5.3.3 Distance models

The central idea of distance models is that the proximity of points in a space is used to represent their empirical similarity, or equivalently that distance represents their dissimilarity. In the vast majority of MDS models, the distance function involved is the familiar Euclidean form, but Euclidean distance is only one special case of a whole family of distance functions, each with its own characteristics and properties (see Appendix A2.1.1.2). Proceeding from the familiar to the less familiar, we shall discuss the basic distance model first, then move on to look in greater detail at the properties of Euclidean and other types of distance.

Given a set of distances it is always possible to reconstruct the configuration of points which generated them. (This procedure is described in Appendix A5.2.2 and forms the basis of classic metric scaling discussed above.) However, such a recovered configuration is not unique, in that several aspects of it are arbitrary and

may be changed at will. (These have been mentioned before (4.1), and are further discussed in Appendix A7.1.) In particular, the actual size or scale of the configuration and the origin of the space are arbitrary. Moreover, the orientation of the axes may be changed and reflected at will. Strictly speaking, it is only the relative distance between points which is significant in interpreting a distance model solution—the origin and axes simply provide a convenient framework to locate the points.

5.3.3.1 Point-point ('two-mode 'unfolding') distance models

When the data consist of a rectangular two-mode matrix, of rankings or ratings, then the distance model can be used to represent both the stimuli *and* the subjects as points. The solution consists of a configuration of p stimulus points and N subject points where each subject is represented as being at a 'maximal' or 'ideal' point, located in such a way that the distances from this point to the stimulus points are in maximum agreement with the subject's preference ratings or rankings.

In external models such as PREFMAP phase III, the stimulus configuration is obtained separately and remains fixed whilst the 'subject' or property points are estimated (see 4.4.2), whereas in internal models, such as MINI-RSA, both sets are estimated simultaneously. As in the case of the vector model, both metric and non-metric versions exist—in the former a linear correlation between the preference data and the subject-stimulus distances is maximised while in the latter a variant of stress involving only the rank order of the data is minimised.

The position of the 'ideal point' is interpreted as the one point in the space where the subject's preferences are at a maximum, and her preference decreases in every direction. This is often termed a 'single peaked preference function', since it assumes that there is only *one* point of maximum preference.

The non-metric version of the distance model is best known under the title of 'unfolding analysis', developed by Coombs (1964, chs. 5–7). The two-dimensional case is illustrated in Figure 5.7 with reference to the same 5-stimulus configuration used in the vector model case (Figure 5.5).

A midline is drawn between each pair of points, dividing the space up into 46 isotonic regions. Every ideal point within one of these regions possesses the same rank order of distances to the five stimuli. This is illustrated in Figure 5.7; thus in region I the corresponding I-scale is ABECD, and in crossing over the midline CE to region II, the I-scale becomes ABCED. Similarly the move from region III to IV represents the transition from DBCEA to DBECA. Notice that some regions are entirely encompassed by midlines (closed isotonic regions), whilst others at the periphery are not (open isotonic regions). Herein is an important distinction between the vector and distance models: the vector model excludes closed regions (see the corresponding Figure 5.5) and can accommodate fewer I-scales than the distance model. The maximum number of I-scales compatible with the two models is illustrated below in Table 5.3 (see Coombs 1964, Tables 7.1 and 12.9).

Normally the points corresponding to the most popular or consensual rankings will lie at the centre of the space, and the least popular ones at the periphery. Research has shown, as Coombs originally suggested, that ideal points within the 'open' isotonic regions are located with less accuracy than those in the closed ones. Moreover, the fewer the midlines constraining a region, the more likely it is that the

