9
Metric and Nonmetric MDS

In the previous chapter, we derived a majorization algorithm for fixed dissimilarities. However, in practical research we often have only rank-order information of the dissimilarities (or proximities), so that transformations that preserve the rank-order of the dissimilarities become admissible. In this chapter, we discuss optimal ways of estimating this and other transformations. One strategy for ordinal MDS is to use monotone regression. A different strategy, rank-images, is not optimal for minimizing Stress, but it has other properties that can be useful in MDS. An attractive group of transformations are spline transformations, which contain ordinal and linear transformations as special cases.

9.1 Allowing for Transformations of the Proximities

So far, we have assumed that the proximities are ratio-scaled values. However, in the social sciences often only the *rank-order* of the proximities is considered meaningful. In such cases, the dissimilarities δ_{ij} are replaced in the Stress function by *disparities*, $\widehat{d}_{ij}$ (d-hats)[1]. Disparities are an admissible transformation of the proximities, chosen in some optimal way. For example, if only the rank-order of the proximities is considered informative,

[1] Other frequently used terminology for disparities is *pseudo distances* (Kruskal, 1977; Heiser, 1990) or *target distances.*

then the disparities must have the same rank-order as the proximities. In this case, we speak of *ordinal* MDS or *nonmetric* MDS. If the disparities are related to the proximities by a specific continuous function, we speak of *metric* MDS. The proximities are in both cases transformed into disparities. In this chapter, we discuss various metric and nonmetric transformations of the proximities, when to use them, and how to calculate them. To simplify the presentation, we assume throughout this chapter that the proximities are dissimilarities, unless stated otherwise.

Stress with d-Hats

Disparities are incorporated in the Stress function as

$$\begin{aligned}
\sigma_r(\widehat{\mathbf{d}}, \mathbf{X}) &= \sum_{i<j} w_{ij}(d_{ij}(\mathbf{X}) - \widehat{d}_{ij})^2 \\
&= \sum_{i<j} w_{ij}\widehat{d}_{ij}^2 + \sum_{i<j} w_{ij}d_{ij}^2(\mathbf{X}) - 2\sum_{i<j} w_{ij}\widehat{d}_{ij}d_{ij}(\mathbf{X}) \\
&= \eta_{\hat{d}}^2 + \eta^2(\mathbf{X}) - 2\rho(\widehat{\mathbf{d}}, \mathbf{X}), \qquad (9.1)
\end{aligned}$$

where $\widehat{\mathbf{d}}$ denotes the $s \times 1$ vector of disparities with $s = n(n-1)/2$. In Section 8.6, we saw how to minimize Stress over the configuration matrix $\mathbf{X}$ by the SMACOF algorithm. We follow De Leeuw (1977), De Leeuw and Heiser (1977), and De Leeuw (1988) in extending this algorithm to include disparities by iteratively alternating an update of $\mathbf{X}$ with an update of $\widehat{\mathbf{d}}$. Clearly, if we optimize over both $\widehat{\mathbf{d}}$ and $\mathbf{X}$, a trivial solution is $\widehat{\mathbf{d}} = \mathbf{0}$ and $\mathbf{X}$=$\mathbf{0}$, which makes (9.1) equal to zero. To avoid this degenerated solution, we norm $\widehat{\mathbf{d}}$ to some fixed length, such as

$$\eta_{\hat{d}}^2 = n(n-1)/2. \qquad (9.2)$$

Metric MDS Models

We now formulate several types or *models* of MDS. In the simplest case (*absolute* MDS), proximities (here dissimilarities) and disparities are related by $p_{ij} = \widehat{d}_{ij}$. Thus,

$$\sigma_r(\widehat{\mathbf{d}}, \mathbf{X}) = \sum_{i<j} w_{ij}(d_{ij}(\mathbf{X}) - \widehat{d}_{ij})^2 = \sum_{i<j} w_{ij}(d_{ij}(\mathbf{X}) - p_{ij})^2, \qquad (9.3)$$

so that each proximity value p_{ij} should correspond exactly to the distance between points i and j in the m-dimensional MDS space.

Absolute MDS is, from an applications point of view, irrelevant, because it is of no interest, for example, to exactly reconstruct from Table 2.1 the European map in its original size. Instead, we settled on *ratio* MDS, where

$\widehat{d}_{ij} = b \cdot p_{ij}$. In this case, the proximities must be dissimilarities. Then, Stress equals

$$\begin{aligned}
\sigma_r(\widehat{\mathbf{d}}, \mathbf{X}) &= \sum_{i<j} w_{ij}(d_{ij}(\mathbf{X}) - \widehat{d}_{ij})^2 \;=\; \sum_{i<j} w_{ij}(d_{ij}(\mathbf{X}) - bp_{ij})^2 \\
&= \sum_{i<j} w_{ij} d_{ij}^2(\mathbf{X}) + b^2 \sum_{i<j} w_{ij} p_{ij}^2 - 2b \sum_{i<j} w_{ij} p_{ij}^2 d_{ij}^2(\mathbf{X}) \\
&= \eta^2(\mathbf{X}) + b^2\eta_p^2 - 2b\rho(\mathbf{X}). \qquad (9.4)
\end{aligned}$$

We see that it is not very difficult to optimize (9.4) over b. Setting the derivative of $\sigma_r(\widehat{\mathbf{d}}, \mathbf{X})$ with respect to b equal to zero yields

$$\begin{aligned}
\frac{\partial \sigma_r(\widehat{\mathbf{d}}, \mathbf{X})}{\partial b} &= 2b\eta_p^2 - 2\rho(\mathbf{X}) = 0, \\
b &= \frac{\rho(\mathbf{X})}{\eta_p^2},
\end{aligned}$$

which gives the update of b for ratio MDS.

It is easy to generate further MDS models from $\widehat{d}_{ij} = f(p_{ij})$ by defining f in different ways. One generalization of ratio MDS is *interval* MDS,

$$\widehat{d}_{ij} = a + b \cdot p_{ij}, \qquad (9.5)$$

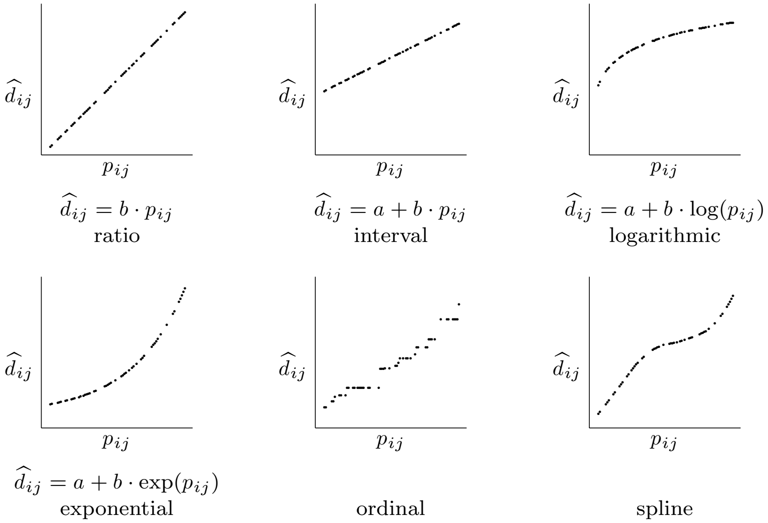
where an additive constant, a, has been added. Ratio and interval MDS are *linear* MDS models, because the $f(p_{ij})$s are linear transformations of the p_{ij}s. This carries certain linear properties of the data into the corresponding distances. If the p_{ij}s are dissimilarities, we require that $b > 0$, because larger dissimilarities should correspond to larger distances. Conversely, if the p_{ij}s represent similarities, then $b < 0$, because a large similarity corresponds to a small distance. In *ratio* MDS, the ratio of any two disparities should be equal to the ratio of the corresponding proximities, because $\widehat{d}_{ij}/\widehat{d}_{kl} = (b \cdot p_{ij})/(b \cdot p_{kl}) = p_{ij}/p_{kl}$. Thus, although it is always possible to assess the ratio of distances in any MDS space and to note that, say, d_{ij} is twice as large as d_{kl}, in ratio MDS such relations should mirror corresponding ratios of the data. In interval MDS, then, the ratio of *differences* ("intervals") of distances should be equal to the corresponding ratio of differences in the data.

