Journal of Classification 3:5-48 (1986)

Journal of
Classification

Metric and Euclidean Properties of Dissimilarity Coefficients

J. C. Gower

Rothamsted Experimental Station

P. Legendre

Université de Montréal

Abstract: We assemble here properties of certain dissimilarity coefficients and are specially concerned with their metric and Euclidean status. No attempt is made to be exhaustive as far as coefficients are concerned, but certain mathematical results that we have found useful are presented and should help establish similar properties for other coefficients. The response to different types of data is investigated, leading to guidance on the choice of an appropriate coefficient.

Résumé: Ce travail présente quelques propriétés de certains coefficients de ressemblance et en particulier leur capacité de produire des matrices de distance métriques et euclidiennes. Sans prétendre être exhaustifs dans cette revue de coefficients, nous présentons certains résultats mathématiques que nous croyons intéressants et qui pourraient être établis pour d'autres coefficients. Finalement, nous analysons la réponse des mesures de ressemblance face à différents types de données, ce qui permet de formuler des recommandations quant au choix d'un coefficient.

Keywords: Choice of coefficient; Dissimilarity; Distance; Euclidean property; Metric property; Similarity.

1. Introduction

In this paper we gather together some results mainly on the metric and Euclidean properties of dissimilarity coefficients, but also some other properties. Only a fraction of the coefficients available in the literature are studied here, although the major types are included. The mathematical apparatus for establishing the results is presented and should help others investigate

The authors wish to thank the referees, one of whom did a magnificent job in painstakingly checking the detailed algebra and detecting several slips.

Authors' Addresses: J.C. Gower, Statistics Department, Rothamsted Experimental Station, Harpenden, Herts. AL5 2JQ, United Kingdom and P. Legendre, Département de Sciences Biologiques, Université de Montréal, C.P. 6128, Succursale A, Montréal, Québec H3C 3J7, Canada.

coefficients that we have not discussed. Previous work has been somewhat diversified but includes papers by Bloom (1981), Faith (1985), Gower (1971, 1985), Hajdu (1981), Legendre, Dallot and Legendre (1985), Legendre and Legendre (1983a), Orlóci (1978), Späth (1980), Wolda (1981). Here we attempt a more unified approach so that this paper may be viewed as an exposé, both of methods for establishing mathematical properties of the coefficients and also of how the information obtained may be used to guide the choice of a coefficient in particular applications.

2. Basic Results

We consider an $n \times n$ dissimilarity matrix **D** with elements d_{ij} where $d_{ii} = 0$ for all i.

Definition 1. **D** is said to be *metric* if the metric (triangle) inequality $d_{ij} + d_{ik} \geqslant d_{jk}$ holds for all triplets (i,j,k).

Some simple but important properties follow from this definition. Consideration of the triplet (i,j,j) shows that $d_{ij} \geqslant 0$ for all pairs (i,j). Consideration of (i,j,i) and (j,i,j) shows that $d_{ij} \geqslant d_{ji}$ and $d_{ji} \geqslant d_{ij}$. Hence all metric dissimilarity matrices are symmetric with non-negative elements. Suppose $d_{ij} = 0$; then considering the triplets (i,k,j) and (j,k,i) yields that $d_{ik} = d_{jk}$ for all k. This is a basic property of metrics that can be strengthened to show that if two items are similar (d_{ij} close to zero) then any third item, k, will have a similar relation to both of them (i.e., d_{ik} and d_{jk} will differ only slightly). Of course Euclidean distance, being a metric, shares this property, suggesting that it might be interesting to investigate what might be meant by the closest Euclidean approximation to **D**. When the metric inequality holds, it is trivial to construct a Euclidean triangle with sides d_{ij}, d_{ik} and d_{jk} but when $n > 3$ it is not true that every metric **D** has a Euclidean representation. A standard counter-example for $n = 4$ is given by the following metric dissimilarity matrix **P**:

$$\begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ \\ \end{array} \begin{bmatrix} 0 & & & \\ 2 & 0 & & \\ 2 & 2 & 0 & \\ 1.1 & 1.1 & 1.1 & 0 \end{bmatrix} \qquad (1)$$

$$\begin{array}{cccc} P_1 & P_2 & P_3 & P_4 \end{array}$$

where P_1, P_2, P_3 form an equilateral triangle of side 2 and P_4 is equidistant (1.1 units) from P_1, P_2 and P_3. The geometry is illustrated in Figure 1.

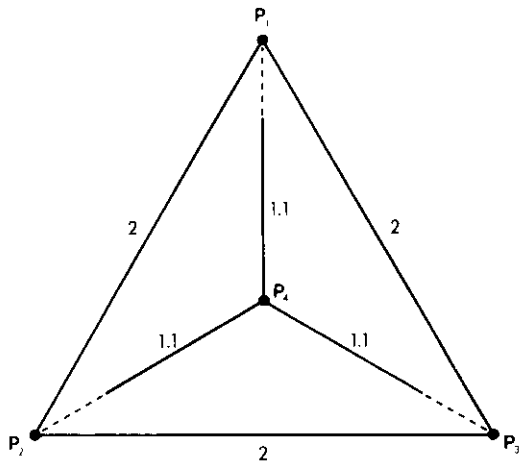

Figure 1. An example of a set of distances that satisfy the metric inequality but which have no Euclidean representation.

It is easy to see that if the configuration is to be Euclidean then the smallest distance P_4 can be from the other vertices is when it is coplanar with them and at their centroid, giving a minimal distance of $(2\sqrt{3})/3 = 1.15$, which is greater than 1.1. Thus **P** is metric but not Euclidean. This type of example will be used often in the following to show that certain metric dissimilarity coefficients can give rise to non-Euclidean configurations.

Failing an exact Euclidean representation, every metric dissimilarity matrix may be represented by a symmetric graph with vertices $P_i (i = 1,2, \ldots, n)$ such that the length of $P_i\ P_j$ is d_{ij}, so that P_i and P_j coincide when $d_{ij} = 0$.

We now state some simple theorems. Proofs are simple and are omitted.

Theorem 1. *If* **D** *is non-metric then the matrix with elements* $d_{ij} + c\ (i \neq j)$ *is metric, where* $c \geqslant \underset{p,q,r}{Max} \mid d_{pq} + d_{pr} - d_{qr} \mid$.

Theorem 2. *If* **D** *is metric then so are the matrices with elements:*

$$\left.\begin{array}{lll} (i) & d_{ij} + c^2 & \\ (ii) & d_{ij}^{1/r} & \text{where } r \geqslant 1 \\ (iii) & d_{ij} \,/\, (d_{ij} + c^2) & \end{array}\right\} \quad i \neq j$$

where c is any real constant.

These results may be used to infer the metric property for other matrices, given that it is true for **D**. Note that the transformations are monotonic and that these results raise the general question of what functions $f(d_{ij})$ preserve the metric property. In section 5 we discuss further some monotonically related coefficients.

Often the validity of the triangle inequality for all triplets (i,j,k) of **D** has to be established *ab inito* and this can be troublesome. Certain procedures and results will now be discussed that have been found helpful when investigating the validity of the metric inequality.

A useful device, suggested independently by several workers, is to try to establish the existence of a fourth item (l, say) that has the property $d_{ij} \geqslant d_{lj}$ and $d_{ik} \geqslant d_{lk}$. It follows that $d_{ij} + d_{ik} \geqslant d_{lj} + d_{lk} \geqslant d_{jk}$, provided the inequality holds for (j,k,l). It is often easy to establish the inequality for a special triplet (j,k,l) when it is difficult for the general triplet (i,j,k). This approach has been found especially useful when **D** comprises a set of dissimilarity coefficients based on binary variables (see below). A bonus is that if the triangle inequality can be shown to be invalid for (j,k,l) then one has a ready-made counter-example showing that d_{ij} is not a metric. For convenience of reference this simple result is stated as a theorem:

Theorem 3. *If for every triplet* (i,j,k) *an* l *can be found such that* $d_{ij} \geqslant d_{lj}$ *and* $d_{ik} \geqslant d_{lk}$ *then* **D** *is metric iff* (j,k,l) *is metric.*

As a variant, which includes theorem 3 as a special case, we note that it is enough to find an l such that $d_{ij} + d_{ik} \geqslant d_{lj} + d_{lk}$ and (j,k,l) is metric.

Further properties of **D** come from investigating whether or not it has a Euclidean (or other) distance representation.

Definition 2. **D** is said to be *Euclidean* if n points $P_i (i = 1,2, \ldots, n)$ can be embedded in a Euclidean space such that the Euclidean distance between P_i and P_j is d_{ij}. This, of course, implies that d_{ij} must be non-negative.

We use the notation **I** for a unit matrix, **1** for a vector of units and $\boldsymbol{\Delta}$ for the matrix with elements $-\frac{1}{2} d_{ij}^2$. Necessary and sufficient conditions for **D** to be Euclidean are given by the following theorem:

Theorem 4. **D** *is Euclidean iff the matrix* $(\mathbf{I} - \mathbf{1s}')\boldsymbol{\Delta}\,(\mathbf{I} - \mathbf{s1}')$ *is positive-semi-definite (p.s.d.) where* $\mathbf{s'1} = 1$.

Gower (1982) gives a discussion and proof of this result, which for the special cases $\mathbf{s} = \mathbf{1}/n$ and $\mathbf{s} = \mathbf{e}_i$ (a vector with 1 in its i-th position, else zero) was proved by Schoenberg (1935). It is easy to show that if the result is true for one choice of **s** (say $\mathbf{s} = \mathbf{e}_i$) it is true for every valid choice of **s**; see Gower (1984b).

An equivalent statement to theorem 4 is that **D** is Euclidean iff $\mathbf{x}' \boldsymbol{\Delta}\, \mathbf{x} \geqslant 0$ for all vectors **x** such that $\mathbf{x}'\mathbf{1} = 0$.

If **D** is Euclidean it is also metric. Further, **D** is metric if and only if every triplet (i,j,k) generates a 3×3 Euclidean matrix. Thus **D** is metric if and only if:

$$(\mathbf{I} - \mathbf{1s}') \begin{bmatrix} 0 & -\tfrac{1}{2} d_{ij}^2 & -\tfrac{1}{2} d_{ik}^2 \\ -\tfrac{1}{2} d_{ij}^2 & 0 & -\tfrac{1}{2} d_{jk}^2 \\ -\tfrac{1}{2} d_{ik}^2 & -\tfrac{1}{2} d_{jk}^2 & 0 \end{bmatrix} (\mathbf{I} - \mathbf{s1}')$$

is p.s.d. for all triplets (i,j,k). Writing $a_i = -\tfrac{1}{2} d_{jk}^2$ and choosing $\mathbf{s} = \mathbf{e}_1$ shows that

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & -2a_k & a_i - a_j - a_k \\ 0 & a_i - a_j - a_k & -2a_j \end{bmatrix}$$

must be p.s.d. Expanding the non-trivial minor, expressing the result in terms of the original distances and allowing for a scaling factor gives the condition

$$2d_{ij}^2 d_{ik}^2 + 2d_{ij}^2 d_{jk}^2 + 2d_{ik}^2 d_{jk}^2 - d_{jk}^4 - d_{ik}^4 - d_{ij}^4 \geqslant 0 \qquad (2)$$

for **D** to be metric. This quantity is the square of four times the area of the triangle $P_i P_j P_k$. The advantage of this form over the simple triangle inequality is that i,j,k enter into (2) symmetrically and therefore their order is irrelevant. Thus we have:

Theorem 5. **D** *is metric iff* d_{ij} *is non-negative for all pairs* (i,j) *and condition (2) is satisfied for all triplets* (i,j,k).

Often similarity s_{ij}, the complement of dissimilarity, is of interest. If $s_{ij} = 1 - d_{ij}$, we define a similarity matrix $\mathbf{S} = \mathbf{11}' - \mathbf{D}$, so that $\boldsymbol{\Delta}$ has elements $-\tfrac{1}{2} d_{ij}^2 = -\tfrac{1}{2}(1 - s_{ij})^2$. Normally, but not necessarily, this would require that $0 \leqslant d_{ij} \leqslant 1$ which implies that $0 \leqslant s_{ij} \leqslant 1$ and, since $d_{ii} = 0$, we also have $s_{ii} = 1$. Now if **S** is p.s.d., any decomposition of the form $\mathbf{S} = \mathbf{XX}'$ gives a matrix **X** of real coordinates such that the squared distance between the i-th and the j-th rows of **X** is $2(1 - s_{ij})$, which is never negative when d_{ij} is non-negative. Hence under these circumstances, the dissimilarity matrix with elements $\sqrt{d_{ij}}$ is Euclidean, because it is generated by the real coordinates given by the rows of **X**. Thus we have:

Theorem 6. *If* $\mathbf{S}$ *is a p.s.d. similarity matrix with elements* $0 \leqslant s_{ij} \leqslant 1$ *and* $s_{ii} = 1$, *then the dissimilarity matrix with elements* $d_{ij} = (1 - s_{ij})^{1/2}$ *is Euclidean.*

Note that theorem 6 gives only a *sufficient* condition for $\sqrt{1 - s_{ij}}$ to be Euclidean. Conditions on $\mathbf{S}$, more simple than those for theorem 4, for the result to be necessary are not known. For example, the similarity matrix

$$\begin{bmatrix} 1 & 0 & .72 \\ 0 & 1 & .72 \\ .72 & .72 & 1 \end{bmatrix} \tag{3}$$

is not p.s.d. but yields a real Euclidean triangle with sides $\sqrt{.28}$, $\sqrt{.28}$ and 1; however this matrix cannot be constructed from any of the definitions of similarity discussed below. Fortunately for many similarity coefficients, $\mathbf{S}$ may be shown to be p.s.d. and hence the question of necessity becomes irrelevant.