	No. of dimensions								
No. of points	2		3		4		5		*Total possible (p!)*
1	1	*1*	1	*1*	1	*1*	1	*1*	1
2	2	*2*	2	*2*	2	*2*	2	*2*	2
3	6	*6*	6	*6*	6	*6*	6	*6*	6
4	18	*12*	24	*24*	24	*24*	24	*24*	24
5	46	*20*	96	*72*	120	*120*	120	*120*	120
6	101	*30*	326	*172*	600	*480*	720	*720*	720
7	197	*42*	932	*352*	$2.56_{10}3$	*1512*	$4.32_{10}3$	$3.60_{10}3$	$5.04_{10}3$
8	351	*56*	$2.31_{10}3$	*646*	$9.08_{10}3$	$3.98_{10}3$	$2.22_{10}4$	$1.42_{10}4$	$4.03_{10}4$
9	583	*72*	$5.12_{10}3$	$1.09_{10}3$	$2.76_{10}4$	$9.14_{10}3$	$9.49_{10}4$	$4.60_{10}4$	$3.63_{10}5$
10	916	*90*	$1.04_{10}4$	$1.74_{10}3$	$7.36_{10}4$	$1.90_{10}4$	$3.43_{10}5$	$1.28_{10}5$	$3.63_{10}6$
MODEL: DISTANCE/*VECTOR*	D	*V*	D	*V*	D	*V*	D	*V*	

Table 5.3 *Total possible number of I-scales, and totals compatible with the distance and vector models for* p *points in* r *dimensions)*