Naturally, f does not have to be linear. In principle, we may choose any function we like. However, some functions have been found to be particularly useful in various contexts of psychology. Among them are the logarithmic function

$$\widehat{d}_{ij} = b \cdot \log(p_{ij}), \qquad (9.6)$$

or, more generally,

$$\widehat{d}_{ij} = a + b \cdot \log(p_{ij}),$$



$\widehat{d}_{ij} = b \cdot p_{ij}$
ratio

$\widehat{d}_{ij} = a + b \cdot p_{ij}$
interval

$\widehat{d}_{ij} = a + b \cdot \log(p_{ij})$
logarithmic

$\widehat{d}_{ij} = a + b \cdot \exp(p_{ij})$
exponential

ordinal

spline

FIGURE 9.1. Transformation plot of several transformations.

and the exponential function

$$\widehat{d}_{ij} = a + b \cdot \exp(p_{ij}). \tag{9.7}$$

Sometimes, we might even consider nonmonotonic functions such as a polynomial function of second degree,

$$\widehat{d}_{ij} = a + b \cdot p_{ij} + c \cdot p_{ij}^2. \tag{9.8}$$

There are no limits to the variety of MDS models that can be constructed in this way. These functions can be viewed in a transformation plot, where the horizontal axis is defined by the proximities (p_{ij}) and the vertical axis is defined by the transformed proximities ($\widehat{d}_{ij}$). Some of the transformations discussed so far are graphed in Figure 9.1.

One problem may occur when fitting some of these models (Heiser, 1990). In the step for finding optimal disparities $\widehat{d}_{ij}$, *negative* disparities can occur. For example, this happens in (9.6) when some p_{ij}s are smaller then 1 and some larger than 1, because $\log(x) < 0$ for $0 < x < 1$. More importantly, in interval MDS, model (9.5), negative disparities can and do occur. Because distances can never be negative, a zero residual in the Stress function is unreachable for negative disparities. Moreover, the majorization algorithm may fail to converge because the inequalities that are used to derive (8.23) are reversed for negative disparities, thereby destroying the convergence proof. This problem can be repaired in two ways: first, on top of the restrictions implied by the model, the disparities are restricted to be positive (which makes updating the disparities more complicated), or, second, the Smacof algorithm is extended to deal with negative disparities. For more details on this issue, we refer to Heiser (1990).

TABLE 9.1. Some MDS models ordered by the scale level of the proximities (from strong to weak).

Transformation	$\widehat{d}_{ij}$
Absolute	p_{ij}
Ratio	$b \cdot p_{ij}$ with $b > 0$
Interval	$a + b \cdot p_{ij}$ with $a \geq 0, b \geq 0$
Spline	A sum of polynomials of p_{ij}
Ordinal	Preserve the order of p_{ij}s in $\widehat{d}_{ij}$s

Nonmetric MDS

All of the models from (9.3) to (9.8) are *metric*; that is, they represent various properties of the data related to algebraic operations (addition, subtraction, multiplication, division). In contrast, *nonmetric* models represent only the ordinal properties of the data. For example, if $p_{12} = 5$ and $p_{34} = 2$, an ordinal model reads this only as $p_{12} > p_{34}$ (assuming here that the data are dissimilarities) and constructs the distances d_{12} and d_{34} so that $d_{12} > d_{34}$.

Ordinal models typically require that

$$\text{if } p_{ij} < p_{kl}, \text{ then } \widehat{d}_{ij} \leq \widehat{d}_{kl}, \tag{9.9}$$

and no particular order of the distances for $p_{ij} = p_{kl}$ (weak monotonicity[2] and the primary approach to ties). Notice that the models (9.3) to (9.7) also lead to distances ordered in the same way as the corresponding proximities. But they are all special cases of (9.9), where no particular function f is required for the monotone relation. In Table 9.1 some common MDS models are ordered by the scale level of the proximities.

Even weaker MDS models are conceivable. If, for example, we had proximities coded as a, b, or c, we only may require that there be three classes of distances, one for each data code. All that the distances represent then is the qualitative distinctness, and the model could be called *nominal* MDS, where the disparities are restricted by

$$\text{if } p_{ij} = p_{kl}, \text{ then } \widehat{d}_{ij} = \widehat{d}_{kl},$$

which is implemented in the program ALSCAL. However, we discourage the use of nominal MDS because when interpreting an MDS solution we usually assume that the closer two points are, the more similar the objects they represent. The nominal MDS model thwarts this interpretation. Moreover, it admits transformations that may radically change the appearance of the

[2] Requiring strong monotonicity or $\widehat{d}_{ij} < \widehat{d}_{kl}$ rather than just $\widehat{d}_{ij} \leq \widehat{d}_{kl}$ does not lead to stronger models in practice, because one can always turn an equality into an inequality by adding a very small number ϵ to one side of the equation.

MDS configuration. Finally, strict equality in empirical proximities is often rather exceptional, and, indeed, it is just the case that is *excluded* in the usual ordinal MDS (primary approach to ties) because of its presumed empirical unreliability.

Ad Hoc MDS Models

In addition to such textbook models of MDS, more complicated models are occasionally necessary in real applications. Typically, they involve a function $\widehat{\mathbf{d}} = f(\mathbf{p})$ that is itself a combination of several component functions. Consider, for example, the case of ordinal MDS in (9.9). We may not be satisfied with simply requiring that the data be mapped by "some" monotonic function into distances. We may also want to insist that this function be negatively accelerated, say, because we have a theory about what is going on behind the data. We then have to restrict the $\widehat{d}_{ij}$s to be negatively accelerating. Such additional restrictions on f come from substantive considerations and, therefore, are without limit in their number and variety.

Smacof *with Admissibly Transformed Proximities*

The Smacof algorithm with transformation of the proximities can be summarized by

```
1.  Set $\mathbf{Z} = \mathbf{X}^{[0]}$, where $\mathbf{X}^{[0]}$ is some (non)random start configuration.
    Set iteration counter $k = 0$. Set $\varepsilon$ to a small positive constant.
2.  Find optimal disparities $\widehat{d}_{ij}$ for fixed distances $d_{ij}(\mathbf{X}^{[0]})$.
3.  Standardize $\widehat{d}_{ij}$ so that $\eta^2_{\hat{d}} = n(n-1)/2$.
4.  Compute $\sigma_r^{[0]} = \sigma_r(\widehat{\mathbf{d}}, \mathbf{X}^{[0]})$. Set $\sigma_r^{[-1]} = \sigma_r^{[0]}$.
5.  While $k = 0$ or $(\sigma_r^{[k-1]} - \sigma_r^{[k]} > \varepsilon$ and $k \leq$ maximum iterations) do
6.      Increase iteration counter $k$ by one.
7.      Compute Guttman transform $\mathbf{X}^{[k]}$ by (8.29) if all $w_{ij} = 1$,
        or by (8.28) otherwise, where $\delta_{ij}$ is replaced by $\widehat{d}_{ij}$.
8.      Find optimal disparities $\widehat{d}_{ij}$ for fixed distances $d_{ij}(\mathbf{X}^{[k]})$.
9.      Standardize $\widehat{d}_{ij}$ so that $\eta^2_{\hat{d}} = n(n-1)/2$.
10.     Compute $\sigma_r(\widehat{\mathbf{d}}, \mathbf{X}^{[k]})$.
11.     Set $\mathbf{Z} = \mathbf{X}^{[k]}$.
```

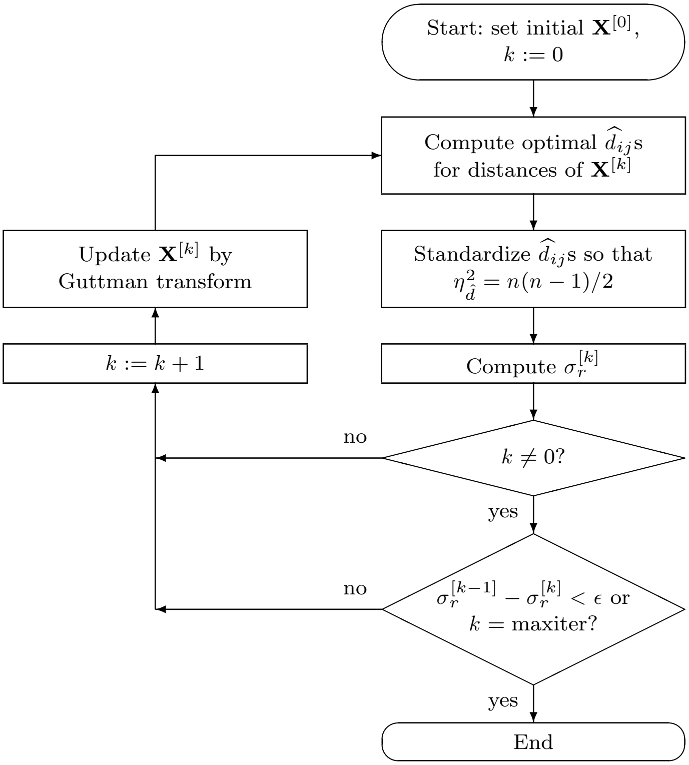

Start: set initial $\mathbf{X}^{[0]}$, $k := 0$
Compute optimal $\widehat{d}_{ij}$s for distances of $\mathbf{X}^{[k]}$
Standardize $\widehat{d}_{ij}$s so that $\eta^2_{\hat{d}} = n(n-1)/2$
Compute $\sigma_r^{[k]}$
$k \neq 0$?
no
yes
$\sigma_r^{[k-1]} - \sigma_r^{[k]} < \epsilon$ or $k =$ maxiter?
no
yes
End
$k := k + 1$
Update $\mathbf{X}^{[k]}$ by Guttman transform

FIGURE 9.2. The flow of the majorization algorithm (SMACOF) for doing MDS with optimal transformations.