From theorem 6 it follows that to show that the matrix with elements $(1 - s_{ij})$ is itself Euclidean, as an alternative to theorem 4, it is sufficient to show that the similarity matrix $\mathbf{\Sigma}$ with elements $1 - (1 - s_{ij})^2$ is p.s.d. Thus $\mathbf{\Sigma} = 2\mathbf{S} - \mathbf{S}*\mathbf{S}$, where $*$ represents the Hadamard (i.e., element-by-element) product. This is a result worth noting but although we have often found it possible to prove $\mathbf{S}$ to be p.s.d. we have never been able to show $\mathbf{\Sigma}$ to be p.s.d. Certainly Table 2 shows that it is not enough for $\mathbf{S}$ to be p.s.d. to guarantee that $\mathbf{\Sigma}$ is p.s.d., although that the converse holds has now been proved by Zegers (1986).

Corollary: *If* $\sqrt{1 - s_{ij}}$ *is not a metric, or is a non-Euclidean metric, then* $\mathbf{S}$ *is not p.s.d. If* $1 - s_{ij}$ *is not a metric, or is a non-Euclidean metric, then* $\mathbf{\Sigma}$ *is not p.s.d.*

There are two theorems giving simple monotonic transformations that carry general dissimilarity matrices into Euclidean distance matrices. These are:

Theorem 7. *If* $\mathbf{D}$ *is a dissimilarity matrix, then there exist constants* h *and* k *such that the matrix with elements*

An error in the original printing has been corrected here according to Legendre & Legendre 1998

$$\left.\begin{array}{lll} (i) & (d_{ij}^2 + 2h)^{1/2} & \text{is Euclidean (Lingoes 1971)} \\ \text{and } (ii) & d_{ij} + k & \text{is Euclidean (Cailliez 1983)} \end{array}\right\} \quad i \neq j$$

For (i), $h \geqslant -\lambda_n$, *the smallest eigenvalue of*

$$\mathbf{\Delta}_1 = (\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\,\mathbf{\Delta}\,(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)$$

For (ii) $k \geqslant \mu_n$, the largest eigenvalue of

$$\begin{bmatrix} 0 & 2\Delta_1 \\ -\mathbf{I} & -4\Delta_2 \end{bmatrix}$$

where Δ_2 is defined as for Δ_1 but with elements $-\frac{1}{2} d_{ij}$ rather than $-\frac{1}{2} d_{ij}^2$.

An error in the original printing has been corrected here according to Legendre & Legendre 1998

Charlton and Wynn (personal communication) discuss the very general transformations that take a Euclidean matrix into other Euclidean matrices.

Many dissimilarity measures are based on an $n \times m$ data-matrix $\mathbf{X}$ and have the form

$$d_{ij} = \sum_{r=1}^{m} f(x_{ir}, x_{jr})$$

where $f(x_{ir}, x_{ir}) = 0$ and $f(x_{ir}, x_{jr}) = f(x_{jr}, x_{ir}) \geqslant 0$. Thus dissimilarity is evaluated for each of the m characters separately and combined assuming independence. When d_{ij} has this form and is metric, then $f(x_{ir}, x_{jr})$ must be metric for every suffix r. This follows from considering data for which $x_{ir} = x_{is}$ for all i and $s \neq r$. Hence to prove that d_{ij} is metric it is necessary and sufficient for each dimension r to satisfy the metric inequality independently of the other dimensions.

Theorem 8. *Let $f(x_{ir}, x_{jr}) = \alpha_{ijr}$, then: if $d_{ij1} = \sum_{r=1}^{m} \alpha_{ijr}$ is metric then so is $d_{ij2} = (\sum_{r=1}^{m} \alpha_{ijr}^2)^{1/2}$, and conversely.*

Proof. If d_{ij1} is a metric, it is metric for each dimension r. Thus from theorem 4 for every triplet (i,j,k) the matrix:

$$\mathbf{A}_r = (\mathbf{I} - \mathbf{N}_3) \begin{bmatrix} 0 & -\frac{1}{2}\alpha_{ijr}^2 & -\frac{1}{2}\alpha_{ikr}^2 \\ -\frac{1}{2}\alpha_{ijr}^2 & 0 & -\frac{1}{2}\alpha_{jkr}^2 \\ -\frac{1}{2}\alpha_{ikr}^2 & -\frac{1}{2}\alpha_{ikr}^2 & 0 \end{bmatrix} (\mathbf{I} - \mathbf{N}_3)$$

is p.s.d., where $\mathbf{N}_3 = (1,1,1)'\,(1,1,1)/3$. Now

$$\mathbf{A} = (\mathbf{I} - \mathbf{N}_3) \begin{bmatrix} 0 & -\frac{1}{2}d_{ij2}^2 & -\frac{1}{2}d_{ik2}^2 \\ -\frac{1}{2}d_{ij2}^2 & 0 & -\frac{1}{2}d_{jk2}^2 \\ -\frac{1}{2}d_{ik2}^2 & -\frac{1}{2}d_{jk2}^2 & 0 \end{bmatrix} (\mathbf{I} - \mathbf{N}_3)$$

$= \sum_{r=1}^{m} \mathbf{A}_r$ and so is p.s.d. But this is the condition for d_{ij2} to be a metric.

TABLE 1

Notation for Occurrences of all 2x2 +/- Combinations for Sampling Units i and j

Unit	i	j	Frequency
Combination	+	+	a_{ij}
	+	−	b_{ij}
	−	+	c_{ij}
	−	−	d_{ij}
Total			m

Conversely, if d_{ij2} is a metric, it is metric for every dimension r, which implies that $\mathbf{A}_r$ is p.s.d. for all dimensions r. It follows that

$$| \alpha_{ijr} | + | \alpha_{ikr} | \geqslant | \alpha_{jkr} |$$

and summing over r shows that d_{ij1} is a metric. ●

It is not known whether theorem 8 remains valid for the general "Minkowski" form $d_{ijr} = (\sum_{r=1}^{m} \alpha_{ijr}^{q})^{1/q}$, although, of course, it does when $\alpha_{ijr} = | x_{ir} - x_{jr} |$.

Corollary: *When d_{ij1} is not a metric then neither is d_{ij2} and conversely.*

3. The Coefficients S_θ and T_θ

Consider two binary variables i and j with the following frequencies for the four combinations, as in Table 1.

We define

$$S_\theta = \frac{a + d}{a + d + \theta\,(b + c)} \quad \text{and} \quad T_\theta = \frac{a}{a + \theta\,(b + c)}$$

where, for simplicity, the suffixes i and j have been dropped. As we see in Table 2, these coefficients are of special importance because for various non-negative values of θ they include many of the well-known similarity coefficients. To avoid the possibility of negative similarity coefficients we shall confine our discussion to non-negative values of θ. The metric and Euclidean properties of the dissimilarity coefficients $1 - S_\theta$, $1 - T_\theta$ and $\sqrt{1 - S_\theta}$, $\sqrt{1 - T_\theta}$ depend on θ. We shall see that for values of θ near zero these coefficients are not metric but for all $\theta \geqslant \theta_M$ they become metric and for values of $\theta \geqslant \theta_E > \theta_M$, $\sqrt{1 - S_\theta}$ and $\sqrt{1 - T_\theta}$ become Euclidean. Thus we shall be concerned with finding the threshold values θ_M and θ_E.

Theorem 9. *$1 - S_\theta$ is metric for $\theta \geqslant 1$ and $\sqrt{1 - S_\theta}$ is metric for $\theta \geqslant 1/3$. If $\theta < 1$ then $1 - S_\theta$ may be non-metric and if $\theta < 1/3$, $\sqrt{1 - S_\theta}$ may be non-metric.*

Proof. The proof is based on theorem 3. We consider three general samples, 1, 2 and 3 (corresponding to i, j and k) and a fourth non-general sample 4 (corresponding to l of theorem 3), with frequencies given in the following table:

$$\left.\begin{array}{lllllllll} 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 3 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 4 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{array}\right\} \qquad (4)$$

$$\begin{array}{lllllllll} \textit{Frequency} & A & B & C & D & E & F & G & H \end{array}$$

We define $m = A + B + C + D + E + F + G + H$.

Writing $\phi = \theta - 1$, these give the following dissimilarities $1 - S_\theta$:

$$\delta_{12} = \theta(B + C + E + F)/\{m + \phi\,(B + C + E + F)\}$$

$$\delta_{13} = \theta(B + D + E + G)/\{m + \phi\,(B + D + E + G)\}$$

$$\delta_{23} = \theta(C + D + F + G)/\{m + \phi\,(C + D + F + G)\}$$

$$\delta_{24} = \theta(C + F)/\{m + \phi\,(C + F)\}$$

$$\delta_{34} = \theta(D + G)/\{m + \phi\,(D + G)\} \quad .$$

It is trivial to show that $\delta_{12} \geqslant \delta_{24}$ and $\delta_{13} \geqslant \delta_{34}$ and hence that $\delta_{12} + \delta_{13} \geqslant \delta_{24} + \delta_{34}$. It follows from theorem 3 that to show that the dissimilarities between the general samples 1, 2 and 3 satisfy the metric inequality, it suffices that the inequality be satisfied by the special samples 2, 3 and 4. We have that

$$\Delta = \delta_{24} + \delta_{34} - \delta_{23} \tag{5}$$

$$= \frac{\theta\,\alpha}{m + \phi\,\alpha} + \frac{\theta\,\beta}{m + \phi\,\beta} - \frac{\theta(\alpha + \beta)}{m + \phi(\alpha + \beta)}$$

where $\alpha = C + F$ and $\beta = D + G$. Δ is certainly non-negative for $\theta \geqslant 1$ (i.e., $\phi \geqslant 0$), so that $1 - S_\theta$ is metric when $\theta \geqslant 1$. The following gives an example of where the metric inequality fails. Take $\alpha = m/2$, $\beta = m/2$ and $A = B = E = H = 0$. Then

$$\Delta = \frac{2\,\theta}{1 + \theta} - 1 = \frac{\theta - 1}{\theta + 1}$$

which is certainly negative for $0 \leqslant \theta < 1$ (recall we are concerned only with non-negative values of θ).

A similar argument establishes when $\sqrt{1 - S_\theta}$ is a metric. Equation (5) is replaced by:

$$\Delta' = \sqrt{\frac{\theta\,\alpha}{m + \phi\,\alpha}} + \sqrt{\frac{\theta\,\beta}{m + \phi\,\beta}} - \sqrt{\frac{\theta(\alpha + \beta)}{m + \phi(\alpha + \beta)}} \quad . \tag{6}$$

Consider those values of α, β for which $\alpha + \beta = k$, a constant. Depending on the value of ϕ, the minimum of the sum of the first two terms on the right-hand side of (6) occurs (i) when $\alpha = k$, $\beta = 0$ and $\alpha = 0$, $\beta = k$ or (ii) when $\alpha = \beta = k/2$. In case (i) $\Delta' \geqslant 0$ and $\sqrt{1 - S_\theta}$ is metric; thus we have only to consider those values of ϕ for which case (ii) gives the minimum. Then

$$\Delta' \geqslant 2\sqrt{\frac{\theta\,k}{2m + \phi\,k}} - \sqrt{\frac{\theta\,k}{m + \phi\,k}}$$

which is non-negative when

$$4(m + \phi\ k) \geqslant 2m + \phi\ k$$

$$i.e. \quad \theta \geqslant \frac{3k - 2m}{3k} \quad .$$

Now $m \geqslant \alpha + \beta = k$ so that Δ' is certainly non-negative for all values of k, and hence $\sqrt{1 - S_\theta}$ is metric, when $\theta \geqslant 1/3$. Even when $\theta < 1/3$, Δ' may still be positive but the example $\alpha = \beta = m/2$, $A = B = E = H = 0$ gives

$$\Delta' = 2\sqrt{\frac{\theta}{1 + \theta}} - 1$$

which is negative when $\theta < 1/3$, showing that non-metric examples can always be constructed for $\theta < 1/3$ and hence that $\theta_M = 1/3$ for the dissimilarity coefficient $\sqrt{1 - S_\theta}$. ●

Theorem 10. *$1 - T_\theta$ is metric for $\theta \geqslant 1$ and $\sqrt{1 - T_\theta}$ is metric for $\theta \geqslant 1/3$. If $\theta < 1$ then $1 - T_\theta$ may be non-metric and if $\theta < 1/3$, $\sqrt{1 - T_\theta}$ may be non-metric.*

Proof. The proof is similar to that of theorem 9 but requires a little more care. With T_θ, the previous frequencies for samples 1, 2, 3 and 4 yield modified formulae for δ_{12}, δ_{13}, δ_{23}, δ_{24}, δ_{34} which gives $\Delta = \delta_{24} + \delta_{34} - \delta_{23}$ as:

$$\Delta = \frac{\theta\ \alpha}{\gamma + D + \theta\ \alpha} + \frac{\theta\ \beta}{\gamma + C + \theta\ \beta} - \frac{\theta(\alpha + \beta)}{\gamma + \theta(\alpha + \beta)} \qquad (7)$$

(where $\gamma = A + B$) from which it is trivial to show that $\Delta \geqslant 0$ when $\theta \geqslant 1$. When $C = D = m/2$ and $A = B = E = F = G = H = 0$ then

$$\Delta = \frac{2\theta}{1 + \theta} - 1$$

which is negative when $0 \leqslant \theta < 1$ (recall we are concerned only with non-negative values of θ).

