# 2 Data: Measures of Similarity and Dissimilarity

*You shall have true and correct weights and true and correct measures, so that you may live long in the land which the Lord your God is giving you*
DEUTERONOMY **25**. 15 (New English Bible)

A very wide range of types of data and of measures of association can be interpreted as providing information about how similar or dissimilar objects are to each other. In turn, the idea of similarity between two objects lends itself readily to interpretation in terms of proximity of two points. Conversely, the more dissimilar the objects are, the more distant the corresponding points should be. The definition of distance measures and the relationship between distances and scalar products is presented in Appendix A2.1. *Readers unfamiliar with these notions are strongly recommended to read this material before proceeding.*

*Empirical measures*
In MDS, a number of different names have been given to empirical measures which are thought of as being estimates of distance (or its antonym, proximity). In this text the generic term 'dissimilarity' will be used to cover any distance measure, and 'similarity' will be used to refer to a proximity measure. The empirical dissimilarity measure between two objects $j$ and $k$ will be denoted: $\delta_{jk}$, and any similarity $s_{jk}$ can readily be converted into a dissimilarity measure by reversing the values.

A large number of (dis)similarity measures can be used in MDS. There are two basic types:

(i) (Dis)similarity data obtained directly in the form of numerical or order estimates ('direct data'). Much attitudinal and survey data are of this form.

(ii) (Dis)similarity data based on aggregating direct data ('aggregate' or 'derived' measures). Usually, association and correlation measures are of this form.

Each type will be considered in turn.

## 2.1 Direct Data

The individual data collection methods have been well discussed by Coombs (1964, ch. 2) in terms of the amount of information which can be obtained by a given method (its channel capacity) and the amount of implied or repeated information (redundancy) it contains, which can be used to check the subject's consistency. His summary is well worth repeating:

> On *a priori* grounds we would expect that the higher the channel capacity the better, but this is certainly not true. A price is paid for data, not only in financial

> terms but in wear and tear on the organism at source. A method with too high a channel capacity may, through boredom and fatigue, result in a decrease in information transmitted, through stereotype of behavior. Furthermore, the potential variety of messages from the organism may not be great, in which case a more powerful method is inefficient ... Ideally a method should be selected which matches the information content in the source but is not such a burden as to generate noise.
>
> (ibid, p. 51)

To a large extent fatigue, stereotyping and inefficiency can be cut down by the use of experimental designs, which provide methods for selecting incomplete sets of judgments, preferably with desirable statistical properties. In Table 2.1 a brief summary of the more commonly used methods is presented, and references which contain information on incomplete designs are asterisked.*

### *2.1.1 Pair comparisons data*

Collecting *similarity ratings of pairs* of stimuli from a set of subjects is a long-established and popular pastime, especially among social scientists, but there is no guarantee that subjects' use of rating categories bears any resemblance to the arithmetic properties which are often ascribed to them. In any event, researchers should take care to ensure that a sufficiently large number of categories is provided (Green and Rao (1970) provide further empirical support for Miller's (1956) 'magical number $7 \pm 2$' as an optimum number) and that subjects use a sufficiently wide range of them. Pairwise ratings are especially prone to 'response sets' or 'response styles' (Cronbach 1946, Rorer 1965) involving highly skewed distributions of ratings (Coxon and Jones 1978a, pp. 68–71, and 1979, T3.3 and T3.4), and wide variation in the number of categories which a subject uses.

*Dominance judgments of pairs* of stimuli are also fairly common. In this case the subject is presented with a pair of objects and asked to indicate which of the two possesses more of a given attribute (is heavier, louder, more prestigious, is preferred, is more sexy etc). The binary data generated in this way can be used to test the transitivity of each single subject's choice (see Kendall 1962, p. 144 et seq.), and are frequently turned into a rank-ordering if the choice is sufficiently consistent.

### *2.1.2 Partitions (sorting) data and hierarchies data*

Any method by which a set of objects is divided into mutually exclusive and exhaustive categories constitutes a partitioning (or 'nominal scale'). The most commonly encountered forms of data collection which produce a partition are:

(i) *Dichotomisation* A set of objects is divided into *two* contrasting groups usually in terms of whether or not the objects concerned does, or does not, possess some specified property.

(ii) *Trichotomisation* Usually, the trichotomy consists of those objects which possess the property, those which do not, and those to which the property does not apply.

*Discussion of practical and methodological issues involved in these and other methods of data collection is contained in Shepard 1972c and in Coxon and Jones 1979, T2.4 and U1.4.

**Table 2.1** *Some direct methods of collecting dissimilarities data*

| *Method* | *Presentation* | *Question Instructions* | *Implied Information on Distances* | *References* |
|---|---|---|---|---|
| 1 PAIR COMPARISONS | All pairs $(j, k)$ | How dissimilar are $j$ and $k$? e.g. on a 7-point rating scale from totally similar (1) to totally dissimilar (7) | $\delta_{jk}$ gives direct estimate of $d_{jk}$ | *David 1963<br>*Spence and Domoney 1974<br>Green and Rao 1970 |
| 2 PARTITIONS (Sorting) | Subject divides stimuli into mutually exclusive and exhaustive categories or groups | Sort/divide stimuli into:<br>(a) fixed number of categories (fixed sorting)<br>(b) as many or as few categories as you wish (free sorting) | For any three stimuli, $(j, k, l)$<br>$d(j, k) < d(j, l)$<br>if $j$ and $k$ are in the same category, and $l$ is in a different category | Miller 1969<br>Anglin 1970<br>Jones and Ashmore 1973<br>Coxon and Jones 1978b, 1979 |
| 3 HIERARCHIES | Subject constructs hierarchical clustering of stimuli | (i) First choose the most similar (MS) pair, then<br>(ii) Either add new stimulus, or begin new similar pair, then<br>(iii) *Either* start new pair, *or* add stimulus, *or* merge existing cluster<br>(see Coxon and Jones 1978b) | $\delta_{jk}$ is given by the level of hierarchy at which $j$ and $k$ are joined | Johnson 1967<br>Coxon and Jones 1978b, 1979<br>Fillenbaum and Rapoport 1971 |
| 4 RANKING | Subject $i$ places stimuli $j, k, l \ldots m$ in rank order in terms of given criterion | What is your order of preference/similarity?<br>Say: $m, l, k, j$ | $d_{im} < d_{il} < d_{ik} < d_{ij}$ | Coombs 1964<br>*Durbin 1951 |
| 5 TRIADS | All triads $(j, k, l)$ of stimuli | (a) Which is the most similar pair?<br>Say: $(k, l)$<br>(b) Which is the most similar (MS) pair, and the least similar (LS) pair?<br>Say: MS is $(j, k)$ and LS is $(k, l)$ | $d_{kl} < d_{jl}$ and $d_{kl} < d_{jk}$<br>$d_{jk} < d_{jl} < d_{kl}$ | *Burton and Nerlove 1976<br>*Cochran and Cox 1951<br>Torgerson 1958 |
| 6 TETRADS | All pairs of pairs $((j, k)$ and $(l, m))$ of stimuli | Which pair is the more similar?<br>Say: $(l, m)$ MS than $(j, k)$ | $d_{lm} < d_{jk}$ | *Cochran and Cox 1951<br>Torgerson 1958 |

*References marked with asterisk contain information on incomplete designs to reduce number of presentations.

(iii) *Fixed sorting* The objects are allocated to a pre-specified number of categories.

(iv) *Free sorting* (or 'own categories') The objects are allocated to a set of categories, but the number of categories is not specified, and can in principle range between one in which all the objects are lumped together in one category and the case where each object forms its own category (Arabie and Boorman 1973).

Such data occur in a wide variety of disciplines, and are especially prevalent in cognitive studies, such as folk taxonomies in anthropology (Tyler 1969), psychosemantics (Miller 1969), subjective classification in sociology (Burton 1972, Coxon and Jones 1978b), and in personal construct theory in psychology (Bannister and Mair 1969). Some examples of free sorting from a number of different disciplines are:

the co-existence of plant species (objects) within chosen field sites (categories);
the sleeping habitat (e.g. trees, which here constitute the 'categories') of a group of monkeys (the 'objects');
the co-location of a set of artefacts (objects) within a set of neolithic graves (categories)
the sorting of a set of words (objects) into piles (categories) in terms of their similarity of meaning;
the co-occurrence of themes (objects) within a set of documents or sentences (categories).

Partitions (nominal) data are usually pre-processed before being scaled. Often, each partition is turned into a matrix of co-occurrence between the objects, where an entry of '1' in $\delta_{jk}$ means that objects $j$ and $k$ both occur in the same category, and '0' otherwise, and the individual matrices are then summed. The analysis of aggregate co-occurrence matrices will be discussed in 2.2.3.3. Another alternative, when interest focuses chiefly on how similar partitions are to each other, is to compare them two at a time, and a dissimilarity measure is then computed for each pair. This is discussed under 2.2.3.5.

Hierarchies data, like partitions produced from the method of sorting and the 'tree-construction' method employed by Fillenbaum and Rapoport (1971, pp. 10–11, 15 et seq.), can be thought of as another way of getting a full set of similarity judgments from a subject without making the task too strenuous. The analysis is similar to that used for comparing partitions (see Coxon and Jones 1978b, ch. 7).

### *2.1.3 Rankings and ratings data*

*Rankings* correspond at the individual level to the ordinal scales of measurement. They are one of the most popular methods of data collection in the social sciences, yielding a considerable amount of information at relatively little cost. The basic form (strict order ranking) has usually been obtained by presenting the subject with a set of objects or stimuli and defining a criterion by which the subject is to make his or her judgments. The subject is then instructed to choose the object which is highest on the specified criterion, followed by the next highest object and so on until a complete order is obtained, with no ties allowed. Several variants exist, especially the *weak order* (where objects may be tied, or treated as equal in terms of the criterion), and the *partial order* (where the subject is allowed to omit some

objects). As we have seen, rankings may also be derived from a set of pair comparison dominance judgments if the subject is sufficiently consistent.

*Ratings* are often obtained by asking the subject to assign a number to each object in such a way that the magnitude of the number reflects her estimate of the extent to which it possesses the attribute defined by the criterion. Common variants of this are where the subject is asked to 'mark each stimulus out of 100' (percentage or 'thermometer' rating, so named because the rating scale is represented to the subject in the form of a thermometer scale), and the 'graphic scale', where the subject positions (say) an arrow to mark her rating of the object, and the investigator then measures the location with a ruler.

The assumption is that each rating judgment is made in relation to the other stimuli which are presented, but in accordance with some absolute standard. This method is known in the psychometric literature as magnitude estimation or direct estimation (Stevens 1966). Quite often, the researcher decides to ignore the numerical information and turn the subject's ratings into a rank order.

