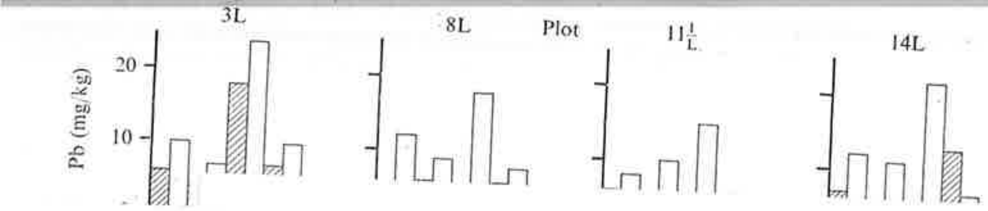

6986

Offprint from
Mathematics in the Archaeological and Historical Sciences
Edinburgh University Press
1971

Proceedings of the Anglo-Roumanina conference
Mamaia 1970. Edited by F.R. Hodson, D.G. Kendall &
P. Tautu.

John C. Gower

Statistical methods of comparing different multivariate analyses of the same data

This paper considers how to compare two sets of distances d_{ij} and d_{ij}^* $(i, j=1, 2, \ldots, n)$ amongst the same n samples. Rather than correlate the $\binom{n}{2}$ distance pairs (d_{ij}, d_{ij}^*) it is suggested that each set of distances be represented by n points P_i, P_i^* $(i=1, 2, \ldots, n)$ which are rotated to best fit defined by minimizing $R^2=\sum_{i=1}^{n} \Delta^2(P_iP_i^*)$. The mathematical technique required is useful with many different multivariate problems. It is illustrated with a new type of analysis using anthropological data on skulls from six hominoid populations with eight recognizable constellations of characters. A canonical variate analysis for each constellation gives eight sets of canonical variate means and each pair (u, v) is rotated to best fit R^2_{uv}. The elements of the 8×8 symmetric R^2 matrix can themselves be treated as distances and represented by points in three dimensions, allowing examination of how the descriptions of the populations are related when analyzed by different constellations of characters. Some of the statistical distributional problems raised by this and similar types of analysis are discussed.

INTRODUCTION

The sort of problem I shall discuss is well illustrated by the long standing dispute over the relative merits of Mahalanobis's D^2 statistic and Karl Pearson's Coefficient of Racial Likeness (CRL) in anthropometry. With k populations there are $\frac{1}{2}k(k-1)$ derived distances for each of the two measures. Many authors have noted that when these $\frac{1}{2}k(k-1)$ pairs of values are plotted against each other, a strong linear relationship expressible as a large positive correlation is often found. From this they deduce that whatever theoretical advantages D^2 may have, the more simply computed CRL is, for all practical purposes, just as good.

That the $k(k-1)$ distances are not independent may well produce a specious correlation, but this is usually overlooked. The effect of non-independence can be seen by considering k populations, $k-1$ of which form a homogeneous set, all of which are very different from the single remaining population. With both D^2 and CRL we shall have $k-1$ long distances and $\frac{1}{2}(k-1)(k-2)$ short distances, and the correlation between D^2 and CRL must therefore be large (*see* figure 1).

This paper describes some preliminary work on a statistic (R^2) for comparing different sets of distances, without using dependent values. The example cited above is only one of many ways that a distance matching problem arises, but before outlining other examples it will be convenient to examine some of the different types of distance used in multivariate analysis.

Many multivariate methods may be regarded as two-step processes. First a set of distances $d_{ij}(i, j=1, 2, \ldots, n)$ between n points, representing samples or populations, is defined. Examples are D^2, CRL, Hiernaux's Δ_g, Penrose's C_Q^2, Sanghvi and Balkrishnan's B and G, and various dissimilarity coefficients all discussed by Gower (1970). Second, these distances are mapped onto a set of n points (preferably in few dimensions) with Euclidean distances d_{ij}^*. The techniques used here include canonical variate analysis, principal components, multi-dimensional scaling, and so on, where d_{ij}^* is chosen to minimize some function of d_{ij} and d_{ij}^*. Alternatively, the mapping is onto a dendrogram (scaled hierarchical representation) using some form of nested cluster analysis. In this case, the dendrogram can be regarded as defining ultra-metric distances d_{ij}^+ ideally chosen to minimize some function of d_{ij} and d_{ij}^+.