For $\sqrt{1 - T_\theta}$, corresponding to (7) we have:

$$\Delta' = \sqrt{\frac{\theta\ \alpha}{\gamma + \beta + \theta\ \alpha}} + \sqrt{\frac{\theta\ \beta}{\gamma + \alpha + \theta\ \beta}} - \sqrt{\frac{\theta(\alpha + \beta)}{\gamma + \theta(\alpha + \beta)}} \quad . \qquad (8)$$

Consider the values of α, β for which $\alpha + \beta = k$, a constant. Depending on the value of ϕ, the minimum of the sum of the first two terms on the right-hand side of (8) occurs (i) when $\alpha = k$, $\beta = 0$ and $\alpha = 0$, $\beta = k$ or (ii) when $\alpha = \beta = k/2$. In case (i) $\Delta' \geqslant 0$ and $\sqrt{1 - T_\theta}$ is metric; thus we have only to consider those values of ϕ for which case (ii) gives the minimum. Then

$$\Delta' \geqslant 2\sqrt{\frac{\theta\ k}{2\gamma + (1+\theta)k}} - \sqrt{\frac{\theta\ k}{\gamma + \theta\ k}}$$

which is non-negative when

$$4(\gamma + \theta\ k) \geqslant 2\gamma + (1+\theta)k$$

$$i.e. \quad \theta \geqslant \frac{k - 2\gamma}{3k} \quad .$$

Thus $\sqrt{1 - T_\theta}$ is metric when $\theta \geqslant 1/3$. The example $C = D = m/2$, $A = B = E = F = G = H = 0$ gives

$$\Delta' = 2\sqrt{\frac{\theta}{1+\theta}} - 1$$

which is negative for $0 \leqslant \theta < 1/3$ (recall we are concerned only with non-negative values of θ). ●

Theorems 9 and 10 establish the thresholds θ_M for $1 - S_\theta$, $\sqrt{1 - S_\theta}$, $1 - T_\theta$ and $\sqrt{1 - T_\theta}$. In particular they show that $1 - S_3$, $1 - S_4$, $1 - S_5$, $1 - S_6$, $1 - S_9$ $(= 2(1 - S_4))$ and $\sqrt{1 - S_8}$ of Table 2 are metric, as well as giving alternative derivations of results that follow from the p.s.d. properties established in section 4.

The Euclidean properties of S_θ and T_θ are investigated in the remainder of this section. We first prove the existence of the Euclidean threshold θ_E.

Theorem 11. *If $\sqrt{1 - S_\theta}$ is Euclidean then so is $\sqrt{1 - S_\phi}$ for all $\phi \geqslant \theta$. If $\sqrt{1 - T_\theta}$ is Euclidean then so is $\sqrt{1 - T_\phi}$ for all $\phi \geqslant \theta$.*

Proof. Both similarity coefficients may be written in the form

$$R_\phi = \frac{x}{x + \phi\ y}$$

which may be manipulated into:

$$R_\phi = \frac{x\ \theta/\phi}{x + \theta\ y}\ [1 - (1 - \frac{\theta}{\phi})\ \frac{x}{x + \theta\ y}\ ]^{-1} \quad .$$

Expanding shows that the similarity matrices $\mathbf{R}_\phi$ and $\mathbf{R}_\theta$ are related by:

$$\mathbf{R}_\phi = \frac{\theta}{\phi}\ \mathbf{R}_\theta * [\mathbf{11}' + \psi\ \mathbf{R}_\theta + \psi^2\ \mathbf{R}_\theta^{(2)} + \psi^3\ \mathbf{R}_\theta^{(3)} + \ldots\ ]$$

where $\psi = 1 - \theta/\phi$ and * denotes the Hadamard product and $\mathbf{R}^{(2)} = \mathbf{R}*\mathbf{R}$ etc. Schur's theorem states that if **A** and **B** are both p.s.d. then so is their Hadamard product (see e.g., Mirsky 1955, p. 421). It follows that, provided $\phi \geqslant \theta$ (i.e., $\psi \geqslant 0$), then $\mathbf{R}_\phi$ is p.s.d. whenever $\mathbf{R}_\theta$ is p.s.d. ●

Theorem 12. *$\sqrt{1 - S_\theta}$ is Euclidean for $\theta \geqslant \theta_E = 1$. $\sqrt{1 - T_\theta}$ is Euclidean for $\theta \geqslant \theta_E = \frac{1}{2}$. $1 - S_\theta$ may be non-Euclidean. $1 - T_\theta$ may be non-Euclidean.*

Proof. We show in section 4 that $S_{\theta=1}$ is p.s.d. and that $T_{\theta=1/2}$ is p.s.d. It follows from theorem 11 that $\theta_E \leqslant 1$ for $\sqrt{1 - S_\theta}$ and that $\theta_E \leqslant 1/2$ for $\sqrt{1 - T_\theta}$. The true values of θ_E cannot be less than the corresponding values of θ_M.

These bounds may be made precise by examining a special example. This consists of a configuration consisting of two regular simplices, side l, each one of $n - 1$ vertices, sharing $n - 2$ of these vertices. The two extra vertices, labeled P_1 and P_2, are distance apart $2h_{n-1}$ where $h_{n-1} = l\sqrt{\frac{(n-1)}{2(n-2)}}$ is the altitude of the $(n-1)$-simplex. The configuration is clearly non-Euclidean if we can arrange that the distance d_{12}, between P_1 and P_2, is greater than $2h_{n-1}$. With S_θ we can choose (see Appendix I) $s_{12} = 0$ and all other similarities equal to $1/(1 + \theta)$, i.e., $l = \sqrt{1 - S_\theta} = \sqrt{\frac{\theta}{1 + \theta}}$. Thus the condition for non-Euclideanarity is:

$$1 > \frac{\theta}{1 + \theta}\ \frac{2(n - 1)}{(n - 2)}$$

or

$$\theta < \frac{n-2}{n} \quad .$$

When $n = 3$, $\theta < 1/3$ gives a non-Euclidean configuration which is consistent with the result that $\theta > 1/3$ gives a coefficient that is metric (and hence is Euclidean for three points). When $n = 4$, then $\theta < ½$ is non-Euclidean (though metric for $\theta > 1/3$). In general this example gives a non-Euclidean configuration for values of θ which depend on n. As n increases a non-Euclidean configuration can be found for θ arbitrarily close to unity. We have already pointed out at the beginning of this proof that a result in section 4 shows that for $\sqrt{1 - S_\theta}$, $\theta_E \leqslant 1$; therefore $\theta_E = 1$ for $\sqrt{1 - S_\theta}$.

For T_θ the same example gives $s_{12} = 0$ and all other similarities equal $1/(1 + 2\theta)$, i.e., $l = \sqrt{1 - T_\theta} = \sqrt{\frac{2\theta}{1 + 2\theta}}$. Thus the configuration is non-Euclidean when

$$1 > \frac{2\theta}{1 + 2\theta} \frac{2(n-1)}{(n-2)}$$

$$i.e.\ \theta < \frac{n-2}{2n} \quad .$$

Thus a non-Euclidean configuration can always be found for $\theta < ½$ and, as stated above, we show in section 4 that $\sqrt{1 - T_\theta}$ is Euclidean for $\theta \geqslant ½$. This shows that $\theta_E = ½$ for $\sqrt{1 - T_\theta}$.

Appendix II gives examples to show that $1 - S_\theta$ and $1 - T_\theta$ need not be Euclidean for any value of θ. ●

The properties of $1 - S_\theta$ and $(1 - S_\theta)^{1/2}$ are special cases of those for $(1 - S_\theta)^{1/t}$, where t is allowed any positive value, and we would expect θ_M and θ_E to decrease with increasing values of t. The above examples are easily adjusted to show that this is nearly, but not quite, so. Δ' of the above equations now becomes:

$$\Delta' = \left(\frac{\theta\,\alpha}{n + \phi\,\alpha}\right)^{1/t} + \left(\frac{\theta\,\beta}{n + \phi\,\beta}\right)^{1/t} - \left(\frac{\theta(\alpha + \beta)}{n + \phi(\alpha + \beta)}\right)^{1/t} \quad . \qquad (9)$$

Depending on the value of ϕ, the minimum of the sum of the first two terms on the right-hand side of (9) occurs (i) when $\alpha = k$, $\beta = 0$ and $\alpha = 0$, $\beta = k$ or (ii) when $\alpha = \beta = k/2$. In case (i) $\Delta' \geqslant 0$ and $(1 - S_\theta)^{1/t}$ is metric. In case (ii) the minimum value of Δ' is

$$\Delta' = 2\left(\frac{\theta\ k}{2n + \phi\ k}\right)^{1/t} - \left(\frac{\theta\ k}{n + \phi\ k}\right)^{1/t}$$

which is non-negative only when

$$\theta \geqslant 1 + \frac{n}{k}\left\{\frac{2 - 2^t}{2^t - 1}\right\} .$$

The term in parenthesis is negative when $t > 1$ and therefore, because $k \leqslant n$, the condition is clearly satisfied for all values of k. The worst possibility is when $k = n$ giving $\theta \geqslant 1/(2^t - 1)$, $t \geqslant 1$ as a condition for $(1 - S_\theta)^{1/t}$ to be metric. The example $\alpha = \beta = n/2$, $A = B = E = H = 0$ is non-metric for $\theta < 1/(2^t - 1)$ establishing θ_M precisely. However when $t < 1$, the term in parenthesis is positive and by taking $k = 1$ and n sufficiently large, non-metric examples can always be constructed for any value of θ.

Similarly for $(1 - T_\theta)^{1/t}$ the same result is obtained, i.e., that the coefficient is metric when $\theta > 1/(2^t - 1)$ provided $t \geqslant 1$, otherwise when $t < 1$ non-metric examples can be found for any θ.

The example in Appendix I may be used to give bounds for Euclidean properties. Thus non-Euclidean configurations of $(1 - S_\theta)^{1/t}$ are given when:

$$1 > 2\left(\frac{\theta}{1 + \theta}\right)^{1/t} \quad \frac{n - 1}{2(n - 2)} > \sqrt{2}\left(\frac{\theta}{1 + \theta}\right)^{1/t} .$$

This gives:

$$\theta \geqslant 1/(2^{t/2} - 1) \tag{10}$$

for Euclideanarity.

Similarly the threshold given by the same example for $(1 - T_\theta)^{1/t}$ is

$$\theta \geqslant 1/(2(2^{t/2} - 1)) . \tag{11}$$

Although when $t = 2$, (10) and (11) agree with the exact bounds discussed above, we cannot ascertain that these bounds are attained for other values of t. Indeed an example in Appendix II shows that the bound of infinity is reached when $t = \log 4/\log 3 = 1.26$ and possibly for some greater value of t less than two.

All the results of this section are summarized in Figure 2.

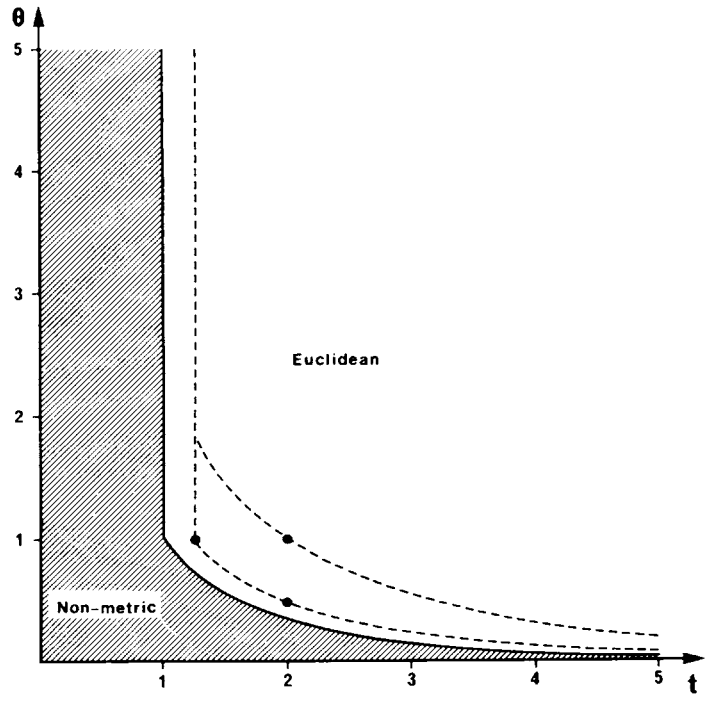

Figure 2. Properties of $(1 - S_\theta)^{1/t}$ and $(1 - T_\theta)^{1/t}$. The solid line gives the boundary between non-metrics and metrics, the same boundary for both coefficients. The dotted lines give the boundaries between metrics and Euclidean metrics, the lower line for $(1 - T_\theta)^{1/t}$ and the upper for $(1 - S_\theta)^{1/t}$. These dotted lines are partly conjectural with only the points for $t = 1$, 2 and ∞ being known. For $t > 2$ they give a lower bound, possibly attained, for both coefficients. When $t < 2$ there is a cut-off at not less than $t = \log 4 / \log 3$. Whether the true curves approach the true cut-off asymptotically, or as shown in the diagram, is not known.

4. Some Special Cases

We may speak loosely of a coefficient S as being non-Euclidean or non-metric. This means that for *some* data the resulting dissimilarities are not Euclidean or are not metric. Conversely when S is said to be Euclidean or metric the corresponding dissimilarities are Euclidean or metric for all data, except possibly when the calculation of the coefficient includes some process for handling missing values (see for instance Gower 1971; Legendre and Legendre 1983a). In this section we shall be concerned with establishing whether or not some well-known dissimilarity coefficients are or are not metric or Euclidean.