Both rankings and ratings give rise to 'rectangular' or 'conditional similarity' data—that is, each subject's rankings or ratings are not considered to be comparable directly to those of other subjects.

### 2.1.4 Triadic data

Triadic data (where the subject selects the most similar pair out of three objects) are very commonly collected by psychologists and others using Kelly's 'repertory grid analysis', which is also used to elicit the constructs which people use in interpreting their social world (Kelly 1955; Bannister and Mair 1969). A common, but dangerous practice is to turn triadic data of this sort into 'vote frequencies', counting the number of times a particular pair is judged more similar than another. Roskam (1970) has shown that this practice can badly distort the information in the data, and should be avoided. Another variant of triadic data occurs where the subject is presented with three objects and then asked to select the most similar pair *and* the least similar pair. In this variant, much more information is obtained than in the first case. Suppose the subject is given the triad $(A, B, C)$ and chooses $AC$ as the most similar and $AB$ as the least similar pair. This implies that $d(A, C) < d(B, C) < d(A, B)$. If the subject is only asked to select the most similar pair we should only be able to infer in this example that $d(A, C) < d(A, B)$ and $d(A, C) < d(B, C)$ but we could infer nothing about the relationship between $(A, B)$ and $(B, C)$ (see 6.1.3).

Although triadic data collection appears to be a powerful and efficient method, and one which is well suited to obtaining personal constructs, it is rarely so in practice. Even when incomplete designs are used for reducing the number of judgments, subjects tend to find the task tedious, and their constructs often become highly stereotyped. Moreover, when several subjects' data are pooled, they cannot usually be represented well by scaling models (see Coxon and Jones 1978a, p. 66 and 1979, T3.1 and T3.11).

### 2.1.5 General issues in direct data collection: numbers and complexity

By and large, the most frequently encountered problems in collecting and scaling

individual sets of data centre around the number of objects and the cognitive complexity of the data collection task. Usually a trading relationship exists between these two characteristics: the more complex the task, the fewer the number of objects that can be employed. Statistical designs for reducing the number of objects to be presented to the subject are certainly useful and are often an elegant solution. But balanced incomplete designs only exist for some types of data collection and for certain numbers of objects, and the user must be prepared to use more practical and rough-and-ready procedures if the research problem really merits it. This topic is discussed further in sections 7.2.1.1 and 7.5.5.3.

The cognitive complexity of tasks poses rather different problems. Experience shows that one's presuppositions about the ease and speed with which a data collection task is completed can be misleading. For instance, ranking turns out to be a far more difficult task than rating, since constant re-ordering and comparison is necessary to yield an ordering, whereas a rating can be made easily, and each object can be judged separately and without repeated comparison.

In a similar way, triadic judgment seems very well suited to eliciting constructs (or bases of judgment) *and* to producing fairly complex (ordered metric) data in a simple format. And so it is, methodologically speaking—except that subjects often find it an extremely wearisome task, and tend to 'lock in' on a single construct (Coxon and Jones 1979, T2.4). By contrast, hierarchy construction—a very complex and time-consuming task—was found to be an interesting and rewarding task yielding rich and reliable data.

Rao and Katz (1971) provide more systematic evidence of such factors in a study of the effectiveness of seven data collection methods, evaluating each method by a simulation method using data from a known configuration. They find that ordering and selection (or 'pick $k$ out of $n$': Coombs 1964, ch. 2) methods such as pair comparisons, triads and tetrads produce better recovery of the original configuration than sorting methods, but that hierarchy construction is superior to the other sorting procedures.

## 2.2 Aggregate Data

Most frequently, the measure of dissimilarity used as input to MDS programs is an aggregate measure (summed over individuals, replications, times, locations etc.), and usually it is also an index of association (typically a measure of correlation or contingency).

As a methodological principle, the inspection of individual differences should always precede aggregation. If the individuals (or other units of analysis) differ systematically among themselves with respect to variables of interest, then such information is lost upon aggregation. Indeed, if data are aggregated, one can never know whether or not such differences even exist. Moreover, if significantly different subgroups do exist then any averaged information will be an artefact and will not reproduce the characteristics of either group accurately. There is always the danger of 'piecemeal distortion' as well. If the data referring to a given unit or individual is complex, then 'local structure' (interrelationships within parts of the data) can be lost entirely when the components are aggregated. It is sometimes argued that individual variation is simply unnecessary noise (or error) which will cancel out on aggregation. Perhaps it will, but such a belief requires a degree of well-behavedness

on the part of error that, however commonly assumed in social science modelling, is scarcely realistic and should at least be investigated.

In any event, a cautious approach is to be preferred. One way is to concentrate attention initially on examining each subject's data structure (meaning here simply 'set of pairwise judgments', or 'rankings', or 'triads', or whatever), and then compare the entire structure of different individuals before examining the aggregate structure of the objects.

In order to simplify matters, we shall assume that the basic data matrix, **X**, from which these summary measures will be computed is a rectangular matrix, with $N$ rows (usually representing individual subjects or units) and $p$ columns (each representing a separate variable or attribute) and whose element $x_{ij}$ gives the value on variable $j$ for individual $i$.

*Level of measurement*

Since each measure of association takes into account the level of measurement of each of the variables involved, it is convenient to distinguish measures intended for interval, ordinal and nominal data (counting dichotomous and co-occurrence measures as special cases of nominal data). The meaning of the word 'association' changes according to the level of measurement, and so it makes sense to compare directly only those measures which summarise data at the same level. Attention will be restricted to *symmetric* measures of association, although asymmetric measures can be represented in scaling models (see below).

*R and Q analysis*

Measures of association also differ as to whether they summarise the similarity between pairs of variables (columns of the data matrix), or between pairs of subjects (rows of the data matrix). The first type is often termed R-analysis, and the second Q-analysis. In some cases a measure can be used in either way, but usually a particular measure is designed for one form of analysis rather than the other.

The list of measures presented in the subsequent sections makes no claim to be exhaustive; measures are chosen either because of their obvious suitability in representing similarity for scaling models, or because they are in common usage among behavioural scientists. More extended treatment of association measures is provided in Galtung (1967, pp. 205–33), Loether and McTavish (1974, pp. 185–262), Blalock (1972, chs. 13, 15 and 18) and Wishart (1978, chs. 28 and 29).

### 2.2.1 Interval level measures

The most commonly encountered measures of similarity for interval (and higher) level data are the product moment (PM) family of coefficients, each of which can be used in either R- or Q-analysis mode. Product moment measures are all basically *vector* measures (see Appendix A2.1) where the similarity between two variables (for R-analysis) or subjects (for Q-analysis) is represented by the combination of

(a) the length of the vectors, and
(b) the product (inner, or scalar product) of the two vectors, represented by the size of the angle separating them.

The most familiar and frequently used product moment measures of similarity are *covariance* (where the scale units of the variables enter into the assessment of

similarity) and the Pearson product moment *correlation coefficient* (where the variables are standardised to unit length, and only the angular separation is considered). In neither case does the measure represent *distance* between the variables, although it is possible to convert vector separation (product moment) measures into distances (Appendix A2.1), and vice versa (Appendix A5.2). Vector and distance representations of similarity are described and compared in Appendix A2.1, where it is shown that since the correlation coefficient is a monotonic transformation of distance, it may be used directly in the basic *non-metric* MDS distance model.*

The basic matrix of scalar products (the product moment matrix) gives the information from which variance, standard deviations, covariances and correlations are produced, and is widely used in descriptive and multivariate statistics. From the data matrix **X**, two related product matrices can be formed:

the *minor* product matrix, $\mathbf{X}'\mathbf{X}$, a symmetric matrix of order $p$, whose entries give the scalar products (sum of squares and cross products) *between the variables* (R-analysis);

the *major* product moment matrix, $\mathbf{X}\mathbf{X}'$, a symmetric matrix of order $N$, or the scalar products between the subjects (Q-analysis or profile analysis).

Both matrices have a number of important and desirable statistical properties in common (see Green and Carroll 1976, p. 227 et seq.). In particular, they are of the same rank, which in the MDS context means that the data can be fully represented in a space of at most $m - 1$ dimensions, (where $m$ is the smaller of $N$ and $p$).

A number of variants of the basic PM matrix are in common use as measures of similarity:

*1 Deviate PM matrix*

Where the original data values have been 'centred', by having the column mean subtracted from each variable value, thus forming the matrix $\mathbf{X}_d$ of 'deviate scores': $x_d = (X_{ij} - \bar{X}_j)$. This has the effect of removing the overall average effect of each variable. (In the case of Q analysis, read 'subject' for 'variable', and 'row' for 'column' in the previous—and subsequent—sentences). The PM matrix formed from the deviate matrix is often termed the matrix of corrected (or deviate) squares and cross products (CSCP).

*2 Dispersion (variance-covariance) matrix*

When each entry in the deviate matrix is divided by $N$ (or by $p$ in Q-analysis), the PM matrix formed from it contains *variances* (averaged sum of squared *deviations*) in the diagonal elements, and *covariances* (averaged sum of corrected cross products) in the off-diagonal elements.

In both the deviate and dispersion PM measures, the units in which the original variables are scaled contribute directly to the overall measure of similarity (so that measurement of height in terms of metres will produce drastically reduced similarity values compared to measurement in centimetres).

*Vector separation and distance are not *linearly* related, so conversion between the two types of measure is necessary in metric scaling, and is sometimes an option provided within computer programs, such as INDSCAL.

*3 Correlation matrix*

If each variable is standardised (i.e. the score is centred and the resulting deviate value is then divided by the standard deviation of the variable) to normal scores $z_{ij} = (X_{ij} - \overline{X}_j)/\sigma_j$, the resulting PM matrix contains ones in the diagonal (representing the total variance), and the Pearson correlation coefficient $r_{ij}$, in the off-diagonal elements.

In calculating the correlation coefficient, each variable has been reduced to a common unit of measurement (its standard deviation) and differences in variance between the variables have hence been removed.

A useful comparison of these PM measures is provided by Skinner (1975), who analyses them in terms of three independent components:

1 *elevation* ($m_j$)—the average or mean effect of the variable $j$, so named because when a set of variable scores is drawn as a profile, the removal of the mean removes the elevation or average height of each variable:

2 *scatter* ($s_j$)—the dispersion, measured by the standard deviation of the variable $j$: and

3 *shape* ($r_{ij}$)—measured by the correlation coefficient, representing simply the angular separation of two variables $i$ and $j$.

In these terms, if original ('raw') data are converted into deviate data, information due to elevation is eliminated. Similarly, standardising the data has the effect of equalising the scatter of all the variables.

