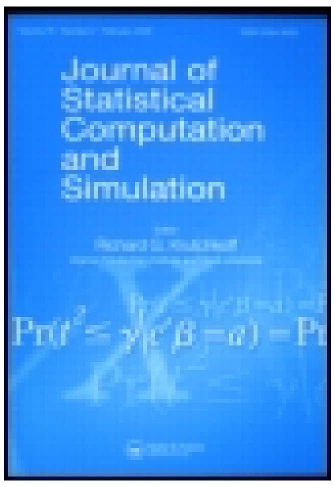

Journal of Statistical Computation and Simulation
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Studies in the robustness of multidimensional scaling: euclidean models and simulation studies

Robin Sibson [a] , Adrian Bowyer [a] & Clive Osmond [a]
[a] School of Mathematics , Utiiversitv of Bath , England

Published online: 20 Mar 2007.

J. Statist. Comput. Simul., 1981, Vol. 13, pp. 273–296
0094-9655/81/1304-0273 $06.50/0

Studies in the Robustness of Multidimensional Scaling: Euclidean Models and Simulation Studies

ROBIN SIBSON, ADRIAN BOWYER and CLIVE OSMOND

School of Mathematics, University of Bath, England

(*Received November 5, 1980*)

This series of papers is devoted to the investigation of the extent to which the accuracy of operation of multidimensional scaling can be put onto a quantitative footing. In this third and final paper, four probabilistic models for the generation of euclidean-distance-like dissimilarity functions are proposed; these models reflect some of the ways in which dissimilarities actually arise, and allow such effects as dependence between dissimilarities to be studied. Using these models, simulation experiments are carried out to assess the response of both classical and ordinal (non-metric) scaling to errors, with procrustes statistics being used to measure accuracy of recovery. A further scaling method, least squares scaling, is discussed briefly and shown to display empirically a useful combination of properties, as is a technique used to preprocess the dissimilarity matrix.

KEY WORDS: Multidimensional scaling, robustness, procrustes statistics, euclidean models, simulation.

1. INTRODUCTION

In this paper the robustness—the response to errors in the data—of both classical scaling and the Shepard–Kruskal method of ordinal ("non-metric") scaling is examined by simulation methods. Our approach to the study of robustness is based on the idea of assessing quantitatively the success of a scaling method in recovering known configurations. On exact data classical scaling does this exactly, and the ordinal method is almost exact; it is when the data contain errors that the exercise becomes an interesting one. Studies of this kind have been carried out by many authors; we give a brief review below. Apart from the scale of the experimentation, two main features distinguish the present study from this earlier work. The first is the use of procrustes statistics to measure degree of fit: it is surprising that this approach has not been used in previous simulation studies. The second distinguishing feature is the way error is introduced into the data: our study includes work on dissimilarities generated from probabilistic models whose structure is designed to capture some features of the way errors actually arise in practice, and in particular to throw some light on the effects of dependence between errors.

The study falls into two major parts. The first treats classical scaling in some detail; the second part compares various different scaling methods on the same data. The scaling methods used are classical scaling, ordinal scaling, and least-squares scaling; a preprocessing transformation of data to specified distributional form is also used.

We follow the notation and terminology of the two previous papers in this series: Sibson (1978), where an account of procrustes statistics may be found; and Sibson (1979a), where an approximate analytical study of the robustness of classical scaling is carried out. The latter paper contains a brief description of the classical scaling method: an account of the ordinal method is given by Kruskal (1964a, b): in this paper we assume familarity with both methods. References to Sibson (1978, 1979a) will sometimes be made in the form, for example, I Theorem 4.3, II Lemma 3.3

2. PREVIOUS WORK AND THE ROLE OF THIS STUDY

Numerous simulation studies of various forms of multidimensional scaling have appeared, almost entirely in the psychometric literature. Some of these studies, such as that of Lingoes and Roskam (1973), have been primarily concerned with the effectiveness of various numerical procedures in reaching the true optimum quickly. This is an important technical problem, but it will not be our concern here. We assume that all the methods can be made to work with adequate efficiency and reliability in practice, although we comment in passing on the "local optimum" problem. We are concerned, as most previous studies have been, with the accuracy of reconstruction effected by a method once it has, by its own criteria, "worked properly". The method most commonly employed to assess this is the evaluation of the (squared) correlation between original and reconstructed distances. Cohen and Jones (1974) point out a number of objections to this: it is insensitive for similar configurations; the displacement of a single point can produce a misleading value; it is unaffected by the addition of a constant to either or both sets of distances; it does not deal satisfactorily with the comparison of configurations of differing dimensionality. We would summarise these objections by saying that the correlation coefficient simply fails to relate properly to either the geometrical or the probabilistic aspects of the problem. Another measure of success of reconstruction which has been used is the Shepard–Kruskal stress achieved at optimality in the ordinal method. This seems to us to reflect a misunderstanding of the role of the optimised stress function, which measures the euclideanness of a set of ordinal data, and not the extent to which the reconstruction matches the original configuration. The approach we adopt, using procrustes statistics, avoids these criticisms. These and other arguments for the use of an approach based on procrustes statistics are presented by Gower (1971b). Gower and Banfield (1974) have explored this approach in a simulation study of classification methods.

Nearly all the parameters and problem-dependent quantities occurring in ordinal scaling have been varied in previous simulation studies to determine their influence. Young (1970) examines the effect of the number of points; Sherman (1972) looks at the effect of varying the Minkowski metric parameter from which the interpoint distances are derived (but see Shepard. 1974, for some cautionary remarks on this); whilst Wagenaar and Padmos (1971) consider the amount of error present in the perturbations. Lingoes and Roskam (1973) compare the Shepard–Kruskal method with the Guttman–Lingoes SSA—I method, but, as pointed out above, mainly with a view to assessing algorithmic efficiency and robustness. Spence (1972) compares the computer programs MDSCAL, SSA-I, and TORSCA; his conclusions are that there is little to choose between the methods, but that a sensible choice of initial configuration in MDSCAL is important if the algorithm is to work properly. In an earlier paper (Spence, 1970) he uses simulation methods to study the local optimum problem. Spence and Graef (1974) and Isaac and Poor (1974) discuss the problem of determining dimensionality in ordinal scaling. Several papers have attempted to provide a measure of comparison for an empirically-obtained stress value against the expected stress value obtained from random rankings in ordinal scaling; such papers are those of Klahr (1969), Spence and Ogilvie (1973), and Stenson and Knoll (1969). Specific geometric configurations are used by Spaeth and Guthery (1969) to test the recovery capability of the methods and to examine the monotonicity criterion. Finally, Cohen and Jones (1974) consider a model, motivated by psychological considerations, in which scaling performs a reconstruction, given that different levels of importance are attached to each dimension by a hypothetical subject.

We choose to approach the problem differently. We do not vary the number of objects, which we keep at fifty. We work exclusively with euclidean, not Minkowski, distance. We generate dissimilarities from various different euclidean models, chosen so as to allow various features of the error response to be examined. Starting from classical scaling we establish the point at which it is inclined to break down and examine how it compares with ordinal scaling and also with least-squares scaling. We consider dimensionality criteria (Sibson, 1979a) in the context of classical scaling.