We shall examine $d_{ij} = 1 - s_{ij}$ and $\sqrt{d_{ij}}$ separately. The results of section 2 can be used systematically and Figure 3 illustrates a sequence of steps

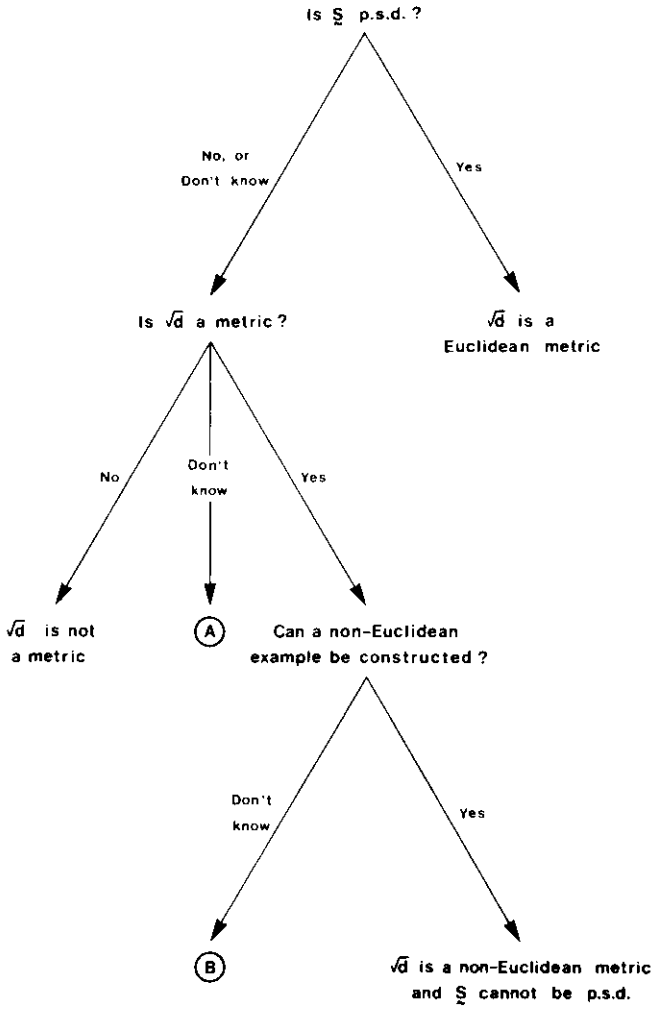

Figure 3. Strategy for determining metric and Euclidean properties of $\sqrt{d_{ij}}$. The answer to each question is one of Yes (Y), No (N), or Don't Know (DK). The possibilities A and B have never arisen but if they did, partial information may be available. For example B would imply that $\sqrt{d_{ij}}$ is a metric that is possibly Euclidean, while with A it may be possible to find an example showing that at least $\sqrt{d_{ij}}$ is not Euclidean. A similar strategy may be used for investigating d_{ij} itself, replacing **S** by $\boldsymbol{\Sigma}$. But with all examples tried, the DK branch has always been given to the question "Is $\boldsymbol{\Sigma}$ p.s.d.?" (see text).

that we have found satisfactory. If **S** can be shown to be p.s.d. then $\sqrt{d_{ij}}$ is metric and Euclidean (theorem 6) and nothing further is required. If **S** is not p.s.d. then we can try to show that $\sqrt{d_{ij}}$ is metric (e.g., theorem 3) or find an example to establish that it is not metric, in which case nothing

more is to be done. If, however, **S** is not p.s.d. but $\sqrt{d_{ij}}$ is a metric then it may also be Euclidean. Fortunately in every such case we have examined we have always found an example to show that $\sqrt{d_{ij}}$ is not Euclidean, thus settling the question, but the theoretical possibility remains that $\sqrt{d_{ij}}$ is Euclidean and this might be established through theorem 4.

To examine the properties of d_{ij} itself we can again use the strategy of Figure 3, now starting with the matrix $\boldsymbol{\Sigma}$ of section 2 (theorem 6). In all the cases examined we have never been able to show directly that $\boldsymbol{\Sigma}$ is or is not p.s.d. but because non-Euclidean examples can be found for every coefficient considered it follows from theorem 6 that $\boldsymbol{\Sigma}$ is not p.s.d. Thus for d_{ij} the first step given in Figure 3 turns out to give no information and we have to proceed directly to the second step.

4.1 Similarity Coefficients for Binary Variables

As in Table 1 we denote presence/absence by +/− and use the standard notation a for the number of (+,+) matches, b for (+,−), c for (−,+) and d for (−, −). Table 2 is an expanded version of a similar table given by Gower (1985). It lists the metric and Euclidean properties of many well-known dissimilarity coefficients and indicates how the given results were obtained. Note that in Table 2, the suffixes in S_1, S_2, S_3 ,... are cardinal numbers, used solely for reference, that do not correspond to values of θ in S_θ of section 3. Table 2 gives, when relevant, synonyms for the reference numbers in terms of S_θ and T_θ.

For both d and $\sqrt{d}$, three properties are listed — (i) metric or not, (ii) Euclidean or not, and (iii) whether or not the similarity matrix is p.s.d. Each result is indicated by a Y (for Yes) or an N (for No). The remarks at the beginning of this section indicate that not all eight combinations can occur. For example if the answer to (iii) is Y then so must it be to (i) and (ii). Similarly if the answer to (i) is N so must it be to (ii) and (iii). This leaves only the four possibilities (YYY), (NNN), (YYN), and (YNN). Of these, although (YYN) is theoretically possible, as we have indicated with the 3 × 3 matrix following theorem 6, it has never occurred with any of the coefficients that we have examined. The other three possibilities all occur in Table 2. The results have all been obtained in one of three ways.

(a) By showing that **S** is p.s.d.

(b) By showing that d_{ij} is not a metric

(c) By showing that although d_{ij} is a metric it is not Euclidean.

To establish (b) requires only a single counter-example and these are listed in Appendix II. To establish (c) requires proof of the metric inequality, usually by theorem 3, followed by a non-Euclidean example — again listed

TABLE 2

Properties of Similarity Coefficients Amongst Binary Variables

Coefficient		Synonyms	$(1-S)$ (1) Metric	$(1-S)$ (2) Euclidean	$(1-S)$ (3) $\underset{\sim}{\Sigma}$ p.s.d.	$\sqrt{1-S}$ (4) Metric	$\sqrt{1-S}$ (5) Euclidean	$\sqrt{1-S}$ (6) $\underset{\sim}{S}$ p.s.d.
[b]S_1	$\frac{a}{b+c}$							
[c]S_2	$\frac{a}{a+b+c+d}$		Y^9	N^a	N_2	Y_6	Y_6	Y^{13}
S_3	$\frac{a}{a+b+c}$	$T_{\theta=1}$	Y^{10}	N^a	N_2	Y_6	Y_6	Y^{13}
S_4	$\frac{a+d}{a+b+c+d}$	$S_{\theta=1}$	Y^9	N^a	N_2	Y_6	Y_6	Y^{13}
S_5	$\frac{a}{a+2(b+c)}$	$T_{\theta=2}$	Y^{10}	N_a	N_2	Y_6	Y_6	Y^{13}
S_6	$\frac{a+d}{a+2(b+c)+d}$	$S_{\theta=2}$	Y^9	N^a	N_2	Y_6	Y_6	Y^{13}
S_7	$\frac{a}{a+\frac{1}{2}(b+c)}$	$T_{\theta=\frac{1}{2}}$	N^a	N_1	N_2	Y_6	Y_6	Y^{13}
S_8	$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$	$S_{\theta=\frac{1}{2}}$	N^a	N_1	N_2	Y^9	N^{12}	N_5
S_9	$\frac{a-(b+c)+d}{a+b+c+d}$	$2S_4-1$	Y^9	N^a	N_2	Y_6	Y_6	Y^{13}
S_{10}	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$		N^a	N_1	N_2	N^a	N_4	N_5
S_{11}	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{c+d}+\frac{d}{b+d}\right)$		N^a	N_1	N_2	N^a	N_4	N_5
S_{12}	$\frac{a}{\sqrt{(a+b)(a+c)}}$		N^a	N_1	N_2	Y_6	Y_6	Y^{13}
S_{13}	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$		N^a	N_1	N_2	Y_6	Y_6	Y^{13}
S_{14}	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$		N^a	N_1	N_2	Y_6	Y_6	Y^{13}
S_{15}	$\frac{ad-bc}{ad+bc}$		N^a	N_1	N_2	N^a	N_4	N_5

[a] The result follows from a counter-example given in Appendix II.

[b] S_1 is referred to in section 5; $(1 - S_1)$ may be negative so its metric and Euclidean properties are irrelevant.

[c] As given in the defining column, self-similarities for S_2 need not be unity so that $(1 - S_2)$ and $\sqrt{(1 - S_2)}$ are not even metrics. The results given in the table are for S_2 defined to have unit self-similarity (see text).

NOTE: Superfices refer to theorems in the text used to establish primary results. Suffices of S are reference numbers, while suffices in columns 1 - 6 refer to columns of this table from which the result follows directly: e.g., N_2 in column (3) means that $\underset{\sim}{\Sigma}$ is not p.s.d. because column (2) states that the coefficient is non-Euclidean.

in Appendix II. To establish (a), i.e., the p.s.d. property of **S**, is the purpose of the following theorem.

Theorem 13. *Matrices of the coefficients* S_2, S_3, S_4, S_5, S_6, S_7, S_9, S_{12}, S_{13}, *and* S_{14} *are p.s.d.*

Proof. The proof requires the use of some standard results of algebra. These are:

I. $\mathbf{X}'\mathbf{X}$ is p.s.d. for all real **X**.

II. Schur's theorem, that if **A** and **B** are p.s.d., then so is **A*B**.

III. **A** is p.s.d. iff all principal minors are non-negative.

IV. The sum of a set of p.s.d. matrices is itself p.s.d.

These are mostly well-known results; references are given by Gower (1971) with short proofs of I and IV and also a simplified proof of II for symmetric matrices, which is all that is required here.

P.s.d. property of coefficients S_2, S_4 **and** S_9: By scoring u for + and v for − in Table 1 and using result I, shows that the matrix with elements $u^2 a_{ij} + uv(b_{ij} + c_{ij}) + v^2 d_{ij}$ is p.s.d. Setting $u = 1$, $v = -1$ immediately shows that $\mathbf{S}_9$ is p.s.d. Setting $u = 1$, $v = 0$ shows that **A**, the matrix with elements a_{ij}, is p.s.d. Because $\mathbf{A}/m$ does not have a unit diagonal, the corresponding dissimilarity is not a metric. $\mathbf{S}_2$ is defined to be $\mathbf{A}/m$ except on the diagonal, where it is defined as unity. Therefore $\mathbf{A}/m$ differs from $\mathbf{S}_2$ only by a non-negative diagonal matrix $diag\,(1 - \frac{x_1}{m}, 1 - \frac{x_2}{m}, \ldots, 1 - \frac{x_n}{m})$ where x_i is the number of occurrences of + for the i-th unit. It follows from IV that $\mathbf{S}_2$ is p.s.d. Setting $u = 0$, $v = 1$ shows that **D** (elements d_{ij}) is p.s.d. and hence (from IV) $\mathbf{A} + \mathbf{D}$ is p.s.d. and so is $\mathbf{S}_4 = (\mathbf{A} + \mathbf{D})/m$. Note that $\mathbf{S}_4$ is the same as $S_{\theta=1}$.

P.s.d. property of coefficients S_3, S_5 **and** S_6: We may write $\mathbf{S}_3 = \mathbf{S}_2 + \mathbf{S}_2^* \sum_{i=1}^{\infty} \mathbf{D}^{(i)}/m^i$ which is p.s.d. from repeated uses of II and IV. Similarly $\mathbf{S}_5 = \sum_{i=1}^{\infty} \mathbf{S}_3^{(i)}/2^i$ and $\mathbf{S}_6 = \sum_{i=1}^{\infty} \mathbf{S}_4^{(i)}/2^i$ which repeated use of II and IV shows to be p.s.d. Note that $\mathbf{S}_3$ and $\mathbf{S}_5$ are the same as $\mathbf{T}_{\theta=1}$ and $\mathbf{T}_{\theta=2}$, respectively, and $\mathbf{S}_6$ is the same as $\mathbf{S}_{\theta=2}$.

P.s.d. property of coefficient S_7: To prove that $\mathbf{S}_7$ is p.s.d. is more delicate. We notice that for $\mathbf{S}_7$, $s_{ij} = \frac{2a_{ij}}{x_i + x_j}$ where x_i is, as above, the number of occurrences of + in the i-th unit. Thus $\mathbf{S}_7 = 2\mathbf{A}^*\mathbf{X}$ where

$x_{ij} = 1/(x_i + x_j)$. Since **A** is p.s.d. it is sufficient to show that **X** is p.s.d. It is easily verified that

$$det\,\mathbf{X} = \frac{\prod_{i \neq j}^{n} \left(\frac{x_i - x_j}{x_i + x_j}\right)^2}{2^n \prod_{i=1}^{n} x_i} .$$

Thus for non-negative $\mathbf{x}_i$ we have that $det\,\mathbf{X} \geqslant 0$ with equality only when $x_i = x_j$. By replacing n by $1, 2, \ldots, n-1$ it is clear that the result is valid for all principal minors of **X** and hence, by III, **X** is p.s.d. and so is $\mathbf{S}_7$. Note that $\mathbf{S}_7$ is the same as $\mathbf{T}_{\theta = 1/2}$.

P.s.d. property of coefficients S_{12}, S_{13} and S_{14}: The scores for the i-th unit, i.e., column 1 of Table 1, may be scaled by an arbitrary factor s_i^{-1}. If we set $u = 1$, $v = 0$ and $s_i = \sqrt{x_i(m - x_i)}$, the 'standard deviation' of the i-th unit, then result I gives that the matrix with elements

$$\frac{a_{ij}}{\sqrt{x_i(m - x_i)\; x_j(m - x_j)}}$$

is p.s.d. Multiplying by **D** shows that $\mathbf{S}_{13}$ is p.s.d. Similarly we may subtract an arbitrary constant k_i from the i-th column. Setting $k_i = x_i/m$, the mean, then result I gives the product-moment correlation matrix for binary variables, which is well-known to have the form $\mathbf{S}_{14}$ of Table 2 and which must be p.s.d. Finally, setting $u = 1$, $v = 0$, $k_i = 0$, $s_i = \sqrt{x_i}$ shows that $\mathbf{S}_{12}$ is p.s.d. ●

4.2 Dissimilarity Coefficients for Quantitative Variables

Table 3 is an expanded version of a similar table given by Gower (1985). It lists many of the standard distance coefficients and states whether they are metric or not, and Euclidean or not. The frequent use of modulus operators is to take count of possible negative values for quantitative variables. Although direct observation of negative values is rare in practice, they can easily arise when data are standardized to have zero mean or are transformed, say, to logarithms (see also section 5.3). It is therefore desirable to examine the coefficients separately, first when all values are non-negative and then when negative values are admissible. This explains the two sets of columns in the table. For dissimilarity coefficients with quantitative variables we have, for the most part, considered only the basic coefficient D_i and not $\sqrt{D_i}$.