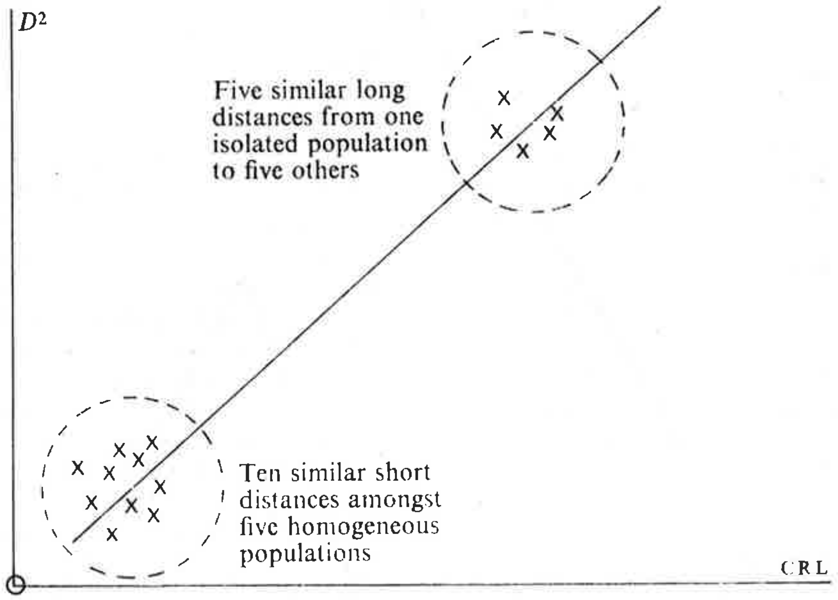

Figure 1. Relation of D^2 to CRL distances between six populations, five forming a homogeneous set and one being an outlier

The derivation, described below, of ultra-metric distances from a dendrogram seems first to have been suggested by Sokal and Rohlf (1962) when defining co-phenetic correlation, but was first set out more formally by Hartigan (1967).

Figure 2 is a simple dendrogram illustrating the hierarchical representation of 5 populations by points A, B, C, D, E, with a scale labelling the branching points or nodes. The distance between any two points is given by the scale value corresponding to the node where they first join. Thus, in figure 2 the distance A, B is 1 and the distance A, D is 7. Also from figure 2, concentrating on the points A and C, we can see that for any point X between A and C (B, for example) that

$$AC = \max(AX, XC) \quad .$$

However, for any point X outside A and C (D, for example) then $AX = XC$ and

$$AC < \max(AX, XC) \quad .$$

Thus, for all points ABC

$$AC \leqslant \max(AB, BC) \quad .$$

This is the main property defining ultra-metric distances. Clearly $AB + BC \geqslant AC$, the triangle rule for Euclidean distances. This is insufficient to show that with a finite set of points ultra-metric distance is a special case of Euclidean distance, but I conjecture that ultra-metric distances are also Euclidean distances. [This conjecture has now been proved true (Buneman and Gower 1971).]