12. End while.

A flowchart of this algorithm is presented in Figure 9.2. Note that when computing the Guttman transform the places of the δ_{ij}s are taken by $\widehat{d}_{ij}$s. The allowed transformation of the disparities determines how the update for the disparities in Steps 2 and 7 should be calculated. In the next sections, we discuss the optimal update for ordinal MDS and MDS with splines.

9.2 Monotone Regression

In ordinal MDS, we have to minimize $\sigma_r(\widehat{\mathbf{d}}, \mathbf{X})$ over both $\mathbf{X}$ and $\widehat{\mathbf{d}}$, where the disparities must have the same order as the proximities p_{ij}; that is,

$$\text{if } p_{ij} < p_{kl}, \text{ then } \widehat{d}_{ij} \leq \widehat{d}_{kl} \tag{9.10}$$

if the proximities are dissimilarities, and an inverse order relationship if they are similarities. We switch to Step 7 in the SMACOF algorithm, where

TABLE 9.2. Pairs, ranks, symbolic proximities, numeric proximities (=ranks), numeric distances of starting configuration **X**, symbolic distances, and target distances for **X**.

Pair	Rank	Sym. p_{ij}	p_{ij}	d_{ij}	Sym. d_{ij}	$\widehat{d}_{ij}$
Humphrey–McGovern	1	p_{HM}	1	7.8	d_{HM}	3.38
McGovern–Percy	2	p_{MP}	2	3.2	d_{MP}	3.38
Nixon–Wallace	3	p_{NW}	3	0.8	d_{NW}	3.38
Nixon–Percy	4	p_{NP}	4	1.7	d_{NP}	3.38
Humphrey–Percy	5	p_{HP}	5	9.1	d_{HP}	5.32
Humphrey–Nixon	6	p_{HN}	6	7.9	d_{HN}	5.32
Humphrey–Wallace	7	p_{HW}	7	7.4	d_{HW}	5.32
McGovern–Nixon	8	p_{MW}	8	2.3	d_{MW}	5.32
Percy–Wallace	9	p_{PW}	9	2.3	d_{PW}	5.32
McGovern–Wallace	10	p_{MW}	10	2.9	d_{MW}	5.32

better-fitting $\widehat{d}_{ij}$s with respect to fixed $d_{ij}(\mathbf{X})$ have to be found, subject to the constraints (9.10). Suppose that the order of the $d_{ij}(\mathbf{X})$s is exactly the same as the order of the proximities p_{ij}. Then, simply choosing $\widehat{d}_{ij} = d_{ij}(\mathbf{X})$ defines the optimal update. If the fixed $d_{ij}(\mathbf{X})$s are *not* in the same order as the proximities, the optimal update is found by *monotone regression* of Kruskal (1964b).

The Up-and-Down-Blocks Algorithm

We discuss the solution of minimizing $\sigma_r(\widehat{\mathbf{d}})$ by monotone regression with Kruskal's up-and-down-blocks algorithm. Consider an example. Rabinowitz (1975) describes a hypothetical experiment where a subject was asked to rank-order all possible pairs of the following politicians from most to least similar: Humphrey (H), McGovern (M), Percy (P), Nixon (N), and Wallace (W). The subject generated the ranking numbers exhibited in the second column of Table 9.2. They are shown in the form of the familiar proximity matrix in Table 9.3.

Now, assume that we have a first configuration **X**, which leads to the distances in Table 9.2. How are the $\widehat{d}_{ij}$s computed? Consider the distances d_{ij} for the pairs Humphrey–McGovern and McGovern–Percy, d_{HM} and d_{MP}. The corresponding proximities are ordered as $p_{HM} < p_{MP}$. Because the proximities are dissimilarities (i.e., the smaller the p-value, the larger the similarity), $d_{HM} \leq d_{MP}$ should hold in a perfect MDS representation. This is obviously not true for the configuration **X**, because it yields $d_{HM} = 7.8$ and $d_{MP} = 3.2$. Thus, the points of **X** must be moved so that d_{HM} becomes smaller and d_{MP} larger. Now, given two numbers, the arithmetical mean yields the number that is closest to both of them in the least-squares sense. Thus, setting $(d_{HM} + d_{MP})/2 = \widehat{d}_{HM} = \widehat{d}_{MP}$ defines target values that satisfy the requirements.

TABLE 9.3. Proximity matrix for politicians.

	H	M	P	W	N
H	–	1	5	7	6
M	1	–	2	10	8
P	5	2	–	9	4
W	7	10	9	–	3
N	6	8	4	3	–

TABLE 9.4. Derivation of the disparities in Table 9.2 by monotone regression.

			Trial Solutions for $\widehat{d}_{ij}$							Final
Pair	p_{ij}	$\widehat{d}_{ij}$	I	II	III	IV	V	VI	VII	$\widehat{d}_{ij}$
Humphrey–McGovern	1	7.8	5.5	3.93	3.38	3.38	3.38	3.38	3.38	3.38
McGovern–Percy	2	3.2	5.5	3.93	3.38	3.38	3.38	3.38	3.38	3.38
Nixon–Wallace	3	0.8	0.8	3.93	3.38	3.38	3.38	3.38	3.38	3.38
Nixon–Percy	4	1.7	1.7	1.7	3.38	3.38	3.38	3.38	3.38	3.38
Humphrey–Percy	5	9.1	9.1	9.1	9.1	8.5	8.13	6.68	5.8	5.32
Humphrey–Nixon	6	7.9	7.9	7.9	7.9	8.5	8.13	6.68	5.8	5.32
Humphrey–Wallace	7	7.4	7.4	7.4	7.4	7.4	8.13	6.68	5.8	5.32
McGovern–Nixon	8	2.3	2.3	2.3	2.3	2.3	2.3	6.68	5.8	5.32
Percy–Wallace	9	2.3	2.3	2.3	2.3	2.3	2.3	2.3	5.8	5.32
McGovern–Wallace	10	2.9	2.9	2.9	2.9	2.9	2.9	2.9	2.9	5.32

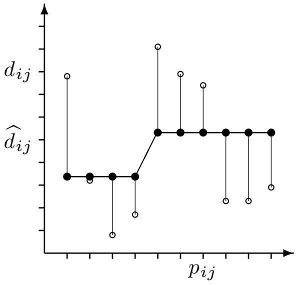


FIGURE 9.3. Shepard diagram of monotone regression as calculated in Tables 9.2 and 9.4. The open points represent pairs of corresponding proximities and distances (p_{ij}, d_{ij}), the solid points disparities $\widehat{d}_{ij}$. The solid line is the best-fitting monotone regression curve.

Beginning with the first pair of distances in Table 9.4, we get a first trial solution for the disparities by setting $5.5 = \widehat{d}_{HM} = \widehat{d}_{MP}$ and $\widehat{d}_{ij} = d_{ij}$ for all remaining distances. This yields the values in column I of Table 9.4. This trial solution, however, satisfies the monotonicity requirement only for its first two elements, and the third disparity value is too small, because $d_{NW} = 0.8$ is smaller than both of the preceding values. So, we create a new block by computing the average of the first three distances (5.5 + 5.5 + 0.8)/3 = 3.93. We then use 3.93 for $\widehat{d}_{HM}, \widehat{d}_{MP}$, and $\widehat{d}_{NW}$, and again hope that everything else is in order, thus setting $\widehat{d}_{ij} = d_{ij}$ for all other distances. This yields the second trial solution for the disparities (column II). This sequence still violates the monotonicity requirement in row 4. Hence, a new block is formed by joining the previous block and d_{NP}. The resulting disparities in column III form a weakly monotonic sequence up to and including row 5. In row 6, a value 7.9 turns up, however, that is smaller than the preceding one, 9.1. So, we join 9.1, 7.9, and all preceding values into one block, average these values, and so on. Table 9.4 shows all of the steps leading to the final disparity sequence of $\widehat{d}_{ij}$s in the last column. This completes monotone regression for the first iteration.

A Shepard diagram is given in Figure 9.3. In the main algorithm, we then have to normalize the $\widehat{d}_{ij}$ such that their sum-of-squares is equal to $n(n-1)/2$. Then, we start the second iteration by computing an update for the configuration $\mathbf{X}$. This gives new distances for which we can compute new disparities by monotone regression, as we have done above.