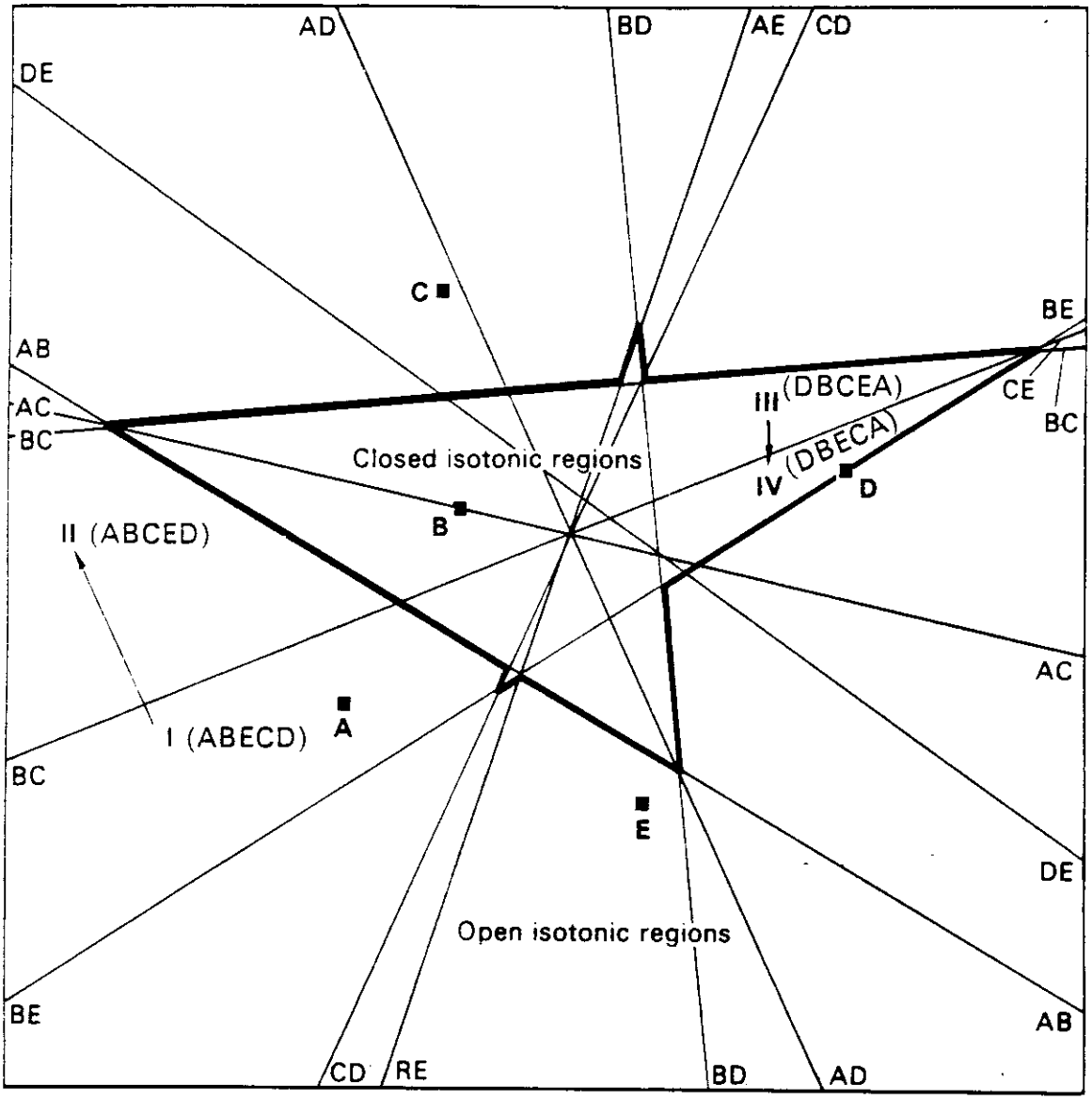
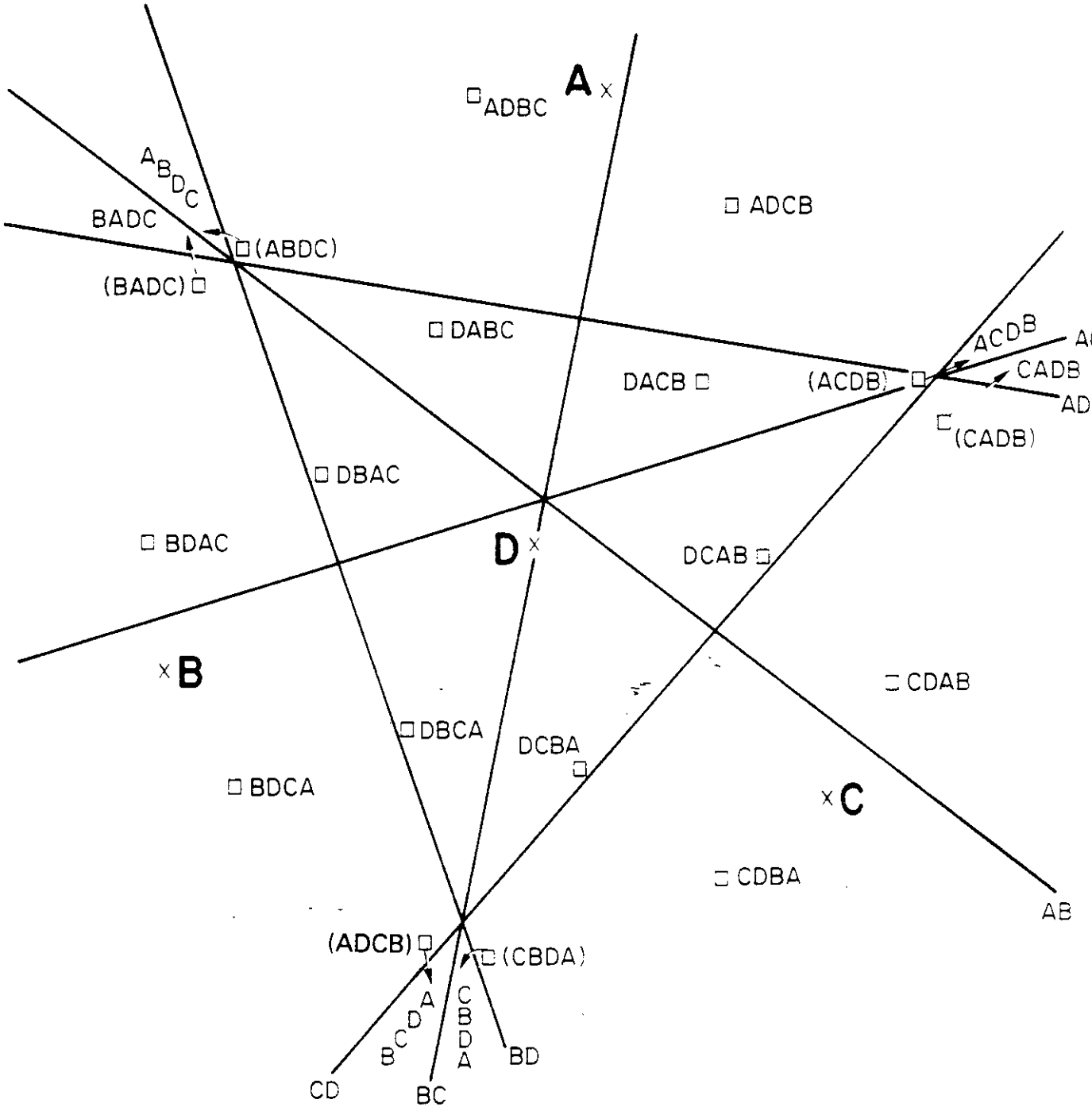