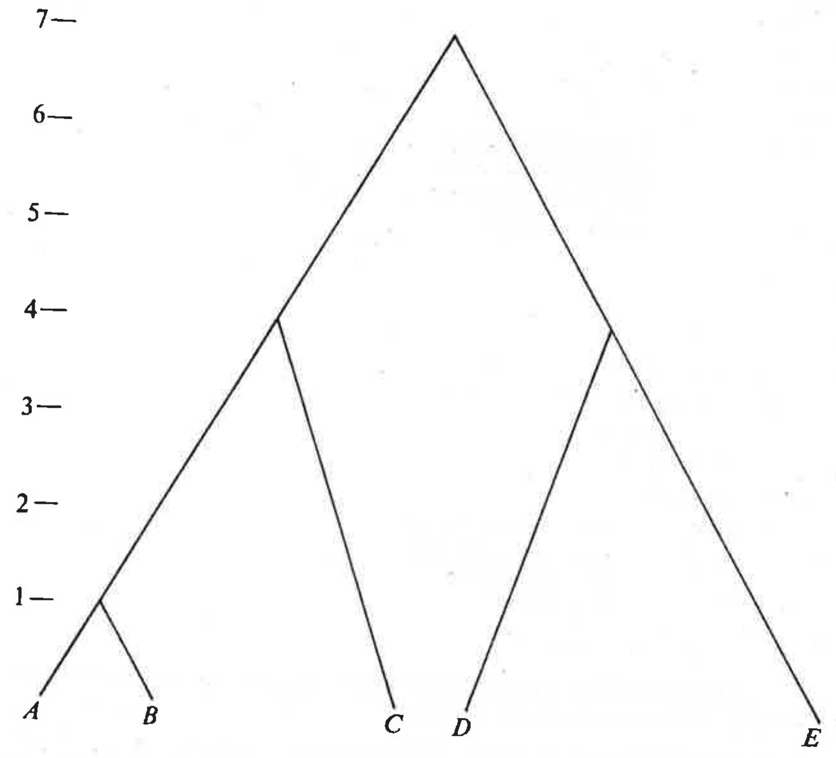

Figure 2. Simple dendrogram illustrating properties of ultrametric distances: $d_{AB}=1; d_{AC}=d_{BC}=4; d_{AD}=d_{CD}=7; d_{AC}=\max(d_{AB}, d_{BC}); d_{AC}<\max(d_{AD}, d_{CD})$

Thus, many forms of multivariate analysis can be regarded as mapping computed distances d_{ij} on to Euclidean distances d^*_{ij} or ultra-metric distances d^+_{ij}. As there are many ways of computing d_{ij} and of analyzing it, we are constantly coming across problems of comparing two different sets of distances pertaining to the same samples or populations. The following are important practical problems of this type.

(1) Comparing different distances derived from the same observations on the same samples (as with D^2 and CRL discussed above). An archaeological example is where an archaeologist (or two different archaeologists) have scored hand-axe data in different ways or decided to examine different measures of distance based on the same hand-axe data.

(2) Comparing different distances (possibly using the same statistical formulae) derived from different observations on the same samples. For example, comparing D^2 as evaluated on one set of variates (say concerned with the jaw-bone) with D^2 evaluated on another set (say from the frontal region of the skull). An analysis of this kind is given below in the section 'Anthropometric example'. In archaeology we may wish to compare different criteria (or graves), first on biological properties (skeletal measurements), and secondly on artifacts.

(3) Comparing the original distances d_{ij} with those obtained by analysis, say d^*_{ij} or d^+_{ij}. That is to say, we want to see how well the distances derived from the analysis agree with the original values. In fact d^*_{ij} or d^+_{ij} are often defined by optimizing a function of d_{ij} with these fitted values.

(4) Comparing distance d^*_{ij} and d^{**}_{ij} derived from two different analyses of the same distances. For example, a set of D^2 values may be expressed in two dimensions by (a) canonical variate analysis, giving d^*_{ij} and (b) non-metric multi-dimensional scaling giving d^{**}_{ij}. Under this heading we can include comparison of d^*_{ij} with d^+_{ij} and comparison of d^+_{ij} with d^{++}_{ij}. An analysis of this kind is given below in the section 'Anthropometric example'.

(5) Comparing distances derived from different samples from the same populations. Thus, we may evaluate D^2 from one set of samples and then re-sample to obtain a second set of D^2 values, pertaining to the same populations. Problems of this kind are theoretically important, for their solution forms the basis of any statistical inference.

Just as some authors have correlated d^*_{ij} with d^{**}_{ij} others have suggested correlating d^+_{ij} with d^{++}_{ij} or with d_{ij}. I have already explained why I think this can be misleading and shall now outline the derivation of an alternative statistic.

Rotational Fits

Rather than concentrate on the distances themselves, consider geometric points $P_i (i=1, 2, \ldots, n)$ that give rise to all the inter-distances d_{ij}. With any Euclidean distance, the coordinates of these points can be evaluated by principal coordinates analysis (Gower 1967). The required coordinates are given by the canonical means in canonical variate analysis and by the data themselves in principal components analysis. Thus, the problem is to compare two sets of distances arising from points P_i and Q_i (say). Clearly we can move the points Q relative to P with translations and rotations. Reflection must also be considered as can be seen from figure 3, which shows how two congruent triangles best fit without reflection. The criterion of best fit adopted is to move the points Q_i relative to the points P_i until the 'residual' sum of squares $R^2 = \sum_{i=1}^{n} \Delta^2(P_i Q_i)$ is minimum. It can be shown that the best fit occurs when the two sets of points have the same centroid and this takes care of translation. To determine the required rotation, we need some matrix algebra.