It is immediate that D_1, D_2, D_3 and D_4 are all metric whether or not negative values are included, because D_4 is the Minkowski metric of which the others are special cases. D_1 and D_2 are in the basic form of the Euclidean metric. The quantities r_k appearing in the definitions of D_2, D_3 and D_4 are arbitrary except that, when this is not otherwise irrelevant, they must be positive. Usually r_k is taken to be either the standard deviation or the range of the k-th variable, but other possibilities exist. Gower (1971) showed that $\sqrt{D_3}$ was Euclidean when r_k was taken as the range, but need not be Euclidean when r_k is taken as the standard error.

The quantities x_{ij} might be restricted to the values 1 and 2 or 0 and 1, which may be regarded as formal scores for the states $-$, $+$ of binary variables. The values 1, 2 substituted into the formulae of Table 3 give results proportional to $1 - S_4$ in every case, except for D_8 (which becomes $(b + c)/(4a + 3(b + c) + 2d)$) and D_9 (which becomes $(b + c)/(2(a + b + c) + d)$). Of course, it is D_1^2, D_2^2, D_4' and D_5^2, not the coefficients themselves, that are proportional to $1 - S_4$. The values of 0, 1 also produce results proportional to $1 - S_4$ for D_1^2, D_2^2, D_3, D_4', but for D_8 give $1 - S_7$ and for D_9 give $1 - S_3$; all other values are indeterminate, leading to terms in zero-divided-by-zero. Apart from the interest of these correspondences, the above results immediately allow some of the non-Euclidean findings of Table 2 to be transferred to Table 3.

Thus binary versions of D_3, D_4 (with $t = 1$), D_6, D_7 and D_{10} are equivalent to $1 - S_4$ and hence these coefficients are non-Euclidean. Similarly D_8 is equivalent to $1 - S_7$ and hence is non-metric, and depending on the scoring used, D_9 is equivalent to $1 - S_3$ and hence is non-Euclidean. Thus D_3, D_4, and D_6 to D_{10} are non-Euclidean and remain so when negative values are permitted, while D_8 is non-metric for positive and negative values.

The indeterminacies induced by allowing zero scores are avoided by ignoring double-zero terms in the coefficients of Table 3 and dividing by the remaining number of matches rather than by p. This process is similar to that used in the definitions of those coefficients of Table 2 that are not functions of the number of negative matches d. The resulting coefficients in Table 3 are then all equivalent to $1 - S_3$ (except D_8 which is equivalent to $1 - S_7$) but we have not further investigated the properties of coefficients defined in this way, although such coefficients are useful (see section 5.1).

It remains to show that D_5, D_6, D_7, D_9, and D_{10} are metrics, at least for positive values. This is done in the following theorem.

Theorem 14. *Coefficients D_5, D_6, D_7, D_9 and D_{10} are metric for positive values of the variables but, with the exception of D_7, are not metric for negative values.*

TABLE 3

Properties of Dissimilarity Coefficients Amongst Quantitative Variables

Coefficient		Positive values only (1) Metric	Positive values only (2) Euclidean	Negative values permitted (3) Metric	Negative values permitted (4) Euclidean
D_1^2	$\frac{1}{p}\sum_{k=1}^{p}(x_{ik}-x_{jk})^2$	Y	Y	Y	Y
D_2^2	$\frac{1}{p}\sum_{k=1}^{p}(x_{ik}-x_{jk})^2/r_k^2$	Y	Y	Y	Y
D_3	$\frac{1}{p}\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert/r_k$	Y	N^+	Y	N_2
D_4^t	$\frac{1}{p}\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert^t/r_k^t, (t\geq 1)$	Y	N^+	Y	N_2
D_5^2	$\frac{1}{p}\sum_{k=1}^{p}\frac{(x_{ik}-x_{jk})^2}{(x_{ik}+x_{jk})^2}$	Y^{14}	Y	N^a	N_3
D_6	$\frac{1}{p}\sum_{k=1}^{p}\frac{\lvert x_{ik}-x_{jk}\rvert}{\lvert x_{ik}+x_{jk}\rvert}$	Y^{14}	N^+	N^a	N_3
D_7	$\frac{1}{p}\sum_{k=1}^{p}\frac{\lvert x_{ik}-x_{jk}\rvert}{\lvert x_{ik}\rvert+\lvert x_{jk}\rvert}$	Y^{14}	N^+	Y^{14}	N_2
D_8	$\frac{\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert}{\sum_{k=1}^{p}(x_{ik}+x_{jk})}$	N^+	N_1	N_1	N_1
D_9	$\frac{\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert}{\sum_{k=1}^{p}\mathrm{Max}(x_{ik},x_{jk})}$	Y^{14}	N^+	N^a	N_3
D_{10}	$\frac{1}{p}\sum_{k=1}^{p}\left(1-\frac{\mathrm{Min}(x_{ik},x_{jk})}{\mathrm{Max}(x_{ik},x_{jk})}\right)$	Y^{14}	N^a	N^a	N_3

NOTE: Explanation as for Table 2, except that + denotes results that follow from Table 2 when the continuous variables are replaced by binary values.

Proof. Appendix II gives counter-examples that show that D_5, D_6, D_9 and D_{10} are non-metric when negative values are permitted.

Coefficients D_5 and D_6: From theorem 8 we see that the metricity of D_5 follows from that of D_6, if we can prove it for D_6, and that it is sufficient to establish this for a single variable of D_6. Suppose a variable takes the values x_1,x_2,x_3 for three sampling units; then without loss of generality we may assume $x_1 \geqslant x_2 \geqslant x_3$. The contributions to D_6 are then

$$d_{23} = \frac{x_2 - x_3}{x_2 + x_3} \quad , d_{13} = \frac{x_1 - x_3}{x_1 + x_3} \quad \textit{and} \quad d_{12} = \frac{x_1 - x_2}{x_1 + x_2} \quad .$$

It is trivial to show that $d_{13} \geqslant d_{23}$ and that $d_{13} \geqslant d_{12}$. It follows that d_{12} and d_{23} are the two shortest sides of the triangle and that the metric property is valid for all permutations provided $d_{12} + d_{23} \geqslant d_{13}$. Simple algebraic manipulation gives

$$d_{12} + d_{23} - d_{13} = \frac{(x_2 - x_3)\ (x_1 - x_3)\ (x_1 - x_2)}{(x_2 + x_3)\ (x_1 + x_3)\ (x_1 + x_2)}$$

which is positive and hence D_6 is metric and, by theorem 8, so is D_5.

Coefficients D_7 and D_{10}: For D_7 and D_{10} it also suffixes to investigate the properties of a single variable. For positive values, D_7 is the same as D_6 and hence is metric. With negative values of x_1,x_2,x_3, it is sufficient in D_7 to consider the case where only one value x_3 (say) is negative. This is because D_7 is invariant to changes of *all* the signs of x_1, x_2 and x_3 so that three negative values have the same effect as three positive values, and hence define a metric, while by changing signs, two negative values give the same result as one negative value. Writing $x_3 = -x_3^*$ the contributions to D_7 become:

$$\frac{x_1 - x_2}{x_1 + x_2} \quad , \frac{x_1 + x_3^*}{x_1 + x_3^*} \quad \textit{and} \quad \frac{x_2 + x_3^*}{x_2 + x_3^*} \quad .$$

In these terms we require only that $x_1 > x_2$ which causes no loss of generality. Thus in this case the two longest sides both equal unity, showing that the metric property is preserved with D_7 even for negative values.

With $x_1 \geqslant x_2 \geqslant x_3 \geqslant 0$, D_{10} gives:

$$d_{23} = \frac{x_2 - x_3}{x_2} \quad , d_{13} = \frac{x_1 - x_3}{x_1} \quad \textit{and} \quad d_{12} = \frac{x_1 - x_2}{x_1}$$

with $d_{13} \geqslant d_{23}$ and $d_{13} \geqslant d_{12}$, showing that d_{13} is the longest side. Thus

$$d_{12} + d_{23} - d_{13} = \frac{(x_2 - x_3)\ (x_1 - x_2)}{x_1\ x_2}$$

which is positive and D_{10} is metric for positive values.

Coefficient D_9: The remaining coefficient, D_9, is less easy to handle because the separate summations of numerator and denominator imply that it does not suffice to establish the metric result for a single variable. The following argument, a variant of theorem 3, starts by noting that every variable must belong to one of the six classes: $A(x_1 \geqslant x_2 \geqslant x_3)$, $B(x_1 \geqslant x_3 \geqslant x_2)$, $C(x_2 \geqslant x_1 \geqslant x_3)$, $D(x_2 \geqslant x_3 \geqslant x_1)$, $E(x_3 \geqslant x_1 \geqslant x_2)$, and $F(x_3 \geqslant x_2 \geqslant x_1)$. The following table gives the sums of every variable in each class for each of three units. The suffixes serve the dual purpose of distinguishing the sums for the different units and indicating their rankings. Thus C_1 is the sum of all variables of type C for the second unit and $C_1 \geqslant C_2 \geqslant C_3$.

Unit	A	B	C	D	E	F
1	A_1	B_1	C_2	D_3	E_2	F_3
2	A_2	B_3	C_1	D_1	E_3	F_2
3	A_3	B_2	C_3	D_2	E_1	F_1
4	A_2	B_2	C_2	D_2	E_2	F_2

The table also shows the sums for a constructed fourth unit, the values always being those of the middle values of the other three units. When all the values in the tables are positive we have that:

$$d_{24} = \frac{B_2 - B_3}{B_2} + \frac{C_1 - C_2}{C_1} + \frac{D_1 - D_2}{D_1} + \frac{E_2 - E_3}{E_2} \quad .$$

$$d_{34} = \frac{A_2 - A_3}{A_2} + \frac{C_2 - C_3}{C_2} + \frac{E_1 - E_2}{E_1} + \frac{F_1 - F_2}{F_1} \quad .$$

Also

$$d_{23} = \frac{A_2 - A_3}{A_2} + \frac{B_2 - B_3}{B_2} + \frac{C_1 - C_3}{C_1} + \frac{D_1 - D_2}{D_1} + \frac{E_1 - E_3}{E_1} + \frac{F_1 - F_2}{F_1} \quad .$$

Now

$$\frac{C_1 - C_2}{C_1} + \frac{C_2 - C_3}{C_2} - \frac{C_1 - C_3}{C_1} = \frac{(C_2 - C_3)\,(C_1 - C_2)}{C_1\,C_2} \geqslant 0$$

and similarly for the terms in E_1, E_2 and E_3. It follows that $d_{24} + d_{34} \geqslant d_{23}$ so that units 2, 3 and 4 satisfy the metric inequality. Further:

$$d_{21} = \frac{A_1 - A_2}{A_1} + \frac{B_1 - B_3}{B_1} + \frac{C_1 - C_2}{C_1} + \frac{D_1 - D_3}{D_1} + \frac{E_2 - E_3}{E_2} + \frac{F_2 - F_3}{F_2} \quad .$$

Because

$$\frac{B_1 - B_3}{B_1} \geqslant \frac{B_2 - B_3}{B_2} \quad \textit{and} \quad \frac{D_1 - D_3}{D_1} \geqslant \frac{D_1 - D_2}{D_1}$$

we have that $d_{21} \geqslant d_{24}$. Similarly $d_{31} \geqslant d_{34}$. Thus

$$d_{12} + d_{13} \geqslant d_{24} + d_{34} \geqslant d_{23}$$

establishing the metric property for D_9. Note that this result does not require that all observations be positive, only that all the sums in tables as above should be non-negative. ●

It remains to show that D_5 is Euclidean for positive values. Consider the similarity matrix with elements

$$s_{ij} = \frac{1}{p} \sum_k \frac{4x_{ik}\, x_{jk}}{(x_{ik} + x_{jk})^2} \quad .$$

It follows from theorem 6 that D_5 is Euclidean if $\mathbf{S}$ is p.s.d. It is sufficient to establish the result for a single variable, i.e., for a matrix with elements

$$s_{ij} = \frac{4x_i\, x_j}{(x_i + x_j)^2} \quad .$$

Now the matrix with elements $x_i\, x_j$ is $\mathbf{x}\,\mathbf{x}'$ which by I is p.s.d. The proof that $\mathbf{S}_7$ is p.s.d. included a proof that $\mathbf{X}$ with elements $1/(x_i + x_j)$ is p.s.d. Because $\mathbf{S} = 4(\mathbf{x}\,\mathbf{x}')^* \mathbf{X}^{(2)}$ it follows from II that $\mathbf{S}$ is p.s.d. and hence that D_5 is Euclidean.

5. Choice of Coefficient

None of the properties listed in the previous sections — or in the present one, for that matter — is conclusive in choosing a coefficient. A coefficient has to be considered in the context of the descriptive statistical study of which it is a part, including the nature of the data, and the intended type of analysis. The purpose of this section is to discuss, within this context, how the known properties of coefficients influence an appropriate choice.

The nature of the data strongly influences the choice of a coefficient. Table 2 lists coefficients appropriate for binary variables, and Table 3 those for quantitative variables. Under certain circumstances, quantitative data may best be treated as binary as when dichotomizing noisy quantitative variables (Legendre and Legendre 1983b), or when the pertinent information, for the purpose one has in mind, depends on a known threshold value. For example, when classifying river areas according to their suitability for growing edible fish as judged by threshold levels of pesticides and heavy metals, the data should be coded in binary form, as falling above or below the tolerated toxicity level, despite the fact that the measurements themselves are quantitative.