3. FOUR EUCLIDEAN MODELS

Our approach to the study of robustness by simulation methods is based on the view that in the majority of applications of scaling methods it is not appropriate to assume that the observed dissimilarities differ from the true interpoint distances by errors that are independent. There are certainly a few cases where the assumption of independent errors is an appropriate one; for example, it is argued by Sibson and Bowyer (1980) that this is so in some problems arising in surveying and photogrammetry. We consider one such model for comparison purposes But these cases are in the minority; in particular, it is clear that when dissimilarities or similarities are calculated from objects-by-attributes data (Sibson, 1972), there will be substantial depence involved Rather than setting up error distributions by decree, we establish simple, but practically relevant, models for the generation of dissimilarities, and then investigate the distributions which arise from these.

The first step in setting up the models is to generate a configuration. We have taken fifty points as representing a typical medium-scale problem, and we have not attempted to investigate by simulation the effects of changing the number of points. In fact our configurations are generated by realising fifty points uniformly and independently on the unit disc of the appropriate dimensionality, but the only feature of real significance is, in our view, that the configurations should be roughly spherical with no special structure. We do not generate an unlimited supply of new configurations, but we do use enough to provide a check against being misled by the behaviour of any particular one.

Our first model arises from the study of binary data, of the type where the coding of the states of each attribute into zero and one is arbitrary. Objects described by such attributes are usually compared by counting the number of attributes in which they differ. The resultant metric is known as Hamming distance in communication theory; its normalised complement with respect to the total number of attributes is a similarity coefficient long familiar in numerical taxonomy as the simple matching coefficient. For a fixed k-dimensional configuration, random Hamming distances can be generated by randomly located hyperplanes each dividing the space into two half-spaces arbitrarily coded as zero and one. If the hyperplanes are realised from a Poisson hyperplane process, then (almost surely) each Hamming distance is well defined and finite, and has a Poisson distribution whose parameter is the product of the euclidean distance between the two points and the intensity of the process, the latter being expressed in suitable units. The Hamming distance is, in geometrical terms, the number of hyperplanes crossed in going from one point to the other. Clearly the mean of a single Hamming distance distribution is proportional to the corresponding euclidean distance, so the relationship between Hamming distance and euclidean distance is roughly linear. It is easy to see that as the intensity of the Poisson process becomes large the relative values of the system of Hamming distances converge to those of the euclidean distances (almost surely). However, the Hamming distances are not independent; any two of them together have a bivariate Poisson distribution (see Mardia, 1970) whose parameters can easily be expressed in geometrical terms. The system as a whole has as its joint distribution a multivariate generalisation of this bivariate Poisson distribution. All realisations of systems of Hamming distances arising from this joint distribution automatically satisfy the metric inequality. Hamming distance is one of a large class of dissimilarity functions which do so (Gower, 1971a). We call this model for the generation of euclidean-like distances the Poisson hyperplane model. In practice it is convenient to condition on the total number of hyperplanes involved, whereupon Poisson distributions become binomial, and it is in this form that we actually realise the model. Figure 3.1 shows the dependence of Hamming distance on euclidean distance for a realisation from this *binomial hyperplane model*. The points lie in a narrow band demonstrating the near-linearity of the two distances. For smaller numbers of hyperplanes the width of the band is correspondingly greater.

In order to be able to assess the effects of the dependence structure in the Poisson hyperplane model we set up a model in which each individual dissimilarity has the same distribution as in the Poisson hyperplane model, but the dependence is removed, thereby producing a model with independent errors. Thus in this model the dissimilarity $d(x,y)$ has a Poisson distribution with parameter $\lambda e(x,y)$ where $e(x,y)$ is euclidean distance, and the $d(x,y)$ are independent; that is, the joint distribution of the $d(x,y)$ is the product of the marginals, these marginals coinciding with those arising in the hyperplane model. In practice it is convenient to match the form of the Poisson hyperplane model actually realised, namely with conditioning on the total number of hyperplanes involved. This requires the use of independent binomial distributions rather than independent Poisson distributions. The dissimilarity-against-distance plots arising from this independent binomial model are visually indistinguishable from those arising in the binomial hyperplane model.

Our third model is chosen to provide some hold over the dissimilarities which arise in problems where the data are in objects-by-attributes form, and where the attributes are binary, but in contrast to the first model, where their states may be coded 0 (="absence")/1(="presence") consistently over the whole system of attributes. A familiar example arises in plant ecology, where the objects are sites and the attributes are plant species, recorded as present absent at each site. A widely-used coefficient in such cases is Jaccard's coefficient, which we consider in the form of *Jaccard distance*, a dissimilarity coefficient (in fact a metric) obtained by dividing Hamming distance by the number of attributes "present" in either or both of the two objects under consideration. Jaccard distance takes values in [0, 1], and so its relationship to euclidean distance certainly cannot be a linear one. A method of generating random Jaccard distances

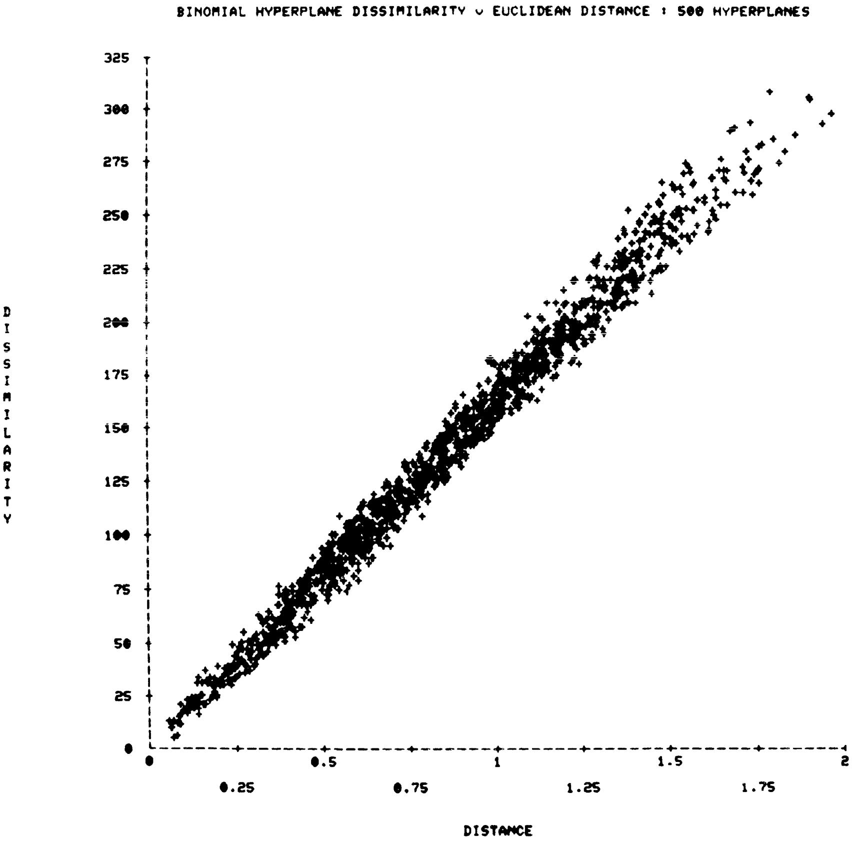

FIGURE 3.1

is as follows. Each attribute is "present" over a region of space interior to a disc, whose radius is drawn from some fixed distribution and whose centre is a point in a realisation of a Poisson point process. The condition that the expected disc content be finite is enough to ensure that Jaccard distances are almost surely well-defined, (except for the rare case of two points each lying in no discs at all, for which we assign the value unity). This model actually has a certain simple-minded appeal as a model for random spatial speciation, at least by comparison with some of the alternatives! Clearly any particular version of the model is characterised by the rate of the Poisson point process and the nature of the radius distribution. It is easy to show that for a fixed radius distribution, the relationship between euclidean distance and expected Jaccard distance is a monotone one, with a decreasing fluctuation about the mean as the intensity increases. The dependence of the monotone relationship on the nature of the radius distribution is calculable in principle, but has no simple form and is better treated as unknown. The model is thus well-adapted to distinguishing between classical and ordinal scaling methods. Figure 3.2 shows a typical dissimilarity-versus-distance plot. plots of the expected dissimilarity against euclidean distance for the constant radius distribution also show this concavity. Again, the form in which the model is realised is a conditioned one, the total number of discs being fixed. We call this the *Jaccard distance model*.