Figure 5.7 *Forty-six rankings compatible with 2-D stimulus configuration of 5 points (vector model)*

subject point be mislocated in a scaling solution. This is well illustrated in Figure 5.8, representing the scaling of the 18 I-scales compatible with a 2-dimensional, 4-point configuration (Coombs, 1964, Figure 7.4, p. 146). Each small square represents the position of a subject as located by the relevant non-metric program (MINIRSA). Note that in the case of the closed regions, the squares are all located within the correct region, although they are deflected to the outer edge. In the case of the open regions, those defined by three lines are correctly located near the centre of the region but those defined by only two lines are, without exception, displaced slightly outside their correct location.

The multidimensional unfolding model is hence clearly more 'tolerant' than the vector model, in the sense that it can accommodate more I-scales (see Table 5.3). So long as the number of stimulus points is large compared to the number of dimensions, the size of the isotonic regions is small, especially towards the centre of the configuration, and they become increasingly well-represented by a point. For this reason, stimuli points in the central part of a configuration are normally the most stable, whilst those at the periphery can usually be moved around fairly freely without affecting the goodness of fit. The variation in judgments about particular



$Stress_2$ = 0.0003 after 45 iterations

N.B. Arrows indicate discrepancy between 'true' locations and scaled ('recovered') location

Figure 5.8 *Actual and scaled location of isotonic regions*

stimuli is also an important factor in assessing the stability of a configuration in an internal scaling model. Highly popular stimuli will tend to be projected into the centre of the subject points (so that they can feature close to most subject's ideal points) and highly unpopular stimuli will be located at the outside of a configuration. Indeed, if a stimulus is sufficiently unpopular it can be located virtually anywhere on the periphery, so long as it is at a maximum distance from the ideal points. An example of this occurs in the analysis of the Delbeke data reported in section 6.2.2 (see Coxon 1974) where virtually all subjects rejected the

stimulus 'no children' in a study of preferences for families of different sizes and sex composition. When scaled, this stimulus was located at greatly varying points, but always at an extreme distance from the centre.

In summary, the properties of the point-point (distance) model of preference which contrast with the vector model are as follows:

(i) *Single peakedness.* It is assumed that each subject has one single point of maximum preference and that preference decreases (symmetrically) from this point.

(ii) *Excellence.* If the distance model holds, then each stimulus must be preferred most by at least one subject.

(iii) There is nothing corresponding to the reversability property of the vector model in the multidimensional unfolding model: some mirror-image pairs of I-scales will exist, but not others. More importantly, the distance model is characterised by the presence of closed isotonic regions, which cannot occur in the vector model.

5.3.3.2 *Euclidean and non-Euclidean distance*

So far, 'distance' and 'Euclidean distance' have been used interchangeably. In fact, a whole family of distance measures can be defined for a given configuration of points. Our interest shifts away from the correct location of points to how we measure the distance between them.

Three types of distance have been found useful in MDS and are represented in various MDS(X) programs: city block, Euclidean and dominance metrics. These are all special instances of the Minkowski *r*-metric family of distance measures which have the form:

General (Minkowski) Distance

$$d_{jk}^{(r)} = \sqrt[r]{\sum_a |x_{ja} - y_{ka}|^r}$$

where x_{ja} is the co-ordinate of the k th point and y_{ka} is the co-ordinate of the j th point on the a th dimension and r is the Minkowski r-metric power.

Each value of r (between 1 and infinity) defines a distinct metric distance. Each can be thought of as a simple composition model—a 'powered additive difference' model which asserts (Beals et al. 1968 pp. 133–5) that:

(i) absolute *differences* on each dimension, a
(ii) which are raised to the same *power* r
(iii) combine *additively* over the dimensions to produce
(iv) the overall distance between a pair of points, j and k.