Smoothed Monotone Regression

A more restrictive version of ordinal MDS is *smoothed* monotone regression (Heiser, 1985, 1989a). Apart from the order restrictions implied by ordinal MDS, we also impose the restriction that the difference between differences of adjacent disparities is never larger than the average disparity. Thus, if the $s = n(n-1)/2$ elements of vector $\widehat{\mathbf{d}}$ are ordered as the proximities, then smoothed monotone regression requires

$$\begin{array}{rcll} |(\widehat{d}_k - 0) - (0 - 0)| & \leq & s^{-1}\sum_{l=1}^{s}\widehat{d}_l, & \text{for } k = 1, \\ |(\widehat{d}_k - \widehat{d}_{k-1}) - (\widehat{d}_{k-1} - 0)| & \leq & s^{-1}\sum_{l=1}^{s}\widehat{d}_l, & \text{for } k = 2, \\ |(\widehat{d}_k - \widehat{d}_{k-1}) - (\widehat{d}_{k-1} - \widehat{d}_{k-2})| & \leq & s^{-1}\sum_{l=1}^{s}\widehat{d}_l, & \text{for } k = 3, \ldots, s. \end{array} \tag{9.11}$$

Thus, the restrictions are imposed on the difference of subsequent differences. The advantage of this internally bounded form of monotone regression is that the steps between two adjacent disparities can never get large. Therefore, the Shepard diagram always shows a smooth relation of p_{ij}s and $\widehat{d}_{ij}$s without irregular steps in the curve. For $k = 1$, the first restriction of (9.11) implies that $\widehat{d}_k$ should be between zero and the average d-hat. Therefore, a smoothed monotone transformation has a first d-hat that is quite close to zero and will be increasing in a smooth way. It can be verified that a quadratically increasing transformation and a logarithmically increasing transformation satisfy the maximal stepsizes as defined in (9.11). Unfortunately, Heiser reports that it is not easy to compute optimal disparities for given distances using smoothed monotone regression. Also, the smoothed monotone regression problem tends to become computationally demanding if n is large (say $n > 25$).

9.3 The Geometry of Monotone Regression

In the previous section, we saw how monotone regression is performed. Here, we give a geometrical explanation of monotone regression. Consider an example. Suppose that we have

$$\mathbf{P} = \begin{bmatrix} - & 1 & 3 \\ 1 & - & 2 \\ 3 & 2 & - \end{bmatrix}, \text{ and } \mathbf{D}(\mathbf{X}) = \begin{bmatrix} - & 1 & 2 \\ 1 & - & 3 \\ 2 & 3 & - \end{bmatrix},$$

and $w_{ij} = 1$ for each pair i, j. Let us reformulate the problem in simpler notation. Denote the unknown $\widehat{d}_{ij}$ as x_l, and the known $d_{ij}(\mathbf{X})$ as a_l. Also, we order the a_l in the rank-order of the proximities. This leads to Table 9.5. Rewriting σ_r accordingly, the monotone regression problem becomes minimizing

$$\sigma_r(\mathbf{x}) = \sum_{l=1}^{s}(x_l - a_l)^2$$

under the restriction that $0 \leq x_1 \leq x_2 \leq \cdots \leq x_s$, where $s = n(n-1)/2$.

TABLE 9.5. Reformulation of the monotone regression problem.

s	Pair i, j	Proximity	$a_s = d_{ij}(\mathbf{X})$	$x_s = \widehat{d}_{ij}$
1	12	p_{12}	$a_1 = 1$	$x_1 = \widehat{d}_{12}$
2	23	p_{23}	$a_2 = 3$	$x_2 = \widehat{d}_{23}$
3	13	p_{13}	$a_3 = 2$	$x_3 = \widehat{d}_{13}$

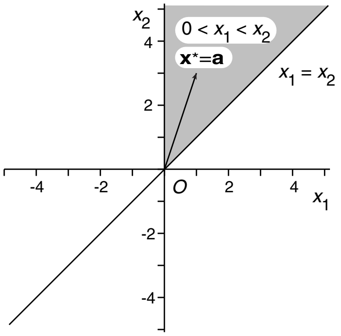


FIGURE 9.4. The area for which $0 \le x_1 \le x_2$.

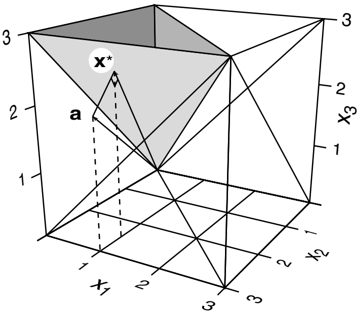


FIGURE 9.5. The area for which $0 \le x_1 \le x_2 \le x_3$.

To see what these restrictions imply geometrically, consider the case where we have the restrictions $0 \le x_1 \le x_2$. The shaded area in Figure 9.4 shows the area in which these inequalities hold. Here, each axis denotes one of the variables x_l. The first part of the inequalities implies that all x_l should be nonnegative, because we do not want the disparities to become negative. The elements $a_1 = 1$ and $a_2 = 3$ fall in the shaded area, so that choosing $x_1^* = a_1 = 1$ and $x_2^* = a_2 = 3$ gives $\sigma_r(\mathbf{x})$ where the order restriction on x_1 and x_2 is not violated. If $\mathbf{a}$ were outside the shaded area, then we would have to find an $\mathbf{x}$ on the border of the shaded area that is closest to $\mathbf{a}$ by the up-and-down-blocks algorithm. The triple of inequalities $0 \le x_1 \le x_2 \le x_3$ of our simple example can be represented graphically as in Figure 9.5. After orthogonal projection on each pair of axes, the area in which the inequalities hold is similar to that of Figure 9.4. The three inequalities combined give the inner part of the *ordered cone* in Figure 9.5. Monotone regression amounts to projecting $\mathbf{a}$ onto this cone. If $\mathbf{a}$ is ordered with increasing values, then it is located inside the cone. In this example, the $\mathbf{x}$ with the shortest distance to $\mathbf{a}$ that is in or on the ordered cone equals $\mathbf{x}^* = [1,\ 2.5,\ 2.5]'$.

Geometrically, monotone regression amounts to finding the $\widehat{\mathbf{d}}$ that is in the ordered cone (defined by the proximities) and as close as possible to the vector of distances.

TABLE 9.6. Calculation of the primary and the secondary approaches to ties in ordinal MDS for given distances d_{ij}.

			Primary Approach				Secondary Approach		
Pair	p_{ij}	d_{ij}	I	II	III	$\widehat{d}_{ij}$	I	II	$\widehat{d}_{ij}$
3,2	1	3	3	2.50	2.50	2.50	3	2.50	2.50
4,1	2	2	2	2.50	2.50	2.50	2	2.50	2.50
3,1	3	6	6	6	4.50	4.50	6	6	4.66
4,2	4	5	3	3	4.50	5	5	4	4.66
2,1	4	3	5	5	5	4.50	3	4	4.66
4,3	5	7	7	7	7	7	7	7	7

9.4 Tied Data in Ordinal MDS

In ordinal MDS, the relevant data information is the rank-order of the proximities. But consider the rank-order of the proximities in the following matrix.

$$\mathbf{P} = \begin{bmatrix} - & 4 & 3 & 2 \\ 4 & - & 1 & 4 \\ 3 & 1 & - & 5 \\ 2 & 4 & 5 & - \end{bmatrix}.$$

We see that the proximities p_{21} and p_{42} have the same ranks; that is, they are *tied*. How should such ties be represented in an MDS configuration? It would seem natural to represent them by equal distances in an MDS solution, but this is known as the *secondary approach to ties*. For our simple example, it means that $\widehat{d}_{21} = \widehat{d}_{42}$, so that $\widehat{d}_{31} \leq \widehat{d}_{21} = \widehat{d}_{42} \leq \widehat{d}_{43}$. In the *primary approach*, tied proximities impose *no* restrictions on the corresponding distances. In other words, it is not necessary to map tied data into equal distances. For our example, the primary approach to ties implies $\widehat{d}_{31} \leq \widehat{d}_{21} \leq \widehat{d}_{43}$ and $\widehat{d}_{31} \leq \widehat{d}_{42} \leq \widehat{d}_{43}$. Nothing is required of the distances representing equal proximities, except that they must be smaller (larger) than the distances corresponding to smaller (larger) proximities. Ties in the data thus can be *broken* in the representing distances.

In Table 9.6, an example is presented of the calculation for the primary and secondary approaches to ties for given distances. The resulting Shepard diagrams are shown in Figure 9.6. In the primary approach, the first estimate of the disparities is obtained by setting $\widehat{d}_{ij} = d_{ij}$ and then reordering these $\widehat{d}_{ij}$ wherever they correspond to tied p_{ij} values so that they increase monotonically. Then, standard monotone regression is applied (see Section 9.2). Finally, the resulting disparities are permuted back into the original order of the distances. The secondary approach to ties follows the same strategy as monotone regression, except that the first disparity estimates for tied data are replaced by their average values.

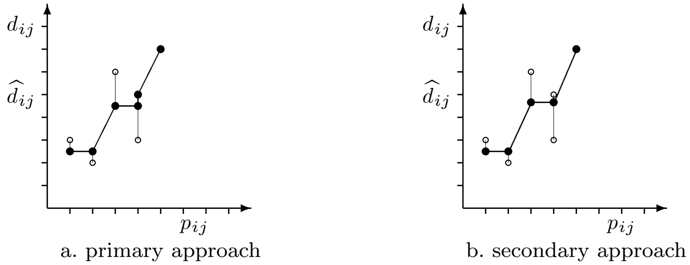


FIGURE 9.6. Shepard diagram of monotone regression calculated with the primary approach to ties (a), or the secondary approach to ties (b) (see Table 9.6). The solid points are the disparities $\widehat{d}_{ij}$, and the open points are the distances $d_{ij}(\mathbf{X})$.

