(\mathbf{X})$ was introduced in Section 11.1. A different solution is to require explicitly $\eta_\delta^2 = c$, with c a positive constant (e.g., $\eta_\delta^2 = n(n-1)/2$), as was imposed in nonmetric MDS by (9.2). This solution is called *explicit* normalization. A third (but historically earlier) solution was pursued by Kruskal (1964a), which is called *implicit normalization*. Here, Stress is expressed in relation to the size of $\mathbf{X}$. More concretely, σ is divided by the sum of the squared distances in $\mathbf{X}$ and the root is taken of the total fraction; that is,

$$\sigma_1(\mathbf{X}) = \left(\frac{\sigma(\mathbf{X})}{\eta^2(\mathbf{X})}\right)^{1/2} = \left(\frac{\sum_{i<j}[\delta_{ij} - d_{ij}(\mathbf{X})]^2}{\sum_{i<j} d_{ij}(\mathbf{X})}\right)^{1/2}.$$

This expression, proposed by Kruskal (1964a) is called *Stress formula 1*. Note that often Stress-1 is expressed using disparities $\widehat{d}_{ij}$ to allow for transformations. Throughout this chapter, we use dissimilarities δ_{ij} instead of $\widehat{d}_{ij}$ for reasons of notational simplicity. Kruskal and Carroll (1969) proved that implicitly or explicitly normalized Stress gives the same configuration up to scaling constant. A different form of implicit normalization is *Stress formula 2*; that is,

$$\sigma_2(\mathbf{X}) = \left(\frac{\sum_{i<j}[\delta_{ij} - d_{ij}(\mathbf{X})]^2}{\sum_{i<j}[d_{ij}(\mathbf{X}) - \bar{d}]^2}\right)^{1/2},$$

with $\bar{d}$ the average distance. This version of Stress was introduced to avoid a particular type of degeneracy in unfolding, that is, solutions where all distances are equal.

The Alienation Coefficient and the Guttman–Lingoes Programs

Another error measure, the *alienation coefficient*, abbreviated as K, is used only in combination with rank-images as target distances. K can be derived from normalized Stress $\sigma_n(\mathbf{X})$ as defined in (11.4) by setting $\delta_{ij} = d_{ij}^*$, where d_{ij}^* denotes the disparity obtained by the rank-image transformation (see Section 9.5). Thus, the alienation coefficient is defined as

$$K = \left(1 - \frac{[\sum_{i<j} d_{ij}^* d_{ij}(\mathbf{X})]^2}{\sum_{i<j}(d_{ij}^*)^2 \sum_{i<j} d_{ij}^2(\mathbf{X})}\right)^{1/2}. \tag{11.6}$$

The quotient term in (11.6) is known as the *monotonicity coefficient*, μ (Guttman, 1981). It is similar to a correlation coefficient, which is easier to see if we rewrite it as

$$\mu = \frac{\sum_{i<j} d^*_{ij} d_{ij}(\mathbf{X})}{[\sum_{i<j}(d^*_{ij})^2 \sum_{i<j} d^2_{ij}(\mathbf{X})]^{1/2}}. \tag{11.7}$$

Hence, μ differs from the usual Pearson correlation coefficient on the variables *distances* and *rank-images* in not subtracting the means from the variables. The regression line, therefore, runs through the origin and not the centroid of the *image diagram*, the plot of all points with coordinates (d_{ij}, d^*_{ij}). Note that in an image diagram all points are exactly on the bisector if and only if the solution is perfect. In that case, $\mu = 1$. Furthermore, μ is equal to Tucker's congruence coefficient of the distances and their rank-images.

For practical purposes, μ has the disadvantage that it takes on values close to 1 even if the MDS solution is far from perfect. We can, however, convert μ into the coefficient of alienation

$$K = (1 - \mu^2)^{1/2},$$

which yields values that vary over a greater range and, thus, are easier to distinguish. K is a measure for the "unexplained" variation of the points in the image diagram, whereas μ^2 is a *coefficient of determination*, that is, a measure for the "explained" variance. The smaller K, the more precise is the representation, or, conversely, the greater K, the worse the fit of the MDS model to the proximities. The squared alienation coefficient is equal to normalized Stress $\sigma_n(\mathbf{X})$ if rank-images are used instead of disparities. The Guttman–Lingoes programs and various other programs (see Appendix A) do ordinal MDS by attempting to minimize K rather than Stress.