The fourth model relates to abuttal data. Data for scaling sometimes arise in the form of a three-valued dissimilarity coefficient whose values are "identical" (precisely between each object and itself), "neighbouring", and "not neighbouring". Such data arise when the objects are really regions rather than points, and it is abuttals between regions which are recorded. Kendall (1971, 1974) has studied such data extensively. One possible method of analysis is to represent regions by points, and then to assign conventionalised regions, with the implied neighbour or contiguity relationship, by way of the Dirichlet tessellation (see Green and Sibson, 1978); this construct assigns to each point the part of the space nearer to it than to any other point. To reconstruct configurations directly from abuttal data is not easy (McGinley, 1977), but can be made so by replacing the original three-valued dissimilarity coefficient by an integer valued one, the graph-theoretic distance or *Wilkinson metric*, which is the minimum number of abuttals traversed along a path from one point to another via abuttals. The distribution of euclidean distance between contiguous points in a planar Poisson process is known (Miles, 1970; Sibson, 1979b), but this knowledge does not extend to points at larger Wilkinson distances. However it appears from simulation that the mean euclidean distance is close to linear with the Wilkinson metric in two dimensions, which for computational reasons is the only currently practicable case. A simulation model may be obtained by taking a fixed configuration of points between which the values are to be calculated, and superimposing on this a realisation of a Poisson process. The Wilkinson metric for the combined configuration may then be calculated from its Dirichlet tessellation. In practice, of course, only finitely many additional points are generated; by generating these over a region considerably larger than that occupied by the original configuration, edge effects are made negligible. As the number of additional points increases it appears that, as in the other models, the relative variability of the Wilkinson metric decreases. Plots of the Wilkinson metric against euclidean distance show their near linear relation, in the way that Figure 3.1 does for the binomial

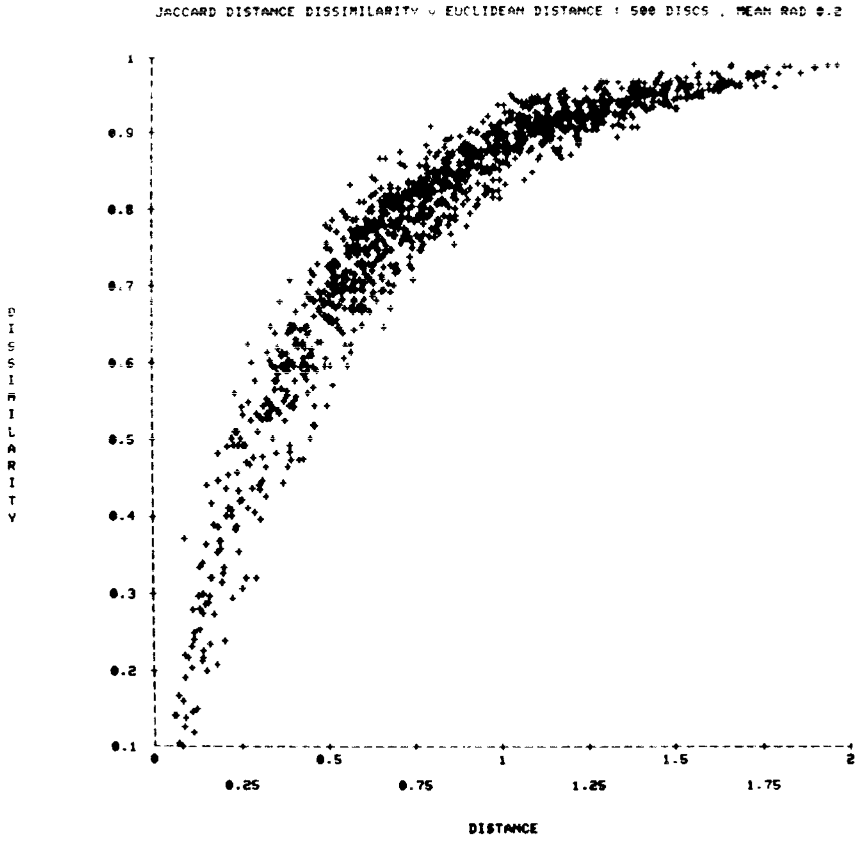

FIGURE 3.2

hyperplane model. Like the other models this *Wilkinson metric model* is realised in conditioned form.

4. PROCRUSTES STATISTICS

We always compare a recovered configuration Y with its parent configuration X by using a procrustes statistic; a detailed account of the theory is given in Sibson (1978) The particular form of statistic employed allows Y to be fitted to X under the action of the group of similarity transformations, that is, the group generated by translation, rotation, reflexion, and uniform scale change. This leads to the statistic

$$G_S(X, Y) = \text{trace } X_0 X_0^T - \frac{(\text{trace } (Y_0 X_0^T X_0 Y_0^T)^{1/2})^2}{\text{trace } Y_0 Y_0^T}$$

where X_0, Y_0 are X, Y translated to have centroid at origin (see I Section 5) and we normalise this to

$$\gamma_S(X, Y) = 1 - \frac{(\text{trace } (Y_0 X_0^T X_0 Y_0^T)^{1/2})^2}{\text{trace } X_0 X_0^T \text{ trace } Y_0 Y_0^T}$$

as in I Section 7. We use $\gamma_S(X, Y)$, which lies in the range $[0, 1]$, for all our comparisons. It is appropriate even with classical scaling to do this, because in practice the approximately linear relation between dissimilarity and distance is usually an unknown one.

5. THE SCALING METHODS

We consider in total three different scaling methods. The first of these is classical scaling of which a brief description can be found in II Section 1; we do not duplicate it here.

The second method is the Shepard–Kruskal ordinal method as described by Kruskal (1964a, b). Many implementations and variants of this method exist. We have used exactly the original form of Kruskal's stress function, with the normalisation factor Σd_{ij}^2; ties were dealt with by the primary treatment in which ties are an expression of ignorance and can be broken without charge; and the global ordering of all values of the dissimilarity coefficient was used. Kruskal's optimisation procedure was used, but in each case a variety of different initial configurations were tried so as to avoid local optimum problems. We believe that the procedures we have adopted make it most unlikely that any part of our study is vitiated by the occurrence of local optima. It is to be noted that the availability of the original configuration for comparison by procrustes analysis with the recovered configuration at each stage further reduces the possibility of the occurrence of undetected local optima.