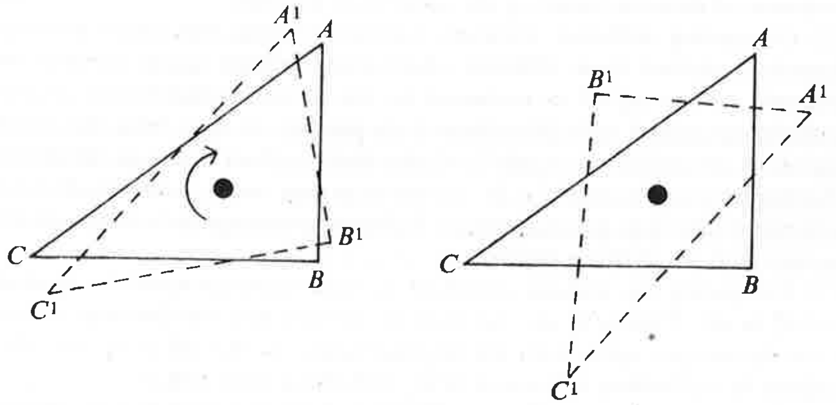

Figure 3. The effect of rotating two congruent triangles in a plane. The two triangles on the left hand side fit exactly. On the right hand side one triangle is a mirror image of the other and the best fit when rotating in a plane is poor; if one is allowed to rotate in three dimensions the fit is again exact

Let X be the $n \times p$ matrix of the coordinates of the points P_i (referred to orthogonal axes and with zero means) and Y the $n \times q$ matrix of the coordinates of the points Q_i (also orthogonal axes and zero means); where we have assumed that the ith row of both matrices refers to the same samples or populations. There is no loss of generality in assuming that $p \geqslant q$ and, to avoid discussing the cases $p=q$ and $p>q$ separately, I shall further assume that $p=q$; if necessary extra zero columns must be appended to Y until it has a total of p columns. Any rotation of Y relative to X can be expressed as an orthogonal matrix H. After rotation, the coordinates of Q_1 are given by the rows of YH. The best estimate of H requires $R=X'Y$ and the solutions to the equation $R=U\Sigma V'$ to be computed where U and V are orthogonal matrices and Σ is a diagonal matrix whose elements are known as the zeros of R. The values of U, V, and Σ are probably best computed by the method recently given by Golub and Reinsch (1970), but in the example given below they have been computed as the latent vectors of RR' and $R'R$. The required rotation is given by $H=VU'$.

After rotation we have

$$R^2 = \text{Trace}\,(XX' + YY' - 2\,YHX') \quad .$$

We have to consider the effect of the arbitrary signs of the latent vectors which are the columns of U and V. These signs determine different reflections of the rotated Y and the best out of the 2^n different possibilities is required. To find the best reflection suppose S is a diagonal matrix whose elements are all $+1$ or -1, in an order to be determined. To change the signs of the columns of V relative to those of U we can write $H=VSU'$. Thus, to minimize R^2 above, we must select S so that the Trace $(YVSU'X')$ is maximum.

This is the same as maximizing Trace $(U'X'YVS)$ (that is, Trace $[\Sigma S]$) which occurs when the elements of S have the same signs as those of Σ. The elements of S determine the signs to be associated with the columns of V and take care of reflection.

A reflection in n dimensions is the same as a rotation in $(n+1)$ dimensions. So it seems that the above method of dealing with reflections could be avoided merely by adding an extra zero column to X and Y, but the arbitrariness of the signs of the latent vectors still remains a problem; so nothing is gained.