Each PM measure can then be broken down into its constituent components, which are related in the following way:

| *Product Moment Measure:* | *Components* | *Comment* |
|---|---|---|
| 1 Original (Raw scores) | $m_i m_j$ + $s_{ij}$ × $r_{ij}$ = Elevation + (scatter × shape) | Mixes all 3 components, merges scatter and shape |
| 2 Variance-Covariance | $s_i s_j$ × $r_{ij}$ (scatter × shape) | Removes elevation, merges scatter and shape |
| 3 Correlation | $r_{ij}$ | Purely shape |

Skinner's treatment is specially relevant in dealing with Q-analysis and profile data, when the researcher wishes to compare subjects in terms of their pattern of scores across a given number of items (test scores, semantic differential concept ratings etc.) Several of the programs in the MDS(X) series give the user the option of centring and standardising the subjects' data where this is appropriate.

### *2.2.2 Ordinal measures of association*

Measures of ordinal (or monotonic) association address the question: To what extent do variables $X$ and $Y$ rank individuals (or whatever) in the same way? Perfect positive association occurs where individuals are ranked in the same order on both variables, and perfect negative association represents a total inversion in ordering. On this, all measures of ordinal association agree. But difficulties arise in giving meaning to intermediate degrees of association, and this is due to two factors:

(i) whether ranks are to be considered as numerical quantities (on which arithmetic operations may legitimately be performed) and

(ii) whether 'the same order' is to be interpreted in a strict or weak sense.

A number of relevant measures are summarised in Table 2.2.

Ironically, the earliest pioneering work on ordinal association (Spearman 1904, see Kendall 1962) produced a measure which is not strictly an ordinal measure at all!

*Spearman's rho* (Table 2.2(1) )

This rank correlation coefficient is the product moment correlation between ranks, considered as integer quantities. The measure compares two rank orderings, and is based upon the (squared) *difference* in rank positions. It thereby measures not only the inversions which occur in two orderings but also the numerical size of the differences. Rho is invariant under linear transformations of the data but it is *not* invariant under monotone transformations, and it is therefore an interval level measure. It varies between $-1$ (when one ranking is the reverse of the other) and $+1$ (perfect agreement).

#### *2.2.2.1 Ordinal measures based on inversions in rankings*

Usually, ordinal variables are weak orderings consisting of a fairly small number of ordered categories, such as high, medium and low levels of motivation, or Likert's five response categories for attitude items (strongly agree, agree, not sure, disagree, strongly disagree). Consequently it usually happens that a large number of subjects share the same ordinal value, i.e. they are 'tied' with respect to the variable concerned. Different measures of ordinal association treat tied data in different ways.

The basic idea underlying genuine measures of ordinal relationships is that of comparing each pair of individuals on the two variables, and seeing how often they are ranked in the same way. An example will help clarify the concepts involved. Suppose we have data for ten individuals on the ordinal variables $X$ and $Y$, (where H, M and L stand for High, Medium and Low respectively):

| | *Variables* | | | *Variables* | |
|---|---|---|---|---|---|
| Individuals | $X$ | $Y$ | Individuals | $X$ | $Y$ |
| 1 | M | M | 6 | H | L |
| 2 | H | M | 7 | H | H |
| 3 | M | H | 8 | M | L |
| 4 | L | L | 9 | L | H |
| 5 | L | M | 10 | H | M |

| *Name* | *Formula* | *Maximum* | *Zero* | *Minimum* | *Advantages* | *Disadvantages* | *Reduces to* |
|---|---|---|---|---|---|---|---|
| 1 SPEARMAN'S RHO ($\rho$) | $1 - \frac{6\Sigma d^2}{n^3 - n}$ | +1 | Both rankings identical | −1 | Takes size of inversions into account | Treats ranks as interval quantities | — |
| 2 GOODMAN AND KRUSKAL'S GAMMA ($\gamma$) | $\frac{C - D}{C + D}$ | +1 | Equal balance between concordant pairs | −1 | Measures *weak* monotonicity (no reversals) | | |
| 3 KENDALL'S TAU ($\tau$) | | | | | | | |
| (a) | $\frac{C - D}{C + D + T_x T_y + T_{xy}}$ | +1 | Equal balance between concordant pairs | −1 | Best defined for *strict* rankings | Ties considered as errors | — |
| (b) | $\frac{C - D}{\sqrt{C + D + T_y}\sqrt{C + D + T_x}}$ | | | | Takes ties into account | Reaches maximum only for square tables | $\tau_a$ when no ties |
| (c) | Corrected version of tau$_a$: (Wilson 1974, p. 352) $\sigma_a\{n - 1)/n\}$ $\{m/(m - 1)\}$ where $n$ is the number of cases and $m$ is the number of cells in longest diagonal | | | | Takes different number of ranks into account | 'Quick and dirty way' of extending $\tau_b$ to non-square tables | $\sigma_b$ when same number of categories |
| 4 WILSON'S $e$ | $\frac{C - D}{C + D + T_y + T_x}$ | +1 | Equal balance between concordant pairs | −1 | Measures *strict* monotonicity | | |

Table 2.2 *Ordinal measures of association*

In comparing pairs of individuals there are five possibilities (see Wilson 1974 for an extended exposition):

1 *Concordant pairs* (C) ($X$ and $Y$ order the individuals in the *same* way). If $i$ is higher (lower) than $j$ on $X$, then $i$ is higher (lower) than $j$ on $Y$ (for instance, as occurs in comparing individuals 7 and 1, or 3 and 5).

2 *Discordant Pairs* (D) ($X$ and $Y$ order the individuals in *opposite* ways). If $i$ is higher (lower) than $j$ on $X$, then $i$ is lower (higher) than $j$ on $Y$ (e.g. as occurs in comparing 3 and 2, and 5 and 6).

3–5 *Tied Pairs* ($X$ and $Y$ share at least one *tied* value)

3 *Tied on* X ($T_x$) (e.g. as occurs in 9 and 5, and 8 and 3)

4 *Tied on* Y ($T_y$) (e.g. as occurs in 8 and 4, and 2 and 5)

5 *Tied on both* X *and* Y ($T_{xy}$) (identical values) (as occurs in comparing 2 and 10).

Measures of ordinal association all have the same basic form:

$$\frac{\text{(numerator)}}{\text{(denominator)}} = \frac{C - D}{N}$$

where $(C - D)$ is the difference between the number of concordant and discordant pairs, and $N$ is the number of pairs which are considered to be relevant to the measure. (The denominator changes from measure to measure).

In terms of the numerator, the measures are either

*weakly monotonic*, allowing ties to count as concordant pairs, so that if $i$ is higher (lower) than $j$ on $X$, then $i$ is at least as high (low) as $j$ on $Y$; or

*strictly monotonic*, insisting that both inequalities *and* ties must be matched, so that:

if $i$ is higher (lower) than $j$ on $X$, then $i$ must be higher (lower) than $j$ on $Y$, and

if $i$ is tied to $j$ on $X$, then $i$ must be tied to $j$ on $Y$.

*Goodman and Kruskal's gamma* (Table 2.2(2); Goodman and Kruskal 1954)
This widely-used index measures *weak monotonicity* between two variables, and was expressly designed for summarising cross-tabulations of data. It is defined as the ratio of the *difference* of concordant and discordant pairs to the *sum* of concordant and discordant pairs. It therefore completely ignores ties on both $X$ and $Y$ and is described by Wilson (1974, p. 331) as the 'measure of the extent to which the data fit a "no reversals" (weak monotonicity) hypothesis'. In the case of 2 × 2 tables, gamma reduces to the Q-coefficient, discussed below under 'dichotomous measures'.

*Kendall's tau measures* (Table 2.2(3); Kendall 1962)
The tau measures have been described as 'coefficients of disarray', and are also based on the difference of the number of concordant and discordant pairs. They differ from gamma in that they expressly take into account *all* pairs, whether tied or not. They are therefore measures of the strong monotonicity hypothesis and can also be interpreted in terms of the number of interchanges necessary to transform one ranking into another. Tau was defined initially in terms of comparing two rank orderings (tau *a*): it was then extended to R-analysis of square cross-tabulations

(tau *b*), and to cross-tabulations with unequal numbers of rows and columns (tau *c*). Tau measures reduce to phi (q.v. *infra*) for the 2 × 2 table and lie in the same range (−1, +1) as rho. But there are some difficulties in interpreting tau when it is zero, and in stating the conditions under which the maximum is attained in the case of non-square tables.

*Wilson's* e: Table 2.2(4); Wilson (1974)
This coefficient resembles the tau family in that it is a test for strict monotonicity, but whereas the denominator of $\text{tau}_b$ (which it most closely resembles) is $(\sqrt{\{C + D + T_x\}}\sqrt{\{C + D + T_y\}})$ the denominator of $e$ is the much simpler quantity, $(C + D + T_y + T_x)$. Whilst $e$ can never be greater than $\text{tau}_b$, its main property is its greater sensitivity to the extent to which data cluster round the main diagonal of a cross tabulation.

The relationship between these properly ordinal measures of association has been investigated—appropriately enough using MDS—by Maimon (1978), who shows the importance of the strict *vs* weak monotonicity distinction and the number of separate categories of the variables involved in accounting for differences between the measures.

### 2.2.3 Nominal level measures

A nominal scale consists simply of the division of a set of objects into a set of mutually exclusive and exhaustive categories, technically referred to as a partition. Three things are relevant to the analysis of such data:

(i) *how many categories there are*—i.e. the 'fineness' of the partition, or the degree of discrimination, from the simplest dichotomy (2-state) to the multiple-state polytomy;

(ii) *how the objects are distributed over the categories*—i.e. the 'shape' of the partition, reflecting how common or how rare the occurrence of frequency of each category is;

(iii) *what the composition of the categories is*—i.e. the 'content' of the partition, indicating which particular objects occur within a category.

Partitions are rarely scaled as they stand. More typically, they are compared two at a time, and some measure of association is defined to summarise their (dis)similarity. In the case of R-analysis, two nominal level *variables* will be compared by means of a 2-way contingency table, where the rows represent the categories of one partition (say, sex), the columns represent the categories of the other partition (say, political affiliation), and the entries in the table consist of the number of subjects who fall in both the row category and in the column category. A large number of measures of association exist for assessing the similarity between the two variables, based upon the information in such tables. Many such measures are suitable for analysis by MDS, and are discussed below.

A particularly interesting special case occurs when the variables are dichotomies, where the contingency table is 2 × 2. Often the 'variables' in this case represent the presence or absence of some property, and the researcher is interested in the extent to which the two properties occur together (for instance, to what extent do two species of plant tend to grow in close proximity in a number of sites? Or, do two

items in an attitude test evoke the same response in a sample of subjects? Or, do two coders agree in their identification of a theme in respondents' answers to a questionnaire?). Sometimes it will be sufficient simply to count how often the two properties occur together ('co-occurrence data', discussed in section 2.2.3.3), but in other cases the extent of disagreement will also be of interest, and a 'matching coefficient' will be necessary to express the overall similarity of the two dichotomies (see section 2.2.3.2).