The third method we have considered has made little previous appearance in the literature; it is the method of least squares scaling, in which the aim is to find a configuration minimising $\Sigma\, w_{ij}(d_{ij}-\delta_{ij})^2$ where d_{ij} is achieved distance, δ_{ij} is dissimilarity and w_{ij} is a weight. The method is in fact related to a specific statistical model: if the errors by which the δ_{ij} differ from the d_{ij} are $N(0, 1/w_{ij})$ and independent, then we are carrying out maximum likelihood estimation. The particular appropriateness of this for the independent binomial model is clear. We consider it here in a general context, as a possible competitor for classical scaling and the Shepard–Kruskal method. Obviously the least squares method is "classical" in spirit, in that it attaches significance to the actual numerical values of the dissimilarities, but in other respects—for example the user defined dimensionality and the need for an iterative approach— it has more in common with the Shepard Kruskal method. The optimisation problem appears to be considerably better behaved than that arising in the Shepard–Kruskal method, Sibson and Bowyer (1980) found that the Fletcher–Reeves algorithm (Fletcher and Reeves, 1964) handles it very successfully. Previous references to least squares scaling are as follows. The least squares goodness of fit criterion is mentioned by Spaeth and Guthery (1969), but there is no indication that they have considered any actual method based upon it. Anderson (1971) also considers this idea, but gives no indication of having implemented a practical method. Chang and Lee (1973), following Sammon (1969) consider the special case where $w_{ij} = 1/\delta_{ij}$, a method Sammon calls "non-linear mapping". The least squares criterion occurs also in the Guttman–Lingoes methods (see Lingoes and Roskam 1973, p. 20) but in a context not quite parallel to the present one. Bloxom (1978) discusses a related but more complicated least squares method requiring a special computational algorithm.

We also give consideration to a preprocessing method of potential value in connection with classical scaling and least squares scaling. This is essentially a technique suggested by Young (1970). The first reference we are aware of for a method of this type is Benzécri (1964) his work is discussed by Shepard (1966). Faced with ordinal data we may obtain a provisional numerical structure by replacing the ranking numbers by suitably chosen quantiles from the distribution we would expect the distances to follow if the configuration were a sample of independent observations from, say, a multivariate normal distribution. The system of distances between independent points is not itself a system of independent random variables, but it is dissociated (McGinley and Sibson, 1975;

Silverman, 1976), and thus many parallel limit theorems apply; in particular, the empirical distribution function of the distances converges to the distribution of a single distance, and this provides the method with some kind of justification. We explore the effects of assigning numerical values to ordinal data under the assumption that the underlying configuration is spherical normal and also under the assumption that it is uniform on the disc.

6. THE SIMULATION EXPERIMENTS AND WHAT THEY SHOW

6.1 Classical scaling

Scaling has been applied to data from all four of the probabilistic models introduced in Section 3. For the binomial hyperplane model these have been related to original configurations lying in 2, 3, 4, 5 and 6 dimensions; for the independent binomial model in 2 and 6 dimensions; but for the others in 2 dimensions only. Extension to higher dimensions for the Jaccard distance model would be straightforward, but the Wilkinson metric model is less easily extended because the computation of general n-dimensional tessellations is not yet available.

We have used a variety of parent configurations, generated as in Section 3, but there seems no great dependence upon them so they are not recorded here individually.

For each model, in each dimensional space, we investigate the procustes statistic at six different levels corresponding to different intensities of the underlying Poisson process. For each model, level, and space we repeat the experiment ten times with different realisations of the Poisson process. There is one exception: for the Wilkinson metric model we use twenty replications.

To compare different numbers of dimensions we make the standardisation that the expected number of hyperplanes intersecting a line segment of given length should be a constant. Thus in six dimensions "1000 hyperplanes" should be interpreted as that number of hyperplanes which will give the same expected number of cuts of a line segment of length l, as would 1000 hyperplanes in two dimensions—generally this number will be higher as the number of dimensions increases.

For the Jaccard distance model the radius of discs that cut the unit disc is given an exponential distribution. We use the values 0.2 and 1.0 for the mean of this distribution.

For the Wilkinson metric model extra points are added to the central disc of radius two, the combination being tessellated in the square $-4 \leqq x, y \leqq 4$. Both of these are precautionary measures to minimise edge

TABLE I

Results from classical scaling experiments. Mean Procrustes Statistic (and sample standard deviation) from 10 values (or 20 for Wilkinson metric)

Model	No. of dimensions	No. of hyperplanes 20	50	100	200	500	1000
	2	0.1145 (0.0348)	0.0495 (0.0167)	0.0284 (0.0111)	0.0144 (0.0060)	0.0054 (0.0028)	0.0033 (0.0020)
	3	0.1642 (0.0623)	0.0727 (0.0160)	0.0376 (0.0106)	0.0206 (0.0057)	0.0082 (0.0032)	0.0039 (0.0010)
Binomial	4	0.2288 (0.0601)	0.0951 (0.0190)	0.0440 (0.0078)	0.0252 (0.0080)	0.0111 (0.0024)	0.0055 (0.0008)
hyperplane	5	0.2806 (0.0620)	0.1166 (0.0177)	0.0578 (0.0120)	0.0303 (0.0047)	0.0136 (0.0018)	0.0060 (0.0008)
	6	0.2974 (0.0534)	0.1387 (0.0299)	0.0699 (0.0097)	0.0375 (0.0047)	0.0142 (0.0025)	0.0069 (0.0005)
		No. of "hyperplanes" 20	50	100	200	500	1000
Independent	2	0.0494 (0.0081)	0.0184 (0.0033)	0.0093 (0.0013)	0.0049 (0.0009)	0.0018 (0.0004)	0.0009 (0.0001)
binomial	6	0.3923 (0.0237)	0.1737 (0.0335)	0.0764 (0.0086)	0.0367 (0.0040)	0.0142 (0.0014)	0.0070 (0.0007)
		No. of discs 20	50	100	200	500	1000
Jaccard		0.1498 (0.0358)	0.0995 (0.0359)	0.0532 (0.0098)	0.0318 (0.0091)	0.0166 (0.0022)	0.0142 (0.0021)
(mean rad 1.0)	2						
distance							
(mean rad 0.2)	2	0.5895 (0.1871)	0.2994 (0.1230)	0.2435 (0.1804)	0.1228 (0.0249)	0 1022 (0.0126)	0.0883 (0.0132)
		$\sqrt{(50+\frac{1}{4}}$ extra points) 7.1	11.2	15.0	18.0	21.2	26.5
Wilkinson							
metric	2	0.0579 (–)	0.0398 (0.0083)	0.0282 (0.0083)	0.0224 (0.0061)	0.0186 (0.0051)	0.0141 (0.0047)

effects. Since the Wilkinson distance increases approximately as the square root of the number of points in the unit disc we may use √(50 + $\frac{1}{4}$ extra points) to measure the level.

Mean values and sample standard deviations for the procrustes statistic are shown in Table I and the corresponding log log plots (Figures 6.1 6.4) provide an illuminating visual representation of the information in the table.