A further complication must be considered when the scales of the two sets of distances are arbitrary. For example, in problem (5) above, d_{ij}^* and d_{ij}^{**} are on the same scales but this is not so in problem (1). To take care of scale changes, we could scale the coordinates in the matrix Y by a factor δ and estimate δ by Δ to get minimum R^2. Luckily this estimation proceeds independently of translation and rotation, to give Δ Trace $(YY')=$Trace $(YHX)=$Trace (Σ). It can be shown that after rotation:

$$\Delta^2\ \text{Trace}\,(YY') + R^2_{\min} = \text{Trace}\,(XX') \quad .$$

This may be used as the basis for an analysis of variance, interpreted as equating the total sum of squares amongst the variates of Y, after scaling, plus the residual sum of squares, to the total sum of squares amongst the variates of X.

Clearly the best system of scaling Y relative to X is not the inverse of the best system of scaling X relative to Y. This unfortunate property needs further investigation, but to avoid it in the numerical problem discussed below, I have scaled both sets of points to have unit total squared distance from their respective centroids, that is, Trace $(XX')=$Trace $(YY')=1$. With this unit scaling, the analysis of variance simplifies to $\Delta^2 + R^2_{\min} = 1$ where $\Delta^2=$Trace (Σ), a value independent of whether we rotate X to Y or vice versa.

A more difficult problem occurs when we do not know how the rows of X match with the rows of Y, that is, we believe that some permutation of the rows Y may match those of X. Suppose P is a permutation matrix such that YP matches X without regard to rotation. A permutation matrix has a single unit in every row and column, and zeros everywhere else. To estimate P, consider the class of doubly stochastic matrices (that is, those which have non-negative elements and whose rows and columns all sum to unity). A permutation matrix is a special case of this class. To maximize the sum of squares

$$R^2 = \text{Trace}\,[(X - YP)'(X - YP)]$$

subject to the n^2+2n linear restrictions $0 \leqslant p_{ij} \leqslant 1$.

$$\sum_{i=1}^{n} p_{ij} = \sum_{j=1}^{n} p_{ij} = 1$$

is a quadratic programming problem which can in theory be solved by the standard methods available. That the optimum must occur on a vertex of the feasible region ensures that a true permutation matrix is found. It seems worth mentioning this version of the problem in case it occurs in an archaeological or historical context.

The problem of seriating the u rows of a matrix X whose elements relate to presence/absence or frequencies of different grave articles can be put into the same framework. We require to find an $n \times n$ permutation matrix P which permutes the rows of X into an optimum for $Y=PX$. A suitable optimality criterion S might be to minimize the sum of squares of the differences between adjacent rows of Y, that is, choose:

$$S=\sum_{i=1}^{n-1} \sum_{j=1} (Y_{ij}-Y_{i+1,j})^2$$

subject to the same linear restrictions as before.

Even more complicated problems would occur if we were to combine permutations with rotation and scaling, and so on, but these problems are not considered any further here.

Anthropometric Example

I am indebted to Mr A. Bilsborough, of the Department of Anthropology, Cambridge, for permission to use the extensive set of data he has collected on skull measurements of ancient human populations. To illustrate the methods discussed above, without giving a thorough anthropometric account here, I have selected six of the hominoid populations, namely:

1. Modern Homo Sapiens
2. Upper Palaeolithic Homo Sapiens
3. Middle East Neanderthal
4. European Würm Neanderthal
5. Late (Pekin) Homo Erectus
6. Australopithecus Africanus;

and measurements from eight different regions of the skull, namely:

1. Upper Face (16)
2. Upper Jaw (15)
3. Articular Region (8)
4. Balance (14)
5. Basicranial Region (12)
6. Cranial Vault (16)
7. Lower Jaw (16)
8. Overall (16).

Each region was characterized by a set of variates referred to as a constellation. The number of variates at each constellation is given in brackets in the above list. For each constellation the Mahalanobis D^2 distances were computed for all 15 population differences, giving eight such matrices in all. For each pair u, v of constellations, the best rotational fit can be found using the canonical variate means as the coordinates representing the populations; this gives the residential fit R^2_{uv}. The analysis is that required for problem (2) above. When all constellations have been rotated to fit all other constellations we have an 8×8 symmetric matrix of R^2 values, shown in table 1.

The present analysis has similarities to the INDSCAL method discussed by Wish and Carroll (1971). In their analysis, each individual constellation is represented relative to the axes of an overall analysis. In the above, each constellation is represented by its own canonical analysis and then combined in a separate constellation analysis. To get a metric equivalent to INDSCAL all that would be required is, first, to provide an overall (canonical) analysis using all the varieties (regardless of what constellation they belonged to), and then to rotate each separate constellation canonical analysis to fit this overall analysis. The R^2 analysis may be taken further, as described in the remainder of this section.