How do ties arise in proximity data? There are several possibilities. Consider the following case. Assume that we want to find out how people perceive cars by studying their similarity impressions. A stack of cards is prepared, where each card shows one pair of cars. The subject is asked to split the stack into two piles, one containing the more similar pairs, the other the more dissimilar pairs. Then, the subject is asked to repeat this exercise for each pile in turn, and repeat again, and so on, until he or she feels that it is not possible to discriminate any further between the pairs of cars in any pile. If the subject stops when each pile has only one card left, then we get a complete similarity order of the pairs of cars and no ties occur. It is more likely, however, that some of the final piles will have more than one card. Most likely, the piles for the extremely similar pairs will be quite small, whereas those for pairs with intermediate similarity will be larger. This means that if we assign the same proximity value to all pairs in a pile, ties will arise for every pile containing one or more cards. However, we would not want to assume that these data are tied because the subject feels that the respective pairs of cars are exactly equal. Rather, the subject stops the card sorting only because the pairs in some piles do not appear to be sufficiently different for a further meaningful or reliable ordering. Hence, the primary approach to ties should be chosen in analyzing these data.

Consider another example, a pilot study on the perception of nations (Wish, Deutsch, & Biener, 1970; Wish, 1971), where the respondents had to judge the degree of similarity between each pair of 12 nations on a 9-point rating scale with endpoints labeled as “very different” and “very similar”, respectively. Here, the proximities for each respondent must have ties, because there are 66 pairs of nations, and, thus, it would require a rating scale with at least 66 categories in order to be able to assign a different proximity value to every stimulus pair. The 9-point rating scale works as a relatively coarse sieve on the true similarities, so the data would be best interpreted as indicators for intervals on a continuum of similarity.

The primary approach to ties is again indicated, inasmuch ties must result due to the data collection method.

A further way for ties to occur is when the proximities are derived from other data. Consider the correlation matrix of intelligence tests in Table 5.1. Several ties occur here, so that with the primary approach to ties, the distances d_{17}, d_{24}, and d_{27}, say, are merely required to be less than d_{37} and greater than d_{26}. However, we can compute the correlation coefficients to more decimal places. Assume that we get, by using three decimal places, $r_{17} = .261$, $r_{24} = .263$, and $r_{27} = .259$. In ordinal MDS, it should then hold that $d_{24} < d_{17} < d_{27}$, and so the MDS solution must satisfy additional properties. But is it worthwhile to place such stronger demands on the solution? Clearly not. The correlations may not even be reliable to three decimal places. Even the value of $r = .26$ should be read as $r \approx .26$. Hence, the secondary approach to ties makes no sense here.

9.5 Rank-Images

A completely different way of computing disparities in *ordinal* MDS is based on *rank-images.* The basic idea is that if a perfect fit exists in ordinal MDS, then the rank-order of the distances must be equal to the rank-order of the proximities. To compute the disparities, a switch is made to a loss function that is different from Stress; that is,

$$\tau(\widehat{\mathbf{d}}) = (\mathbf{R}_p\widehat{\mathbf{d}} - \mathbf{R}_d\mathbf{d})'(\mathbf{R}_p\widehat{\mathbf{d}} - \mathbf{R}_d\mathbf{d}), \tag{9.12}$$

where we assume for simplicity that all the weights w_{ij} are one in the Stress function. $\mathbf{R}_p$ is a permutation matrix (that has only a single one in each row and column, and zeros elsewhere) such that $\mathbf{R}_p\mathbf{p}$ is the vector of proximities ordered from small to large. Similarly, $\mathbf{R}_d$ is a permutation matrix that orders the distances $\mathbf{d}$ from small to large. $\mathbf{R}_p$ is known, the vector of distances $\mathbf{d}$ is known, and thus $\mathbf{R}_d$ is known. The only unknown vector is the vector of disparities $\widehat{\mathbf{d}}$ that we intend to find. To find the minimum of (9.12) we use the fact that $\mathbf{R}'\mathbf{R} = \mathbf{I}$ for any permutation matrix $\mathbf{R}$. Equation (9.12) is a quadratic function in $\widehat{\mathbf{d}}$, so that its minimum can be found in one step by setting the gradient (first derivative)

$$\nabla\tau(\widehat{\mathbf{d}}) = 2\mathbf{R}_p'\mathbf{R}_p\widehat{\mathbf{d}} - 2\mathbf{R}_p'\mathbf{R}_d\mathbf{d} = 2\widehat{\mathbf{d}} - 2\mathbf{R}_p'\mathbf{R}_d\mathbf{d}$$

equal to zero for all elements: $\nabla\tau(\widehat{\mathbf{d}}) = \mathbf{0}$ implies $\widehat{\mathbf{d}} = \mathbf{R}_p'\mathbf{R}_d\mathbf{d}$ (and $\mathbf{R}_p\widehat{\mathbf{d}} = \mathbf{R}_d\mathbf{d}$). If the proximities are already ordered increasingly, then $\mathbf{R}_p = \mathbf{I}$ and the rank-image transformation amounts to setting the disparities equal to the ordered distances.

A flaw of using rank-images for ordinal MDS is that convergence of the overall algorithm cannot be guaranteed. This is caused by the switch from

TABLE 9.7. Derivation of rank-image disparities from the politicians data given in Tables 9.2 and 9.3.

Pair	$\mathbf{R}_p\mathbf{p}$	$\mathbf{R}_p\mathbf{d}$	$\mathbf{R}_d\mathbf{d}$	$\mathbf{R}_p\widehat{\mathbf{d}}$
Humphrey–McGovern	1	7.8	0.8	0.8
McGovern–Percy	2	3.2	1.7	1.7
Nixon–Wallace	3	0.8	2.3	2.3
Nixon–Percy	4	1.7	2.3	2.3
Humphrey–Percy	5	9.1	2.9	2.9
Humphrey–Nixon	6	7.9	3.2	3.2
Humphrey–Wallace	7	7.4	7.4	7.4
McGovern–Nixon	8	2.3	7.8	7.8
Percy–Wallace	9	2.3	7.9	7.9
McGovern–Wallace	10	2.9	9.1	9.1

the Stress loss function to the loss function (9.12). This could be solved by trying to minimize the same function (9.12) for updating the configuration $\mathbf{X}$. However, because $\mathbf{R}_d$ is dependent on the distances and thus on $\mathbf{X}$, it is very hard to minimize (9.12) over $\mathbf{X}$. Nevertheless, we can still use rank-images in the SMACOF algorithm, although convergence is no longer guaranteed. As De Leeuw and Heiser (1977) remark: "It is, of course, perfectly legitimate to use the rank-images ... in the earlier iterations (this may speed up the process, cf. Lingoes & Roskam, 1973). As long as one switches to [monotone regression] in the final iterations convergence will be achieved" (p. 742). Lingoes and Roskam (1973) do exactly this in their MINISSA program, because they claim that "the rank-image transformation is more robust against trivial solutions and local minima" (Roskam, 1979a, p. 332).

As an example of the calculation of rank-images, we again use the data on the similarity of politicians from Table 9.2. The proximities are already ordered from small to large, so that $\mathbf{R}_p = \mathbf{I}$. The disparities according to the rank-image transformation are given in Table 9.7.

In Guttman (1968) and in some computer programs, rank-images are denoted by d^*_{ij} (d-star) as opposed to $\widehat{d}_{ij}$ (d-hat) obtained by monotone regression. Here, we retain the notation of $\widehat{d}_{ij}$ for a disparity, even if the disparity is a rank-image.

9.6 Monotone Splines

Quite flexible transformations are obtained by using splines. We show that special cases of (monotone) splines include interval transformations, polynomial transformations, and ordinal transformations. In this section, we limit ourselves to the class of monotone splines, which are also called *I-splines* (integrated splines) in the literature. Whenever we refer to a spline in the sequel, we mean a monotone spline. One of its main characteristics is that the resulting transformation is smooth. For a good review of applications of monotone splines in statistics, we refer to Ramsay (1988). For more general references on splines, see De Boor (1978) and Schumaker (1981).

There are three reasons for wanting a smooth transformation in MDS. First, ordinal MDS can result in a crude transformation. For example, in Figure 9.3 the rank-order of ten different proximities was transformed in only two different disparities. Such crude transformations neglect much of the variation in the proximities. A second reason is that we want to retain more than ordinal information of the data. For example, if the proximities are correlations, we may want to consider more than just the rank-orders of the correlations (as in ordinal MDS), but less than the interval information (as in interval MDS). Third, degenerate solutions (see Chapter 13) can be avoided by imposing smooth transformations. In general, a spline transformation yields a much smoother transformation curve than an ordinal transformation. Compare, for example, the nonsmooth ordinal transformation in Figure 9.1 and the smooth spline transformation in the same figure. Thus, splines can be used to obtain smooth transformation curves, while keeping the ordinal information of the proximities intact.

