5.3.2 Scalar product models

In the MDS(X) series, all the scalar products (or vector or factor) models assume that the data consist of (or can be reduced to)* a rectangular *two-mode* matrix consisting of a set of (preference) ratings or rankings of a set of p stimuli made by a

*In the MDPREF vector model, input may be a set of pair comparison dominance matrices.

set of N subjects. (In MDPREF this matrix is termed the 'first score matrix'.) For convenience the entries in this matrix are usually denoted s_{ij} to mean the similarity between subject i and object j, or more usually the preference score given by subject i to object j.

The vector solution consists of a configuration of p stimulus points in a user-chosen number of dimensions, and each of the N subjects' set of preference ranks or ratings is represented as a vector, located so that the projections of the stimuli on the vector are in maximum agreement (correlate as highly as possible) with that subject's preferences. The external form of this analysis, i.e. where the stimulus configuration is obtained separately and remains fixed whilst the subject vectors are estimated, was discussed in section 4.4.1.

The purpose of these models is to represent both the stimuli and the subjects in a common 'joint space'. Each subject's preferences are represented as a vector—a projection down, or collapsing of, the stimulus space onto a single dimension—just like the properties embedded in a stimulus space. Interest will chiefly focus therefore on two things:

(i) how well the subject's preferences can be accommodated by the model, and hence represented in the stimulus space (this can be assessed by the correlation of the projections with the original data) and

(ii) how the vectors relate to each other, since the main purpose may be to investigate individual differences in a set of rankings/ratings.

Differences between rankings are signalled in the vector model principally by angular separation. On the one hand, as we saw earlier, the direction in which a vector points is highly significant, for it indicates the manner in which the subject mixes or trades off the characteristics of the stimuli in producing her preferences, and this is measured by the cosine of the angle which the vector makes with the dimensions of the space. By the same token, if we are interested in how one subject vector relates to another, we inspect the angular separation between them—the linear correlation, or cosine of the angle between the two vectors. In inspecting a vector model solution, the first point of interest is how the subject vectors are dispersed around the unit-circle (or sphere).*

If the vector ends are located in a small sector, this indicates high consensus or agreement in subjects' preferences, whereas the more unevenly they are distributed round the circle, the greater the dissensus. The researcher will presumably become interested in whether distinguishably different 'points of view' exist, suggested by small sectors with a high density of vector ends and empty sectors between sectors. If there are different categories of subjects we may also want to know whether the average direction differs significantly between the categories, and statistical tests and procedures for analysing directional data have been developed and are available. (They are discussed in Mardia 1972, and in the MDS context in Coxon and Jones 1979, pp. 128–36 as well as in the MDS(X) documentation for the MDPREF program.)

*By convention, subject vectors are normalised to have the same (unit) length in MDPREF. Though this is not a necessary restriction of the model, it makes for greater simplicity if vectors are of standard length. In two dimensions, vector ends will therefore lie around a unit circle, in three (and higher) dimensions, they lie around a (hyper) sphere.