	1	2	3	4	5	6	7
1. Upper Face							
2. Upper Jaw	1·0012						
3. Articular Region	1·0753	0·5766					
4. Balance	1·0530	0·6324	1·0997				
5. Basicranial Region	0·3485	0·5736	0·6533	0·5486			
6. Cranial Vault	0·8332	0·5596	0·8034	0·2582	0·3466		
7. Lower Jaw	1·0275	0·8155	0·4385	0·5952	0·5541	0·3309	
8. Overall	0·8498	0·2147	0·5483	0·4580	0·3504	0·4155	0·5075

Table 1. Values of R^2 between eight constellations, based on best rotational fits of canonical variate means for six hominoid populations

These R^2 values may now themselves be regarded as squared distances. (In this example these distances turned out to be Euclidean but I have been unable to prove that this is necessary.) The R^2 distances were analyzed in two ways to give low-dimensional representations of the eight different constellations.

The first analysis using principal coordinates (Gower 1967) gave figure 4. Psychologists refer to this type of analysis as a metric multi-dimensional scaling.

The second analysis using non-metric multi-dimensional scaling (Kruskal 1964) gave figure 5.

Both methods needed three dimensions to express the distances of table 1 adequately (two and three dimensions accounted respectively for 62% and 84% of the total squared distance from the centroid with principal coordinates and gave respective stresses of 0·146 and 0·044 with multi-dimensional scaling). The third dimension is represented in the figures by a horizontal arrowed line of appropriate length (to the right if positive, to the left if negative). It is hard to say how figures 4 and 5 compare. Their general agreement becomes clearer when both figures are referred to their principal axes and it would ease such comparisons if all multi-dimensional coordinates of maps produced by whatever method of analysis were presented in this way. To get more information on this comparison, the coordinate values depicted in figures 4 and 5 were used as starting points for an analysis [of the type discussed in problem (4) above] rotating one analysis of the distances of table 1 (principal coordinates) to fit best another analysis of the same distances (non-metric scaling). This was done (*see* figure 6) and gave an R^2 value of 0·066. Although