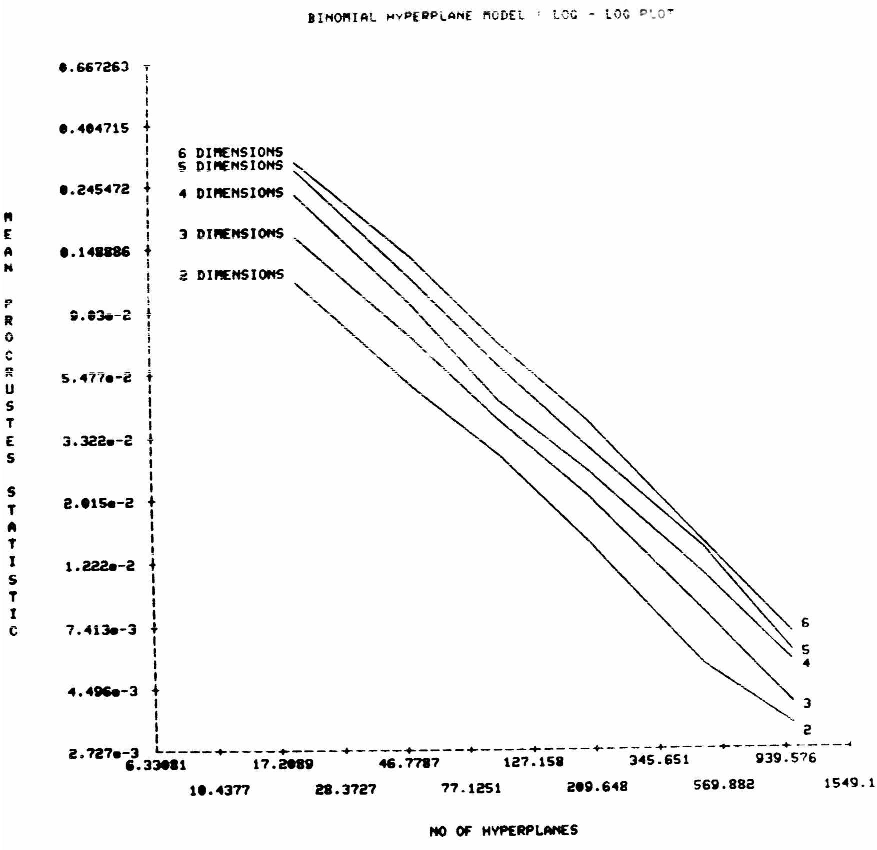

FIGURE 6.1

The distribution of the procrustes statistic is known to be approximately general χ^2 in type, see II, and so will have some skewness. We plot and record the mean value over the ten or twenty replications. The procrustes statistic increases with increasing dimensions; it decreases as the rates of the Poisson process increase; it is always smaller for the larger of the two disc radii in the Jaccard distance model.

Binomial hyperplane model—the log/log plots are remarkably linear, with slope close to −1. Thus the dominant term in the procrustes statistic is constant/no. of hyperplanes, where the constant depends upon the dimensionality, the standardisation used to compare dimensions, and the procrustes statistic normalisation.

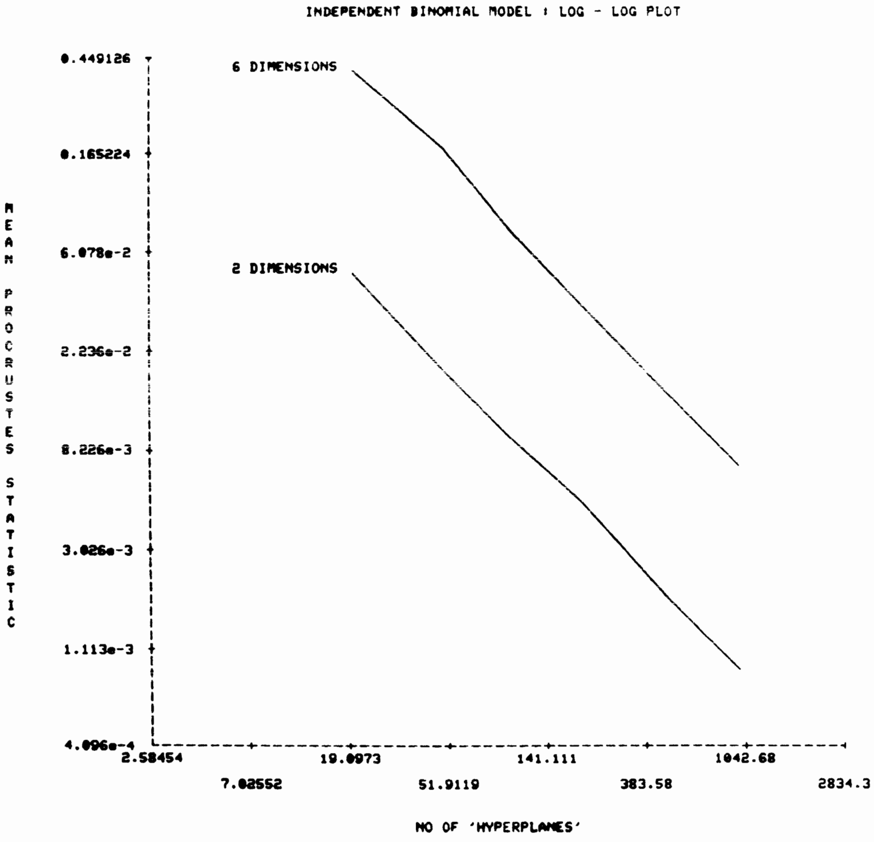

FIGURE 6.2

Independent binomial model—here again there is striking "constant/no. of hyperplanes" behaviour for both numbers of dimensions chosen. However the constant is quite different from that in the binomial hyperplane model. For two dimensional configurations the difference is a factor of about two. This represents one of the most interesting results.

The covariance structure has the effect of reducing the available information, certainly in the case of two dimensions. For six dimensions there is little difference.

Jaccard distance model—here the log/log plot is much flatter than before. There is the suggestion that classical scaling will never be able exactly to reproduce the original configuration, and that the procrustes statistic will not fall beneath a certain level dependent upon the disc radius distribution.