Minimizing S-Stress

The S-Stress loss function of Takane, Young, and De Leeuw (1977),

$$\sigma_{AL}(\mathbf{X}) = \sum_{i<j} (d^2_{ij}(\mathbf{X}) - \delta^2_{ij})^2, \tag{11.8}$$

is minimized by ALSCAL (see Appendix A). This loss function sums the differences of squared dissimilarities and squared distances. One of the reasons for using squared distances is that $\sigma_{AL}(\mathbf{X})$ is differentiable everywhere, even if $d_{ij}(\mathbf{X}) = 0$ for some pair i, j. Squaring distances and dissimilarities causes S-Stress to emphasize larger dissimilarities more than smaller ones, which may be viewed as a disadvantage of S-Stress. A fast Newton–Raphson procedure to minimize S-Stress was proposed by Browne (1987). An alternative algorithm was presented by Glunt, Hayden, and Liu (1991). For the full-dimensional case of $m = n - 1$, Gaffke and Mathar (1989) developed an algorithm that always yields a global minimum.

Maximum Likelihood MDS and MULTISCALE

The MULTISCALE loss function of Ramsay (1977) is based on the sum of the squared difference of the logarithm of the dissimilarities and the distances; that is,

$$\sigma_{MU}(\mathbf{X}) = \sum_{i<j} \left[\log(d_{ij}(\mathbf{X})) - \log(\delta_{ij})\right]^2 .$$

This loss function is used in a *maximum likelihood* (ML) framework. The likelihood is the probability that we find the data given $\mathbf{X}$. This probability is maximized in ML-MDS. For ML estimation, we need to assume independence among the residuals and a *lognormal* distribution of the residuals. In many cases, these assumptions are too rigid. However, if they do hold, then σ_{MU} has the advantage that confidence regions of the points can be obtained and that different models can be tested. If the residuals are assumed to be *normally* distributed, then MULTISCALE reduces to minimizing Stress. An advantage of using a logarithm in σ_{MU} is that the large dissimilarities do not determine the solution as much as when Stress is minimized. Conversely, dissimilarities close to zero are relatively important for the solution. The MULTISCALE program is discussed in Appendix A.

Further Algorithms and Developments

Groenen, De Leeuw, and Mathar (1996) discussed a least-squares loss function for MDS that includes Stress, S-Stress, and MULTISCALE as special cases. They used

$$\sigma_G(\mathbf{X}) = \sum_{i<j} w_{ij} \left[f(\delta_{ij}^2) - f(d_{ij}^2(\mathbf{X}))\right]^2 ,$$

where $f(z)$ is an increasing scalar function. For example, choosing $f(z) = z^{1/2}$ gives Stress, $f(z) = z$ gives S-Stress, and $f(z) = \log(z)$ gives the MULTISCALE loss function. They derive several properties of the gradient and hessian (the matrix of second derivatives) of this function. For example, it can be shown that S-Stress is differentiable everywhere (Takane et al., 1977) and that at a local minimum Stress has no zero distances (and thus is differentiable) if $w_{ij}\delta_{ij} > 0$ for all i, j (De Leeuw, 1984). Kearsley, Tapia, and Trosset (1998) provide an algorithm for the Stress and S-Stress versions of σ_G based on a globalized Newton's method, which they claim uses fewer iterations than the majorizing algorithm and yields lower Stress solutions.

To minimize Stress, Luengo, Raydan, Glunt, and Hayden (2002) have elaborated on the so-called spectral gradient algorithm. In a small comparison study, Groenen and Heiser (2000) found that the spectral gradient algorithm was the fastest algorithm, outperforming SMACOF and KYST. This may be of importance for MDS with a large number of objects.

A special case of σ_G occurs in applications in chemistry, where the objective is to find stable molecules. The energy function used is essentially equal to $\sigma_G(\mathbf{X})$ with $f(z) = z^6$ and $\delta_{ij} = 1$ for all i, j. The gradient becomes so steep that this problem turns out to be combinatorial in nature (see, e.g., Xue, 1994).