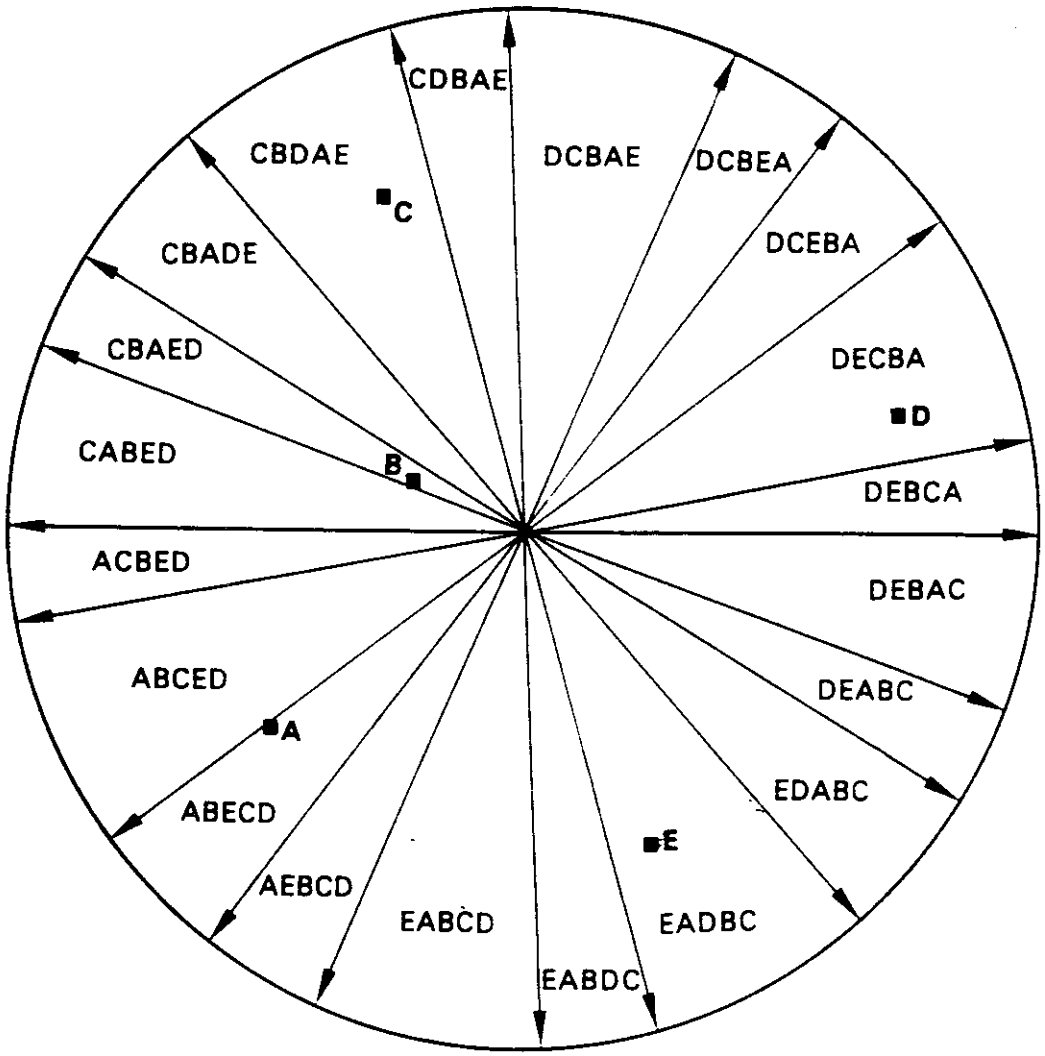
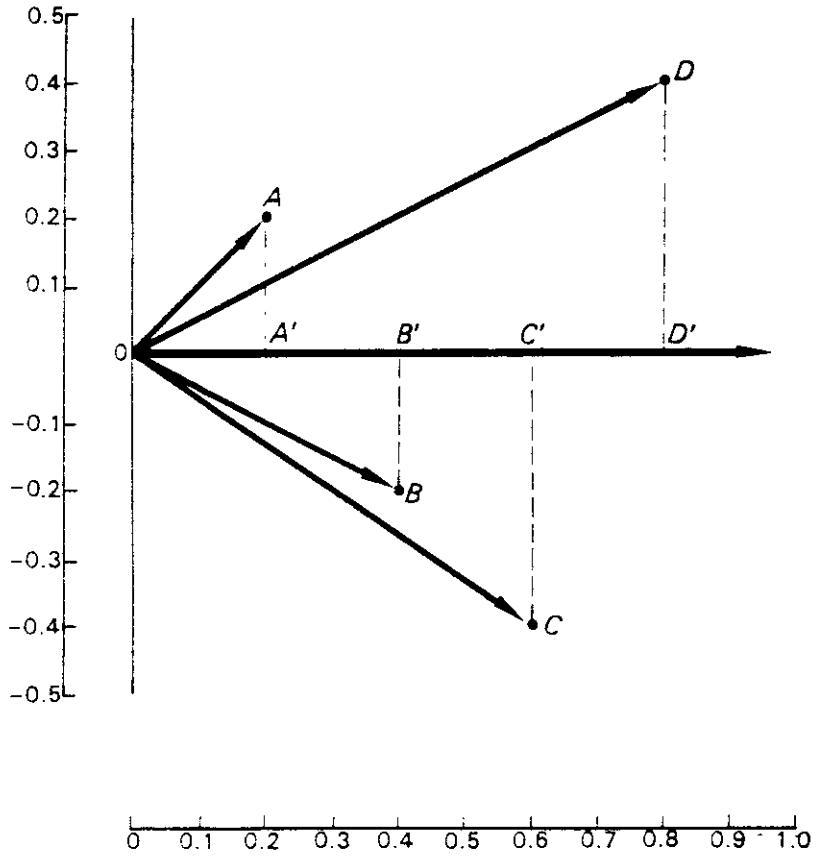


Figure 5.5 *Twenty rankings compatible with 2-D stimulus configurations of 5 points (vector models)*

Although a total of $p!$ (i.e. $p(p-1)(p-2)\ldots\times 1$) rankings of p objects is possible, only a limited number of these can be accommodated within a stimulus configuration. We therefore need to enquire both how many rankings can be accommodated in a configuration of p points in r dimensions and how they are represented therein. As an example, take the 5-point stimulus configuration given in Figure 5.5. There are $5! = 120$ possible rank orderings of 5 stimuli, but only 20 of these can be represented perfectly in a given vector configuration of 5 points in two dimensions, and one half of these will simply be mirror-images of each other formed by reversing the direction of the vector. The 20 rankings compatible with this configuration are given in the figure. Notice that there is an orderly interlocking between the rankings, akin to that shown by Coombs (1964, p. 87 et seq.) in the context of discussing the unidimensional unfolding (distance) model for preferences. As one moves around the circle, only adjacent stimuli are interchanged in the rankings (beginning in the north-easterly position and moving clockwise: *DCBEA*, *DCEBA*, *DECBA*, *DEBCA*, and so forth).