In Q-analysis of nominal data, attention is focussed principally on comparing the structure of the individual partitions, taken two at a time. For instance, suppose a sample of subjects has been asked to sort a set of six objects into classes or categories of their own choosing, and the researcher wishes to examine how similar her subjects' classifications are. An important step will be to form a contingency table (as in R-analysis) but one which has the first subject's categories as the rows, and the second subject's categories as columns. The entries in the table will be the objects which are common to both the row category and the column category.

As in R-analysis, a number of measures exist for summarising the (dis)similarity between the two partitions, and these are discussed in section 2.2.3.4. We have written as if each partition in the Q-analysis comes from a separate subject. There is in fact no reason why the partitions should have been produced by individuals at all. Equally well, the subjects could be different times, occasions, methods, locations etc. (e.g., how reliable is a particular subject's classification system over a number of retests; or in different circumstances; or with different interviewers? How similar are census classifications of occupations in different countries, or over a number of revisions within the same country? How similar are the psychiatric diagnoses of a set of patients by doctors who have been trained in different traditions?)

*2.2.3.1 Chi-square based measures*

The most commonly-used family of measures compares the observed frequency of objects in each category with that which would be expected by chance. In the case of a cross-tabulation of two variables, the number expected by chance to be in class $i$ of variable $X$ and class $j$ of variable $Y$ ($f_{ij}$) is defined by statistical independence

$$f_{ij} = f_i f_j / N$$

where $f_i$ is the total number in class $i$, $f_j$ is the total in class $j$, and $N$ is the total number of cases. Put in terms of proportions, independence is defined more simply as:

$$P_{ij} = P_i P_j$$

(or its equivalent $P_{ij} - P_i P_j = 0$).

The chi-square coefficient itself is a much-used test for statistical independence. But the value of chi-square is proportional to the number of cases, so it cannot serve as a measure of association for comparing groups of different sizes. Several attempts have been made to devise a measure of association based upon chi-square which will vary between 0 and 1 (the 'direction' or sign (+, −) of a relationship is meaningless in the case of nominal data, since the classes may be arranged in any

order), and will not depend on the number of subjects or the number of categories. The attempts to norm chi-square have not been entirely successful, but the measures presented in Table 2.3 are used fairly frequently in scaling analysis.

($\chi^2$ is the chi-square value; $N$ is the total number of objects;
$r$ is the number of rows in the 2-way table;
$c$ is the number of columns, MIN is the smaller of $r$ and $c$).

*2.2.3.2 Measures for dichotomies*

Dichotomous variables simply differentiate the presence and the absence of an attribute. Paradoxically, a dichotomy can be considered as being a nominal, ordinal or interval level variable:

(i) The categories of a nominal level variable can always be converted into a set of dichotomies (thus the 4-fold religious categorisation 'Protestant, Catholic, Jew, other' can be converted into three* dichotomies: Protestant/not; Catholic/not; Jew/not.

(ii) Presence/absence can be thought of as a particularly simple ordering.

(iii) Since there is only a single difference (presence/absence) the numbers (1, 0) (or any linear transformation of them) can represent the two states quite legitimately, and indeed it is a common practice in regression and related linear models to follow this convention, calling them 'dummy variables'.

In constructing measures of association between two dichotomous variables, attention has been concentrated chiefly upon the question of 'matching'. This can best be illustrated by inspecting the 2 × 2 frequency table:

| | | *Property Y* Yes | No | |
|---|---|---|---|---|
| *Property X* | Yes | $a$ | $b$ | $(a + b)$ |
| | No | $c$ | $d$ | $(c + d)$ |
| | | $(a + c)$ | $(b + d)$ | $N(= a + b + c + d)$ |

Cells $a$ and $d$ represent positive and negative matches respectively: '$a$' gives the number of individuals or objects who possess both property $X$ and property $Y$ (the positive matches) '$d$' signifies those who possess neither $X$ nor $Y$ (the negative matches).

Cells $b$ and $c$ represent mismatches, individuals who possess one, but not the other property.

The large number of measures of association differ in large part in terms of (i) whether the 'negative matches' should enter into the assessment of similarity (i.e. are those who do not have either attribute even relevant in comparing properties?) and (ii) what weight should the matches and the mismatches have in defining the degree of similarity?

*N.B. One category must be omitted in converting nominal scale to dichotomies, since the response to the omitted category is perfectly predictable from knowing the response on the others. For instance, knowing that someone is not Protestant, not other and not Jewish implies that the person is Catholic since the categorisation must be exclusive and exhaustive to qualify on a nominal scale.

| *Name* | *Formula* | Limits | | | *Advantages* | *Disadvantages* | *Reduces to* |
|---|---|---|---|---|---|---|---|
| | | *Maximum* | *Zero* | *Minimum* | | | |
| CHI SQUARE ($\chi^2$) | $\Sigma \frac{(f_o - f_e)^2}{f_e}$ | $N$(MIN-1) | Statistical Independence (SI) | 0 | Zero in case of independence Serves as basis for other measures | Non normed Dependent on $N$ | — |
| PHI ($\varphi$) | $\sqrt{\left(\frac{\chi^2}{N}\right)}$ | $\sqrt{}$ (MIN-1) | SI | 0 | Normed, Independent of $N$ Reaches maximum for 2 × 2 tables | Can exceed 1 as maximum when MIN > 1 | PM Correlation for 2 × 2 table |
| PEARSON'S CONTINGENCY COEFFICIENT $C$ | $\sqrt{\left(\frac{\chi^2}{\chi^2 + N}\right)}$ | Depends on $r$ and $c$ | SI | 0 | Normed, Independent of $N$ | Can neither exceed, nor reach, 1 as maximum | |
| TSCHUPROW'S COEFFICIENT $T$ | $\sqrt{\left(\frac{\chi^2}{N\sqrt{(r-1)(c-1)}}\right)}$ | Depends on $r$ and $c$ | SI | 0 | Normed, Independent on $N$ Reaches maximum (1) for square tables | Cannot reach 1 as maximum in non-square tables | Phi (in 2 × 2 case) |
| CRAMER'S COEFFICIENT $V$ | $\sqrt{\left(\frac{\chi^2}{N(\text{MIN}-1)}\right)}$ | Unity | SI | 0 | Normed, Independent of $N$ Reaches maximum, even for non-square tables. Independent of number of rows and columns | | $T$, if $r = c$ in 2 × 2, and 2 × $k$ cases |

Table 2.3 *Chi-square based measures of association*

The superabundance of measures of dichotomous association is due in large part to their importance in numerical taxonomy, where the basic crucial operation is often the comparison of pairs of OTUs (operational taxonomic units) or individual organisms who share a number of attributes to a greater or lesser extent (Sokal and Sneath 1963; Sneath and Sokal 1973) and in the social sciences, where dichotomous variables abound, though here use has generally only been made of a very restricted number of coefficients (but see Lazarsfeld and Henry 1968 and Rasch 1960 for a wide variety of models based upon the covariation of dichotomous data).

Sokal and Sneath (1963) present a particularly useful classification of measures of (dis)similarity for dichotomous data, which in simplified form provides the basis for Table 2.4. The measures all take the form of a ratio between the number of 'matches' (numerator) and the elements considered to be the relevant reference set (the denominator). They differ in two major respects:

(i) *How 'matching' is to be defined in the numerator*—in particular whether the negative matches (cell d) are to be excluded (I), included (II), or whether the numerator should take into account matched *and* unmatched pairs (III).

(ii) *What weight is to be given to the relative preponderance of matched and unmatched pairs.* Here there is greater variety, with quantities such as marginal totals (e and h) and the sum of cross products (g) entering the definition of the denominator.

Brief comments on the properties of some of these measures are given below:

*Measure number:*

1 Represents the conditional probability that a pair of objects will both have a randomly chosen variable, and is one of the simplest and longest used coefficients, which excludes negative matches.

2 A curious measure which by implication treats negative matches as generically different from positive matches.

3 The 'simple matching coefficient', which includes negative matches.

7 This measure was defined originally for polytomies, and allows missing data.

8 and 9 Despite their apparently simple interpretation, both these measures have the unfortunate property of being normed between 0 and *infinity*, unlike the other measures which all have an upper limit of unity.

14 Unlike the other measures, this bases association on the preponderance of matches over mismatches.

15 Used extensively in social science data analysis, and based like phi on the determinant of the table. It shares the unfortunate property with phi that if no negative matches occur (i.e. if $d = 0$), then an association of zero results.

16 Phi has been extensively discussed, and is widely used as the dichotomous equivalent of the Pearsonian PM correlation coefficient.

refs ← Gower & Legendre '86 → $\frac{a-(b+c)+d}{a+b+c+d}$ (1) | $\frac{a+d}{a+\sqrt{2}(b+c)+d}$

Everitt + Rabe-Hesketh '97 (Am. & Prac. Data).

| DENOMINATOR† | NUMERATOR<br>(I) EXCLUDES negative matches ($d$) | (II) INCLUDES negative matches ($d$) |
|---|---|---|
| (a) $m$ and $u$ equally weighted | 1 *Jaccard* $a/(a+b+c)$<br>2 *Russell and Rao* $a/(a+b+c+d)$ | 3 *Sokal* $(a+d)/(a+b+c+d)$ matching coeff. |
| (b) $m$ given twice weight of $u$ | 4 *Dice* $2a/(2a+b+c)$ | 5 (*s.n.*)* $2(a+d)/(2(a+d)+b+c)$ |
| (c) $u$ given twice weight of $m$ | 6 ~~(s.n.)~~ Sokal + Sneath '63 * $a/(a+2(b+c))$ | 7 *Rogers and Tanimoto* $(a+d)/(a+d+2(b+c))$ |
| (d) $u$ only | 8 *Kulczynski* $a/(b+c)$ | 9 (*s.n.*)* $(a+d)/(b+c)$ |
| (e) marginal totals | 10 *Kulczynski* $\frac{1}{2}\{a/(a+c)+a/(a+b)\}$<br>12 *Ochiai* $a/\sqrt{\{(a+c)(a+b)\}}$ | 11 (*s.n.*)* $\frac{1}{4}\{a/(a+c)+a/(a+b)+d/(b+d)+d/(c+d)\}$<br>13 (*s.n.*)* $ad/\sqrt{\{(a+c)(a+b)(b+d)(c+d)\}}$ |

| DENOMINATOR† | NUMERATOR<br>(III) BALANCE of matched and unmatched pairs |
|---|---|
| (f) $m$ and $u$ equally weighted | 14 *Hamann* $(a+d)-(b+c)/(a+b+c+d)$ |
| (g) sum of cross products | 15 *Yule's Q* $(ad-bc)/(ad+bc)$ |
| (h) marginal totals | 16 *Pearson's Phi* $(ad-bc)/\sqrt{\{(a+c)(a+b)(b+d)(c+d)\}}$ |

**sine nomine* (unnamed coefficient)
†*Note* $m$ signifies matched pairs ($a$ and $d$); $u$ signifies unmatched pairs ($b$ and $c$).

