A General Coefficient of Similarity and Some of Its Properties

J. C. Gower

Biometrics, Vol. 27, No. 4. (Dec., 1971), pp. 857-871.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197112%2927%3A4%3C857%3AAGCOSA%3E2.0.CO%3B2-3

Biometrics is currently published by International Biometric Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ibs.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Mon Apr 2 11:33:35 2007

BIOMETRICS 27, 857-74
December 1971

A GENERAL COEFFICIENT OF SIMILARITY AND SOME OF ITS PROPERTIES

J. C. GOWER

Rothamsted Experimental Station, Harpenden, Herts., U. K.

SUMMARY

A general coefficient measuring the similarity between two sampling units is defined. The matrix of similarities between all pairs of sample units is shown to be positive semi-definite (except possibly when there are missing values). This is important for the multi-dimensional Euclidean representation of the sample and also establishes some inequalities amongst the similarities relating three individuals. The definition is extended to cope with a hierarchy of characters.

1. INTRODUCTION

A similarity coefficient measures the resemblance between two individuals based on either or both of two logically distinct kinds of information pertaining to v variables and allowing for possible missing information.

First there is information on the existence, or not, of the variables. In taxonomy, where similarity coefficients are often used, this may be the only kind of information used to build up a taxonomic classification. The taxonomist has the problem of deciding whether a character occurring in one group of organisms also occurs in another group; this is the so-called homology problem. A missing character should not be confused with missing information because it is known that the character definitely does not exist. Missing information can occur, for example, with incomplete fossil material or with poor descriptions in the literature, from which the existence or otherwise of a character cannot be inferred.

The other type of information pertains to observed values of qualitative or quantitative properties of existing characters. An absent character cannot have any associated properties and this suggests that the two types of information might be viewed hierarchically, a topic returned to in section 4.

A common simple situation occurs when all information is of the presence/absence type (or from 2-level qualitative characters). This gives the familiar 2×2 association table shown in Table 1, where presence is denoted by $+$ and absence by $-$.

Many different coefficients have been derived from Table 1. Yule's early work on this subject was reviewed by Yates [1952]. More recently Sokal and Sneath [1963] discussed numerous association coefficients, not all of which have yet been used. We are not concerned here with recommending

TABLE 1

NUMBERS OF CHARACTERS OCCURRING IN, OR ABSENT FROM, TWO INDIVIDUALS: a (+, +) COMMON TO BOTH INDIVIDUALS; b (−, +) AND c (+, −) OCCURRING IN ONLY ONE INDIVIDUAL; AND d (−, −) ABSENT FROM BOTH

	Individual 1		
	+	−	Totals
Individual 2 +	a	b	$a + b$
−	c	d	$c + d$
Totals	$a + c$	$b + d$	v

what coefficients should be used in different circumstances but merely wish to describe a general coefficient that includes several existing ones as special cases, and can therefore be used under many different circumstances. It is particularly suitable for including in computer programs because it can cope with a variety of different data-types without any reprogramming and also because the positive semi-definite property established in section 3 is a prerequisite for certain types of statistical and numerical analyses (Gower [1966]).

This coefficient has been used since 1960 in various computer programs. To find out how it has behaved the reader is referred to the asterisked references given at the end of this paper.

2. THE DEFINITION OF SIMILARITY

2.1. *Terminology*

Dichotomous, *qualitative*, and *quantitative* variates are distinguished. The term dichotomous is reserved for characters that are either present or absent and whose absence in both of a pair of individuals is not taken as a match; when both levels of a two-level qualitative variate are to be treated on a par, the levels will be termed *alternatives*. A discussion of some of the considerations governing the choice of scoring the two levels of a response as dichotomous or as alternatives is deferred until section 4. Qualitative characters may have many levels (e.g. black, green, yellow, blue) but unlike the levels of quantitative characters they do not form an ordered set, although for convenience in computing, coded numerical values may be given.