Although the scalar products model has been described as a point (stimulus) and vector (subject) representation, formally the model is expressed entirely in terms of vectors—a set of vectors drawn from the origin of the space to the location of each



Subject vector **y**	Stimulus point **x**	Scalar product = **yx'**	Projection
(1,0)	*A* (0.2, 0.2)	0.2	*OA'* = 0.2
	B (0.4,−0.2)	0.4	*OB'* = 0.4
	C (0.6,−0.4)	0.6	*OC'* = 0.6
	D (0.8, 0.4)	0.8	*OD'* = 0.8

Figure 5.6 *Projections and scalar products*

stimulus, and a set of unit-length *subject* vectors. The key to understanding the formula for the model is knowing that the scalar product of the stimulus vector with the (unit-length) subject vector is the same as the vertical projection of that stimulus point on the subject vector. This property is illustrated in Figure 5.6, where the subject vector is drawn along the first dimension to simplify the arithmetic.

Let $\mathbf{x}_j$ represent the vector from the origin to the location of stimulus j in r-dimensional space, and $\mathbf{y}_i$ represent the (unit-length) vector for subject i, then the preference value which stimulus j has for subject i is estimated as the scalar product of the vector concerned:

$$\hat{s}_{ij} = \mathbf{y}_i \cdot \mathbf{x}'_j = \sum_{a=1}^{r} y_{ia} x_{ja}$$

or in matrix form:

$$\hat{\mathbf{S}} = \mathbf{Y}\mathbf{X}'$$

The matrix of preference scores estimated by the model is termed the 'second-score matrix', and the purpose of the vector model is to obtain a stimulus configuration **X** and subject vectors **Y**, so that the discrepancy between the original 'first-score' data (s_{ij}) and the estimated 'second-score' values ($\hat{s}_{ij}$) is as small as possible. (In the case of a non-metric version, the monotonically transformed data will be compared to the estimated values.) Carroll (1972, p. 124 et seq.) and the MDS(X) documentation describe the stress-like index of agreement, C_1 used to measure the goodness of fit. The method of solution involves factoring two product matrices formed from the first score matrix.*

The main properties of the vector model (cf. Roskam 1968, p. 28) may be summarised as follows:

(i) *Increasing utility*. A subject's preference (or similarity rating) increases continuously in the direction of the vector: the further out an object projects on it, the more it is preferred.

(ii) *Mediocrity*. An object may always occupy a position between the extremes of all the subject's preferences, i.e. never be either most or least preferred (see object B in Figure 5.5, for example).

(iii) *Reversability*. If a given ordering occurs, the opposite ordering may also occur. Indeed, the orderings compatible with a given stimulus configuration divide into two opposite halves, producing the characteristic 'spokes of a wheel' isotonic regions (sector of the space where the same rank ordering of stimuli is implied) seen in Figure 5.5.

The vector model differs considerably in these respects from the distance (unfolding) model of preference discussed in the next section. The differences and the related issues of interpretation of configurations produced by programs implementing the models are discussed in Chapter 6.

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

*The first score matrix **S** is approximated in the user-chosen dimensionality a, by a least squares approximation $\mathbf{S} = \mathbf{YX}'$ (of rank a) using the Eckart-Young factorising procedure. The eigenvectors of the minor product matrix **S'S** provide estimates of the stimulus configuration **Y**, and the eigenvectors of the major product matrix **SS'** provide the estimates of the subject vectors **X**, when the rows are normalised to unity. The eigenvalues of both product matrices are the same and indicate the concentration of variation in the principal axes (see Appendix A5.2.2).

6.2.2 Internal mapping by the point-vector model (MDPREF)

Concisely: MDPREF (MultiDimensional PREFerence Scaling) provides:

internal analysis of two-way preference data in the form either of a row-conditional matrix or of a set of paired comparisons matrices
by a scalar products (point-vector) model,
using a linear transformation of the data.

Note that MDPREF is an *internal* form of analysis, positioning stimuli points and subject vectors simultaneously from the data, and is a *linear* procedure: data are assumed to be at the interval level of measurement. As a form of two-mode factor analysis, MDPREF is becoming increasingly popular for analysing preference data, personal constructs rankings* and semantic differential ratings data. Since the solution is analytic rather than iterative, it is a computationally cheap and efficient procedure and the results are often a good deal more stable than for other internal two-mode models such as multidimensional unfolding (MINIRSA, q.v.).