Characterization of Monotone Splines

What does a spline transformation look like? In general, the transformation is a smooth monotone increasing curve. The conceptual idea is that it is not possible to map *all* proximities into disparities by one simple function (such as the linear transformation in interval MDS). Then, splines can be used to specify such simple mappings for several intervals. The additional restriction on the separate transformation of each interval is that they should be smoothly connected and monotone increasing. We discuss later that interval and ordinal transformations are two extreme cases of monotone spline transformations. Hence, other spline mappings can be seen as more restrictive than ordinal mappings and less restrictive than linear mappings.

The endpoints (extrema) of the intervals are called *knots*. Because splines are required to be smooth, the endpoint of one interval coincides with an extremal point of the adjacent interval, so that a knot ties together the two intervals. The size of the intervals is characterized by the *knot* sequence of the knots t_i. As before in this chapter, we string out the $s = n(n-1)/2$ proximities in the vector $\mathbf{p}$ and index its elements by i, where $i = 1, 2, \ldots, s$. We also assume that the elements in $\mathbf{p}$ are ordered increasingly. Two knots are reserved, one for the smallest value of the proximities $t_0 = p_{\min}$ and the other for the largest value $t_m = p_{\max}$. The other knots, if present, are called *interior knots*, because they must be greater than t_0 and smaller than t_m. Thus, the ordered knot sequence of the m knots $t_0 = p_{\min}, t_1, t_2, \ldots, t_m = p_{\max}$ defines the intervals $[t_0, t_1]$, $[t_1, t_2]$, $\ldots$, $[t_{m-1}, t_m]$, so that every observed value p_{ij} falls into one of these intervals.

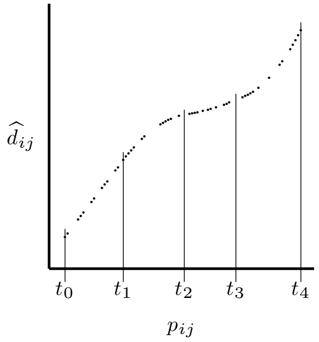


FIGURE 9.7. Example of a spline transformation with three interior knots.

Usually, the interior knots are placed at $K-1$ quantiles so that the K intervals are equally filled with proximities. Figure 9.7 shows an example of a spline transformation with three interior knots that define four intervals.

The smoothness within an interval is guaranteed by choosing the transformation as a polynomial function of the proximities. Examples of polynomial functions are $f(p) = 3p^2 - 2p + 1$ (a second degree polynomial), and $f(p) = 6p - 3$ (a first degree polynomial). In general, a polynomial function of degree r is defined as $f(p) = \sum_{k=0}^{r} a_k p^k$, where a_k are weights and $p^0 = 1$. The degree r of the polynomial is specified by the *order* of the spline, or the *degree* of the spline. Because the entire spline transformation must be smooth, we must also have smoothness between the intervals at the knots. The smoothness at the knots is also determined by the order of the spline in the following way: at knot t_i, the first $r-1$ derivatives of two polynomials of the adjacent intervals $[t_{i-1}, t_i]$ and $[t_i, t_{i+1}]$ must be equal. For a spline of order 1, this property implies that the lines are joined at each interior knot, so that the transformation is continuous. A quadratic spline has—apart from continuity—equal first derivatives at each interior knot. A third-order spline has continuity up to the second derivatives at the interior knots, and so on. Note that a spline of order 0 is not even continuous.

It remains to be seen how a spline transformation can be computed. Suppose that we specify a spline of degree r with k interior knots. It turns out that the spline transformation (with the properties outlined above) can be computed by using a special $s \times (r+k)$ matrix $\mathbf{M}$ that can be derived from $\mathbf{p}$. The spline transformation is defined simply as $\widehat{\mathbf{d}} = \mathbf{Mb}$ for any vector of nonnegative weights $\mathbf{b}$. Viewed this way, finding a spline transformation is nothing more than solving a multiple regression problem for optimal weights $\mathbf{b}$. These weights are used to predict the fixed distances $\mathbf{d}$ by the weighted sum $\widehat{\mathbf{d}} = \mathbf{Mb}$. We restrict $\mathbf{b}$ to be nonnegative, which ensures that the transformation is monotone increasing.

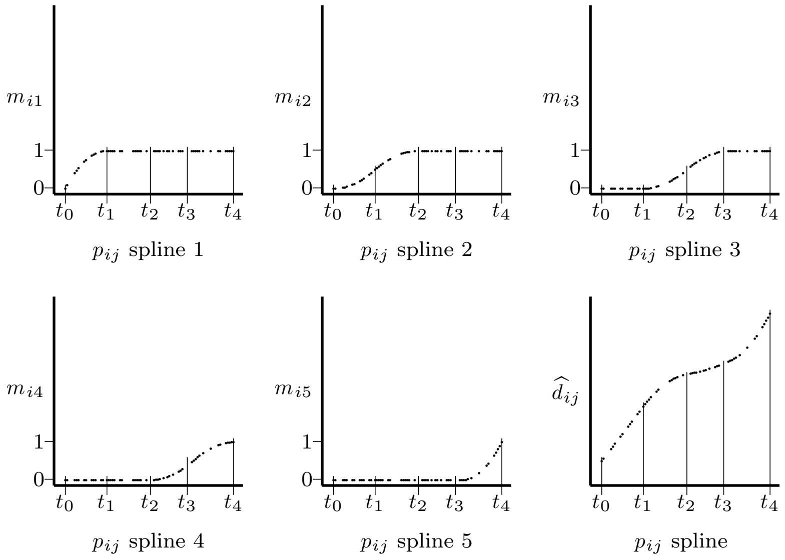


FIGURE 9.8. Separate columns of spline bases **M** belonging to a monotone spline of order 2 with three interior knots. The plot on the bottom right is a transformation resulting from a weighted sum of the previous five columns of **M** plus an intercept.

Specifying the Matrix M

The crux of a spline transformation lies in how the matrix **M** is set up. It turns out that for monotone splines each column is a piecewise polynomial function of **p**, in such a way that any linear combination satisfies the required smoothness restrictions at the knots. To accomplish this, matrix **M** has a special form. The elements of column j of **M** in the first $\max(0, j-r)$ intervals are equal to 0, and the elements in the last $\max(0, k-j+1)$ intervals are equal to 1. The remaining intervals contain a special polynomial function of degree r, which we specify below for splines of orders zero, one, and two. Figure 9.8 shows an example of the columns of **M** as a function of **p** for $k = 3$ and $r = 2$. The first column $\mathbf{m}_1$ has elements equal to 1 in the last three intervals, the second column $\mathbf{m}_2$ has elements 1 in the last two intervals, the third column $\mathbf{m}_3$ has 0s in the first interval and 1s in the last interval, the fourth column $\mathbf{m}_4$ has 0s in the first and second intervals, and the fifth and final column has 0s in the first, second, and third intervals. The values in the intervals that are not 0 or 1 are a quadratic function in p_{ij} that is continuous and has equal derivatives at the knots.

We now come to explicit expressions for splines of orders zero, one, and two. The columns of **M** for an order-zero spline are defined by an indicator function that is 0 if p_i is smaller than knot j and 1 otherwise; that is, the spline basis $\mathbf{M}$ is an $s \times k$ matrix with elements

$$m_{ij} = \begin{cases} 0 & \text{if } t_0 \leq p_i < t_j, \\ 1 & \text{if } t_j \leq p_i < t_{k+1}. \end{cases}$$

If the number of interior knots k is 0, then $\mathbf{M}$ is not defined in a zero-order spline, because all values p_i fall in the same interval $[t_0, t_1]$, so that $m_i = 1$ for all i. Clearly, for our purpose, the transformation $\widehat{d}_{ij} = 1$ for all i, j is not acceptable, because it ignores the variability in the observed proximities.

The columns of $\mathbf{M}$ of an order-one spline are defined by a piecewise linear function; that is,

$$m_{ij} = \begin{cases} 0 & \text{if} \quad t_0 \leq p_i < t_{j-1}, \\ \dfrac{p_i - t_{j-1}}{t_j - t_{j-1}} & \text{if} \quad t_{j-1} \leq p_i < t_j, \\ 1 & \text{if} \quad t_j \leq p_i \leq t_{k+1}. \end{cases}$$

For a monotone spline of order two, we can write a direct formulation of the elements of $\mathbf{M}$; that is,

$$m_{ij} = \begin{cases} 0 & \text{if} \quad t_0 \leq p_i < t_{j-2}, \\ \dfrac{(t_{j-2} - p_i)^2}{(t_{j-1} - t_{j-2})(t_j - t_{j-2})} & \text{if} \quad t_{j-2} \leq p_i < t_{j-1}, \\ 1 - \dfrac{(t_j - p_i)^2}{(t_j - t_{j-1})(t_j - t_{j-2})} & \text{if} \quad t_{j-1} \leq p_i < t_j, \\ 1 & \text{if} \quad t_j \leq p_i < t_{k+1}, \end{cases}$$

after Ramsay (1988). Note that for $j = 1$ we have reference to $t_{j-2} = t_{-1}$, which we define as $t_{-1} = t_0$. Equivalently, for $j = k + 1$ we define knot $t_j = t_{k+1}$. The lower-right plot in Figure 9.8 plots the proximities against $\widehat{\mathbf{d}} = \mathbf{Mb}$ for some given vector $\mathbf{b}$. For the calculation of monotone splines of higher order and for more general information on splines, we refer to Ramsay (1988) and De Boor (1978).