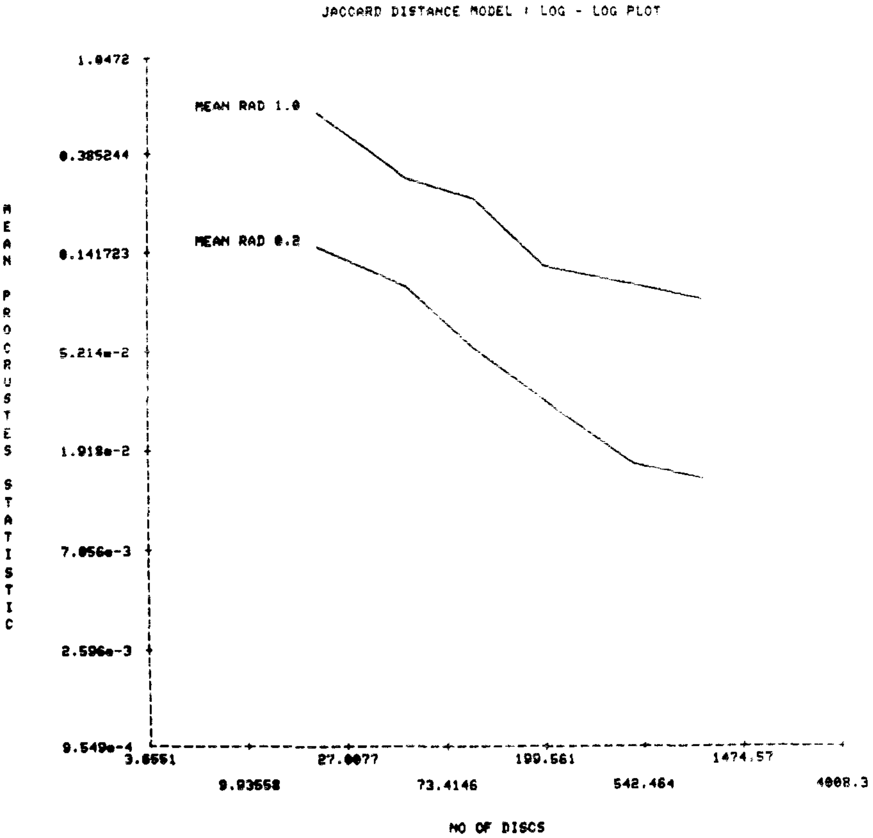

FIGURE 6.3

Wilkinson metric model—the log/log plot is again quite linear with slope about -1: This has been achieved only after the transformation of the number of extra points. It would seem that the procrustes statistic could be made arbitrarily small but more slowly, the behaviour being $1/\sqrt{}$ (No. of extra points).

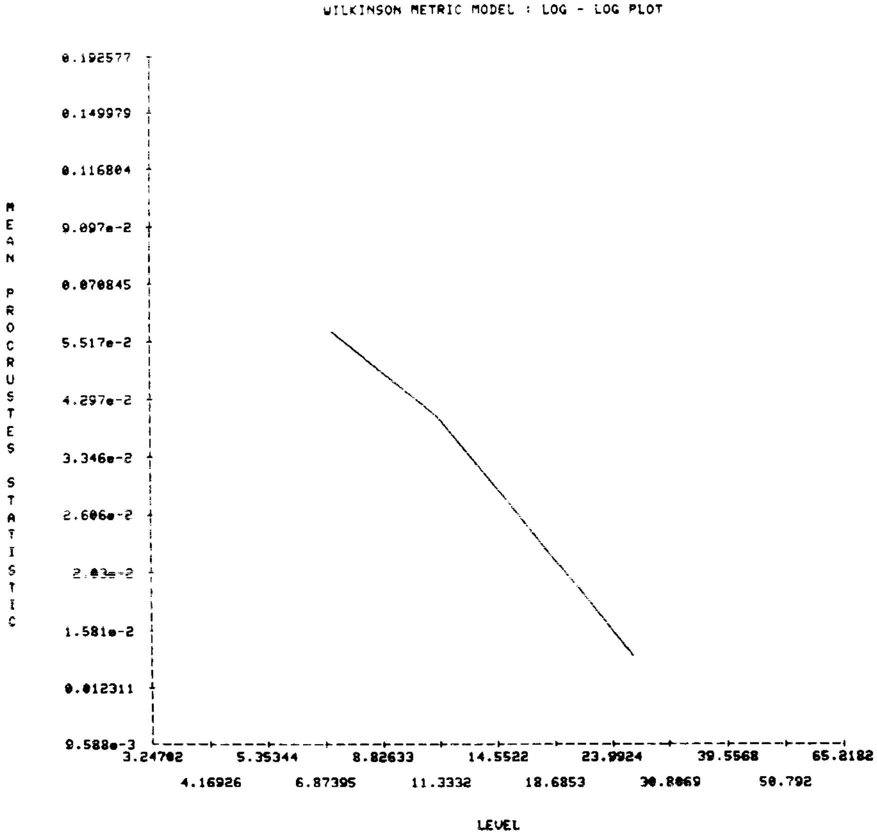

FIGURE 6.4

Eigenvalue Spectra

One extra difference between the binomial hyperplane and independent binomial models is that the former produces a more clearly determined dimensionality of configuration. The covariance structure is at work to provide this effect. Lingering perturbed zero eigenvalues for the independent binomial model occur at both choices of dimensionality. The information contained within these higher eigenvalues is available to be exploited by ordinal scaling (see Section 6.2). For the Jaccard distance model there are more positive perturbed eigenvalues as the rate of the Poisson disc process increases. However, the third eigenvalue decreases in loading so that it is possible to estimate the dimensionality of the parent configuration as the rate increases. These effects are caused by the nonlinearity between Jaccard and Euclidean distances, which forces many dissimilarities to be close and thus classical scaling tries to reproduce in higher and higher numbers of dimensions The Wilkinson metric model gives a clearly two-dimensional spectrum for which the perturbed zero eigenvalues drop away quite slowly.

Subsidiary techniques

The trace criterion, of II Section 2, for determining the dimensionality of the configuration suggests that the sum of genuine positive eigenvalues ought to be approximately equal to the sum of all the eigenvalues. Since we know the dimensionality of the original configuration it is possible to examine the success of this criterion by simply looking at the eigenvalue spectrum after classical scaling. We find, as might be expected, that the success is much dependent upon the linearity of the dissimilarity with euclidean distance. Thus the trace criterion works well for the independent binomial and binomial hyperplane models, especially at high numbers of hyperplanes: it works fairly well for the Wilkinson metric model: it is very poor for the Jaccard distance model, where the addition of discs forces more and more of the perturbed zero eigenvalues to become positive. So its use can only be recommended in the case when confidence can be placed in the linearity of the dissimilarity with euclidean distance.

Much the same applies to the magnitude criterion. As in II Section 2 we reject as spurious any positive eigenvalue whose magnitude does not substantially exceed that of the largest negative eigenvalue. When the perturbed zero eigenvalues are mainly positive the criterion will be useless, and this happens when there is much non-euclideanness. So in practice we find that the correct dimensionality will be found only in the hyperplane, and to a lesser extent, Wilkinson models.

6.2 Comparison of different scaling methods

Thirty dissimilarity matrices were derived, six from each of the probabilistic models except the Jaccard distance model which contributed twelve. These dissimilarity matrices were then used as input to the scaling methods which were thus compared on the same data. We considered only two-dimensional parent configurations. Although this involves only a moderate number of dissimilarity matrices, other simulations have been undertaken and the results seem similar enough to regard the values presented as typical.

The scaling methods used were classical scaling, ordinal scaling, least squares scaling with weights all 1, least squares scaling with weights $1/\delta_{ij}$

TABLE II

Results from comparisons of scaling methods. Procrustes Statistic for the method applied to a particular dissimilarity matrix.