2.2. *The calculation of similarity*

Two individuals i and j may be compared on a character k and assigned a score s_{ijk}, zero when i and j are considered different and a positive fraction, or unity, when they have some degree of agreement or similarity. There are many ways of calculating s_{ijk}, some of which are described below. Sometimes no comparison is possible because information is missing, or in the case of dichotomous variables a character is non-existent in both i and j. The possibility of making comparisons can be represented by a quantity δ_{ijk}, equal to 1 when character k can be compared for i and j, and 0 otherwise. When $\delta_{ijk} = 0$, s_{ijk} is unknown but is conventionally set to zero. The similarity between i and j is defined as the average score taken over all possible comparisons:

$$S_{ij} = \sum_{k=1}^{v} s_{ijk} \Big/ \sum_{k=1}^{v} \delta_{ijk} . \tag{1}$$

When $\delta_{ijk} = 0$ for all characters, S_{ij} is undefined. When all comparisons are possible $\sum_{k=1}^{v} \delta_{ijk} = v$, the total number of characters; otherwise it is the number of characters over which the comparison is made. An alternative, but exactly equivalent, form to (1) is

$$S_{ij} = \sum_{k=1}^{v} s_{ijk}\, \delta_{ijk} \Big/ \sum_{k=1}^{v} \delta_{ijk} . \tag{2}$$

This is in the form of a weighted average but it will not be interpreted in that fashion until weighted similarity coefficients are discussed in section 4; at present δ_{ijk} indicates only when comparisons are possible. The scores s_{ijk} are assigned as follows:

(a) *For dichotomous characters* the presence of the character is denoted by + and its absence by −. When there are no unknown values of character k, four different combinations of its values may occur for two individuals and the score and validity assigned to each combination is given in Table 2.

(b) *For qualitative characters* we set $s_{ijk} = 1$ if the two individuals i and j agree in the kth character and $s_{ijk} = 0$ if they differ.

(c) *For quantitative characters* with values $x_1, x_2, \cdots, x_n$ of character k for the total sample of n individuals we set $s_{ijk} = 1 - |x_i - x_j|/R_k$. Here R_k is the range of character k and may be the total range in the population or the range in the sample.

When $x_i = x_j$ then $s_{ijk} = 1$, and when x_i and x_j are at opposite ends of their range, s_{ijk} is a minimum (0 when R_k is determined from the sample). With intermediate values, s_{ijk} is a positive fraction.

TABLE 2

SCORES AND VALIDITY OF DICHOTOMOUS CHARACTER COMPARISONS

	Values of character k			
Individual i	+	+	−	−
j	+	−	+	−
s_{ijk}	1	0	0	0
δ_{ijk}	1	1	1	0

Thus S_{ij} defined by (1) ranges between 0 and 1; a value of 1 means that the two individuals differ in no character whereas 0 means they differ maximally in all their characters.

2.3. *Relationship with other similarity coefficients*

If all characters are dichotomous we have the similarity coefficient used, for example, by Sneath [1957] where the comparison of negative characters between two individuals is not considered a valid match. This coefficient is denoted by S_J in Sokal and Sneath [1963] and in terms of Table 1 is defined by $S_J = a/(a + b + c)$. The treatment for qualitative variates has also been proposed by Silvestri *et al.* [1962]. When all characters are qualitative with two levels (i.e. alternatives) we have the simple matching coefficient denoted by S_{SM} in Sokal and Sneath [1963], defined by $S_{SM} = (a + d)/v$.

The treatment in 2.2(c) for quantitative characters resembles the mean character difference of Cain and Harrison [1958], which is, however, a distance rather than a similarity. We have normalised the units of measurement of each quantitative variate by dividing by the range and not the standard deviation, because the range is easier to calculate and the standard deviation has little meaning for the heterogeneous populations where similarity coefficients are usually employed (see also the Appendix).

3. POSITIVE SEMI-DEFINITE PROPERTY OF THE SIMILARITY MATRIX

With n individuals, the $n \times n$ matrix S can be formed whose element S_{ij} is the similarity, as described in section 2, between individuals i and j. We often require to represent the n individuals of a sample as a set of points in Euclidean space. Gower [1966] has discussed this problem and shown that a convenient representation can be obtained by taking the distance between the ith and jth individuals as proportional to $(1 - S_{ij})^{\frac{1}{2}}$. The coordinates of points with these distances are the elements of the latent vectors of S scaled so that their sums of squares equal the latent roots. Thus to get a real Euclidean representation with distances $(1 - S_{ij})^{\frac{1}{2}}$ it is sufficient for S to be positive semi-definite (p.s.d.). It is shown in the Appendix that when there are no missing values S is p.s.d. The law relating the lengths of the three sides of a triangle must therefore hold and we have

$$(1 - S_{ij})^{\frac{1}{2}} + (1 - S_{ik})^{\frac{1}{2}} \geq (1 - S_{jk})^{\frac{1}{2}}. \tag{3}$$