Table 2.4 *Measures of association between dichotomies* (based upon Sokal and Sneath 1963, pp. 125 et seq.)

*2.2.3.3 Co-occurrence measures*

Co-occurrence measures are all based upon the 'abundance matrix' (Kendall 1971a, p. 219) whose entries $s_{jk}$ assess the similarity between objects $j$ and $k$ in terms of the frequency with which objects $j$ and $k$ occur in (or are allocated to) the same category. Miller (1969, p. 171 et seq.) has shown that the simple measure: $(N - s_{jk})$, where $N$ is the total number of subjects or partitions, obeys the axioms of a distance metric (see A2.1.1) and is frequently used in MDS as the basic dissimilarity measure between objects. In some applications, the user may be advised to modify this simple measure by taking into account the size of the category in which each pair of objects occurs. Burton (1975, and see Coxon and Jones 1979, U2.10) has defined a family of four co-occurrence measures which differ in this respect:

*M1* Each individual co-occurrence contributes equally to the overall measure (the basic measure).

*M2* Each co-occurrence is weighted by the size of the category in which it occurs (thus emphasising gross discriminations).

*M3* Each co-occurrence is weighted *inversely* by the size of the category (emphasizing fine discriminations).

*M4* An information theoretic measure which also emphasises fine discriminations, and in addition, takes into account the number of times in which a pair of objects are sorted into *different* groups.

(These measures are defined in Appendix A.2). Burton (1975) examined how well measures M1, M2 and M4 can be scaled using the basic MDS model, and the effect which the differences between the measures have upon the structure of the final configuration of points. As in other applications (Burton and Romney 1975; Coxon and Jones 1979, U2.11), M2 (which emphasizes gross distinctions) gives the best fit, but is liable to collapse points into large clusters. M4 is less readily representable by the basic MDS model, but is probably the most satisfactory measure for MDS analysis, due both to its greater resistance to 'degeneracy' (i.e. the tendency to collapse points into clusters, ignoring significant information in order to get a better fit) and because, like M3, it attempts to take into account both the tendency for objects to occur together in some partitions *and* for them to be separated in others. In the sense that it balances concordant and discordant pairs, M4 resembles the ordinal measures of association.

*2.2.3.4 The index of dissimilarity between distributions*

Very commonly, data analysts wish to compare two distributions of a categorical variable—such as the incidence of a number of diseases in two countries, or in two social classes. A particularly simple measure of how (dis)similar two distributions are is provided by the 'index of dissimilarity' (see Blau and Duncan 1967, p. 43 et seq.) which is based simply on the percentage difference for each category and illustrated below.

Each distribution is first converted into percentage form (for comparability). Then for each category one calculates the difference (e.g., $12.3 - 25.8 = -13.5$ for category A). The absolute differences—ignoring the sign—are then added to form the basic index (here, 60.2). Clearly, if the percentage distributions had been

| Category | Group: I | Group: II | % diff. | Absolute % diff. (AD) |
|---|---|---|---|---|
| A | 12.3% | 25.8% | −13.5 | 13.5 |
| B | 5.8 | 13.2 | −7.4 | 7.4 |
| C | 54.2 | 40.3 | 13.9 | 13.9 |
| D | 22.4 | 6.2 | 16.2 | 16.2 |
| E | 5.3 | 14.5 | −9.2 | 9.2 |
| *Total* | 100.0 | 100.0 | 0.0 | 60.2 |
| $N$ = | 532 | 821 | | |

INDEX OF DISSIMILARITY (ID) BETWEEN GROUPS I AND II = $\Sigma$ AD/2 = 30.1

identical the value of the index would be 0 indicating 'zero dissimilarity'. If the index is halved, forming the index of dissimilarity (ID), it has a particularly simple interpretation: it represents the percentage of cases which would have to be moved between categories in order to change one distribution into the other—a notion encountered already in ordinal and other nominal measures of association. In this example, 30.1 per cent of cases would need to change categories if the distributions were to become identical. The index of dissimilarity is a distance measure: it is zero only if the distributions are identical, it is symmetric, and for comparisons between three distributions it obeys the triangle inequality. Moreover, its value is unaffected by re-arrangement of the order of the categories and is therefore appropriate to nominal data.

The ID measure has been frequently used to analyse 'flow data'—for example, mobility between occupational groups, migration between regions, input/output analysis between economic sectors, volume of diplomatic correspondence between countries, settlement of plant species in different sites. In many cases the flow is asymmetric in the sense that not as many objects move from category $a$ to category $b$ as do from category $b$ to category $a$. A common way of analysing these data is first to convert the raw frequency ('turnover') table into a row-percentaged table (for assessing 'outflow' movement from a given row category into the column categories) and secondly into a column-percentaged table (for assessing 'inflow' into a given column category from the row categories). This is illustrated in Table 2.5 for some (fictional) migration data between four cities.

The index of dissimilarity can now be used to summarise these rather complex flow data. In the case of outflow each pair of rows is compared by calculating the index (how dissimilar are cities $j$ and $k$ in terms of the destinations of their inhabitants?)—in this instance, cities B and C are most alike in their pattern of outmigration (ID = 7.8), and cities A and D are least alike (ID = 54.9). For inflow data, each pair of columns is compared (how dissimilar are cities $j$ and $k$ in terms of their in-migration or recruitment?). The values of the ID coefficients for both outflow and inflow are given in the lowest table.

The ID coefficient is used extensively as a prelude to MDS analysis (Blau and

I RAW AND CONDITIONAL PERCENTAGED CROSS TABULATION

(a) *Raw frequencies*

| | | To city A | B | C | D | Total |
|---|---|---|---|---|---|---|
| *From* city | A | 58 | 22 | 41 | 19 | 140 |
| | B | 30 | 38 | 14 | 23 | 105 |
| | C | 25 | 44 | 19 | 22 | 110 |
| | D | 7 | 51 | 12 | 51 | 121 |
| | Total | 120 | 155 | 86 | 115 | $N = 476$ |

(b) *Row-percentaged data ('outflow')*

| | A | B | C | D | Total |
|---|---|---|---|---|---|
| A | 41.4 | 15.7 | 29.3 | 13.6 | 100.0 per cent |
| B | 28.6 | 36.2 | 13.3 | 21.9 | 100.0 |
| C | 22.7 | 40.0 | 17.3 | 20.0 | 100.0 |
| D | 5.8 | 42.1 | 10.0 | 42.1 | 100.0 |

(c) *Column-percentaged data ('inflow')*

| | A | B | C | D | Total |
|---|---|---|---|---|---|
| A | 48.4 | 14.2 | 47.6 | 16.6 | |
| B | 25.0 | 24.5 | 16.3 | 20.0 | |
| C | 20.8 | 28.4 | 22.1 | 19.1 | |
| D | 5.8 | 32.9 | 14.0 | 44.3 | |
| Total | 100.0 | 100.0 | 100.0 | 100.0 | per cent |

II INDEX OF DISSIMILARITY VALUES
Above diagonal: based on row percentages ('outflow')
Below diagonal: based on column percentages ('inflow')

| | A | B | C | D | |
|---|---|---|---|---|---|
| A | — | 28.7 | 30.7 | 54.9 | |
| B | 34.7 | — | 7.8 | 27.0 | |
| C | 9.5 | 33.4 | — | 24.2 | |
| D | 38.5 | 13.8 | 34.0 | — | |

Table 2.5 *Calculation of index of dissimilarity*

Duncan 1967, p. 67 et seq.; Macdonald 1972, p. 213 et seq.), especially for asymmetric data of this sort.* It is further discussed in section 5.1.1.1 below.

*2.2.3.5 Similarity between pairs of partitions*

Although Q-analysis of sortings and partitions is a fairly recent development in MDS analysis, it is receiving increasing attention, especially following the

*Blau and Duncan further discuss the possibility of leaving out the diagonal elements corresponding to the 'stayers' in calculating an index of dissimilarity on the grounds that it is intended to assess the dispersion or flow, not the stability, between distributions.

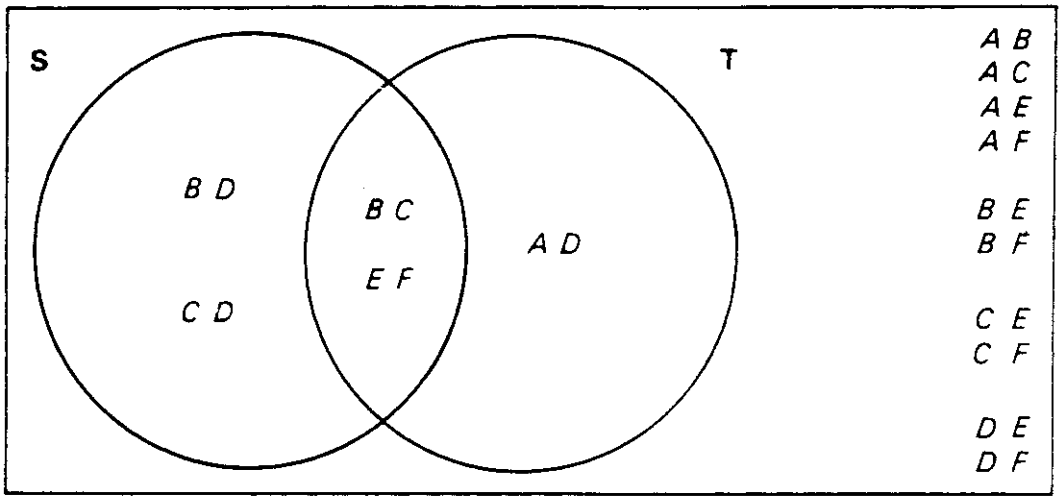
important methodological work of Boorman and Arabie (Arabie and Boorman 1973). Although the literature is at a fairly technical level, the basic concepts are surprisingly simple. The idea is to provide a measure of dissimilarity between any two partitions by examining how many moves it will take to change one partition into the other. As an example, take the following three partitions:

$$\begin{aligned} S &= \{A \quad \mid BDC \mid FE\} \\ T &= \{AD \quad \mid \quad BC \mid FE\} \\ U &= \{ACF \mid \quad B \mid E \mid D\} \end{aligned}$$

(The order of the categories, and of the objects within a category, is arbitrary). It is fully obvious that $S$ and $T$ are very similar, both in the number and in the composition of their categories, whilst they both differ substantially from $U$. The question is how different they are, and a quantitative answer depends entirely upon how a 'move' or 'change' is defined. If the move consists of single elements, then moving $D$ out of $(BDC)$ and into conjunction with $A$ will turn $S$ into $T$. However, if we wish to preserve information about what objects are linked together in the same category, then we shall at least need to move three *pairs* of elements—$(BD)$, $(CD)$, and $(AD)$—to change $S$ into $T$. And so on: the definition of a move depends upon how much structure one wishes to preserve intact in moving from one partition to another.