In the case of Euclidean distance, the power is 2, so differences are squared, and the final distance measure deflates the value by taking the square root.*

*Carroll and Wish 1974, p. 412 et seq. argue persuasively that the final r-th root may often be usefully ignored, and when this is done a wider range of models qualify as metrics. In the Euclidean case, a number of models are more simply expressed and best understood by treating *squared* distances (i.e. ignoring the final square root). Carroll and Wish (ibid. p. 413) and Shepard (1974, p. 405 et seq.) discuss even more general distance measures, some of which do not even satisfy the triangle inequality.

6.2.3 *Internal mapping by the distance model* (MINIRSA)

Concisely: Multidimensional Unfolding provides: MINIRSA (Rectangular Space Analysis) or

internal analysis of two-way data in a row-conditional format of a (dis)similarity measure
by a Euclidean distance model
using a monotonic transformation of the data.

The basic multidimensional point-point distance ('unfolding') model was described in 5.3.3.1, and the algorithm is described in the MDS(X) documentation and in Roskam (1979, pp. 300–4). MINIRSA positions each subject *as a point* (the 'ideal point' or single point of maximum preference) in a joint space with the stimuli, also represented as points, such that the rank order of the distances from the ideal point to each of the stimuli is as close as possible to being in the same rank order as the subject's preferences (or other similar) data.

In the current MDS(X) series a metric variant of multidimensional unfolding is possible using the 'quasi-internal' options* of PREFMAP-III.

*INIT (0), FIT (0) and S-PHASE (3). See 7.5.5.2.

The *non-metric* internal analysis of rectangular data by multidimensional unfolding is implemented by the program MINIRSA. When used with few data or on data with little variation, however, the solution is not likely to be well constrained, and the algorithm is particularly subject to local minima. It also tends to be expensive in terms of computer time, requiring a large number of iterations to achieve satisfactory improvement. Despite these inherent problems of non-metric unfolding, MINIRSA is generally a useful program so long as the data are sufficient (at least 30 subjects and 8 stimuli is a rule of thumb for a 2-dimensional solution), with variation in the ranks of each stimulus.

Once again, the program can be illustrated by reference to the Bollen-Delbeke data. MINIRSA minimises $stress_2$ and the overall values for 3 dimensions was 0.069, and 0.124 for 2 dimensions. The 2-dimensional configuration is presented in Figure 6.10.