De Leeuw and Groenen (1997) considered the problem of finding those dissimilarity matrices for which a given $\mathbf{X}$ is a local minimum (or has a zero gradient) for Stress. This problem is called *inverse MDS*. If this set of dissimilarities is large, then the local minimum is not very informative. After all, many dissimilarity matrices have $\mathbf{X}$ as a (possible) local minimum. Groenen et al. (1996) discuss the problem of inverse MDS for the loss function $\sigma_G(\mathbf{X})$.

An overview of various algorithmic approaches in MDS is given by Mathar (1997).

11.3 Using Weights in MDS

So far, we have used the weights w_{ij} only to indicate nonmissing dissimilarities. Choosing $w_{ij} = 1$ indicates that for object pair ij a dissimilarity has been observed, whereas $w_{ij} = 0$ is used for pairs ij where a dissimilarity is "missing". As zero weights lead to zero error terms in the Stress loss function, the distance that corresponds to a missing data value cannot be assessed in terms of fit. Hence, it contributes nothing to the Stress, whatever its value. But this also means that this distance cannot be interpreted directly, but only in terms of what is implied by the distances that represent given data. If the number of missing dissimilarities gets large or if they form special block patterns (as in Table 6.1, e.g.), we should take care in interpreting distances that "represent" missing data. Then, one should emphasize the interpretation of distances that represent observed data values.

Using Particular Weighting Schemes

Instead of using w_{ij}'s that are zero or one in the minimization of Raw Stress, we can apply any positive value for w_{ij}. Heiser (1988a) exploited this powerful idea and distinguished several weighting schemes of which we discuss a few below.

Consider the S-Stress loss function. Instead of (11.8), S-Stress may also be written as

$$\sigma_{AL}(\mathbf{X}) = \sum_{i<j} (\delta_{ij} + d_{ij}(\mathbf{X}))^2 (\delta_{ij} - d_{ij}(\mathbf{X}))^2,$$

which shows that each S-Stress error term consists of two factors: the square of the ordinary Stress residual $(\delta_{ij} - d_{ij}(\mathbf{X}))^2$ and a weighting term $(\delta_{ij} +$

$d_{ij}(\mathbf{X}))^2$ that is also dependent on $d_{ij}(\mathbf{X})$. Assume that the residuals are reasonably small. Then, $(\delta_{ij} + d_{ij}(\mathbf{X}))^2$ can be approximated by replacing $d_{ij}(\mathbf{X})$ by δ_{ij} so that

$$(\delta_{ij} + d_{ij}(\mathbf{X}))^2 \approx 4\delta_{ij}^2.$$

Therefore, the minimization of S-Stress can be approximated by minimizing Stress choosing $w_{ij} = 4\delta_{ij}^2$. This approximation shows that optimizing S-Stress tends to lead to small errors for the large dissimilarities and large errors for the smaller dissimilarities. In other words, large dissimilarities are much better represented than the small ones.

McGee (1966) proposed the idea of *elastic scaling*. This form of MDS fits relative residuals so that the proper representation of small dissimilarities is equally important as fitting large dissimilarities. The loss function minimized in elastic scaling is

$$\sigma_{EL}(\mathbf{X}) = \sum_{i<j}(1 - d_{ij}(\mathbf{X})/\delta_{ij})^2 = \sum_{i<j}\delta_{ij}^{-2}(\delta_{ij} - d_{ij}(\mathbf{X}))^2.$$

Thus, choosing $w_{ij} = \delta_{ij}^{-2}$ makes minimizing raw Stress do the same as McGee's elastic scaling.

An MDS method popular in the pattern recognition literature is called *Sammon mapping* after Sammon (1969). The loss function can be expressed as

$$\sigma_{SAM}(\mathbf{X}) = \sum_{i<j}\delta_{ij}^{-1}(\delta_{ij} - d_{ij}(\mathbf{X}))^2,$$

which is identical to raw Stress for $w_{ij} = \delta_{ij}^{-q}$, with $q = -1$. The objective is somewhat similar to that of elastic scaling of McGee (1966), although larger dissimilarities still are somewhat more emphasized in the MDS solution.