To show the usefulness and versatility of the Boorman-Arabie group of measures, the pairwise ('Pairbonds') dissimilarity measure between $S$ and $T$ will serve as an example. In all, there are 15 possible pair-linkages between the 6 elements: $\{A, B, C, D, E, F\}$. When these are enumerated and illustrated as a Venn Diagram, (Figure 2.1a) it can be seen that $S$ and $T$ *agree* that $(BC)$ and $(EF)$ go together and that a further 10 pairs do not go together. The disagreement between $S$ and $T$ is limited to three pairs: the pairs $(BD)$ and $(CD)$ (which occur in $S$ but not in $T$) and the pair $(AD)$ (which occurs in $T$ but not in $S$). These three pairs together define the distance, or Pairbonds dissimilarity measure, between $S$ and $T$: $d_{\text{pairbonds}}(S, T) = 3$. The representations of the Pairbonds measure in Figure 2.1 emphasise some important parallels with measures we have already discussed. First, Pairbonds is clearly another matching measure between two partitions, but is one which counts the *mismatches* in the pairs involved (i.e. the count $(b + c)$ in Figure 2.1c). Secondly, the Venn Diagram makes it clear that Pairbonds is in set-theoretic terms the symmetric difference of the pairs involved in the partitions $S$ and $T$. Indeed, Flament (1963: pp. 14–17) and Restle (1959) before him, discuss precisely this measure, prove that it is a metric, and Flament shows that it may usefully be employed to compare the dissimilarity of two graphs such as communication and friendship networks, and the dissimilarity of any two binary (0, 1) relations. (In this context, Pairbonds provides a natural Q-analysis comparison to the R-analysis measures of pairwise co-occurrence discussed in the previous section). Thirdly the formula for the Pairbonds dissimilarity measure, like many others discussed by Boorman, has a familiar form, akin to the cosine rule discussed in Appendix A2.1:

$$d(S, T) = m(S) + m(T) - 2m(S \text{ and } T)$$




(a) Venn diagram of pairs

T:

| Partition | *AD* | *BC* | *EF* |
|---|---|---|---|
| *A* | *A* | – | – |
| S: *BDC* | *D* | *BC* | – |
| *FE* | – | – | *FE* |

(b) Intersection table of objects

Pairs in T

| Pairs in S | Y | N |
|---|---|---|
| Y | 2 | 2 |
| N | 1 | 10 |

$N = 15$

(c) Matching table of pairs

Figure 2.1 *Pairbonds measure of dissimilarity between partitions*

In the case of Pairbonds, the measure $m$ consists simply of the number of pairs involved:

$$m(S) = 4 \quad \text{i.e.} \quad (BD), (CD), (BC), (EF)$$
$$m(T) = 3 \quad \text{i.e.} \quad (BC), (EF), (AD)$$

The table of the intersection between $S$ and $T$, which is similar to the contingency table in R-analysis, is given in Figure 2.1b. It is produced simply by cross-classifying the two partitions, and writing in the cell $(i, j)$ the elements which are in both category $i$ of $S$ and category $j$ of $T$. (Taken together, the entries in the table incidentally comprise a partition which is 'finer' than either $S$ or $T$, in the sense that both $S$ and $T$ can be built up from it). There are two pairs, namely $BC$ and $FE$, which occur in the intersection table, hence

$$m(S \text{ and } T) = 2$$

Thus: Pairbonds: $d(S, T) = m(S) + m(T) - 2m(S \text{ and } T)$
$= 4 + 3 - 2 \times 2$
$= 3$

A further eleven measures of dissimilarity between partitions are defined by Arabie and Boorman (1973) in a similar manner. They also investigate in a very

instructive manner the behaviour of these measures when submitted to MDS analysis. We shall examine some of their conclusions in Chapter 4. Examples of the use of the measures are provided in Arabie and Boorman (*op. cit*) and in Coxon and Jones (1979).

Unfortunately, partitions do not provide any information on the *relationship* between the categories, but it is a simple matter to supplement nominal information if the data are obtained directly from the subjects, either by asking them to place the categories in order (thus producing a weak ordering), or to continue merging categories in terms of their relative similarity, which produces a hierarchical clustering of the objects (see above). Boorman and Olivier (1973) have used the partition measures as a basis for constructing measures of similarity between hierarchies of this sort, and these are employed in Coxon and Jones (1978b) to scale differences between subjective hierarchies.

## 2.3 Summary

Two main criteria have been used to distinguish the profusion of measures of association appropriate for scaling analysis:

(i) direct *vs* derived (aggregate) measures;
(ii) the level of measurement of the data or of the coefficients.

The distinction between direct and aggregate data is very important, and the two types of data differ in terms of information loss and the form of data produced. Indirect measures are summaries calculated from original data, and are produced by aggregating over one facet of the data (over subjects in the case of R-analysis, and over objects in Q-analysis). By contrast, direct measures preserve individual data intact and make it possible, at least in principle, to detect systematic individual differences. Secondly, aggregate measures take the form of a coefficient of (dis)similarity between each pair of objects (or subjects) and are almost without exception 'well-behaved' measures, obeying the triangle inequality. Since aggregate measures produce a square symmetric matrix of (dis)similarity coefficients, they can all be analysed using the basic MDS model. By contrast, the properties of direct measures are *not* known in advance, and it is quite likely that, as they stand, some of the data may be inconsistent with any numerical representation. On the other hand, most direct forms of data can be analysed by specifically designed programs, which allow the researcher to examine how well each set of individual data fits the overall configuration.

The question of the level of measurement of the data is also important, but is not always crucial. It is of course sensible for the researcher to choose a measure of association which matches as closely as possible the level of measurement of the variables concerned, and it is advisable to use more than one measure in order to assess the extent to which results are dependent upon the properties of particular measures. But one of the main uses of non-metric MDS is to see whether, whilst making very conservative claims for the level of measurement of the data, it is possible to find a legitimate transformation (re-scaling) which will yield a much higher level, better behaved, set of values.

Normally, the most general rule is to preserve as much information from the original data as possible when choosing data for input to MDS. In the case of

direct data, it is important to choose a program which matches the type of data as closely as possible. For aggregate measures, it is important to choose a dis(similarity) measure whose properties are well understood, and which match the level of measurement of the data as closely as possible.

| | *Type of Data* | |
|---|---|---|
| *Level of Measurement* | *Direct* (Method of data collection) | *Derived Aggregate* (Measure of (dis)similarity) |
| INTERVAL | Ratings, flow rates, latencies | Product moment measures, (especially covariance, correlation) |
| ORDERED METRIC | Tetrads | |
| | Triads | |
| *Strict* ORDINAL: | Rankings | Tau measures, and Wilson's e |
| *Weak* | Rankings, with ties | Gamma |
| HIERARCHICAL | Hierarchies (rooted trees) | Boorman-Olivier family of measures |
| *Polytomies* | Partitions | Chi-square based measures (esp. Cramer's V) |
| NOMINAL | | Co-occurrence measures (esp. Burton's Z) |
| | | Boorman-Arabie family of measures |
| *Dichotomies* | Pair comparisons | Matching coefficients |
| | | Symmetric difference |

Table 2.6 *Types of (dis)similarity measures for scaling*

# APPENDIX A2.1 DISTANCE MEASURES AND SCALAR PRODUCTS

### A.2.1.1 Distance measures

The correspondence between dissimilarity and distance has been analysed rigorously in mathematics in terms of the notion of a 'metric' or general distance measure (of which Euclidean distance is a special case). Most measures of association used in statistics and data analysis, for example, satisfy the requirements of such a general distance measure, though this does not necessarily mean that they can be directly represented in a Euclidean space. Indeed, one purpose of non-metric MDS is to see whether we can *re-scale* data into a set of quantities which are capable of such a representation.

There are two basic properties (or axioms) which a measure must satisfy to count as a metric, and a further one is also normally required. These are listed below. ($d(A, B)$ should be read: the distance between points $A$ and $B$):

**Properties of a Distance Measure**

(i) *Non-negativity and equivalence*

$d(A, B) \geqslant 0$ for all points $A, B$

and

$d(A, B) = 0$ if and only if $A$ coincides with $B$

(ii) *Symmetry*

$d(A, B) = d(B, A)$ for all points $A, B$

(iii) *Triangle inequality*

$d(A, C) \leqslant d(A, B) + d(B, C)$ for all points $A, B, C$

A measure that satisfies properties (i) and (ii) but not (iii) is often termed a 'semi-metric': one that satisfies all three axioms is referred to as a 'metric'. Several comments on these properties are appropriate.

*Non-negativity* might seem to exclude covariances, correlations etc., which can take on negative values. This difficulty can be overcome fairly easily.

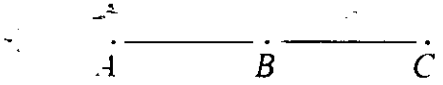
*Symmetry* might seem to exclude asymmetric dependence measures, such as regression coefficients, although asymmetry *can* be represented in a spatial manner (see 5.1.1.1).

*Triangle inequality* is the most restrictive axiom, and can best be illustrated by the fact that in Euclidean space a point $B$ must either lie *on* the line $AC$, in which case $d(A, C) = d(A, B) + d(B, C)$, or else it must lie *off* the line $AC$, in which case the sum: $d(A, B) + d(B, C)$ must exceed $d(A, C)$.

**The Triangle Inequality Axiom**

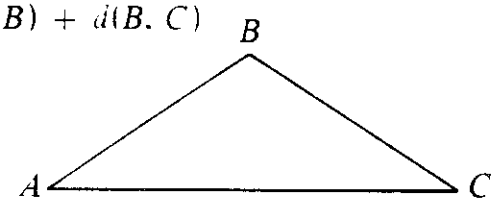
(a) *Triangle equality* ($B$ lies on line $AC$)

$d(A, C) = d(A, B) + d(B, C)$




(b) *Triangle strict inequality* ($B$ lies off line $AC$)

$d(A, C) < d(A, B) + d(B, C)$




This axiom clearly excludes the possibility that

$$d(A, C) > d(A, B) + d(B, C)$$

A large number of measures used in empirical research satisfy these three axioms but it is by no means clear simply by inspection whether they do or do not.