Let the proximity matrix $\mathbf{P}$ be given by

$$\mathbf{P} = \begin{bmatrix} 0 & 1.0 & 1.5 & 3.2 \\ 1.0 & 0 & 2.0 & 3.8 \\ 1.5 & 2.0 & 0 & 4.5 \\ 3.2 & 3.8 & 4.5 & 0 \end{bmatrix},$$

or in vector notation $\mathbf{p}' = (1.0, 1.5, 2.0, 3.2, 3.8, 4.5)$. Let the knots be given by $t_0 = 1.0$, $t_1 = 3.0$, $t_2 = 4.5$, so that the number of interior knots k equals 1. For these data Table 9.8 shows $\mathbf{M}$ for a zero-order spline, a first-order spline, and a second-order spline.

The spline basis $\mathbf{M}$ of $\mathbf{p}$ is invariant under linear transformation of $\mathbf{p}$. It turns out that by choosing the two extrema as knots, we obtain a row

TABLE 9.8. Example of spline bases **M** for a zero-order spline, a first-order spline, and a second-order spline. The knots are $t_0 = 1.0$, $t_1 = 3.0$, $t_2 = 4.5$.

	$r = 0$	$r = 1$		$r = 2$		
$\mathbf{p}$	$\mathbf{m}_1$	$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$
1.0	0	0.00	0.00	0.00	0.00	0.00
1.5	0	0.25	0.00	0.44	0.04	0.00
2.0	0	0.50	0.00	0.75	0.14	0.00
3.2	1	1.00	0.13	1.00	0.68	0.02
3.8	1	1.00	0.53	1.00	0.91	0.28
4.5	1	1.00	1.00	1.00	1.00	1.00

of 0s for the smallest proximity and a row of 1s for the largest proximity, as can be verified in the examples of Table 9.8. This implies that whatever the weights **b**, the disparity of the smallest proximity will be 0. This is not desirable in MDS, because the smallest proximity does not necessarily have to be represented by a zero distance. Therefore, we include a positive intercept in our spline transformation; that is, $\widehat{\mathbf{d}} = b_0\mathbf{1} + \mathbf{Mb}$. For MDS, we need the intercept, so that the disparity corresponding to the smallest proximity can be transformed into any nonnegative value.

Special Cases of Monotone Splines

Let us look at two special cases of monotone splines. The first case is a spline with order larger than zero ($r > 0$) and no interior knots ($k = 0$), so that there are only two knots, one at the smallest value of **p** and one at the largest value of **p**. For this case, monotone splines have the property that the row sum of **M** is equal to $c\mathbf{p}$ (with $c > 0$ an arbitrary factor); that is, $cp_i = \sum_j m_{ij}$. An example of a transformation plot for this case is given in Figure 9.9a. A second property is that such spline transformations are equivalent to transformations obtained by polynomial regression of the same degree. If we deal with a first-order spline, then **M** consists of one column only that is linearly related to p. Therefore, a first-order spline with two knots and an intercept is equivalent to an interval transformation, as can be seen in Figure 9.9b.

The second special case of a monotone spline occurs if exactly $k = n - 1$ interior knots and the order $r = 0$ are specified. If an intercept is included, then this is equivalent to performing monotone regression. A small example clarifies this statement. Let the proximities be $\mathbf{p}' = (1, 2, 3, 4, 5)$ and the knots be at $\mathbf{t}' = (0.5, 1.5, 2.5, 3.5, 4.5, 5.5)$. Then, the matrix **M** of a

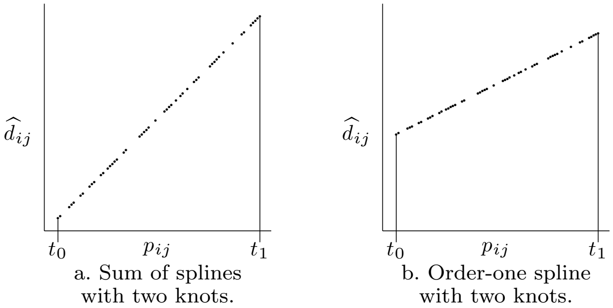


FIGURE 9.9. Special cases of spline transformation: (a) all weights $b_i = 1$, two knots, and order larger than zero; (b) spline with two knots and order one, which is equal to an interval transformation if an intercept is included.

zero-order spline is equal to

$$\mathbf{M} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

For monotone splines, we require weights $\mathbf{b}$ to be larger or at most equal to 0, so that $\mathbf{Mb}$ plus an intercept $b_0\mathbf{1}$ (with $b_0 \geq 0$) is always larger than 0. The matrix multiplication plus the intercept results in

$$\begin{aligned} \widehat{\mathbf{d}} &= b_0\mathbf{1} + \mathbf{Mb} = b_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \\ &= \begin{bmatrix} b_0 \\ b_0 + b_1 \\ b_0 + b_1 + b_2 \\ b_0 + b_1 + b_2 + b_3 \\ b_0 + b_1 + b_2 + b_3 + b_4 \end{bmatrix}. \end{aligned} \tag{9.13}$$

But the restrictions $b_j \geq 0$ for $j = 0$ to 4 in (9.13) imply that $0 \leq \widehat{d}_1 \leq \widehat{d}_2 \leq \widehat{d}_3 \leq \widehat{d}_4 \leq \widehat{d}_5$, which is exactly the same restriction as in monotone regression. Thus, a zero-order monotone spline transformation with appropriately chosen knots is exactly equal to a monotone regression transformation.

Therefore, a monotone spline transformation can be seen as a general transformation with linear and ordinal transformations as extreme cases.

Solving the Nonnegative Least-Squares Problem for Monotone Splines

How can we calculate the disparities $\widehat{\mathbf{d}}$ for a monotone spline transformation? Remember that the disparities for splines with intercept are defined by $\widehat{\mathbf{d}} = \mathbf{Mb}$, where $\mathbf{M}$ here is augmented with a column of 1s for the intercept and the weight vector $\mathbf{b}$ is augmented with element b_0 for the intercept. We have to find weights b_j such that they are as close as possible to the (fixed) distance vector $\mathbf{d}$, subject to the constraints that $b_j \geq 0$. Thus, we have to minimize

$$\tau(\mathbf{b}) = (\mathbf{d} - \widehat{\mathbf{d}})'(\mathbf{d} - \widehat{\mathbf{d}}) = (\mathbf{d} - \mathbf{Mb})'(\mathbf{d} - \mathbf{Mb}), \qquad (9.14)$$

subject to $b_j \geq 0$. Minimizing $\tau(\mathbf{b})$ over $\mathbf{b}$ is a *nonnegative least-squares* problem. It can be solved by alternating least squares (ALS), which, in this case, amounts to the following strategy. First, start with an initial weight vector $\mathbf{b}$, with $b_j \geq 0$. Then, fix all weights except b_j. Then, compute $\mathbf{r} = \mathbf{d} - \sum_{l \neq j} b_l \mathbf{m}_l$, where $\mathbf{m}_j$ denotes column j of matrix $\mathbf{M}$. Problem (9.14) simplifies into

$$\tau(b_j) = (\mathbf{r} - b_j\mathbf{m}_j)'(\mathbf{r} - b_j\mathbf{m}_j) = \mathbf{r}'\mathbf{r} + b_j^2\mathbf{m}_j'\mathbf{m}_j - 2b_j\mathbf{m}_j'\mathbf{r},$$

which reaches its unconstrained minimum at $b_j = \mathbf{m}_j'\mathbf{r}/\mathbf{m}_j'\mathbf{m}_j$. If $b_j < 0$, then we set $b_j = 0$. Then, we update the next weight, while keeping the other weights fixed, compute the unconstrained minimum (if negative, then set it to zero), and so on, until we have updated all of the weights once. These steps define one iteration of the alternating least-squares algorithm, because every weight b_j has been updated once. Iterate over this process until the weights $\mathbf{b}$ do not change anymore. It can be proved that this alternating least-squares algorithm always reaches a global minimum of the nonnegative least-squares problem. A different strategy for solving (9.14) under nonnegative constraints is described in Lawson and Hanson (1974, p. 161).