By Theorem 2 of the Appendix, the matrix, with elements S_{ij}^r, where r is any positive integer, is p.s.d. because S is p.s.d. Consequently points can be found in Euclidean space with distances $(1 - S_{ij}^r)^{\frac{1}{2}}$ and the triangle law becomes

$$(1 - S_{ij}^r)^{\frac{1}{2}} + (1 - S_{ik}^r)^{\frac{1}{2}} \geq (1 - S_{jk}^r)^{\frac{1}{2}}. \tag{4}$$

The results (3) and (4) are true for the general similarity defined in section 2 and hence also for any of the more restricted, commonly used definitions contained in the general definition. Because correlation matrices are p.s.d. these results are also true for correlation coefficients.

4. WEIGHTING AND HIERARCHIC CHARACTERS

The decision to weight or not to weight character scores has become a controversial problem for taxonomists; in general those in favour of using numerical methods prefer not to weight but the traditional taxonomist holds that taxonomic classifications have always been constructed by recognising that certain characters are more important than others. At least part of the difficulty seems to arise from the fact that, with a new set of organisms completely unrelated to any known group, no *a priori* weighting would be acceptable, but once this set has been classified it becomes clear that certain characters are better than others for identification. In any subsequent reclassification, or when classifying related groups, these characters might be regarded as more important and assigned greater weights than the others.

There is no problem in incorporating weights in the similarity coefficient of equation (1) or its equivalent, (2). How to decide on a rational set of weights is more difficult. The most simple weighting gives a constant weight w_k (say) to each character and, if all comparisons are possible, could be represented by (2) with $\delta_{ijk} = w_k$. It is convenient, however, to distinguish δ_{ijk} from more direct weighting and write, corresponding to (1),

$$S_{ij} = \sum_{k=1}^{v} s_{ijk}w_k \Big/ \sum_{k=1}^{v} \delta_{ijk}w_k . \tag{5}$$

Arguments similar to those given in the Appendix show that equation (5) defines a p.s.d. similarity matrix provided there are no missing values and $w_k \geq 0$.

Alternatively weights may be regarded as a function of the *result* of the values of a character being compared. Thus differences in a character may be considered more important than agreement, or agreement between rare character states might be given more weight than agreement between common states. The similarity coefficient then takes the form

$$S_{ij} = \sum_{k=1}^{v} s_{ijk}w_k(x_{ik}, x_{jk}) \Big/ \sum_{k=1}^{v} \delta_{ijk}w_k(x_{ik}, x_{jk}), \tag{6}$$

where $w_k(x_{ik}, x_{jk})$ indicates that the weight for character k is a function of the character values x_{ik} and x_{jk} for individuals i and j, and that the functional form is allowed to differ from character to character. Burnaby [1970] suggested calculating $w_k(x_{ik}, x_{jk})$ from the Shannon information in the sample values of the kth character. For 0/1 data this is a function of p_k, the proportion of 1's. Goodall [1966] proposed a probabilistic similarity coefficient based on the p_k's. Gower [1970] discussed various points to be considered before basing weights on the observed values of p_k.

In equation (6) the indicator δ_{ijk} is redundant as it can be completely absorbed in $w_k(x_{ik}, x_{jk})$ by defining $w_k(x_{ik}, x_{jk}) = 0$ when either or both of x_{ik}, x_{jk} are missing, or if character k is dichotomous and both of x_{ik}, x_{jk} are negative.

Similarity matrices derived from (6) need not be p.s.d. as can be seen by considering three individuals a, b, and c each with two alternative qualitative characters taking the values $a(-, -)$, $b(-, +)$, $c(+, -)$. Define $w_k(-, -) = 3$ and $w_k(-, +) = w_k(+, -) = 1$. This gives a similarity matrix

$$\begin{array}{c} a \\ b \\ c \end{array}\begin{bmatrix} 1 & \frac{3}{4} & \frac{3}{4} \\ \frac{3}{4} & 1 & 0 \\ \frac{3}{4} & 0 & 1 \end{bmatrix}$$

with determinant $-\frac{1}{8}$.