It may be appropriate to treat different variables differently; at least three coefficients have been described that handle mixtures of different kinds of variable (Estabrook and Rogers 1966; Gower 1971; Legendre and Chodorowski 1977). Other mixed coefficients are easily constructed by combining coefficients from Tables 2 and 3, either with, or without, differential weighting. Some coefficients may at times give negative values of similarity (S_9, S_{14}, and S_{15}). Translated into dissimilarities, these coefficients produce values larger than one, as with D_1. Most analyses are unaffected, but suitable scaling is needed when these coefficients are combined with other binary similarity coefficients. The relationships between binary and quantitative coefficients, described in section 4.2, can be used with profit to handle mixtures of binary and quantitative variables. This can be done most easily with coefficients, like D_2 to D_7 and D_{10}, where each variable-comparison is divided by a normalizing factor, depending upon this variable alone, before summing. The equivalence between binary and quantitative coefficients tells us what binary coefficient is mimicked by each quantitative coefficient, in such combinations of variables.

The method of analysis itself may limit the choice of coefficient. Thus a matrix of resemblance is often analyzed by a clustering or ordination method which may be either ordinal or metric. Metric methods often have a geometric rationale that implies that a metric and possibly a Euclidean coefficient should be chosen, thus disfavoring non-metric coefficients (Gower 1984a). Table 2 (column 5) and Table 3 (columns 2 and 4) list the coefficients that are fully Euclidean. Notice however that the treatment of

missing values (e.g., Gower 1971; Legendre and Legendre 1983a) may destroy the metric and Euclidean properties listed in Tables 2 and 3. The definitions of metric **D** and Euclidean **D** given in section 1 require that the coefficients be Euclidean and/or metric for *all* data. However it is a general observation that constructed non-Euclidean matrices tend to be pathological and only slightly non-Euclidean. Thus in matrix (1), section 2, the distance of P_4 from the other vertices is constrained to have length between 1 and 1.15 units if it is to be metric but not Euclidean. A simple monotonic transformation often restores Euclideanarity. For example theorem 7 shows that simple additive constants can always be found to make a coefficient Euclidean, and Table 2 shows that it often suffices to replace d by $\sqrt{d}$; the close of section 5.1 indicates further results of this kind. Well-defined measures of non-metricity and non-Euclideanarity need development and their values ascertained for data, using the various coefficients.

Even with those methods, like classical scaling/principal coordinates analysis, that seem to require a Euclidean coefficient, a modest departure may be inconsequential. Cailliez and Pagès (1976), and Sibson (1979) have shown that ignoring 'small' imaginary dimensions is often acceptable. Nevertheless to know that a matrix is p.s.d. is helpful (Table 2, columns 3 and 6) since some algorithms for spectral decomposition would otherwise fail.

Although criteria have been proposed for choosing among clustering methods (e.g., Baker 1974; Blashfield 1976; Cunningham and Ogilvie 1972; Everitt 1974; Fisher and Van Ness 1971; Hubert 1974; Jardine and Sibson 1968; Legendre and Legendre 1983a; Rand 1971; Sibson 1971; Williams *et al.* 1971a, 1971b), it is often considered that different methods focus on complementary aspects of the geometry of the set of objects. High resolution and linearity of the measure of resemblance (see section 5.3) are desirable properties when clustering.

5.1 Families of Coefficients

Binary coefficients (Tables 2 and 4) may be classified by the way they deal with negative matches. In Table 4, the coefficients S_1, S_2, S_3, S_5, S_7, S_{10}, and S_{12}, do not involve d and hence ignore negative matches; they are sometimes termed asymmetric, or asymmetrical. The other coefficients are symmetric in a and d and so treat positive and negative matches equally. Indeed, our +/− notation may not refer to presence/absence but to qualitative values of equal status, such as black/white; in taxonomy the difference is often unclear, as when "white" indicates the absence of a gene that controls the state "black." In ecological applications, coefficients that do not regard double absence as an indication of similarity are relevant when the absence of a species from two sites may correspond to extreme but opposite conditions, both prohibiting growth. However the double absence may

represent the same unfavorable condition at both sites. That those who use these coefficients must decide what is appropriate in each instance, and for which variables or species, is basic to the process of choosing a coefficient appropriate to a given problem (section 5.4).

Similarly, some quantitative coefficients (Table 3) exclude double zeros from the comparison in both the numerator and the denominator (D_8, D_9). In other instances (D_5, D_6, D_7 and D_{10}), double zeros lead to indeterminacy of the coefficient. When frequencies of zero represent the same kind of common absence of condition that, with binary variables, leads one to reject negative matches, these coefficients are appropriate, but it is necessary then to skip double zeros in their computation. This is equivalent to interpreting 0/0 as zero in the numerator, and counting p as the number of variables in the comparison not presenting double-zeros, as described in section 4.2. Care has to be taken that instability is not introduced by pairs of values which, although non-zero, are very small. Finally double zeros may be excluded in D_1 to D_4 simply by defining p as the number of variables not presenting a double-zero, for this pair of objects, but we have not investigated whether the resulting dissimilarities remain metric.

The coefficients $(1 - T_\theta)^{1/t}$ and $(1 - S_\theta)^{1/t}$ of section 3 may be written

$$(1 - T_\theta)^{1/t} = (\theta/\ (x + \theta))^{1/t}$$

and

$$(1 - S_\theta)^{1/t} = (\theta/\ (y + \theta))^{1/t}$$

where $x = a/\ (b + c)$ and $y = (a + d)/\ (b + c)$. Thus the first coefficient, which includes S_3, S_5 and S_7 as special cases, decreases its value with x, and increases with θ and with t. All the coefficients of this set are monotonically related and will give the same result when used with order-invariant methods. These are methods like single, complete, or proportional-link linkage cluster analysis, or any form of non-metric multidimensional scaling, that use only the ordinal and not the absolute values of the data. Another way of putting this is that for order-invariant methods, all the coefficients of the class are equivalent to x, which is also S_1, sometimes used itself as a coefficient (Kulczynski 1928). Similar remarks apply to the set of coefficients $(1 - S_\theta)^{1/t}$, which includes S_4, S_6 and S_8, that are equivalent to y. This series includes $S_9 = (2S_4 - 1) = (1 - 2/\ (y + 1))$ which is also monotonically related to y. Yet Figure 2 shows that the metric and Euclidean properties of both general coefficients vary substantially with θ and t. In Table 6, coefficients within each of these two families will be treated as equivalent, except for S_8 because of its non-Euclidean behavior.

TABLE 4

Properties of Similarity Coefficients for Binary Variables

Coefficient		Negative matches[a]	Results from the OCCAS		
			Resolution	Non-linearity	C.V.
S_1	$\frac{a}{b+c}$	E	0.773	1.377	178.3
S_2	$\frac{a}{a+b+c+d}$	E	0.056	0.016	28.9
S_3	$\frac{a}{a+b+c}$	E	0.083	0.024	28.9
S_4	$\frac{a+d}{a+b+c+d}$	I	0.056	0.016	28.9
S_5	$\frac{a}{a+2(b+c)}$	E	0.075	0.039	52.7
S_6	$\frac{a+d}{a+2(b+c)+d}$	I	0.062	0.028	44.9
S_7	$\frac{a}{a+\frac{1}{2}(b+c)}$	E	0.086	0.030	35.0
S_8	$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$	I	0.043	0.012	28.5
S_9	$\frac{a-(b+c)+d}{a+b+c+d}$	I	0.111	0.032	28.9
S_{10}	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	E	0.073	0.026	36.2
S_{11}	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{c+d}+\frac{d}{b+d}\right)$	I	0.054	0.015	28.6
S_{12}	$\frac{a}{\sqrt{(a+b)(a+c)}}$	E	0.081	0.026	32.0
S_{13}	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	I	0.076	0.023	29.7
S_{14}	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	I	0.109	0.029	26.4
S_{15}	$\frac{ad-bc}{ad+bc}$	I	0.133	0.118	88.3

[a] Negative matches included (I) or excluded (E).

5.2 Quantitative Dissimilarity Coefficient Types

Another criterion governing the choice of a coefficient has been developed by Legendre, Dallot and Legendre (1985). It consists of

classifying the coefficients of resemblance for quantitative variables into three types, according to how they weight a given difference for variables with different ranges of variation. This classification is, of course, only relevant for data tables that are dimensionally homogeneous (e.g., financial currencies, frequencies (counts), standardized data). Such tables mostly contain only non-negative values, and the following argument is developed in terms of frequencies. The problem of negative values found in transformed data is examined in section 5.3.

Type-1. Suppose the same difference is found, between two objects, for a variable bearing high frequencies, and also for one with much smaller frequencies. In type-1 coefficients both variables contribute equally to the distance. The coefficients D_1, D_8 and D_9 of Tables 3 and 5 belong to this group as will any coefficient whose denominator is constant or is a separate summation from the numerator. Thus with D_8, using the simple numerical example shown in Table 5 ("positive values only"), it is easy to verify that each variable's difference contributes 10/290 to the sum of terms making up D_8. With some coefficients, other than those studied in the present paper, this property holds only when the two vectors corresponding to the sampling units being compared have the same "importance," the importance corresponding to different concepts, depending upon the coefficient. It is the length of the vector in the chord distance and in the geodesic metric, while it is the sum of the values in the vector, in Renkonen's (1938) percentage similarity, all of which belong to this sub-class of type-1 coefficients. Table 5 shows that D_8 can behave badly when negative values are permitted; its use should be limited to positive values, unlike D_1 which can handle negative values quite nicely.

Type-2a. In coefficients of this class, a difference found between two sampling units for a variable with high values contributes less to the dissimilarity (more to the similarity) than the same difference found between these units for variables with smaller values (except when the difference is zero). The Canberra metric, D_6, as well as its associated forms D_5 and D_7, belong to this type; so does D_{10} (Table 5, positive values only). When negative values are allowed, only D_7 remains of type-2a, since the four others can be indeterminate.

Type-2b. *In this class are found the dissimilarities for which an equal difference receives a weight inversely proportional to the variability of the* variable in the whole set of sampling units under study. This is so with D_4 and its special cases, D_2 and D_3, where the measure of variability is the range of variation of each variable over all units. D_2 to D_4 are not affected by negative values. Coefficients of type-2b behave similarly to those of type-2a but instead of the weight of a variable depending only on the pair of sampling units being compared, it is defined over all pairs of unit-comparisons.

TABLE 5

Properties of Dissimilarity Coefficients for Quantitative Variables

Coefficient	Positive values only: Test example 1			Type	Negative values permitted[a]: Test example 2				Type
	100	40	20		90	-100	5	0	
	90	30	10		80	- 90	-5	-10	
r^b =	100	40	20		100	100	20	20	
$D_1^2 = \frac{1}{p}\sum_{k=1}^{p}(x_{ik}-x_{jk})^2$	10^2	10^2	10^2	1	10^2	10^2	10^2	10^2	1
$D_2^2 = \frac{1}{p}\sum_{k=1}^{p}(x_{ik}-x_{jk})^2/r_k^2$	$\frac{10^2}{100^2}$	$\frac{10^2}{40^2}$	$\frac{10^2}{20^2}$	2b	$\frac{10^2}{100^2}$	$\frac{10^2}{100^2}$	$\frac{10^2}{20^2}$	$\frac{10^2}{20^2}$	2b
$D_3 = \frac{1}{p}\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert/r_k$	$\frac{10}{100}$	$\frac{10}{40}$	$\frac{10}{20}$	2b	$\frac{10}{100}$	$\frac{10}{100}$	$\frac{10}{20}$	$\frac{10}{20}$	2b
$D_4^t = \frac{1}{p}\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert^t/r_k^t$	$\frac{10^t}{100^t}$	$\frac{10^t}{40^t}$	$\frac{10^t}{20^t}$	2b	$\frac{10^t}{100^t}$	$\frac{10^t}{100^t}$	$\frac{10^t}{20^t}$	$\frac{10^t}{20^t}$	2b
$D_5^2 = \frac{1}{p}\sum_{k=1}^{p}\frac{(x_{ik}-x_{jk})^2}{(x_{ik}+x_{jk})^2}$	$\frac{10^2}{190^2}$	$\frac{10^2}{70^2}$	$\frac{10^2}{30^2}$	2a	$\frac{10^2}{170^2}$	$\frac{10^2}{190^2}$	$\frac{10^2}{0^2}$	$\frac{10^2}{10^2}$	indet.
$D_6 = \frac{1}{p}\sum_{k=1}^{p}\frac{\lvert x_{ik}-x_{jk}\rvert}{\lvert x_{ik}+x_{jk}\rvert}$	$\frac{10}{190}$	$\frac{10}{70}$	$\frac{10}{30}$	2a	$\frac{10}{170}$	$\frac{10}{190}$	$\frac{10}{0}$	$\frac{10}{10}$	indet.
$D_7 = \frac{1}{p}\sum_{k=1}^{p}\frac{\lvert x_{ik}-x_{jk}\rvert}{\lvert x_{ik}\rvert+\lvert x_{jk}\rvert}$	$\frac{10}{190}$	$\frac{10}{70}$	$\frac{10}{30}$	2a	$\frac{10}{170}$	$\frac{10}{190}$	$\frac{10}{10}$	$\frac{10}{10}$	2a
$D_8 = \frac{\sum\lvert x_{ik}-x_{jk}\rvert}{\sum(x_{ik}+x_{jk})}$	$\frac{10}{290}$	$\frac{10}{290}$	$\frac{10}{290}$	1	$\frac{10}{-30}$	$\frac{10}{-30}$	$\frac{10}{-30}$	$\frac{10}{-30}$	1
$D_9 = \frac{\sum\lvert x_{ik}-x_{jk}\rvert}{\sum\mathrm{Max}(x_{ik},x_{jk})}$	$\frac{10}{160}$	$\frac{10}{160}$	$\frac{10}{160}$	1	$\frac{10}{5}$	$\frac{10}{5}$	$\frac{10}{5}$	$\frac{10}{5}$	1
$D_{10} = \frac{1}{p}\sum_{k=1}^{p}\left[1-\frac{\mathrm{Min}(x_{ik},x_{jk})}{\mathrm{Max}(x_{ik},x_{jk})}\right]$	$\frac{10}{100}$	$\frac{10}{40}$	$\frac{10}{20}$	2a	$\frac{10}{100}$	$\frac{10}{-90}$	$\frac{10}{5}$	$\frac{10}{0}$	indet.