From the user's point of view the main difference between the external (PREFMAP-IV) and internal (MDPREF) vector models is in the way of interpreting the solutions. In the external case, it will be recalled, subjects are located within a *fixed* reference configuration and the location of subjects' vectors could with some confidence be referred to or interpreted in terms of the stimulus locations. For internal analysis this is not true: the stimulus points are located in such a way that as many as possible of the subjects' data are fit well and the stimulus configuration can only be 'read' by direct reference to the location of the subject vectors.

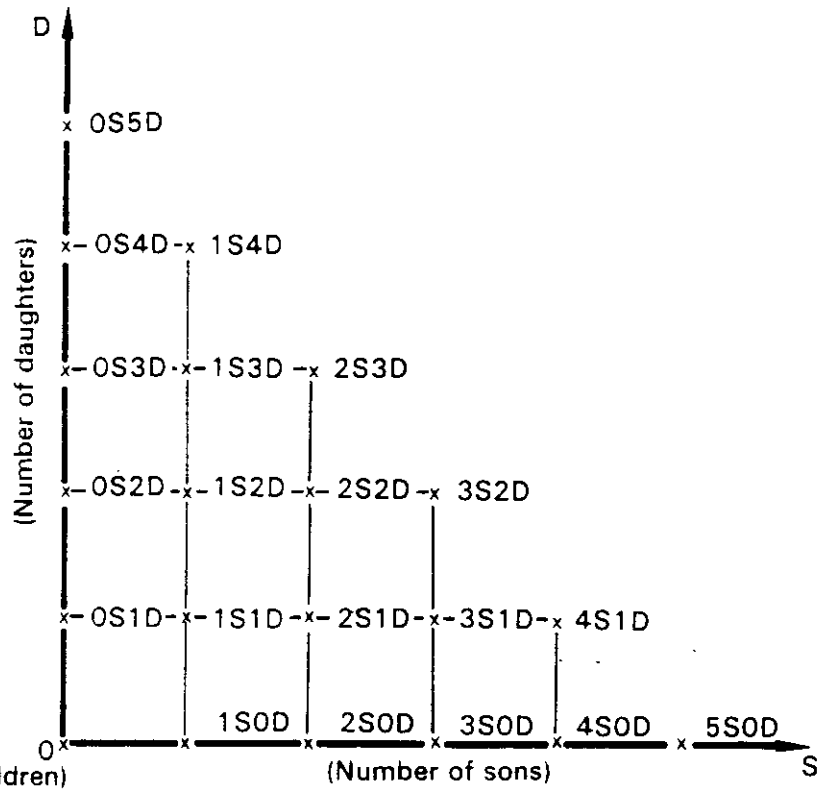
The MDPREF algorithm is described in the MDS(X) documentation and in Carroll (1972, pp. 123–9). Basically, the algorithm forms the major and minor product moment matrices from the original rectangular data matrix (called the 'first score matrix' in the MDPREF terminology) and obtains the latent roots of those matrices. (These give a good estimate of the 'true'—or at least the lowest acceptable—dimensionality of the data.) The location of the stimulus points and subject vectors are then found by producing a factoring or decomposition which gives a 'second score' matrix which best fits the data in the number of dimensions chosen by the user. The model has already been discussed above in 5.3.2.

Issues in interpretation and application of the MDPREF program are best illustrated by reference to what is now a quite well-known data set: the Bollen-Delbeke data on family size and composition preferences. It will also be used to illustrate the corresponding distance model analysis (MINIRSA).

In 1960 Bollen collected data from psychology students at the Catholic University of Louvain on their preferences for families, which differed in terms of the number and sex of the children. In all, 21 such stimuli, (family size compositions) were defined—all possible compositions from no children up to families of size five. These are illustrated in Figure 6.7. Each subject was then given each of the 210 pairs of stimuli, e.g. (3 sons, 2 daughters) *vs* (1 son, 1 daughter), and asked which he or she preferred. The data for 80 subjects (40 male, 40 female)

*In many ways MDPREF resembles (and is indeed superior to) the INGRID program developed by Slater (1960) for the analysis of repertory grid data, frequently used in personal constructs analysis in psychology and sociology. See Tagg (1979).

formed the basis for subsequent analysis, and were analysed by Bollen and later by Delbeke (1968) and Coxon (1974). The data exist both in the original form as a set of 80 pair-comparison dominance (0, 1) matrices and as a set of 80 preference



(a)