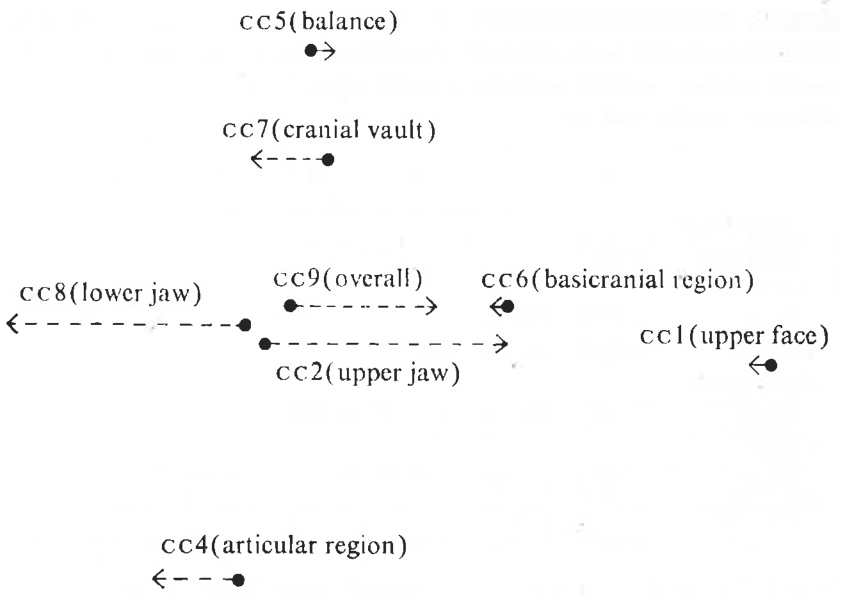

Figure 4. 3-dimensional principle coordinate fit of the R^2 matrix

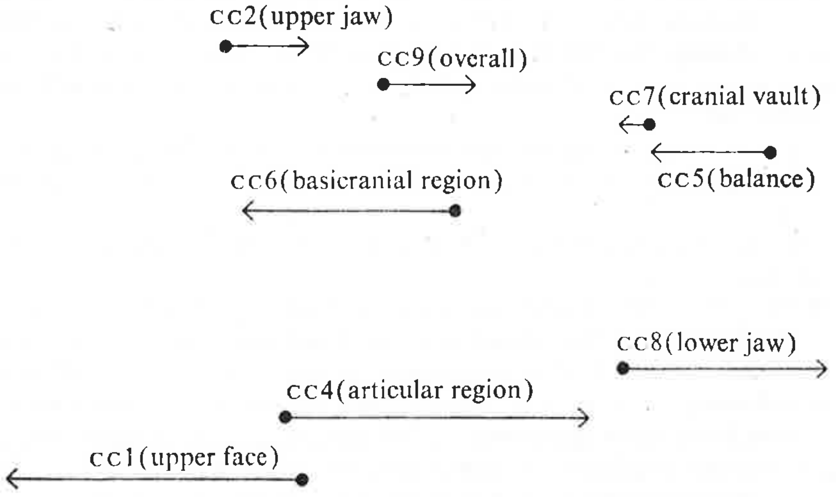

Figure 5. 3-dimensional MDSCAL ($r=3$) fit of the R^2 matrix. This figure does not obviously agree with figure 4

as yet we know nothing of the sampling properties of R^2, this value seems satisfactorily small.

The multi-dimensional scaling program when asked to give a fit in r dimensions also gives a fit in all lower dimensions, using the coordinates found at the sth stage as starting values for the solution in one fewer dimension at the $(s+1)$th stage. With this example I set $r=3$ and 4, and so got two different 3-dimensional solutions (and two 2-dimensional solutions too). Figure 5 is the solution for $r=3$. Although ideally the 3-dimensional solutions for $r=3$ and $r=4$ should fit each other exactly after a suitable rotation, they will differ because of the effects of whatever stopping rule is used to define convergence, and also because convergence may be to different local optima. In fact $r=3$ gave a 3-dimensional fit with stress 0·044 and $r=4$ gave a 3-dimensional fit with stress 0·042 which, although close, did not give obviously similar representations. When the two results were rotated to best fit, it was clear that both solutions were much the same, but the value 0·091 of R^2 was worse than the R^2 for figure 6, indicating that $R^2=0·066$ is satisfactory and that principal coordinates and non-metric scaling have given, effectively, the same results with these data. The 3-dimensional solutions for $r=3, 4, 8$ are compared with the principal coordinates solution in table 2. It is remarkable that each of the three MDSCAL solutions fit the principal coordinates solution better than any of the other 3-dimensional MDSCAL solutions.