The MULTISCALE loss function of Ramsay (1977) can be written as

$$\sigma_{MU}(\mathbf{X}) = \sum_{i<j}\log^2(d_{ij}(\mathbf{X})/\delta_{ij}),$$

showing that the squared logarithm of the relative error is minimized. Provided that the relative error is close to one, $\log(a)$ can be approximated by $a - 1$; that is,

$$\sigma_{MU}(\mathbf{X}) = \sum_{i<j}\log^2(d_{ij}(\mathbf{X})/\delta_{ij}) \approx \sum_{i<j}(1 - d_{ij}(\mathbf{X})/\delta_{ij})^2 = \sigma_{EL}(\mathbf{X}).$$

Thus, the objective of MULTISCALE and elastic scaling scaling coincides in that errors are corrected for the size of the dissimilarities.

The examples above show that choosing w_{ij} as a power of δ_{ij} leads to (approximations) of other loss functions. For this reason, Buja and Swayne

(2002) incorporated the weights $w_{ij} = \delta_{ij}^q$ in their GGVIS software (see Appendix A). Figure 11.1 shows solutions for ratio MDS of the facial expression data of Table 4.4 using several values of q. The middle panels show the standard solution with $q = 0$ and all weights being one as $w_{ij} = \delta_{ij}^0 = 1$. The Shepard diagram in the middle-right panel shows that the size of the errors does not depend on the size of the dissimilarities. Note that the solution for $q = 0$ is the same as Figures 4.8 and 4.9 up to a rotation.

In contrast, for $q = -5$, the large dissimilarities show much error and thus are not well represented. For example, the two worst fitting large dissimilarities are between faces 12 and 13 ("Knows plane will crash" and "Light sleep") and faces 3 and 7 ("Very pleasant surprise" and "Anger at seeing dog beaten"). Both distances are too small in this representation. In this case, the small dissimilarities have little error, and thus can be safely interpreted.

The reverse situation occurs for $q = 5$ where the large dissimilarities are fitted with almost no error and there is quite some error in the representation of the smaller errors. The Shepard plot shows three or four bad-fitting small dissimilarities, which turn out to be connected with face 12. However, face 12 is located so far away because it has several large dissimilarities with other faces (2, 3, 4, 5, 8, 9, and 13) that are all large and represented with almost no error. This compromise is typical for choosing large q. Hence, only large distances can be properly interpreted and small distances should be interpreted with care. If the dissimilarities have some clustering, then choosing a large q may reveal a clearer clustering structure than choosing all $w_{ij} = 1$.

Summarizing, to emphasize the representation of small dissimilarities, choose a large negative q. For a proper representation of the large dissimilarities, choose a large q. If you want to use relative errors to penalize small deviations for small dissimilarities equally heavy as large deviations for large dissimilarities, choose $q = -2$. To measure the error directly without any modification, choose $q = 1$.

Using Weights on Substantive Grounds

All of the above schemes for picking weights w_{ij} had in common that the weights were specified on the basis of general and rather formal considerations. We conclude this discussion about using weights in MDS by pointing out that weights can also be picked on a substantive basis. One particular choice for w_{ij} would be to set it equal to the empirically assessed reliability of the proximity p_{ij}. This means that highly reliable proximities have more impact on the MDS solution than unreliable ones.

The problem, of course, is that reliabilities are seldom collected, because to collect one set of proximities is typically demanding enough. Estimating reliabilities from other information is not that simple either. Consider, for example, the Morse code data in Table 4.2. We may come to the conclusion