4.1. *Hierarchic systems of characters*

In the introduction it was pointed out that when a character exists, its quantitative and/or qualitative properties can be observed. We may also observe the existence, or not, of subsidiary characters and their properties, and so on. The situation is similar to recording multi-phase information in sample surveys. Kendrick and Proctor [1964] discussed the case of primary and secondary characters, requiring that similarity coefficients should be designed so that secondary character results should never be allowed to outweigh agreements between primary characters. They demonstrated that this is not a property of existing coefficients (nor is it a property of similarity defined by equation (1)) by considering the following example:

Individual X	$+$	$x_i(i = 1, 2, \cdots, m)$	$q_i(i = 1, 2, \cdots, n)$
Individual Y	$+$	$y_i(i = 1, 2, \cdots, m)$	$p_i(i = 1, 2, \cdots, n)$
Individual Z	$-$		$p_i(i = 1, 2, \cdots, n)$
Individual W	$-$		$q_i(i = 1, 2, \cdots, n)$

The individual W was not part of the original example but has been introduced here for further discussion. Here $+$ represents a primary character present in X and Y but absent in Z and W. This primary character has m secondary character values x_i, y_i observed for X and Y; there are no secondary character values for Z and W because they lack the primary character and therefore missing values are assumed here. In addition, there are n other characters with values q_i for X and W and values p_i for Y and Z. All values are assumed alternative levels of two-level qualitative variables. The sets (x_i) and (y_i) are supposed to have s matches out of the m comparisons, and the sets (p_i) and (q_i) have k matches out of the n comparisons. We have the following similarities from equation (1):

$$\begin{aligned} S_{XY} &= (1 + s + k)/(1 + m + n), \\ S_{WY} = S_{XZ} &= k/(1 + n), \\ S_{WZ} &= (1 + k)/(1 + n). \end{aligned} \tag{7}$$

Here S_{WZ}, agreeing on the primary character of W and Z, is always greater than S_{WY} which differ, but S_{XY} is not greater than S_{XZ} when $(1 + s)/m \leq k/(n + 1)$. To avoid this difficulty Kendrick and Proctor suggested setting $w_k = m + 1$ in equation (5), that is weighting each primary character by one more than the number of its associated secondary characters. This gives

$$\begin{aligned} S'_{XY} &= (m + 1 + s + k)/(2m + 1 + n), \\ S'_{WY} = S'_{XZ} &= k/(m + 1 + n), \\ S'_{WZ} &= (m + 1 + k)/(m + 1 + n). \end{aligned} \tag{8}$$

Again $S'_{WZ} > S'_{WY}$ but now we also have $S'_{XY} > S'_{XZ}$ so that comparisons amongst the secondary characters can never reverse the results of matches amongst the primary.

The most unsatisfactory thing about this form of weighting is that the value of m is somewhat arbitrary; by a sufficiently diligent search we might be able to increase m to any desired value. An alternative scheme is to give each primary character unit weight but adjust its score by the similarity among its associated characters. In the above example this gives

$$\begin{aligned} S''_{XY} &= (s/m + k)/(n + 1), \\ S''_{WY} = S''_{XZ} &= k/(n + 1), \\ S''_{WZ} &= (1 + k)/(n + 1). \end{aligned} \tag{9}$$

This method of weighting also ensures that $S''_{XY} > S''_{XZ}$ and also $S''_{WZ} > S''_{WY}$; it is simpler and has certain advantages in programming. The general form for S_{ij} can now be written

$$S_{ij} = \sum_{k=1}^{v} s_{ijk} S_{ij}^{(k)} \Big/ \sum_{k=1}^{v} \delta_{ijk} . \tag{10}$$

Summation is over the v primary characters, which can be of any type (dichotomous, qualitative, quantitative) and the score for each primary character k is multiplied by the similarity $S_{ij}^{(k)}$ between its associated secondary characters. If $S_{ij}^{(k)} = 0/0$ we conventially assign $S_{ij}^{(k)} = 1$. On a computer the subroutine for calculating S_{ij} can also be used to calculate $S_{ij}^{(k)}$. Clearly when secondary characters themselves have subsidiary characters, or even whole hierarchies of characters, the subroutine for formula (10) requires recursive programming. Williams [1969], discussing Kendrick and Procter's ideas, has suggested a form of weighting similar to (10) where the secondary character agreements modify those of the primary characters which get unit weight.

Another property seen from the similarities in (9) is that $S''_{WZ} \geq S''_{XY}$. With the weighting given in (8) $S'_{WZ} \lessgtr S'_{XY}$ as $(m + 1 + k)/(m + 1 + n) \lessgtr s/m$. It seems perverse that, all other things being equal, matches between non-existent primary characters should give higher similarities than matches between existing characters. Without secondary characters and with the primary character treated as qualitative, both similarities will be equal.