#### *A2.1.1.1 Euclidean distance*

The most common way of representing dissimilarity in MDS is in terms of Euclidean distance, which involves a surprisingly restrictive set of further assumptions. The user should be aware of what these are. The mathematical definition of Euclidean distance is as follows:

$$d_{jk} = \sqrt{\left\{\sum_a (x_{ja} - x_{ka})^2\right\}}$$

where $x_{ja}$ refers to the co-ordinate of point $j$ on dimension $a$.

The formal and substantive assumptions of Euclidean distance have been investigated from a measurement theory viewpoint by Beals et al. (1968). Four definitional characteristics of Euclidean distance are of special relevance in elucidating the assumptions implicit in the use of the distance model (Tversky and Krantz 1970, p. 4):

(i) *Decomposability* The distance between the objects can be decomposed into a contribution from each of the dimensions ($a$) of the space.

(ii) *Intra-dimensional subtractivity* $(x_{ja} - x_{ka})$
Each contribution to the distance between two points is composed of the difference in scale values within each dimension.

(iii) *Inter-dimensional additivity* (the summation over $a$)
The distance measure combines these differences additively from each dimension.

(iv) *Metric* (the squaring of the differences)
All the differences in (ii) are transformed by the same power-function.

As the authors show, it is possible to state the empirical conditions which data must satisfy if at least the first three of these characteristics are to be justifiable. Users should note that there is no guarantee that a given set of ordinal similarity data can be embedded in a metric space and that the metric and dimensional assumptions of the distance model are quite distinct; it is quite possible that some data will satisfy the latter but not the former set of assumptions.

#### *A2.1.1.2 Minkowski metrics*

The Euclidean metric is a special case of a more general family of distance measures, referred to under as Minkowski $r$ (or $L_p$ or power) metrics of the form:

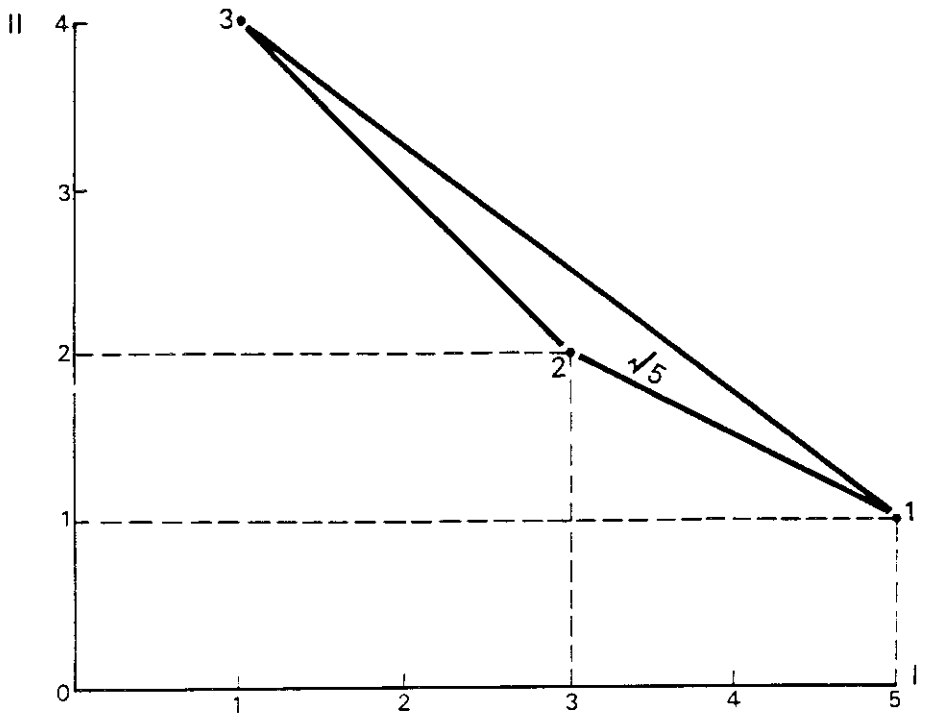
$$d_{jk}^{(r)} = \sqrt[r]{\sum_{a=1} |x_{ja} - x_{ka}|^r}$$

Each value of $r$ ($\geqslant 1$) substituted in this formula defines a distinct metric, all of which obey the triangle inequality. Clearly, if $r = 2$ then the Euclidean metric results. Other values used in scaling include the city-block (or 'Manhattan' or 'taxi-cab') metric (where $r = 1$), and the dominance metric (where $r = \infty$). The properties of these metrics and their applications in MDS are discussed under 5.3.3.2. Carroll and Wish (1974a, p. 412 et seq.) give an extended treatment and also consider the case where $r < 1$, and other metric families such as Riemannian metrics of constant curvature which have also been used in MDS.

### A.2.1.2 Distance and vector representation of data

Consider the following data matrix:

$$\mathbf{X} = \begin{bmatrix} 5 & 1 \\ 3 & 2 \\ 1 & 4 \end{bmatrix}$$




Distance between variables 1 and 2

$$d(1,2) = \sqrt{(5-3)^2 + (1-2)^2}$$
$$= \sqrt{(4+1)}$$
$$= 2.236$$

Matrix of distances

| *Variable* | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 0 | $\sqrt{5}$ | $\sqrt{25}$ |
| 2 | $\sqrt{5}$ | 0 | $\sqrt{8}$ |
| 3 | $\sqrt{25}$ | $\sqrt{8}$ | 0 |

Figure A2.1 *Euclidean distances between three variables*

The information in this matrix can be thought of *either* as giving the co-ordinates for locating two (column) elements in three (row) dimensional space, *or* for locating three (row) elements in two (column) dimensional space. For simplicity of exposition we shall assume that it locates three variables describing two subjects, and that we wish to assess the similarity of the variables (R-analysis).

If a distance measure of similarity is required, then this can be calculated using the Euclidean distance formula illustrated in Figure A2.1.

Two operations often performed on raw data scores are:

(i) *centring*—creating the 'deviate score' by removing the overall effect of a variable by subtracting the mean.

(ii) *standardising*—creating the 'normal score' by dividing the deviate score by the standard deviation of the variable, thus reducing all variables to a common unit of measurement.

Centring the variables has the geometric effect of removing the origin of the space

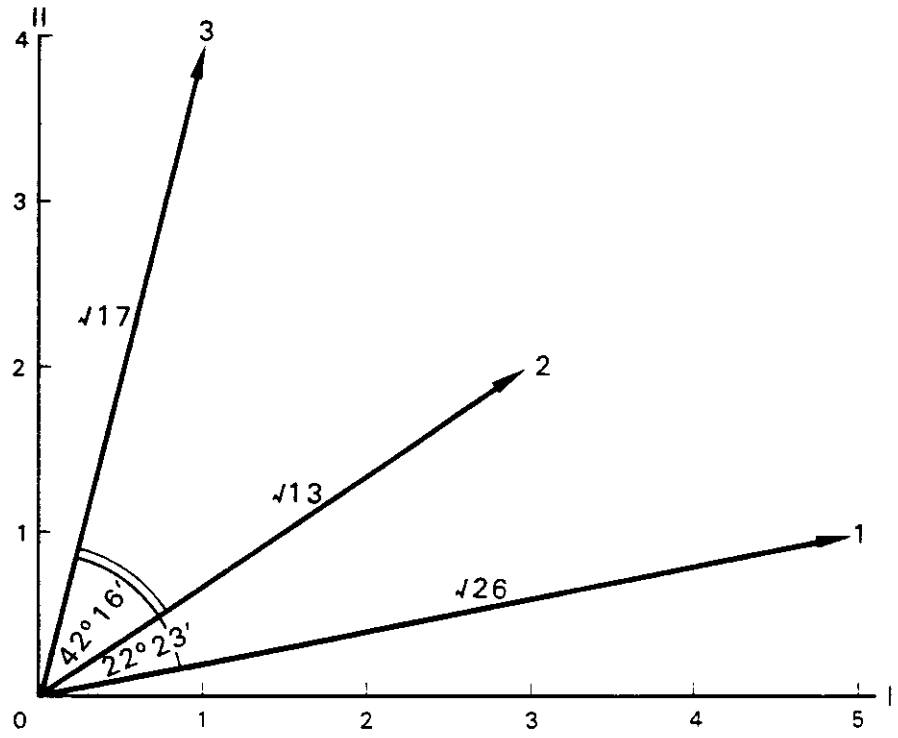



Figure A2.2 *Vector representation of similarity*

(0, 0) to the centroid (centre of gravity) of the points defined by the means of the variables, i.e. to (3, 2) in this case, but it does not affect the distances in any way. By contrast, standardising *does* affect distances since the axes are now differentially stretched to a common unit.

The vector representation of the same data is given in Figure A2.2. The scalar products measure of similarity between the variables is produced by forming the (major) product moment matrix, $\mathbf{A} = \mathbf{XX}'$, which in this case is:

$$\underset{\mathbf{X}}{\begin{bmatrix} 5 & 1 \\ 3 & 2 \\ 1 & 4 \end{bmatrix}} \quad \underset{\mathbf{X}'}{\begin{bmatrix} 5 & 3 & 1 \\ 1 & 2 & 4 \end{bmatrix}} = \underset{\mathbf{A}}{\begin{bmatrix} 26 & 17 & 9 \\ 17 & 13 & 11 \\ 9 & 11 & 17 \end{bmatrix}}$$

This is often referred to as the matrix of 'crude sums of squares and cross products' (CSSCP) in the multivariate analysis literature. The *minor* product matrix $\mathbf{X}'\mathbf{X}$ (of order two) provides the scalar products between the two individuals, aggregated over the variables). The entries in the product moment matrix are readily interpretable. The diagonal entry, $a_{ii}$, gives the squared length of the vector draw from the origin to the point $i$ (call it $l_i^2$). The symmetric off-diagonal elements $a_{ij}$ give the scalar product between vector $i$ and $j$, which is related to the angular separation between the vectors. Explicitly it is the product of the length of each vector and of the cosine of the angle separating them: (see van der Geer 1971, pp. 19–21)

$$a_{ij} = l_i l_j \cos \theta_{ij}$$

e.g. in the case of variables 1 and 2:

$$\begin{aligned} a_{12} &= \sqrt{26}\sqrt{13} \cos (22°23') \simeq (5.100)(3.606)(0.924) \\ &= 17.000 \end{aligned}$$