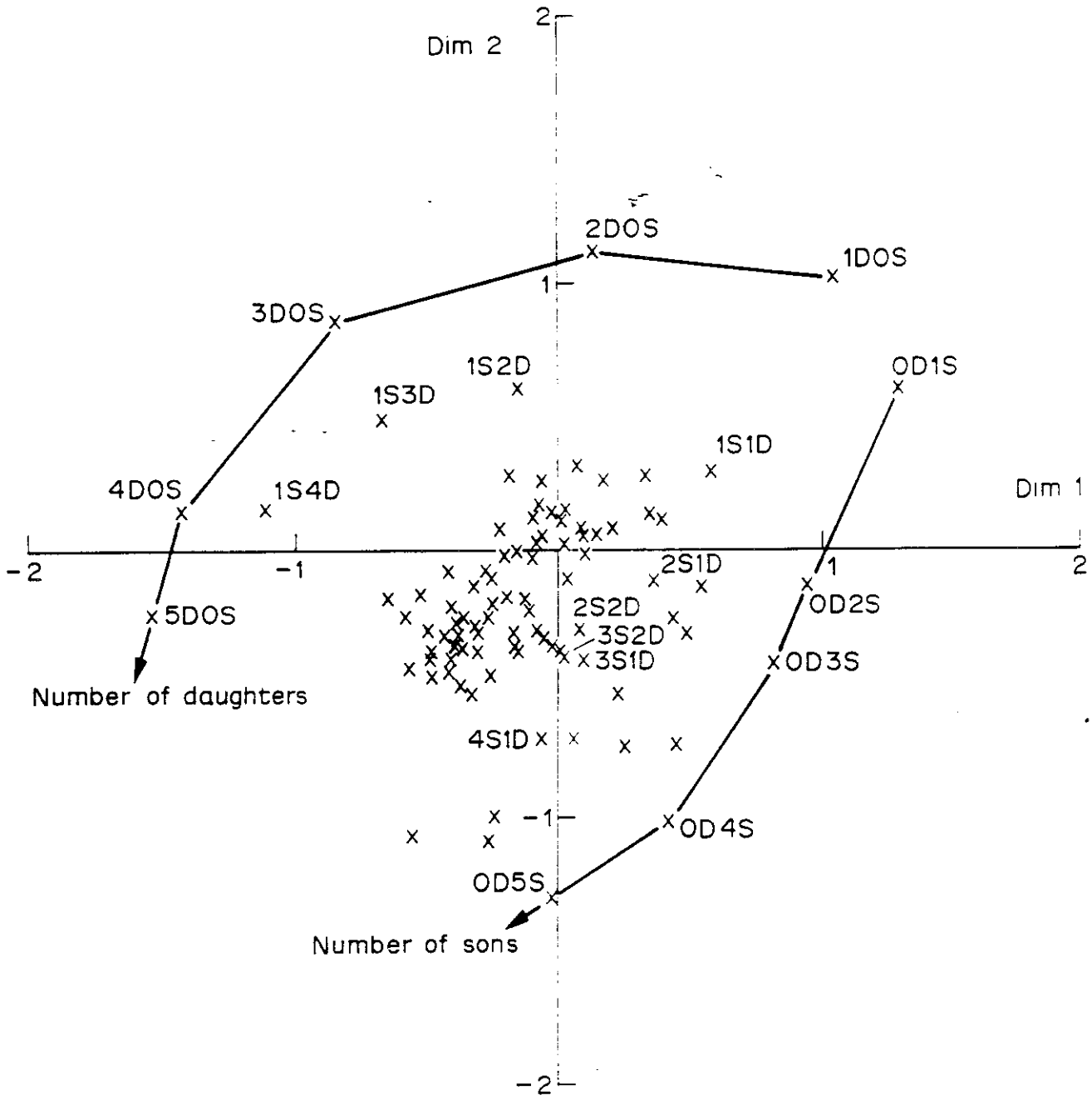


Figure 6.10 *MINIRSA analysis of family size and composition data*

Because it is an internal analysis, it is important to begin by examining the subject points—located in a main cluster at the centre of the configuration (representing preference for largish mixed families), with stragglers to the SE (representing those who prefer a preponderance of boys).

Significant information is additionally conveyed by the occurrence of empty spaces in a configuration. In the vector model, we saw that the fact that parts of the unit circle (or sphere) do not have subject vector ends indicates that no preferences increase in these directions. In the distance case, we need to take note of the *regions* within which ideal points do not occur. This can be done explicitly by constructing the isotonic regions from the stimulus configuration (see Figure 5.5) and then looking at the regions in which ideal points are concentrated and at the regions which do not contain any ideal points.

The stimulus configuration for these data is quite different from the MDPREF one, but can be interpreted in a similar manner. First, the stimuli do form a distorted version of the defining lattice of Figure 6.7a, though the component dimensions (drawn in the figure) are not quite linear or at 90° (uncorrelated); rather, they are pulled in towards one another to enfold the ideal points. Nonetheless, a linear PROFIT fits both 'number of sons' and 'number of daughters' properties with a correlation of around 0.93. The main distortion in the configuration is the way in which the larger family-sizes equally mixed in composition virtually collapse onto one point, and the program deals with them by locating them as close to as many ideal points as possible. Here is a further example of how internal analysis is likely to contort a stimulus structure to satisfy the subjects' data better.

What of the three characteristics of the data? How can these be read out of the configuration?

(i) *All-boy families are preferred to all-girl families.* This is evident in the way in which a number of ideal points congregate closely to the Number of Sons line, whilst none are located very close to the Number of Daughters line.

(ii) *Mixed sex composition is preferred.* The main concentration within the swarm of subject points lies almost exactly at 45° counter-clockwise inclination, which represents the mixed composition line of the stimuli.

(iii) *Within a given family size, a preponderance of boys is preferred.* This also is discernible in the location close to the Number of Sons line.

Although the form of representation is different, the same content can be read out of both the vector and the point representation, but as we have seen, the researcher has to pay especial attention to the *joint* space and the assumptions of the model: the stimulus configuration is not likely to be accurately recovered where there is not enough variation in the data. In any event, it would be advisable to obtain a separate estimate of the stimulus configuration and then map preferences into it, particularly in the case of data based on human judgments of this sort.