Table 5 (cont'd)

Coefficient	Negative matches[c]	Results from the OCCAS: Positive values only			Negative values permitted		
		Resol.	Non-lin.	C.V.	Resol.	Non-lin.	C.V.
D_1(B)[d]	I or E	0.077	0.025	31.9	0.073	0.017	23.8
D_2	I or E	0.059	0.019	31.9	0.063	0.021	33.2
D_3	I or E	0.054	0.019	35.9	0.056	0.021	38.5
D_4(t=4)	I or E	0.064	0.023	35.4	0.068	0.023	33.8
D_5	E	0.017	0.022	131.7	1.640	3.189	194.4
D_6	E	0.029	0.028	94.8	2.044	5.621	275.0
D_7	E	0.029	0.028	94.8	1.617	5.573	344.6
D_8	E	0.069	0.043	62.0	4.843	2.151	250.9
D_9	E	0.065	0.041	63.7	0.917	1.028	112.1
D_{10}	E	0.030	0.025	84.5	-	-	-

[a] Double zeros are not used, because most of these coefficients exclude them.
[b] The range r is fixed arbitrarily.
[c] Negative matches included (I) or excluded (E).
[d] D_1(B): form of coefficient D_1 bounded between 0 and 1 by dividing all values by the maximum found in all three OCCAS.

TABLE 6

Choice of a Q-mode Coefficient of Resemblance: Decision-making Process

1) Binary data (or treated as such) ----------→ see 2

2) Negative matches included: reject S_{15} (problem of linearity); for metric scaling, prefer metric coefficients S_4, S_6 and S_9, or the Euclidean SQRT(1-S) form of S_4, S_6, S_9, S_{13} and S_{14}.

2) Negative matches excluded: reject S_1 (S larger than 1; problem of linearity); for metric scaling, prefer metric coefficients S_2, S_3 and S_5, or the Euclidean SQRT(1-S) form of S_2, S_3, S_5, S_7 and S_{12}.

1) Quantitative data ----------→ see 3

3) Positive values only ----------→ see 4

4) Double-zeroes included ----------→ see 5

5) Type-1 data: D_1

5) Type-2 data: D_2 to D_4

4) Double-zeroes excluded: reject D_5 (problem of linearity ---→ see 6

6) Type-1 data: D_1, D_8, D_9; skewed data, with a few extreme values: avoid D_1 that would give undue importance to those data, by squaring.

6) Type-2 data: choose among type-2a coefficients D_6 (same as D_7) and D_{10}, or among type-2b coefficients D_2 to D_4.

3) Negative values permitted: same coefficients for double-zeroes included or excluded ----------→ see 7

7) Type-1 data: D_1

7) Type-2 data: D_2 to D_4

1) Mixture of binary and quantitative data: choose among D_2 to D_7, and D_{10}, that divide each variable-comparison by a normalizing factor, before summing; go to quantitative data, above.

With standardized variables (where about half the values are negative) that call for a type-1 coefficient, D_1 is the only function, of those listed in Tables 3 and 5, that seems appropriate. The "patterns of sensitivity" established by Faith (1985) point out properties of coefficients similar to the types described above.

5.3 Resolution and Linearity

Recognizing that resemblance functions may behave differentially with different data sets, Hajdu (1981) has developed "ordered comparison case series" (OCCAS) that he applied to a variety of measures of resemblance. Each OCCAS is made of artificial quantitative data comparing two sampling units that vary linearly from one case to the next. Thus if for the first unit the two variates take values p,q then a second unit is considered where the variates take values of the form $rx + u$, $sx + v$ so that dissimilarity may be considered as a function $d(p,q,r,s,u,v,x)$. The effects of "linear changes" given by varying x can then be studied. If one requires linear changes of this kind to imply linear changes in dissimilarity, then only certain coefficients will be acceptable. Hajdu considers a range of equally-spaced x-values $x_1,x_2, \ldots, x_t$ and for fixed p,q,r,s,u,v defines the differences:

$$\delta_i = \mid d(p,q,r,s,u,v,x_{i+1}) - d(p,q,r,s,u,v,x_i) \mid \quad \textit{for } i = 1, \ldots, (t-1) \ .$$

Departure from linearity (termed *non-linearity* in the following) is defined as the standard deviation of δ_i and a quantity termed *resolution* is the sample mean of δ_i. When the OCCAS results are monotonically related to the amount of change, then this mean becomes $(d(x_t) - d(x_1))/\ (t-1)$. The ratio non-linearity/resolution (x 100) defines a *coefficient of variation* which should be small for a good resemblance function.

High resolution is a property that seems desirable especially in cluster analysis, and probably more so with order-varying cluster methods like UPGMA, UPGMC, WPGMA, WPGMC, or flexible clustering. Hajdu (1981) argues that a coefficient with high resolution makes the clustering structure more stable and less likely to be modified following small changes in data values. Linearity, on the other hand, seems desirable mainly for metric scaling. With nonmetric scaling, linearity is likely to reduce stress, although the resulting ordination would be little affected.

In practice linearity and resolution are evaluated and averaged over sets of values p,q,r,s,u,v; the different OCCAS used, the values we considered, and those of x_i are given in Table 7. Note that Hajdu did not consider negative values or binary variables. The values of resolution and non-linearity, as well as the coefficient of variation for coefficients D_2, D_3 and D_4, listed in Table 5, have been computed with equal ranges (r_k) of 120 for

TABLE 7

Ordered Comparison Case Series (OCCAS) Used to Test the Response of Different Coefficients to Data Series (see text)

(a) Quantitative data, positive values only:

OCCAS1 (9 comparisons)		
	object 1	p=100 and q=0
	object 2	r= -10, u=100 and s=10, v=0,(x=1...9)
OCCAS2 (9 comparisons)		
	object 1	p=100 and q=0
	object 2	r= -10, u=95 and s=0, v=5,(x=1...9)
OCCAS3 (12 comparisons)		
	object 1	p=100 and q=0
	object 2	r=0, u=50 and s=10, v=0,(x=1...12)

(b) Quantitative data, negative values permitted:

OCCAS1 (9 comparisons)		
	object 1	p=45 and q= -55
	object 2	r= -10, u=50 and s=10, v= -50,(x=1...9)
OCCAS2 (9 comparisons)		
	object 1	p=45 and q= -55
	object 2	r= -10, u=50 and s=0, v= -50,(x=1...9)
OCCAS3 (12 comparisons)		
	object 1	p=45 and q= -70
	object 2	r=0, u= -5 and s=10, v= -45,(x=1...12)

(c) Binary data:

	a	b	c	d
OCCAS1 (9 comparisons)	(45 → 5)	(5 → 45)	50	50
OCCAS2 (9 comparisons)	(90 → 10)	(10 → 90)	0	50
OCCAS3 (9 comparisons)	(90 → 10)	(5 → 45)	(5 → 45)	50

both variables. With uneven ranges, non-linearity increases (thus affecting also the coefficient of variation), although much less so in D_3, which then should be preferred for metric scaling. This is probably because a change of range, as created artificially here, corresponds to making some ratios very small compared to others in the sum, an effect that is magnified by squaring (in D_2) or higher powers (in D_4). This family of coefficients deserves more detailed study.

According to the results in Table 5, D_1 to D_4 are almost equally suitable when a coefficient including double-zeros is sought, with preference for D_1 and D_3. When double-zeros are to be excluded, D_8 and D_9 should be

preferred for clustering (high resolution), while all seem equally good for ordination (linearity). D_5 should be avoided when both clustering and ordination are to be attempted (high coefficient of variation).

The dissimilarities, computed from OCCAS with negative values, are quite variable (dissimilarity values range from -29 to $+31$ in these OCCAS); D_5 to D_9 often fall outside the [0,1] range; all five can become negative. Their high coefficients of variation (Table 5) are another reason for excluding them from studies containing negative values. Thirdly, pathological situations can make coefficients D_5 to D_{10} indeterminate; in this respect, D_7 is the best one of this group since only double-zeros can bring it to indeterminacy. This leaves us with coefficients D_1 to D_4, that can be applied in both the symmetrical and asymmetrical cases. We have pointed out, however, that negative values are unlikely to correspond to a problem in which double-zeros are to be excluded. Negative-value problems usually consist either of raw values of variables that can go on either side of a non-absolute zero (which calls probably for one of the type-2b coefficients, D_2 to D_4), or of standardized variables, where a type-1 coefficient seems the most appropriate, as already mentioned.

The results for binary coefficients are summarized in Table 4. For binary variables, Hajdu's scheme is extended to cover many, not just two variates. For this purpose, three new OCCAS are made up. They are described in Table 7c by the values attributed to the a,b,c and d components that make up these coefficients. In all three tests, $(a + b + c)$ equals 100 and d equals 50. In the first case series, a decreases as b increases by steps of 5, leaving $(a + b)$ constant; c and d each have the constant value of 50. The second OCCAS follows the same scheme, except that here c equals zero. In the third OCCAS, a decreases by steps of 10 while b and c increase equally by steps of 5.

The most striking feature is that S_1 can take values larger than one, which corresponds to a negative dissimilarity, making it unsuitable for metric scaling. If it were not for that, its high non-linearity would also be sufficient to rule it out as an "interesting" coefficient, although it would produce the same clustering topology as the coefficients of the T_θ family (section 5.1) if subjected to an order-invariant clustering method, as was pointed out above. All the other coefficients seem suitable for clustering or ordination, except perhaps S_{15} which should be avoided in metric scaling for its lack of linearity in certain cases (OCCAS3) or its lack of resolution in other cases (OCCAS2).

5.4 Summary of the Decision-Making Process

We have seen above that various elements have to be considered in the choice of a coefficient of resemblance: its mathematical properties, its behavior when confronted with data sets, the nature of the data, the use that

will be made of the resemblance matrix, even the degree of confidence that the user attaches to the various variables. These remarks are summarized in Table 6.

In that table dissimilarities D_2 to D_4 are always presented as equivalent. In practice, further considerations may narrow down a choice among them. On one hand we have seen that a unit exponent of this metric may be preferred when values in the data set are very variable, creating ratios that pertain to different orders of magnitude, because these differences are magnified by squaring (D_2) or higher powers (D_4). On the other hand, one may prefer using D_2 (Table 3) for metric scaling. We have seen, however, that non-Euclideanarity presents little difficulty when imaginary dimensions can be ignored or a simple monotonic transformation used.

The reader who finds Table 6 enlightening could apply its principles with profit to his own set of preferred resemblance coefficients. This could help to develop further criteria facilitating the decision-making process of the choice of a coefficient.

Appendix I

Construction of an Example with $s_{12} = 0$, $t_{12} = 0$ and $s_{ij} = 1/(1+\theta)$, $t_{ij} = 1/(1+2\theta)$ for all $(i, j) \neq (1, 2)$.

Sample Number															
1	1	1	.	.	.	.	1	0	0	.	.	.	.	0
2	0	0	.	.	.	.	0	1	1	.	.	.	.	1
3															
.															
.															
.					**X**							**X**			
.															
n															

The above data matrix has n rows and 2^{n-1} columns. The first row has 2^{n-2} units and 2^{n-2} zeros; the second row is the complement of the first. The matrix **X** has $n-2$ rows and 2^{n-2} columns, the entries in the columns being all the binary numbers from zero to $2^{n-2}-1$, in any order. Thus every row of **X** has 2^{n-3} zeros and 2^{n-3} units and every pair of rows has 2^{n-4} matches of zeros and 2^{n-4} matches of units.

Thus $s_{12} = t_{12} = 0$ and $s_{1i} = s_{2i} = 1/(1+\theta)$, $t_{1i} = t_{2i} = 1/(1+2\theta)$ for $i \neq 1, 2$. Also $s_{ij} = 1/(1+\theta)$ and $t_{ij} = 1/(1+2\theta)$ for $(i, j) \neq (1, 2)$ because every pair of rows of **X** has equal numbers of (0,0),(0,1)(1,0) and (1,1) matches and mismatches. This establishes that the configuration assumed in theorems 11 and 12 is attainable.

Appendix II: Counter-examples

We give here some simple examples that demonstrate that various coefficients are not Euclidean or are not metric.

II.1 Non-metric Examples for Binary Variable Coefficients

S_7 and S_8

Unit Numbers	***1***	***2***	***3***
Variable I	−	+	+
Variable II	+	−	+

This gives $d_{12} = 1$, $d_{13} = 1/3$, $d_{23} = 1/3$ which do not satisfy the metric inequality.

S_{10}

Unit Numbers	***1***	***2***	***3***	***Frequency***
Variable I	+	−	+	u
Variable II	−	+	+	v
Variable III	−	−	−	w

This example gives the following for the numbers of the different combinations occurring for each comparison:

Combination	***a***	***b***	***c***	***d***
Comparison (1,2)	0	u	v	w
Comparison (1,3)	u	0	v	w
Comparison (2,3)	v	0	u	w

With $u = v = 1$ we have that $d_{12} = 1$, $d_{13} = 1/4$, $d_{23} = 1/4$ so that d_{ij} is not metric. With $u = 1$, $v = 2$, we have that $d_{12} = 1$, $d_{13} = 1/3$, $d_{23} = 1/6$; this shows that $\sqrt{d_{ij}}$ is not metric.

S_{11}

Using the same example as for S_{10} gives:

$$d_{12} = 1/4\ (2 + \frac{u}{u+w} + \frac{v}{v+w})$$

$$d_{13} = 1/4\ \left(\frac{v}{u+v} + \frac{v}{v+w}\right)$$
$$d_{23} = 1/4\ \left(\frac{u}{u+v} + \frac{u}{u+w}\right) \quad .$$

For any non-zero values of u, v, w this gives $d_{13} + d_{23} - d_{12} = -1/4$ so that d_{ij} is not a metric. Taking $u = 12$, $v = 1$, $w = 10$ gives $\sqrt{d_{12}} = .8118$, $\sqrt{d_{13}} = .2048$ and $\sqrt{d_{12}} = .6059$ which is (just) non-metric.