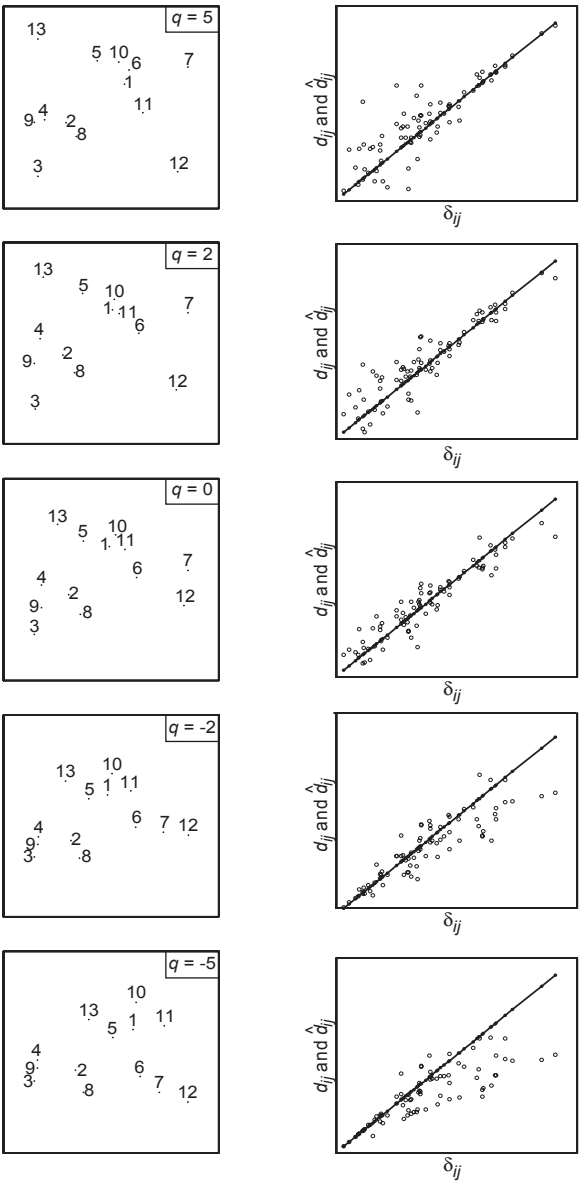

FIGURE 11.1. Ratio MDS of facial expression data of Table 4.4 where $w_{ij} = \delta_{ij}^q$ for q is $-5, -2, 0, 2$, and 5. The left panels show the configurations, the right panels the corresponding Shepard plots.

that these data are essentially symmetric, symmetrize the data, and use the degree of asymmetry as a measure of the unreliability of the confusion probability for each pair. This approach sounds plausible but a closer study of the asymmetries in Chapter 23 reveals that the asymmetries are clearly not just random. Other solutions to obtain reliabilities from the data we have could be considered. For example, one may feel that confusing a signal with itself relates to reliability, and then compute a reliability measure for a pair of signals on the basis of their individual reliabilities. Obviously, many such measures could be considered, and there are many ways to collect reliabilities directly such as, for example, simply replicating the proximity observations at least twice. What is and what is not a good reliability estimate must be decided within the substantive context of the particular data.

Note also that proximities are often data that are collapsed over individuals. This is true too for the Morse code data in Table 4.2. But different individuals can agree on the similarity of some pairs, and disagree on others. This information could also be used to weight the data so that the respondents' common perceptual space relies more on data where interindividual agreement is relatively high.

11.4 Exercises

Exercise 11.1 Compute, by hand, the alienation coefficient for the p_{ij} and d_{ij} in Table 9.2, p. 206.

Exercise 11.2 Consider the data in Table 1.3, p. 10. One may attempt to weight these data somehow to account for possible differences in their reliability. For example, the students who generated these similarity ratings were certainly less familiar with (what was then) "Congo" than with the U.S.A. or the U.K.

(a) Develop a scheme that generates reliability estimates for each proximity in Table 1.3 on the basis of simple ratings of the different nations in terms of their assumed familiarity to the students in this experiment. (Hint: One way of rating the reliability of the proximity p_{ij} is to multiply the familiarity ratings for i and for j.)

(b) Use these estimates to weight the proximities, and redo the (ordinal) MDS with these weights.

(c) Discuss any differences (configuration, Stress, pointwise Stress, interpretation) of the weighted MDS solution and the "unweighted" (or, rather, unit-weights) solution in Figure 1.5.

Exercise 11.3 There are many ways to generate weights δ_{ij} for proximities p_{ij}.

(a) MDS is often used to analyze the structure of correlation matrices (see, e.g., Tables 1.1, 5.1, and 20.1). Discuss some ways to sensibly weight correlations for potentially more robust MDS analyses of such data.