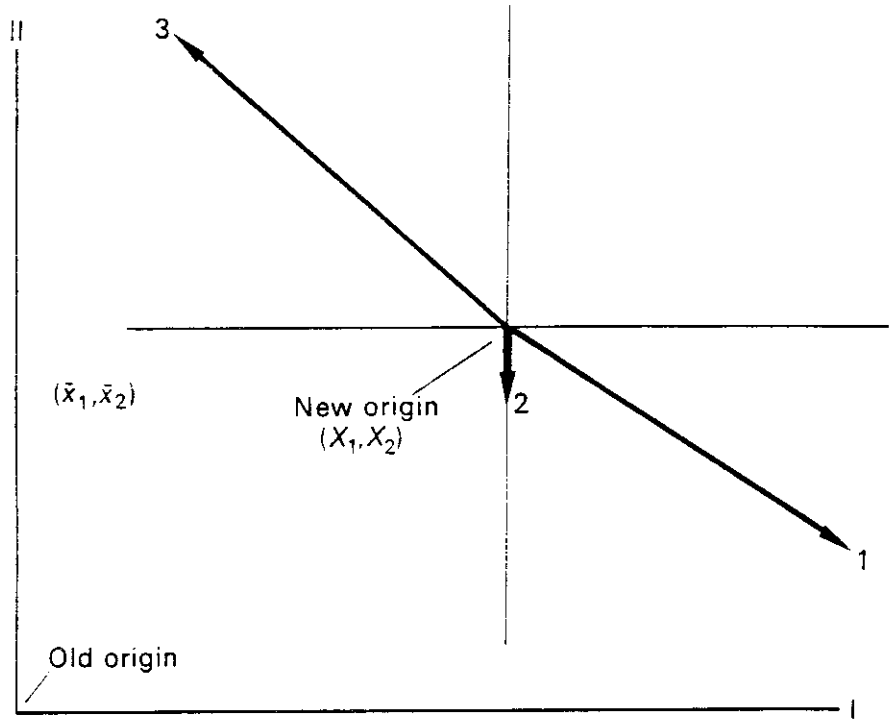



Figure A2.3 *Vector representation from centroid origin*

Although centring has no effect upon distances, it dramatically alters the measure of similarity based upon scalar products, since the lengths of the vectors and their angular separation is now assessed from a different origin. The effect of centring is illustrated in Figure A2.3. The product moment matrix **B** formed from the centred deviate matrix $\mathbf{X}_d$ is as follows:

$$\mathbf{B} = \mathbf{X}_d\mathbf{X}_d' = \begin{bmatrix} 2 & -\frac{4}{3} \\ 0 & -\frac{1}{3} \\ -2 & \frac{5}{3} \end{bmatrix} \begin{bmatrix} 2 & 0 & -2 \\ -\frac{4}{3} & -\frac{1}{3} & \frac{5}{3} \end{bmatrix}$$

$$= \begin{bmatrix} 5.78 & 0.44 & -6.22 \\ 0.44 & 0.11 & -0.56 \\ -6.22 & -0.56 & 6.78 \end{bmatrix}$$

Notice that the scalar products between 1 and 3 and between 2 and 3 are now negative. Obviously, centring variables does *not* leave vector separation measures unchanged. Consequently when vector or factor model solutions are presented in MDS the origin of the configuration is fixed, and may not be relocated at will. By contrast, the origin in Euclidean distance model configurations is arbitrary and may be moved.

There are two especially useful variants of the relation: $a_{ij} = l_i l_j \cos \theta_{ij}$, which apply when the variables are centred, and normalised.

(i) If the $\mathbf{X}_d\mathbf{X}_d'$ matrix is multiplied throughout by the constant $1/N$ (where $N$ is the number of individuals), the resulting matrix consists of the dispersion matrix, $\Sigma$, which features centrally in multivariate analysis, whose diagonal elements $\sigma_{ii}$ are the variances, and whose off-diagonal elements $\sigma_{ij}$ are the covariances between the variables $i$ and $j$. In this case, the relation is:

$$a_{ij} = \sigma_i\sigma_j \cos \theta_{ij} = \sigma_{ij} \quad \text{(covariance)}.$$

Again, both the length of the variables (in this case, their dispersion) and their angular separation contribute to the covariance measure, which is therefore sensitive to the scaling (measurement units) of the original variables.

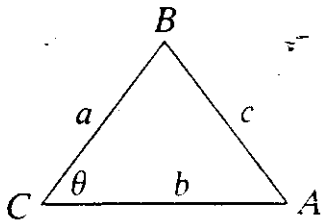
(ii) If in addition the variables are standardised (to zero mean and unit variance $z_{ij} = (x_{ij} - \bar{X}_j)/\sigma_j$), then the dispersions all become unit length, which is tantamount to saying that the original variable scaling units are arbitrary. Then the form of the relation becomes especially simple. Since $\sigma_i = \sigma_j = 1$ for standardised variables, then it reduces to:

$$a_{ij} = (1)(1) \cos \theta_{ij} = \cos \theta_{ij} = r_{ij} \quad \text{(correlation)}.$$

Hence Pearson's correlation coefficient preserves *only* the angular separation *in normalised axes* (shrunk or expanded in order to equalise dispersions) between the variables, and the scale units of the original values in no way contribute to the measure of similarity.

### A.2.1.3 Coverting scalar products into distances

Conversion of scalar products into Euclidean distances involves a simple application of the cosine rule, which states that in a non-right-angled triangle, $c^2 = a^2 + b^2 - 2ab \cos \theta$.




Thinking of $CB$ and $CA$ as vectors, $\theta$ as the angle separating them, and $AB$ as the distance $d(A, B)$ corresponding to the angular separation, then the rule may be rewritten as:

$$\begin{aligned} d(A, B) &= l_a^2 + l_b^2 - 2l_a l_b \cos \theta_{ab} \\ &= l_a^2 + l_b^2 - 2 \times \text{(scalar product between } a \text{ and } b) \\ &= a_{ii} + a_{jj} - 2a_{ij} \quad \text{(in terms of the product moment matrix of scalar products)} \end{aligned}$$

For example, in Figure A2.2,

$$\begin{aligned} d_{12}^2 &= 26 + 13 - (2)(17) \\ &= 39 - 34 = 5 \\ d_{12} &= \sqrt{5}. \quad \text{which is the quantity calculated in Figure A2.1.} \end{aligned}$$

A special application of the cosine rule, and remembering that standardised variables are unit length, shows that correlations are (inversely) monotonic with distances.

The cosine rule:

$$d_{ij}^2 = l_i^2 + l_j^2 - 2l_i l_j \cos \theta_{ij}$$

Since $r_{ij} = \cos \theta_{ij}$,

then $d_{ij}^2 = 2 - 2r_{ij}$

and $d_{ij} = \sqrt{(2 - 2r_{ij})} = \sqrt{2(1 - r_{ij})}$.

Clearly, the relationship between distance and correlation is a decreasing one (in effect, $1 - r_{ij}$ forms the *dis*similarity coefficient), and it is *non*-linear (because of the square root). But there *is* a monotonic relationship between distance and correlation, including negative correlation values.

The conversion of Euclidean distances into scalar product form is a slightly more complicated matter, and is taken up in Appendix A5.2.

## APPENDIX A2.2 CO-OCCURRENCE MEASURES OF SIMILARITY

An individual partition of a set of $N$ objects or elements can be represented as a square symmetric (0, 1) matrix S of order $N$, where $S_{ij} = 1$ if objects $i$ and $j$ occur in the same category of the partition, and $S_{ij} = 0$ otherwise.

Thus, the partition $I = \{3, 1 \mid 4, 2, 5\}$ can be represented by the co-occurrence or 'incidence' matrix:

$$\mathbf{S}^{(I)} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

Clearly, co-occurrence is a metric, obeying the triangle inequality since $i$ cannot occur with $j$, and $j$ occur with $k$ without $i$ also occurring with $k$.

When there is a set of $r$ partitions, the S matrices are simply added together to form the aggregate co-occurrence matrix. It also represents a metric, since the sum of metrics is a metric. The four measures referred to in the text (2.2.3.3) and defined in Burton (1975) differ basically in how the size of the category is taken into account before the individual matrices are aggregated.

*M1* (the basic measure, cf. Miller 1969, which is called *F* in Burton 1975)
Each co-occurrence contributes equally, so the aggregate matrix S is the simple sum of the (1, 0) individual matrices.

*M2* (called *G* in Burton 1975)
The entries in the individual co-occurrence matrix $\mathbf{S}^{(I)}$ are *the number of elements in the category* from which the pair is drawn. In this case, the individual co-occurrence matrix corresponding to partition $I$ would be:

$$\mathbf{S}^{(I)} = \begin{bmatrix} 2 & 0 & 2 & 0 & 0 \\ 0 & 3 & 0 & 3 & 3 \\ 2 & 0 & 2 & 0 & 0 \\ 0 & 3 & 0 & 3 & 3 \\ 0 & 3 & 0 & 3 & 3 \end{bmatrix}$$

(N.B. diagonal elements are ignored)

Hence, the larger the category in which a pair of objects occur, the higher their similarity is considered to be.

*M3* The entries in $\mathbf{S}^{(I)}$ *are* the *reciprocal of the number* of elements in the category from which the pair is drawn. In this case,

$$\begin{bmatrix} (-) & 0 & \frac{1}{2} & 0 & 0 \\ 0 & (-) & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & (-) & 0 & 0 \\ 0 & \frac{1}{3} & 0 & (-) & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & (-) \end{bmatrix}$$

Clearly, this measure *deflates* similarity by the size of category, on the reasoning that the more unusual co-occurrences, or the more fine discriminations denote greater similarity.

*M4* (called *Z* in Burton 1975)
This information-theoretic measure, which is akin to *M3* in emphasising the similarity of pairs from small categories, is based upon the 'surprisal value' of each category. This is defined in terms of

(i) the probability that two objects $j$ and $k$ will be found in the *same* category, $a$.

$$p_a^{(I)} = n_a(n_a - 1)/n$$

(where $n_a$ is the number of objects in category $a$, and $n$ is the total number of pairs) and

(ii) the probability that $j$ and $k$ will be found in,

$$Q^{(I)} = 1 - \sum_a p_a^{(I)}$$

The contribution which each pair of objects $(j, k)$ makes is defined as its surprisal value.

$-\log_2 (p_a^{(I)})$ if $j$ and $k$ are in the same group
and $-\log_2 (Q^{(I)})$ if $j$ and $k$ are in different groups

Since surprisal is negatively related to the size of the group, *M4* also emphasizes finer discriminations, but (unlike *M3*) makes explicit allowance for pairs occurring in different categories. In the present case, since the probability of two objects being in the same category is 0.1 for category 1, and 0.3 for category 2, and the probability for being in different categories is 0.6, the matrix of similarity (surprisal values) is:

$$\begin{bmatrix} — & 0.74 & 3.32 & 0.74 & 0.74 \\ 0.74 & — & 0.74 & 1.74 & 1.74 \\ 3.32 & 0.74 & — & 0.74 & 0.74 \\ 0.74 & 1.74 & 0.74 & — & 1.74 \\ 0.74 & 1.74 & 0.74 & 1.74 & — \end{bmatrix}$$

Diagonal values are usually defined by convention to be slightly larger than the maximum element, to preserve the positivity axiom of a metric.