S_{12} Using the same example as for S_{10}, and setting $u = v = 1$ gives: $d_{12} = 1$, $d_{13} = d_{23} = 1 - 1/\sqrt{2}$ which do not satisfy the metric inequality.

S_{13} Using the same example as for S_{10}, and setting $u = v = 1$, $w = 2$ gives: $d_{12} = 1$, $d_{13} = d_{23} = 1 - 1/\sqrt{3}$ which do not satisfy the metric inequality.

S_{14} Using the same example as for S_{10}, and setting $u = v = 1$, $w = 2$ gives: $d_{12} = 4/3$, $d_{13} = d_{23} = 1 - 1/\sqrt{3}$ which do not satisfy the metric inequality.

S_{15} Using the same example as for S_{10}, and setting $u = v = w = 1$ (or indeed any non-zero value) gives: $d_{12} = 2$, $d_{13} = d_{23} = 0$. Hence neither d_{ij} nor $\sqrt{d_{ij}}$ are metrics.

II.2 Non-Euclidean Examples for Binary Variable Coefficients

To show that a metric is non-Euclidean we have, with only one exception, successfully constructed the example discussed at the beginning of section 2, where three points form the vertices of an equilateral triangle side l (say) and a fourth point is equidistant (say, distance m) from the other three and where $m < l\sqrt{3}$. It would be interesting to know whether or not the construction of this example is a necessary condition for a non-Euclidean coefficient. We show here this construction for the general coefficients S_θ and T_θ.

S_θ	***Unit Numbers***	***1***	***2***	***3***	***4***	***Frequency***
	Variable I	+	+	+	+	q
	Variable II	+	−	−	−	1
	Variable III	−	+	−	−	1
	Variable IV	−	−	+	−	1

This example gives the following for the numbers of the different combinations occurring for each comparison:

Combinations	*a*	*(b + c)*	*d*
Comparisons (12),(13),(23)	q	2	1
Comparisons (14),(24),(34)	q	1	2

$$\text{Thus } l = d_{12} = d_{13} = d_{23} = 2\theta/(q + 1 + 2\theta)$$
$$\text{and } m = d_{14} = d_{24} = d_{34} = \theta/(q + 2 + \theta) \ .$$

When $\theta < 1$, $2m < l$ so that construction is then not even metric let alone Euclidean. When $\theta \geqslant 1$ the construction is metric but is non-Euclidean provided $l > m\sqrt{3}$. This yields

$$q > 2\theta(\sqrt{3} + 1) - (5 + 2\sqrt{3})$$

which for any given θ is clearly satisfied by choosing q sufficiently large. Thus non-Euclidean constructions of $1 - S_\theta$ exist for all values of θ, showing that S_4, S_6 and S_8, which correspond to $S_{\theta=1}$, $S_{\theta=2}$ and $S_{\theta=1/2}$, do not generate Euclidean dissimilarities. Also because $1 - S_9 = 2(1 - S_4)$, it follows that S_4 and S_9 have the same metric and Euclidean properties, and therefore that $1 - S_9$ is also non-Euclidean.

T_θ

Using the same example as for S_θ gives:

$$l = d_{12} = d_{13} = d_{23} = 2\theta/(q + 2\theta)$$
$$\text{and } m = d_{14} = d_{24} = d_{34} = \theta/(q + \theta) \ .$$

This construction is always metric but the condition $l > m\sqrt{3}$ now yields

$$q > 2\theta(\sqrt{3} + 1)$$

which again is clearly satisfied for any θ by choosing q sufficiently large. This shows that S_3, S_5 and S_7, which correspond to $T_{\theta=1}$, $T_{\theta=2}$ and $T_{\theta=1/2}$, do not generate Euclidean dissimilarities.

$(1 - S_\theta)^{1/t}$

The example used for S_θ and T_θ gives

$$l = [2\theta/(q + 1 + 2\theta)]^{1/t}$$
$$m = [\theta/(q + 2 + \theta)]^{1/t}$$

which gives a non-Euclidean configuration when

$$\sqrt{3} < [\frac{2(q+2+\theta)}{(q+1+2\theta)}]^{1/t}$$

Indeed, for fixed θ the right-hand-side may be made arbitrarily close to $2^{1/t}$ by choosing q sufficiently large. Hence for all θ a non-Euclidean example may be constructed whenever $\sqrt{3} < 2^{1/t}$, i.e.,

$$t < \frac{\log 4}{\log 3}.$$

$(1 - T_\theta)^{1/t}$ The example for T_θ gives:

$$l = [\frac{2\theta}{q+2\theta}]^{1/t}$$
$$m = [\frac{\theta}{q+\theta}]^{1/t}$$

which as in the previous example is non-Euclidean when $t < \frac{\log 4}{\log 3}$ for all values of θ, provided q is chosen sufficiently large. Note that in these two examples the value of $t = \log 4/\log 3$ gives a lower bound on t for non-Euclidean configurations for all θ; they do not show that a higher bound does not exist, although theorem 12 shows that the bound must be less than $t = 2$.

Of the results established above, we have already seen that S_7 and S_8 are not even metric, so that to establish that they are also non-Euclidean is superfluous. The remaining non-Euclidean metrics need the special counter-example that is now described.

S_2

Unit Numbers	***1***	***2***	***3***	***4***	***Frequency***
Variable I	+	+	+	+	q
Variable II	+	+	−	+	3
Variable III	+	−	+	+	3
Variable IV	−	+	+	+	3
Variable V	−	−	−	−	1

This gives

$$l = d_{12} = d_{13} = d_{23} = 7/(q+10)$$
and $$m = d_{14} = d_{24} = d_{34} = 4/(q+10)\ .$$

Because $7 > 4\sqrt{3}$ it follows that $1 - S_2$ is not Euclidean.

The one exception where we have been unable to achieve the non-Euclidean construction used in all the above examples is to show that $\sqrt{d_{ij}}$ is non-Euclidean for S_8. This coefficient requires the example of Appendix I and the proof of theorem 12. That we have failed to find an example with the usual construction does not show that one does not exist.

II.3 Non-metric/Non-Euclidean Examples for Continuous Variable Coefficients

D_5 Consider a variable taking values $x_1 = 5$, $x_2 = 10$ and $x_3 = -9$ for three units. Then $d_{12} = 1/3$, $d_{13} = 7/2$ and $d_{23} = 19$ which is clearly non-metric. Thus D_5 is not necessarily metric or Euclidean for negative values.

D_6 The example for D_5 gives the same results for D_6 which is therefore not necessarily metric or Euclidean for negative values.

D_9 Clearly for a single variable any pair of negative values gives a negative denominator and hence a negative and non-metric coefficient.

D_{10} Consider a variable taking values $x_1 = -10$, $x_2 = -5$ and $x_3 = 10$. Then $d_{12} = -1$, $d_{13} = 2$, $d_{23} = 3/2$ in which d_{12} is not even positive.

For positive values of the variable consider the example:

Unit Numbers	***1***	***2***	***3***	***4***
Variable I	a	b	b	b
Variable II	b	a	b	b
Variable III	b	b	a	b

where $b \geqslant a$.

Then $d_{12} = d_{13} = d_{23} = (2/3)\ (1 - a/b)$
and $d_{14} = d_{24} = d_{34} = (1/3)\ (1 - a/b)$.

The values of a and b are immaterial. The example is of the usual kind and is non-Euclidean because $2 > \sqrt{3}$. Thus D_{10} is not necessarily Euclidean, even for positive values.

References

BAKER, F.B. (1974), "Stability of Two Hierarchical Grouping Techniques. Case 1: Sensitivity to Data Errors," *Journal of the American Statistical Association, 69*, 440-445.

BLASHFIELD, R.K. (1976), "Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods," *Psychological Bulletin, 83*, 377-388.

BLOOM, S.A. (1981), "Similarity Indices in Community Studies: Potential Pitfalls," *Marine Ecology Progress Series, 5*, 125-128.

CAILLIEZ, F. (1983), "The Analytical Solution to the Additive Constant Problem," *Psychometrika, 48*, 305-308.

CAILLIEZ, F., and PAGES, J.-P. (1976), *Introduction à l'analyse des données*, Paris: Société de Mathématiques appliquées et de Sciences humaines.

CHARLTON, J.R.H., and WYNN, H.P. (1985), "Metric Scaling and Infinitely Divisible Distributions: Schoenberg's Theorem," Personal Communication.

CUNNINGHAM, K.M., and OGILVIE, J.C. (1972), "Evaluation of Hierarchical Grouping Techniques: A Preliminary Study," *Computer Journal, 15*, 209-213.

ESTABROOK, G.F., and ROGERS, D.J. (1966), "A General Method of Taxonomic Description for a Computed Similarity Measure," *BioScience, 16*, 789-793.

EVERITT, B. (1974), *Cluster Analysis*, London: Heinemann Educational Books.

FAITH, D.P. (1985), "Distance Methods and the Approximation of Most-Parsimonious Trees," *Systematic Zoology, 34*, 312-325.

FISHER, L., and VAN NESS, J.W. (1971), "Admissible Clustering Procedures," *Biometrika, 58*, 91-104.

GOWER, J.C. (1971), "A General Coefficient of Similarity and Some of its Properties," *Biometrics, 27*, 857-871.

GOWER, J.C. (1982), "Euclidean Distance Geometry," *Mathematical Scientist, 7*, 1-14.

GOWER, J.C. (1984a), "Multivariate Analysis: Ordination, Multidimensional Scaling and Allied Topics," in *Handbook of Applicable Mathematics, Vol. VI: Statistics, Part B*, Ed. E. Lloyd, Chichester: John Wiley and Sons, 727-781.

GOWER, J.C. (1984b), "Distance Matrices and Their Euclidean Approximation," in *Data Analysis and Informatics, 3*, Eds. E. Diday, M. Jambu, L. Lebart, J. Pagès and R. Tomassone, Amsterdam: North-Holland, 3-21.

GOWER, J.C. (1985), "Measures of Similarity, Dissimilarity, and Distance," in *Encyclopedia of Statistical Sciences, Vol. 5*, Eds. S. Kotz, N.L. Johnson and C.B. Read, New York: John Wiley and Sons, 397-405.

HAJDU, L.J. (1981), "Graphical Comparison of Resemblance Measures in Phytosociology," *Vegetatio, 48*, 47-59.

HUBERT, L. (1974), "Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarchical Clustering Procedures," *Journal of the American Statistical Association, 69*, 698-704.

JACCARD, P. (1901), "Etude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura," *Bulletin de la Société vaudoise des Sciences Naturelles, 37*, 547-579.

JARDINE, N., and SIBSON, R. (1968), "The Construction of Hierarchic and Non-Hierarchic Classifications," *Computer Journal, 11*, 177-184.

KULCZYNSKI, S. (1928), " Die Pflanzenassoziationen der Pieninen," *Bulletin international de l'Académie polonaise des Sciences et des Lettres, Classe des Sciences mathématiques et naturelles, Série B*, Supplément II (1927), 57-203.

LEGENDRE, P., and CHODOROWSKI, A. (1977), "A Generalization of Jaccard's Association Coefficient for Q Analysis of Multi-State Ecological Data Matrices," *Ekologia Polska, 25*, 297-308.

LEGENDRE, P., DALLOT, S., and LEGENDRE, L. (1985), "Succession of Species Within a Community: Chronological Clustering, with Applications to Marine and Freshwater Zooplankton," *American Naturalist, 125*, 257-288.

LEGENDRE, L., and LEGENDRE, P. (1983a), *Numerical Ecology*, Developments in Environmental Modelling, Vol. 3, Amsterdam: Elsevier Scientific Publishing Company.

LEGENDRE, L. and LEGENDRE, P. (1983b), "Partitioning Ordered Variables into Discrete States for Discriminant Analysis of Ecological Classifications," *Canadian Journal of Zoology, 61*, 1002-1010.

LINGOES, J.C. (1971), "Some Boundary Conditions for a Monotone Analysis of Symmetric Matrices," *Psychometrika, 36*, 195-203.

MIRSKY, L. (1955), *Introduction to Linear Algebra*, Oxford: Oxford University Press.

ORLOCI, L. (1978), *Multivariate Analysis in Vegetation Research*, Second Edition, The Hague: Dr. W. Junk B.V.

RAND, W.M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association, 66*, 846-850.

RENKONEN, O. (1938), "Statistisch-ökologische Untersuchungen über die terrestische Käferwelt der finnischen Bruchmoore," *Annales Zoologici Societatis Zoologicae-Botanicae Fennicae 'Vanamo', 6*, 1-231.

SCHOENBERG, I.J. (1935), "Remarks to Maurice Fréchet's article 'Sur la définition axiomatique d'une classe d'espaces vectoriels distanciés applicables vectoriellement sur l'espace de Hilbert'," *Annals of Mathematics, 36*, 724-732.

SIBSON, R. (1971), "Some Observations on a Paper by Lance and Williams," *Computer Journal, 14*, 156-157.

SIBSON, R. (1979), "Studies in the Robustness of Multidimensional Scaling: Perturbational Analysis of Classical Scaling," *Journal of the Royal Statistical Society, Series B, 41*, 217-229.

SPATH, H. (1980), *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, translated by Ursula Bull, Chichester: Ellis Horwood Ltd., and New York: John Wiley and Sons.

WILLIAMS, W.T., CLIFFORD, H.T., and LANCE, G.N. (1971a), "Group-size Dependence: A Rationale for Choice Between Numerical Classifications," *Computer Journal, 14*, 157-162.

WILLIAMS, W.T., LANCE, G.N., DALE, M.B., and CLIFFORD, H.T. (1971b), "Controversy Concerning the Criteria for Taxonometric Strategies," *Computer Journal, 14*, 162-165.

WOLDA, H. (1981), "Similarity Indices, Sample Size and Diversity," *Oecologia (Berl.), 50*, 296-302.

ZEGERS, F.E. (1986), "Two Classes of Element-Wise Transformations Preserving the Positive Semi-Definite Nature of Coefficient Matrices," *Journal of Classification, 3*, 49-53.