(b) Consider the similarity judgments on facial expressions described in Section 4.3. The respondents may make these judgments with different degrees of confidence. How could this information be collected and incorporated into the MDS analysis?

(c) Even the similarities on the colors in Table 4.1 could be weighted. One possible way is to assume that primary colors (red, blue, green) generate more reliable judgments. Devise a method to generate weights on that basis.

Exercise 11.4 Consider the data in Table 4.1, p. 65. Their Shepard diagram in Figure 4.2 exhibits a slightly nonlinear trend. Find a transformation on the similarities that linearizes the relationship of these data to their MDS distances. Justify this transformation in terms of psychophysics, if possible. Redo the MDS analysis with the rescaled data and a linear MDS model.

Exercise 11.5 Dissimilarities may be related to nonlinear manifolds that are embedded in very high-dimensional space. For example, a constant face that an observer looks at from different angles in space corresponds to different points in the space of its image pixels on the retina. This space has thousands of dimensions, but the points that represent the faces still lie on some nonlinear manifolds (with the angles as parameters) within this space. MDS does not necessarily uncover such manifolds, because of "using greedy optimization techniques that first fit the large-scale (linear) structure of the data, before making small-scale (nonlinear) refinements" (Tenenbaum, 1998, p. 683). One suggestion to solve this problem is to use a "bottom-up" approach that computes distances for points in small local environments only, and then build up large distances by concatenating such distances over geodesics within the manifolds (given that these manifolds are densely packed with points).

(a) Construct a so-called Swiss roll of points in 3D as in the left panel of Figure 11.2. A Swiss roll can be made as follows. Generate two uniformly distributed vectors $\mathbf{u}$ and $\mathbf{v}$ of n points (say, choose $n = 1000$). Then, the coordinates are $x_i = \frac{1}{2} v_i \sin(4\pi v_i)$, $y_i = u_i - \frac{1}{2}$, and $z_i = \frac{1}{2} v_i \cos(4\pi v_i)$.

(b) Compute Euclidean distances for the points in your manifold, and then use metric MDS in an attempt to recover the original Swiss roll configuration.

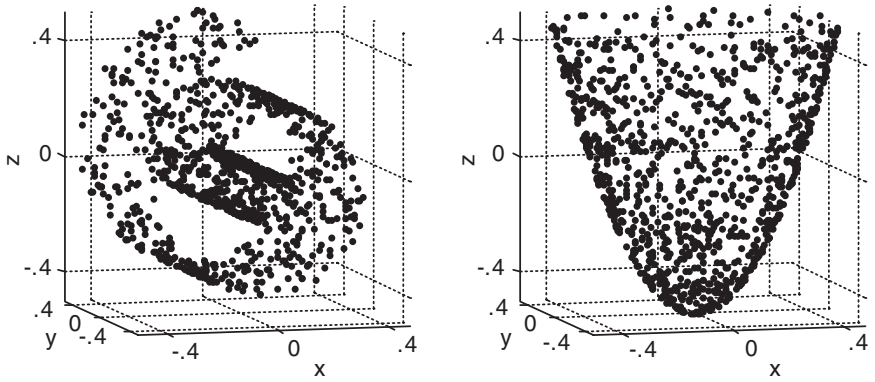

FIGURE 11.2. Manifolds described in Exercise 11.5. The left panel shows the "Swiss roll" manifold, the right panel a "bowl".

(c) Now focus predominantly on small distances by a suitable weighting pattern, and repeat the MDS analyses with small or even zero weights on large distances. Check to what extent this approach manages to unroll the Swiss roll into a plane. [Shepard and Carroll (1966) call this the "intrinsic" dimensionality of the manifold.] Compare the resulting MDS configuration to the one obtained in Exercise (b) above.

(d) Repeat (a) to (c), but now for a "bowl" of points in 3D. A bowl is generated similarly as the Swiss roll in (a), except that $x_i = \frac{1}{2}v_i^{1/2}\cos(2\pi u_i)$, $y_i = \frac{1}{2}v_i^{1/2}\sin(2\pi u_i)$, and $z_i = v_i - \frac{1}{2}$. Can you "flatten" the bowl-like manifold by appropriate weighting into a 2D MDS configuration?