Summary information on rank scores for different family size compositions*

Stimuli	Code	Range	Mean	Variance
0 children		0–4	0.2	0.4
1 son	1S0D	0–12	3.6	7.3
2 sons	2S0D	2–17	8.1	13.1
3 sons	3S0D	5–16	9.7	6.6
4 sons	4S0D	4–18	10.0	9.6
5 sons	5S0D	0–20	8.7	20.6
1 daughter	0S1D	1–11	2.6	4.4
2 daughters	0S2D	2–13	5.6	7.7
3 daughters	0S3D	2–12	6.2	5.0
4 daughters	0S4D	1–13	5.8	5.9
5 daughters	0S5D	0–13	4.8	11.8
1 son, 1 daughter	1S1D	4–19	11.0	17.5
2 sons, 1 daughter	2S1D	10–20	15.1	6.1
3 sons, 1 daughter	3S1D	10–19	15.9	3.5
4 sons, 1 daughter	4S1D	5–20	14.8	12.6
1 son, 2 daughters	1S2D	6–20	12.5	7.5
2 sons, 2 daughters	2S2D	13–20	17.9	2.3
3 sons, 2 daughters	3S2D	12–20	18.4	4.6
1 son, 3 daughters	1S3D	2–18	12.0	8.7
2 sons, 3 daughters	2S3D	9–20	17.1	5.6
1 son, 4 daughters	1S4D	1–18	10.1	17.5

(b)

*Highest preference is a rank of 20, and lowest preference has a rank of 0.

Figure 6.7 *Delbeke family size and composition data: summary*

rankings, formed by summarising the rows of each dominance matrix to produce a 'vote count' preference order, since the subjects were remarkably consistent.* A basic summary information on the rankings data is provided in Figure 6.7b. Several points should be noted. First, the 'no children' stimulus had so little variation (virtually everyone preferred it least) that it was removed from subsequent analysis. Its inclusion led to the structure in scaling solutions being distorted, and the universal dislike of no children meant that its location was highly unstable, varying from run to run—being located at any position so long as it was maximally distant from ideal points in the distance model, and maximally opposite in direction to the preferred regions in the vector case.

Secondly, there is marked preference for large, mixed families, with the composition of 3 sons, 2 daughters having highest overall preference. The data, it should be remembered, were collected from unmarried Catholic students before the changes in attitude to and practice of birth control following Vatican II. Other characteristics are also evident, and are recognisable in the scaling solutions in differing ways, depending on the model.

(i) For each given single-sex family size, all-boy families are preferred to all-girl families, and the difference in average preference of boys increases systematically with the overall size of family.

(ii) For every family size, a mixed-sex composition is preferred.

(iii) A preponderance of boys is preferred in mixed-composition families.

MDPREF was applied to both the preference scores and the pair-comparison data. Solutions were sought in two and three dimensions, for men and women subjects separately, though only that for the male subjects is reported here. The preference score and pair comparisons data produce almost identical solutions in each case. Inspection of the roots of the first-score matrices strongly suggests that a two-dimensional solution is adequate—the percentage of variation accounted for by the first four dimensions is 75 per cent, 14 per cent, 6 per cent, 1 per cent, so 89 per cent is concentrated in the first two dimensions and little is gained by adding any subsequent dimensions. The two-dimensional configuration for males is presented in Figure 6.8. (The female data form the test data for the MDPREF program in the MDS(X) version.)

First, let us concentrate on the location of subject vectors, recalling that for a two-dimensional MDPREF solution their termini (end points) are normalised to unit length and will therefore lie on a circle. However, virtually all subjects' vectors occupy just over one-quarter of the circle, indicating a quite high degree of consensus in their preferences. In internal analysis the position of subject vectors should take priority in interpreting the joint space. It is often helpful to begin by locating an average subject and reading back along from the vector end, through the origin of the space to the other side of the circle, noting how the stimuli project onto it. (The location of the dimensions is of course arbitrary in a vector model so they do not need interpretation, but the origin is significant as the centroid of the stimulus points and the point through which all subject vectors pass.) In this case

*See Coxon (1974, p. 197, ftn 11) for details of tests of consistency. The average coefficient of consistence was 0.94.