With observed secondary character values associated with the positive primary characters, more information is available and just as we want S_{XY} to exceed S_{XZ} it seems natural to want S_{XY} to exceed S_{WZ}. That is, it would be preferred to have $S_{XY} \geq S_{WZ} \geq S_{XZ} = S_{WY}$ so that, all other things being equal, a positive match amongst primary characters gives greater similarity than a negative match, which itself is greater than a primary character mismatch. This cannot be achieved with any of these coefficients but is approximated with formula (10) when primary characters are treated as dichotomous. This would leave the results in (9) unchanged except for S''_{WZ} which becomes k/n. We now have $S''_{WZ} \gtrless S''_{XZ}$ as $k/n \gtrless s/m$ which, although an improvement on (8), is not perfect.

Yet another possibility, which exactly fulfills the requirements, is to use equation (6) setting $w_k(x_{ik}, x_{jk}) = 1 + S_{ij}^{(k)}$ for the kth primary character and defining when $S_{ij}^{(k)} = 0/0$ to be zero. Also treat all primary characters as dichotomous. This gives for W, X, Y, and Z the following values:

$$\begin{aligned} S'''_{XY} &= [1 + (s/m) + k]/[1 + (s/m) + n], \\ S'''_{WY} = S'''_{XZ} &= k/(1 + n), \qquad\qquad (11) \\ S'''_{WZ} &= k/n. \end{aligned}$$

Thus this last definition seems to fulfill best some of the intuitive ideas of how to deal with primary and secondary characters and also partially justifies coding primary characters as dichotomous, so excluding negative matches. In fact, if all the characters had been treated as dichotomous it would not have affected these results but merely changed the interpretation of m and n to refer to the number of valid comparisons between the sets (x), (y) and (p), (q), respectively; s and k refer to positive matches only.

The whole question of whether it is reasonable to disregard negative matches is still unresolved. Taxonomists usually regard the classifications they build as approximations to some ideal genetic classification. Sometimes genes repress the formation of a character so that absence signifies the existence of a gene. Sometimes the levels of a qualitative character are clearly of equal status, as when they are black or white, and it would be difficult to justify disregarding either of the matches. At other times the levels might be black and not-black, and it is more reasonable to regard a not-black match as containing little useful information. Clearly the negative match question has no unique answer and each situation must be judged separately. The merit of the coefficients discussed here is that they give the option of treating each two-level character either as dichotomous or as having two levels of equal status.

Taxonomists have objected to the idea of primary and secondary characters on the grounds that it is not easy to say what characters are primary and what are secondary; also because they regard, on genetic grounds, all characters to be equally useful *a priori* for classification purposes. The dictionary definition of taxonomy is that it is 'the science or technique of classification'; there is no restriction to biological classification. When dealing with inanimate objects, genetic arguments are invalid and it might then, if not with biological material, be valid to consider hierarchies of characters.

5. DISCUSSION

The similarity coefficient described in section 2 has been used in programmes for hierarchial cluster analysis since 1960 and more recently in principal co-ordinate analysis and other ordination programmes. It has been found sufficiently flexible to cope with nearly all forms of character coding so far encountered, and unlike many coefficients currently in use does not require any recoding for multistate or quantitative characters.

The p.s.d. property is important on two counts. First, just as with a correlation matrix, it allows numerical methods which operate only on p.s.d. matrices to be used with confidence, provided there are no missing values. Second, it aids interpretation of those methods of cluster and ordination analysis which are based on Euclidean metrics. Many of these methods will also operate on non-Euclidean metrics, but interpretation of the results is often difficult.

The coefficient for hierarchies of characters, discussed in section 4, shares these advantages but I am more hesitant in recommending it because I myself have never used, or felt the need for, such a coefficient. However, there is current interest in this type of data and the coefficient described here is thought to be better than those previously described. McNeil [1971] has recently used the coefficient and reported his experiences.

ACKNOWLEDGMENTS

This paper was revised (after lying dormant for four years) and some new material included, while the author was at the University of North Carolina at Chapel Hill. The work was partially supported by the National Institutes of Health, Institute of General Medical Sciences, Grant No. 12868-07.