9.7 A Priori Transformations Versus Optimal Transformations

In data analysis it is not uncommon to preprocess the data to make their distribution more “normal.” The researcher may want to preprocess his or her dissimilarity data with a similar goal in mind. It may appear more attractive from a theoretical point-of-view not to optimally transform dissimilarities into d-hats by “some” monotonic function, but to apply a fixed a priori transformation on them. One such choice was suggested by Buja and Swayne (2002): they recommend using a power transformation of the

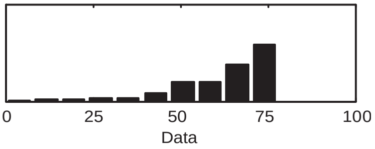
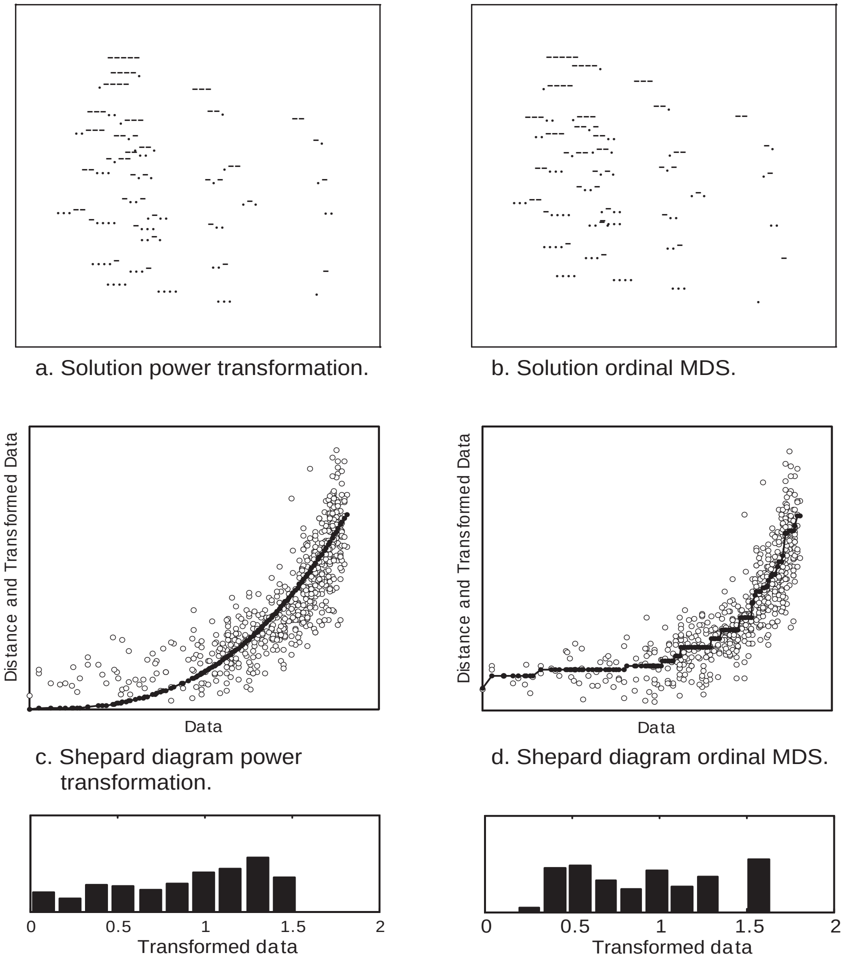


FIGURE 9.10. Distribution of the Morse code dissimilarities.

dissimilarities; that is, $\widehat{d}_{ij} = \delta_{ij}^q$ with q any positive or negative value. A positive value of q yields a convex transformation that stretches the larger dissimilarities and shrinks the smaller ones. In the case where the distribution of the dissimilarities is negatively skewed (thus with relatively many large values and few small values), then a positive q will make the $\widehat{d}_{ij}$s more evenly distributed. For negative q, the power transformation has a concave form thereby shrinking the larger dissimilarities and stretching the smaller ones. For dissimilarities that have a positively skewed distribution (i.e., data with few large and many small values), a negative q stretches the larger values and shrinks the smaller ones. The larger (or smaller) q, the stronger the shrinking and stretching. Values of q close to zero or exactly equal to zero are not very informative as all d-hats become the same; that is, $\widehat{d}_{ij} = \delta_{ij}^0 = 1$ for all ij. These d-hats can be seen as totally uninformative because they do not depend on the data (see also Section 13.3).

Let us consider the Morse code data from Section 4.2. To apply MDS, we first have to symmetrize the similarities in Table 4.2. To apply the power transformation, we also need to transform the similarities into dissimilarities. This was done by setting $\delta_{ij} = \max_{ij}((s_{ij} + s_{ji})/2) - (s_{ij} + s_{ji})/2$ thereby ensuring that the smallest δ_{ij} is zero and the largest is equal to $\max_{ij}((s_{ij} + s_{ji})/2)$. Note that Buja and Swayne (2002) also extensively discuss the Morse code data but use a different way of constructing the dissimilarities. Figure 9.10 shows the distribution of the dissimilarities obtained this way. This distribution has a tail to the left (a negatively skewed distribution), so that there are more large dissimilarities than small dissimilarities. A power transformation using $q = 3.1$ yields the distribution in Figure 9.11e which seems to be reasonably evenly distributed. One way to find out how to choose the value of q is simply trying out different values and see which one gives the best Stress value. In our case, the optimal value for q was 3.1.

We now compare the power transformation to an ordinal MDS on these data. Figure 9.11 exhibits the results for both analyses with the left panels showing the results of a power transformation and the right panels the ordinal MDS results. The Stress-1 for the power transformation is .2290 and for ordinal MDS 0.2102 indicating that only a little information is lost by switching from ordinal to a power transformation. Looking at the distributions of the $\widehat{d}_{ij}$s, the ordinal transformation seems to be better able to stretch the smaller values than the the power transformation. Thus, the



a. Solution power transformation.

b. Solution ordinal MDS.

c. Shepard diagram power transformation.

d. Shepard diagram ordinal MDS.

e. Distribution of d-hats of the power transformation.

f. Distribution of d-hats of ordinal MDS.

FIGURE 9.11. MDS of the Morse code data using the power transformation with $q = 3.1$ (left panels) and an ordinal MDS (right panels).

gain in fit is due to the better ability of the ordinal MDS to properly represent the smaller dissimilarities, because their errors in the Shepard plot are smaller for ordinal MDS than for the power transformation. The solutions in Figures 9.11a and 9.11b are highly similar. Close inspection reveals small differences in location, perhaps most notably so for points "– – – – ." (9), "." (e), and "–" (t).

The example shows that a power transformation using only a single parameter can yield an MDS solution that is close to ordinal MDS. Clearly, a power transformation is a more parsimonious function than an ordinal transformation. A strong point of Buja and Swayne (2002) is to consider the distribution of the $\widehat{d}_{ij}$s. We conjecture that good transformations tend to give d-hats that are evenly distributed. For dissimilarities that have an "irregular" shape (e.g., a bimodal shape), we expect that the power transformation will not be able to yield a solution close to an ordinal one. For regularly shaped but skewed distributions, we expect the power transformation to work fine.

Applying the power transformation in MDS is easily done in the GGVIS software discussed extensively in Buja and Swayne (2002) (see also, Appendix A). In an interactive way, GGVIS allows you to determine the optimal q. Note that SYSTAT has a special option to find the optimal q by the program itself.

9.8 Exercises

Exercise 9.1 Consider the dissimilarity data in Exercise 2.4 and the MDS coordinates for these data in Exercise 3.2. For convenience, they are both reproduced in the table below.

Dissimilarities

Color	Red	Orange	Green	Blue
Red	-	1	3	5
Orange	1	-	2	6
Green	3	2	-	4
Blue	5	6	4	1

MDS Coordinates

Dim.1	Dim. 2
0	2
0	0
4	0
6	6

(a) String out the dissimilarities for the different pairs of colors in a column vector.

(b) Compute the MDS distances from the points' coordinates, and append a column with these distances to the vector of dissimilarities from above.

(c) Derive the $\widehat{d}_{ij}$s for the data-distance pairs, proceeding as we did above in Table 9.4.

(d) Plot a Shepard diagram for the data, distances, and $\widehat{d}_{ij}$s.

(e) Find the rank-images of the distances.

(f) Make a scatter plot of the distances vs. the rank-images. What does that plot tell you about the MDS solution?

Exercise 9.2 Discuss the Lingoes–Roskam conjecture that rank-images are less prone to degenerated solutions than monotone regression in Kruskal's sense. What is the rationale for this conjecture?

Exercise 9.3 Consider the notion of primary and secondary approaches to ties in ordinal MDS.

(a) List arguments or describe circumstances where the primary approach makes more sense than the secondary approach.

(b) Collect and discuss arguments in favor of the secondary approach.

Exercise 9.4 Consider the transformation plots in Figure 9.1. Sketch some monotone functions that satisfy the primary approach to ties. How do they differ from functions for the secondary approach to ties? (Hint: Consider Figure 3.3.)