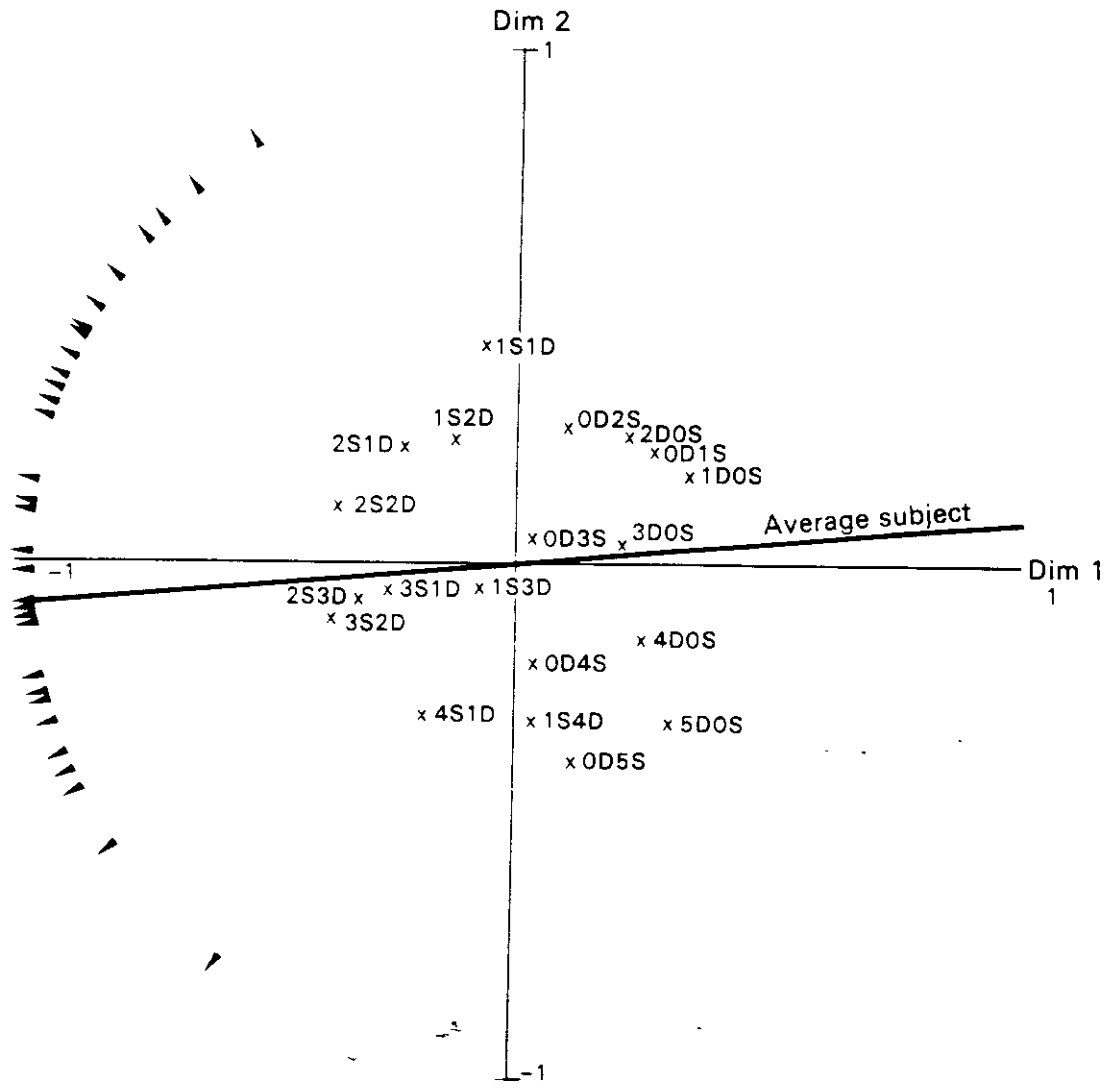


Figure 6.8 *MDPREF analysis of family size and composition data 2-D, males only*

the most popular stimulus (3S2D) is the first, most preferred one, followed closely by other fairly large, mixed compositions, then into larger single-sex families, and finishing with the least preferred single son and single daughter compositions. (The projections mirror fairly accurately the mean ranks given in Figure 6.7b.)

Secondly, what do the differences in vectors signify? The subject at one extreme (located in the NNW direction) clearly prefers much smaller families and is not by any means as concerned with the balance between sexes. By contrast, the subject at the other extreme (in the SSW direction) is greatly in favour of very large family size whether mixed in composition or not. On the whole, virtually everyone prefers mixed to unmixed composition, and the greatest variation is on the size of the family composition.

The structure of the stimulus configuration, then, should be interpreted by reference to the subjects' vectors in internal analysis and it is important to realise that the 'natural' lattice-like structure of the stimuli (see Figure 6.7a) cannot be

discerned in the stimulus configuration. (Try connecting the sons' axis, denoted on Figure 6.8 as OD1S, OD2S, OD3S, OD4S, OD5S, and the daughters' axis and you will note that they follow a roughly parallel and non-linear sequence, with a major shift in direction at OD2S and 2DOS. In no way can the lattice structure be discerned in this configuration.) Nonetheless, a very coherent structure *is* evident, which happens to be part of a radex (see 4.5), and this is illustrated in Figure 6.9. A semi-circle can be drawn which divides the *mixed* from the single sex or *unmixed* family compositions, and a set of lines can be drawn emanating from the 'centre' which divide the space into sectors corresponding exactly in this case to the overall *family size*. Another way of describing the radex (Shepard 1978, p. 57 et seq.) is to view the structure as having polar co-ordinates, with latitude (distance from the centre of the radex located in an approximate way in Figure 6.9) corresponding to the degree of mixedness, and incidentally to the degree of preference, and the angular position around the perimeter corresponding to the size of the family. Moreover, the three characteristics noted above from the initial data analysis can