Model	Method	No. of hyperplanes					
		20	50	100	200	500	1000
	classical	0.1574	0.0501	0.0170	0.0077	0.0039	0.0021
Binomial	ordinal	0.1396	0.0434	0.0148	0.0063	0.0032	0.0019
hyperplane	least squares (1)	0.1446	0.0442	0.0148	0.0062	0.0032	0.0019
	least squares $(1/\delta_{ij})$	0.1465	0.0436	0.0145	0.0059	0.0031	0.0019
		No. of "hyperplanes"					
		20	50	100	200	500	1000
	classical	0.0486	0.0177	0.0088	0.0042	0.0017	0.0008
independent	ordinal	0.0191	0.0076	0.0038	0.0018	0.0007	0.0004
binomial	least squares (1)	0.0172	0.0065	0.0032	0.0016	0.0006	0.0003
	least squares $(1/\delta_{ij})$	0.0304	0.0077	0.0035	0.0017	0.0006	0.0003
		No. of discs					
		20	50	100	200	500	1000
Jaccard	classical	0.6601	0.3700	0.2166	0.1427	0.1002	0.0877
distance	ordinal	0.6682	0.3804	0.1120	0.0479	0.0115	0.0081
(exponential	least squares (1)	0.8995	0.2405	0.0588	0.0296	0.0180	0.0169
radius	least squares $(1/\delta_{ij})$	0.8970	p.2740	0.0585	0.0279	0.0157	0.0148
distribution	preprocessing (normal)	0.6597	0.3127	0.1727	0.0951	0.0394	0.0292
mean 0.2)	preprocessing (uniform)	0.6597	0.3124	0.1665	0.0711	0.0216	0.0147

		No. of discs					
		20	50	100	200	500	1000
	classical	0.8998	0.8664	0.7603	0.7775	0.7668	0.7463
Jaccard	ordinal	0.8999	0.8298	0.7461	0.7359	0.6921	0.5922
distance	least squares (1)	0.9200	0.8595	0.9033	0.8433	0.9525	0.9850
(constant	least squares ($1/\delta_{ij}$)	0.9162	0.9007	0.8772	0.8291	0.9753	0.9606
radius	preprocessing (normal)	0.9000	0.7975	0.7355	0.7172	0.7118	0.6717
0.2)	preprocessing (uniform)	0.9000	0.7972	0.7354	0.7170	0.7117	0.6713
		$\sqrt{(50+\frac{1}{4}$ extra points$)}$					
		7.1	15.8	21.2	25.5	29.2	30.8
	classical	0.0579	0.0222	0.0095	0.0101	0.0111	0.0110
Wilkinson	ordinal	0.0506	0.0175	0.0076	0.0074	0.0095	0.0100
metric	least squares (1)	0.0493	0.0160	0.0076	0.0076	0.0093	0.0097
	least squares ($1/\delta_{ij}$)	0.0458	0.0155	0.0070	0.0069	0.0089	0.0091

("non-linear mapping"). In addition we used the pre-processing technique on all of the Jaccard dissimilarities, followed by classical scaling. Here it was assumed that the parent configuration resembled a sample from a spherical normal distribution, or alternatively a sample from a uniform distribution on the disc, which is the actual parent distribution.

To avoid the problem of local optima, ordinal scaling was used both with random starting configurations and with the configuration output by classical scaling. The least squares methods were always used with the configuration output from classical scaling as a starting configuration.

For the Jaccard distance model one set of six dissimilarity matrices was derived from a constant radius distribution of 0.2, the other set of six were from an exponential disc radius distribution of mean 0.2. The convention for recording Wilkinson metric model levels is maintained Where there are n extra points added in the disc of radius 2 the level is recorded as $\sqrt{(50+\frac{1}{4}n)}$.

Preprocessing method: following the idea of III Section 5 we replace the ith ordered dissimilarity by the appropriate quantile of the interpoint distance distribution. Under the assumption that the configuration is spherically normally distributed with unit variances the distribution of squared interpoint distances is approximately a $2\chi_2^2$ distribution. Under the assumption that the configuration is uniformly distributed on a disc of radius one, the interpoint distance density is (see e.g., Bartlett, 1964) $4r/\pi\{\cos^{-1}\frac{1}{2}r-\frac{1}{2}r\sqrt{(1-\frac{1}{4}r^2)}\}$ $(0\leqq r\leqq 2)$. Where there happen to be ties the transformed values may be averaged. There is no point in considering trace or magnitude criteria because we have imposed the dimensionality of the reconstruction.

The results are summarised in Table II. We look at them first from the point of view of the probabilistic models.

Binomial hyperplane model—all methods of scaling produce reconstructions yielding procrustes statistics of similar low value. In particular classical scaling compares well with the others. For higher numbers of hyperplanes the least squares method with weights $1/\delta_{ij}$ is superior; this fits the theory of Section 5 concerning maximum likelihood estimation.

Independent binomial model—in this case classical scaling is markedly inferior to the other methods, which are able to exploit the information contained in the perturbed zero eigenvalues that persist for this model. The least squares methods work even better than ordinal scaling; and the maximum likelihood theory is seen to be valid for higher numbers of hyperplanes where weights $1/\delta_{ij}$ are again appropriate.

Jaccard distance model—variable radius of disc—classical scaling works very badly compared to ordinal scaling. This is no surprise since the dissimilarity bears little resemblance to euclidean distance. The least squares methods fail for the same reason when the number of discs is large, but are surprisingly good for moderate numbers. The performance of classical scaling can be significantly improved by either of the preprocessing transformations, especially that which assumes, correctly, an underlying uniform distribution for the configuration.

Jaccard distance model—constant radius of disc—all dissimilarities corresponding to points at distance greater than 0.4 will be equal and have value unity. The least squares methods are unable to cope with this sparsity of differentiation and consistently produce bad reconstructions. Classical scaling is slightly better and can be still improved by the preprocessing transformations; although there is little to choose between them. Ordinal scaling never produces a significantly worse reconstruction: it is again clearly best for higher numbers of discs. However none of the methods are able to produce reconstructions yielding low procrustes statistics due to the nature of the dissimilarity function.

Wilkinson metric model—classical scaling is a worthy competitor for the other methods: although not superior to the other methods it is never much worse. The least squares methods are generally superior, especially when the weights are $1/\delta_{ij}$.

We may also look at the results from the point of view of the scaling methods.

Classical scaling—compares well with ordinal scaling for the most euclidean-like models except where there is much information in the higher eigenvalues (independent binomial). For the least euclidean model it compares unfavourably, as would be expected.

Least squares scaling—generally slightly superior to ordinal scaling for the more euclidean models but certainly inferior for the Jaccard distance model. The use of weights $1/\delta_{ij}$ (that is, Sammon's (1969) "nonlinear mapping") is usually to be recommended.

Ordinal scaling—even when there is useful information in the numerical values of the dissimilarities, ordinal scaling is never significantly worse than methods which take advantage of the numerical values, provided that it is given a sensible starting configuration. About one-half of the runs from random starts had not converged adequately within 50 iterations. It is only on the Jaccard distance, constant radius model that ordinal scaling really fails to obtain a reasonable reconstruction, and even there its overall performance is better than that of other methods once the number of discs becomes large. On the Jaccard distance, exponential radius model the superiority of ordinal scaling over all alternative methods is very apparent at larger numbers of discs.

Preprocessing techniques—these seemed quite successful in improving the performance of classical scaling. However, the marked superiority of the (correct) uniform assumption as against the (incorrect) spherical assumption in our experiments suggests that the technique may be quite sensitive to the details of the underlying structure. Under the assumption of normality there are too many large dissimilarities after transformation and the dissimilarity/distance plot becomes blurred at higher distances, lying in a more convex shape. The corresponding plot under the uniform assumption produces the familiar cigar-shaped scatter.