UN COEFFICIENT GENERAL DE SIMILARITE ET QUELQUES UNES DE SES PROPRIETES

RESUME

L'auteur définit un coefficient général pour mesurer la similarité entre deux unités d'échantillonnage; il montre que la matrice des similarités entre toutes les paires d'unités d'échantillonnage est semi-définie positive (sauf, éventuellement, quand il y a des données manquantes). Ceci est important pour représenter l'échantillon dans un espace euclidien multidimensionnel et aussi pour établir quelques inégalités entre les similarités reliant trois individus. L'auteur étend la définition pour couvrir le cas de caractères hiérarchisés.

REFERENCES

Burnaby, T. P. [1970]. On a method for character weighting a similarity coefficient, employing the concept of information. *Math. Geol. 2*, 25–38.

Cain, A. J. and Harrison, G. A. [1958]. An analysis of the taxonomists judgment of affinity. *Proc. Zool. Soc. Lond. 131*, 85–98.

*Eddy, B. P. and Carpenter, K. P. [1964]. Further studies on Aeromonas. II. Taxonomy of Aeromonas and C27 strains. *J. Appl. Bact. 27*, 96–109.

Ferrar, W. L. [1941]. *Algebra: A Text-book of Determinants, Matrices and Algebraic Forms.* Oxford University Press.

Goodall, D. W. [1966]. A new similarity index based on probability. *Biometrics 22*, 882–907.

Gower, J. C. [1966]. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53*, 315–28.

Gower, J. C. [1970]. A note on Burnaby's character weighted similarity coefficient. *Math. Geol. 2*, 39–45.

Kendrick, W. B. and Proctor, J. R. [1964]. Computer taxonomy in the fungi imperfecti. *Canad. J. Bot. 42*, 65–88.

McNeil, J. [1971]. The hierarchical ordering of characters as a solution to the dependent character problem in numerical taxonomy. *Taxon* (to be submitted).

Mirsky, L. [1955]. *Introduction to Linear Algebra.* Oxford University Press.

*Rayner, J. H. [1965]. Multivariate analysis of montmorillonite. *Clay Minerals 6*, 59–70.

*Rayner, J. H. [1966]. Classification of soils by numerical methods. *J. Soil Sci. 17*, 79–92.

*Sheals, J. G. [1964]. The application of computer techniques to Acarine taxonomy: a preliminary examination with species of the *Hypoaspis-Androlaelaps* comples (Acarina). *Proc. Linn. Soc. Lond. 176*, 11–21.

Silvestri, L., Turri, M., Hill, L. R., and Gilardi, E. [1962]. A quantitative approach to the systematics of actinomycetes based on overall similarity. Symposia of the Society for General Microbiology, XII, Microbial Classification.

*Smith, I. W. [1963]. The classification of 'bacterium salmonicida.' *J. Gen. Microbiol. 33*, 263–74.

Sneath, P. H. [1957]. The application of computers to taxonomy. *J. Gen. Microbiol. 17*, 201–26.

Sokal, R. and Sneath, P. H. [1963]. *The Principles of Numerical Taxonomy.* W. H. Freeman, San Francisco and London.

Williams, W. T. [1969]. The problem of attribute weighting in numerical classification. *Taxon 18*, 369–74.

Yates, F. [1952]. George Udny Yule: 1871–1951. *Obit. Not. Roy. Soc. 8*, 309–23.

APPENDIX

Properties of positive semi-definite matrices

In this Appendix it is shown that the matrix S defined in section 2 is p.s.d. A few properties of p.s.d. matrices are required. Except for Theorem 2, all the results given below are well known, but they are listed here for ease of reference. (Because the proofs of Theorems 1 and 4 are very short, they are included for completeness.)

Definition 1. An $n \times n$ real matrix $\mathbf{A}$ is p.s.d. when for every real $n \times 1$ vector $\mathbf{x}$, $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$.

Theorem 1. If both $\mathbf{A}$ and $\mathbf{B}$ are p.s.d. then so is $\mathbf{A} + \mathbf{B}$. This follows immediately from the equation $\mathbf{x}' (\mathbf{A} + \mathbf{B}) \mathbf{x} = \mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{x}'\mathbf{B}\mathbf{x} \geq 0$. Consequently any sum of p.s.d. matrices is p.s.d.