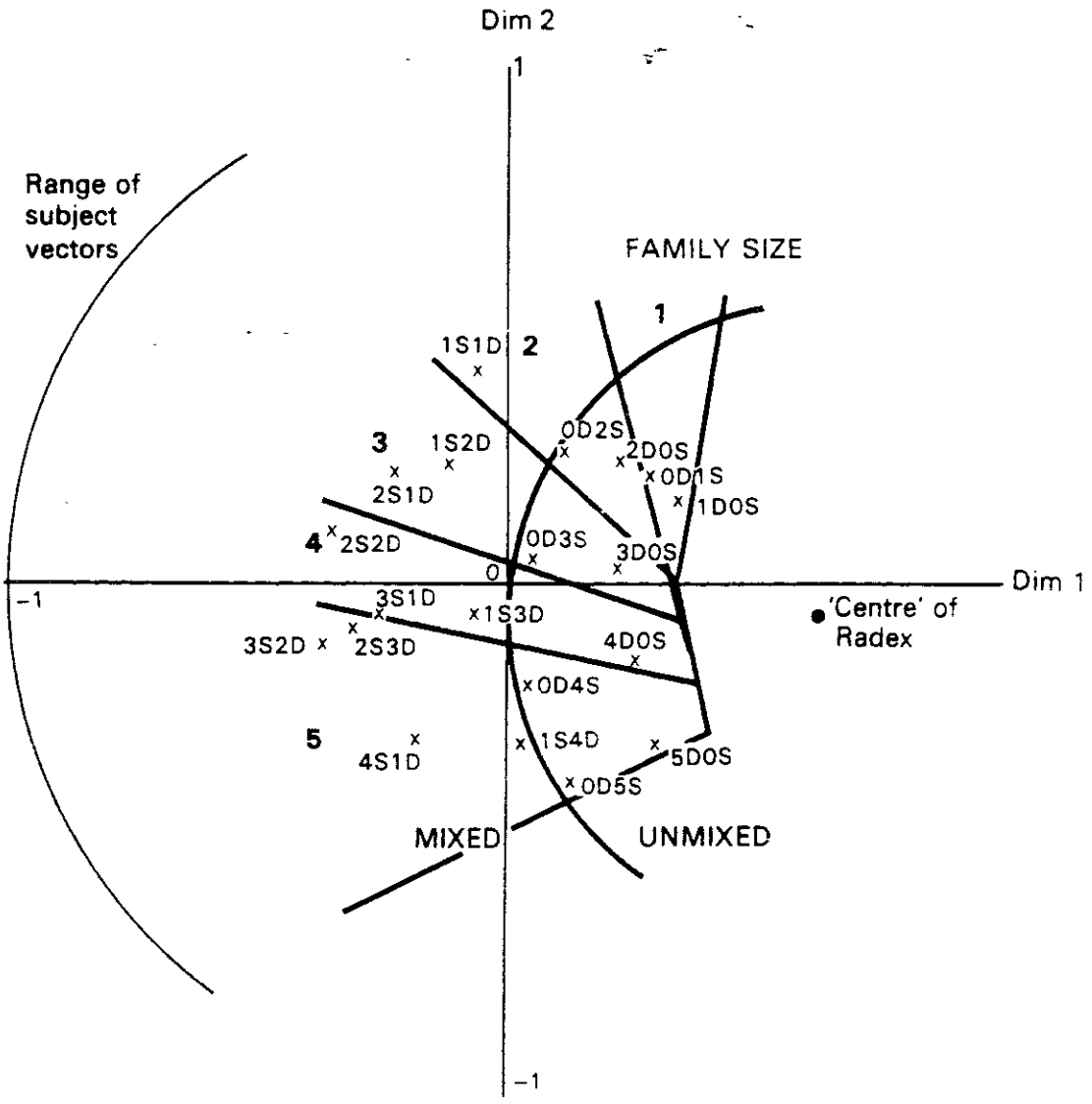


Figure 6.9 *Radex structure of family preference MDPREF solution*

in most cases also be 'read out' in a manner consistent with the form of the vector model:

(i) *All-boy families are preferred to all-girl families, for given single-sex family-sizes.* Concentrating on single-sex pairs of fixed family-size, the points representing all-boy families are systematically to the left of (more preferred than) all-girl families of the same size.

(ii) *A mixed sex-composition is preferred with a given family-size.* Within each given family-size the points representing mixed families are consistently to the left of (preferred to) those representing unmixed family composition.

(iii) However, the inference that *within* a given type of mixed composition family, a preponderance of boys is preferred does not seem to be detectable.