The time taken by the algorithms varies as follows

$$\text{classical} \ll \underset{\substack{\text{(random}\\ \text{start)}}}{\text{ordinal}} < \underset{\substack{\text{(classical}\\ \text{start)}}}{\text{least squares}} < \underset{\substack{\text{(classical}\\ \text{start)}}}{\text{ordinal}} < \underset{\substack{\text{(random}\\ \text{start)}}}{2 \times \text{ordinal}}$$

As a less serious footnote we mention the procrustes statistics obtained when we reconstruct a map of Great Britain from the A.A. Handbook familiar road distance table:

Classical scaling	0.02965
Ordinal scaling	0.03194
Least squares scaling (1)	0.02843
Least squares scaling ($1/\delta_{ij}$)	0.02824

There is not much to choose between the methods in this case, although classical scaling is more liable to misplace a few individual points badly. Ordinal scaling from a random start is prone to produce local minimum solutions with individual towns placed on the wrong side of the main North-South axis, or, in one extreme case, with the whole of Scotland reflected about this axis!

7. CONCLUSIONS

We have gained insight into the use of procrustes statistics, developing a feel for the absolute values they take and the variability to be expected. We have investigated the effect of dependence among dissimilarities and seen that it may substantially influence accuracy of reconstruction. The behaviour of the procrustes statistic after classical scaling has been clearly related to the euclideanness of the dissimiliarity function. The comparative experiments show the superiority of ordinal scaling, especially for non-euclidean dissimilarity functions, from the relative value of the procrustes statistics for the various methods. Both least squares scaling and the preprocessing technique have been shown to possess useful properties.

Acknowledgement

The provision of research assistance and facilities by the Social Science Research Council is acknowledged with gratitude.

References

Anderson, A. J. B. (1971). Ordination methods in ecology. *Journal of Ecology* **59**, 713–726.

Bartlett, M. S. (1964). The spectral analysis of two-dimensional point processes. *Biometrika* **51**, 299–311.

Benzécri, J. P. (1964). Analyse factorielle des proximités. *Publications de l'Institute de Statistique de l'Université de Paris I*, **13**, 235–282.

Bloxom, B. (1978). Constrained Multidimensional Scaling in *N* Spaces. *Psychometrika* **43** 397–408.

Chang, C. L. and Lee, R. C. T. (1973). A heuristic relaxation method for non-linear mapping in cluster analysis. *I.E.E.E. Trans on Systems, Man and Cybernetics* , 197–200.

Cohen, H. S. and Jones, L. E. (1974). The effects of random error and sub-sampling of dimensions on recovery of configurations by non-metric multidimensional scaling. *Psychometrika* **39**, 69–90

Fletcher, R and Reeves, C M. (1964). Function minimization by conjugate gradients *Computer Journal* 7, 149–153.

Gower, J C (1971a) A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–871

Gower, J. C. (1971b). Statistical methods of comparing different multivariate analyses of the same data. In *Mathematics in the Archaeological and Historical Sciences* (F R. Hodson, D. G. Kendall and P. Tautu, eds) pp. 138–149, Edinburgh: University Press.

Gower, J. C. and Banfield, C. F. (1974). Goodness of fit criteria for hierarchical classifications and their empirical distributions. In *Processing of the 8th International Biometric Conference 1974, Constanza, Romania* (eds. L. C. A. Corsten and T. Postelnicu), Romanian Academy of Sciences Press, Bucharest (1975). 347–361.

Green, P. J. and Sibson, R. (1978). Computing Dirichlet tessellations in the plane. *Computer Journal* **21**, 168–173.

Isaac, P. D. and Poor, D. D. S. (1974). On the determination of appropriate dimensionality in data with error. *Psychometrika* **39**, 91–109.

Kendall, D. G. (1971). Construction of maps from "odd bits of information." *Nature* **231**, 158–159.

Kendall, D. G. (1974). The recovery of structure from fragmentary information. *Philosophical Transactions of The Royal Society A. Mathematical and Physical Sciences* **279**, 168–173.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. *Psychometrika* **29**, 1–27.

Kruskal, J. B. (1964b). Non-metric multidimensional scaling: a numerical method. *Psychometrika* **29**, 115–129.

Lingoes, J. C. and Roskam, E. E. (1973). A mathematical and empirical analysis of two multidimensional scaling algorithms. *Psychometrika* **38**, Monograph Supplement 1–93.

McGinley, W. G. and Sibson, R. (1975). Dissociated random variables. *Math. Proc. Camb. Phil. Soc.* **77**, 185–188.

McGinley, W. G. (1977). *Some Optimisation Problems in Data Analysis*. Ph.D. thesis, University of Cambridge

Mardia, K. V. (1970). *Families of Bivariate Distributions*. Griffin, London.

Miles, R. E. (1970). On the homogeneous planar Poisson process. *Mathematical Biosciences* **6**, 85–127.

Sammon, J. W. (1969) A nonlinear mapping for data structure analysis. *I.E.E.E. Trans. on Computers* **18,** 401–409

Shepard, R N (1966). Metric Structures in Ordinal Data. *Journal of Mathematical Psychology* **3,** 287 315.

Shepard, R N (1974). Representation of structure in similarity data: problems and prospects. *Psychometrika* **39,** 323–355.

Sherman, C. R. (1972). Nonmetric multidimensional scaling A Monte Carlo study of the basic parameters. *Psychometrika* **37,** 323 355.

Sibson, R. (1972). Order invariant methods for data analysis. *Journal of the Royal Statistical Society, Series B (Methodological).* **34,** 311 349.

Sibson, R. (1978). Studies in the robustness of multidimensional scaling: procrustes statistics. *Journal of the Royal Statistical Society, Series B (Methodological),* **40,** 234 238.

Sibson, R. (1979a). Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *Journal of the Royal Statistical Society, Series B (Methodological)* **41,** 217 229.

Sibson, R. (1979b). The Dirichlet tessellation as an aid in data analysis. *Scandinavian Journal of Statistics* **7,** (1980), 14 20.

Sibson, R and Bowyer, A (1980). Trilateration and scaling methods in surveying and photogrammetry (to appear)

Silverman, B. W (1976). Limit theorems for dissociated random variables *Advances in Applied Probability* **8,** 806 819.

Spaeth, H J and Guthery, S. B. (1969). The use and utility of the monotone criterion in multidimensional scaling *Multivariate Behavioural Research* **4,** 501 515

Spence, I. (1970). Local minimum solutions in nonmetric multidimensional scaling. *Proc. of the Soc. Stats. Section of the American Statistical Association* **13,** 365–367.

Spence, I. (1972). A Monte Carlo evaluation of three nonmetric multidimensional scaling algorithms. *Psychometrika* **37,** 461–486.

Spence, I. and Ogilvie, J. C. (1973). A table of expected stress values for random rankings in nonmetric multidimensional scaling. *Multivariate Behavioural Research* **8,** 511–517.

Spence, I and Graef, J (1974). The determination of the underlying dimensionality of an empirically obtained matrix of proximities, *Multivariate Behavioural Research* **9,** pp. 331–341.

Stenson, H. H. and Knoll, R. L. (1969). Goodness-of-fit for random rankings in Kruskal's nonmetric scaling procedure. *Psychological Bulletin* **71,** 122 126.

Wagenaar, W. A. and Padmos, P. (1971) Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. *British Journal of Mathematical and Statistical Psychology* **24,** 101–110.

Young, F. W. (1970). Nonmetric multidimensional scaling: recovery of metric information. *Psychometrika* **35,** 455–473.