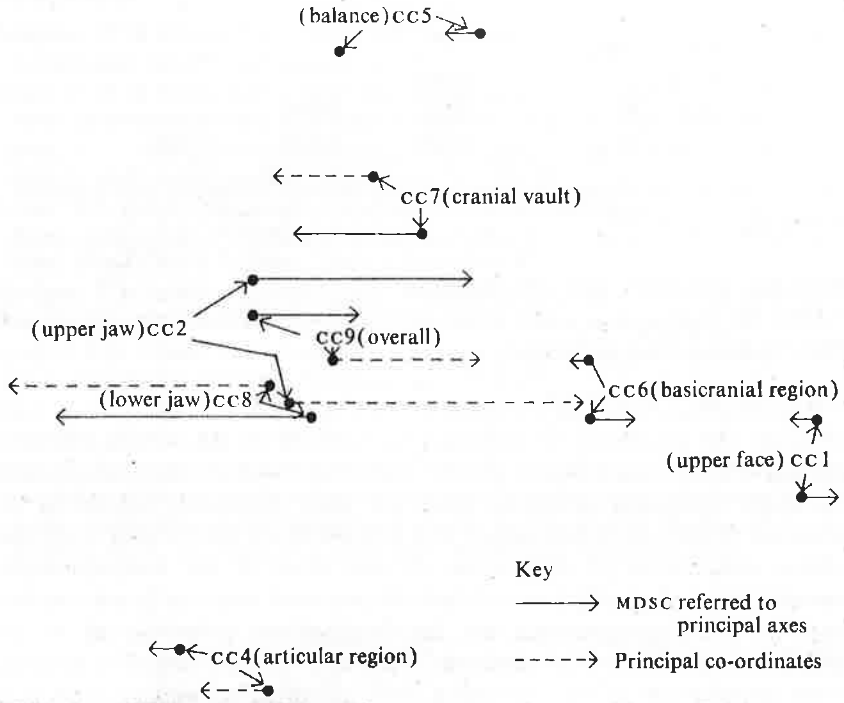

Figure 6. MDSCAL ($r=3$) rotated to fit 3-dimensional principal coordinates solution. The two solutions are seen to agree very well

From figures 4, 5, or 6 it seems that different constellations of characters give different interpretations of the differences between the six populations in terms of D^2. To be more precise we would like to be able to construct confidence regions about each point. Statements like 'the balance and cranial vault regions of the skull determine a set of D^2 values more alike than do the

	1	2	3	4
1. Principal coordinates	—			
2. MDSCAL ($r=3$)	0·066	—		
3. MDSCAL ($r=4$)	0·027	0·091	—	
4. MDSCAL ($r=8$)	0·043	0·120	0·048	—

Table 2. R^2 values obtained when rotating various 3-dimensional analyses to best fit

lower jaw and upper face' can be made.

Thus the analysis has aided visual comparison of two representations and put a figure on their agreement.

Statistical Problems

Although the rotational fit technique as outlined in the section above is applicable mathematically to all the problems listed in the introduction, specifically statistical problems have not been discussed. Individual R^2 values are interesting in themselves, and to some extent can be used to express relative magnitudes of differences, as was done in the anthropometric example above. Rotating pairs of multi-dimensional maps to fit one another simplifies visual comparisons, but the distributional properties of R^2 are needed for a truly objective analysis. When it is realized that each problem in the introduction poses a different statistical distributional problem it is clear that much work remains to be done. Problem (5) seems most likely to yield to analytical treatment. Assuming multi-normal populations with equal covariance matrices, the distribution of R^2 is required when sample values of the canonical variate means are rotated to fit the true values of these means. The next step would be to find the distribution when canonical variate means obtained from two different samples are rotated to fit. Problem (2) requires an extension to consider the effect of using different or overlapping sets of variates for the different analyses. The latter problem would still be meaningful when assessing distances between individual samples drawn from a single multi-normal population.

Whether any of these distributional problems can be solved remains to be seen, but large sample asymptotic χ^2 approximations should be available. I am less hopeful of any analytical solution for distributional problems involving ultrametric distances, where it seems that the best hope is to get information from Monte Carlo sampling experiments. I hope to tackle some of these problems soon.

ACKNOWLEDGMENTS

I thank Mr A. Bilsborough for the data and Mr W. J. Krzanowski for programming the rotational fit technique and for seeing most of the calculations through the computer. The mathematical derivation of the results quoted in the section of this paper on 'Rotational fits' was largely worked out while the author was visiting the Bell Telephone Laboratories, Murray Hill, N.J.

REFERENCES

Buneman, P. & Gower, J.C. (1971) The representation of ultrametric distances in Euclidean space (in preparation).

Golub, C.H. & Reinsch, C. (1970) Handbook series linear algebra. Singular value decomposition and least squares solutions. *Numer. Math.*, **14,** 403–20.

Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53,** 325–38.

Gower, J.C. (1970) Measures of taxonomic distance and their analysis. *Proc. Symp. Assessment of Biological Affinity and Distance between Human Populations, Utrecht* 1969. London: Oxford University Press.

Hartigan, J.R. (1967) Representation of similarity matrices by trees. *J. Amer. statist. Ass.*, **62,** 1140–58.

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

Sokal, R.R. & Rohlf, F.J. (1962) The comparison of dendrograms by objective methods. *Taxon.*, **11,** 33–40.

Wish, M. & Carroll, J.D. (1971) Multi-dimensional scaling with differential weighting of dimensions emphasizing interesting applications. *Mathematics in the Archaeological and Historical Sciences*, pp. 150–67 (eds Hodson, F.R., Kendall, D.G., & Tăutu, P.). Edinburgh: Edinburgh University Press.