(The same characteristics will be recognisable, but in a different and apparently unrelated representation, when the distance (unfolding) MDS solution is given later in section 6.2.3.)

A final caution in the use of MDPREF. The program allows users to remove either the row effects (individual response-style) and or the column effects (removing the 'mean utility' or consensus) of the first-score data matrix. The removal of row effects rarely has important consequences, but the removal of column effects will produce a much wider dispersal of subject vectors, and individual differences then become the major focus of the analysis (see Heiser and de Leeuw 1979, p. 28 et seq.). On the other hand, *double*-centring—removing both column and row effects—actually turns the vector model into a distance model and typically leads to an over-estimation of the appropriate dimensionality (see Carroll, 1970, p. 278). It should be used with considerable caution, if at all.

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.

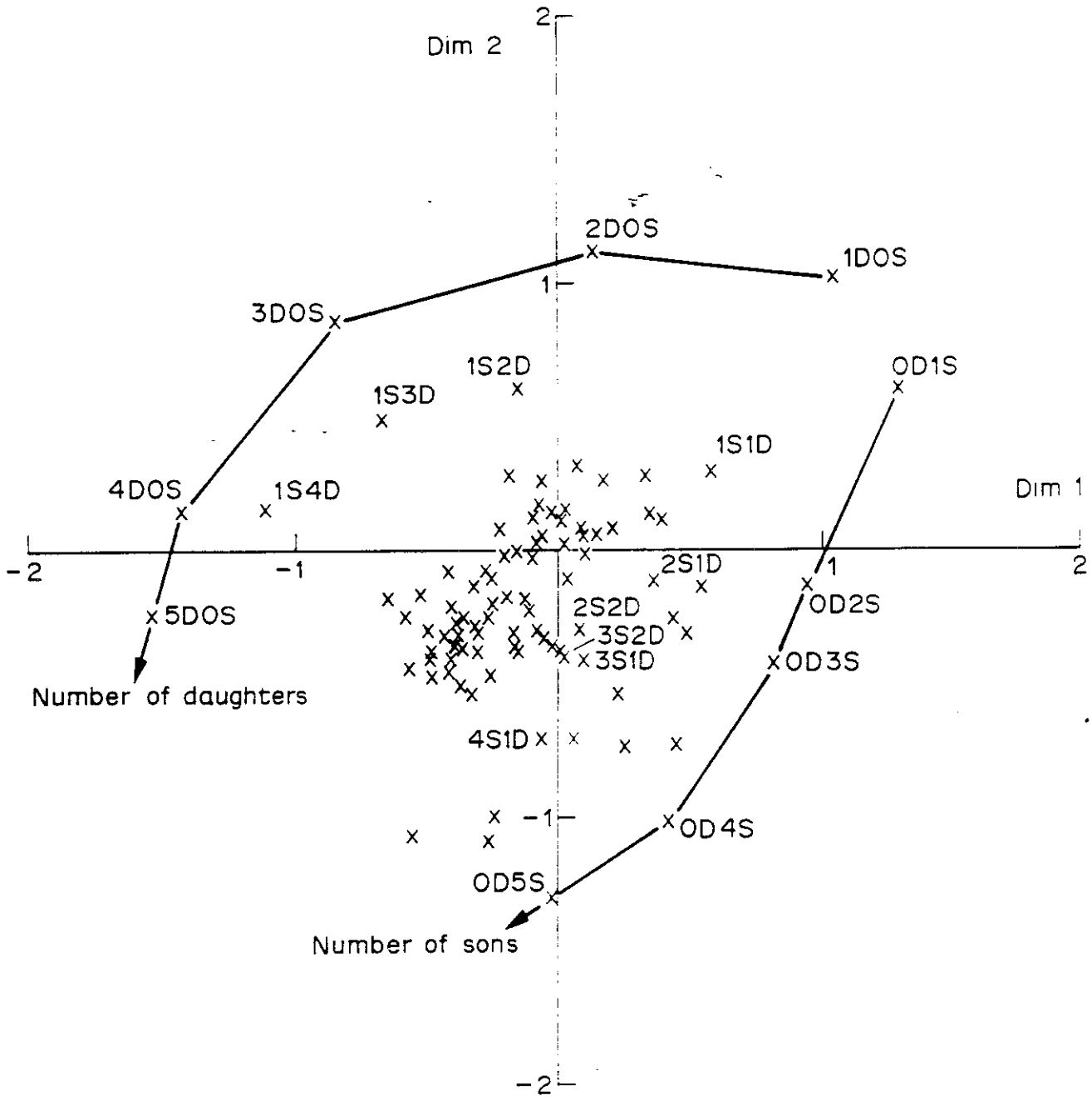


Figure 6.10 *MINIRSA analysis of family size and composition data*

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.