Definition 2. If $\mathbf{A}$ and $\mathbf{B}$ are matrices with elements a_{ij}, b_{ij} $(i, j = 1, 2, \cdots, n)$, a matrix with elements $a_{ij} \times b_{ij}$ may be defined. This type of matrix product

* These are not directly referred to in the text but use the similarity coefficient described in section 2.

will be written **A*B**; the result is an $n \times n$ matrix not to be confused with the Kronecker product.

Theorem 2. If **A** and **B** are p.s.d. and symmetric then so is $\mathbf{C} = \mathbf{A}^*\mathbf{B}$.

To prove this theorem, let **B** have latent roots and column vectors $\lambda_1, \lambda_2, \cdots, \lambda_n; \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$. Because **B** is symmetric, we may write

$$\mathbf{B} = \lambda_1\mathbf{v}_1\mathbf{v}_1' + \lambda_2\mathbf{v}_2\mathbf{v}_2' + \cdots + \lambda_n\mathbf{v}_n\mathbf{v}_n',$$

where each vector is normalised to have unit sum of squares. We have $\lambda_i = \mathbf{v}_i'\mathbf{B}\mathbf{v}_i \geq 0$ because **B** is p.s.d. Equating the elements of the rth row and sth column on both sides of the previous equation gives,

$$b_{rs} = \sum_{i=1}^{n} \lambda_i v_{ir} v_{is}$$

where v_{ir} is the rth element of $\mathbf{v}_i$. Also

$$\begin{aligned}\mathbf{x}'\mathbf{C}\mathbf{x} &= \sum_{r,s}^{n} a_{rs} b_{rs} x_r x_s \\ &= \sum_{r,s}^{n} a_{rs}\left(\sum_{i=1}^{n} \lambda_i v_{ir} v_{is}\right) x_r x_s \\ &= \sum_{i=1}^{n} \lambda_i \left[\sum_{r,s}^{n} a_{rs}(v_{ir}x_r)(v_{is}x_s)\right]. \qquad \text{(A1)}\end{aligned}$$

The expression in square brackets cannot be negative because **A** is p.s.d.; we shall write it as ρ_i^2 so that (A1) becomes

$$\mathbf{x}'\mathbf{C}\mathbf{x} = \sum_{i=1}^{n} \lambda_i \rho_i^2 .$$

This cannot be negative because no λ_i is negative, proving that **C** is p.s.d. For a more general statement and proof of this theorem see Mirsky ([1955] p. 421).

Theorem 3. A set of necessary and sufficient conditions for a $n \times n$ symmetric matrix **A** to be p.s.d. is that all the principal leading minors $\Delta_{pp}(p=1, 2, \cdots, n)$ of **A** must be non-negative. For a proof of this result see e.g. Ferrar ([1941] p. 138).

Theorem 4. All sums of squares and products (SSP) matrices $\mathbf{X}'\mathbf{X}$ are p.s.d. We have

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})\mathbf{x} = (\mathbf{X}\mathbf{x})'(\mathbf{X}\mathbf{x}) = \sum_{i=1}^{n} u_i^2 \geq 0,$$

where $(u_1, u_2, \cdots, u_n)$ is the row vector $(\mathbf{X}\mathbf{x})'$.

Proof that the similarity matrix is p.s.d

We now prove that various special cases of the similarity matrix defined in section 2 are p.s.d. and then combine these results to show that the general matrix is p.s.d., assuming no missing values.

In the association data of Table 1, + and − may be given arbitrary numerical scores. Writing Table 1 in the form below and considering various SSP matrices derived from it, we see that two elementary matrices are p.s.d.

Individual i	Individual j	Frequency
+	+	a_{ij}
+	−	b_{ij}
−	+	c_{ij}
−	−	d_{ij}
Total		v

(i) Scoring + as 1 and − as 0 and forming the $n \times n$ SSP matrix of scores for the n individuals, gives a p.s.d. matrix with general element a_{ij} .

(ii) Scoring + as 0 and − as 1 and forming the $n \times n$ SSP matrix gives a p.s.d. matrix with general element d_{ij} .

In the following we shall drop the suffices i, j and refer to the matrices $\mathbf{a}_0$, $\mathbf{d}_0$ etc. with general elements a, d etc. The ith diagonal term a_{ii} of $\mathbf{a}_0$ gives the number of + responses for the ith character, a number not greater than v. Therefore the matrix diag $(v - a_{ii})$ is p.s.d. and when added to $\mathbf{a}_0$ shows that the matrix $\mathbf{a}$ with constant diagonal term v and off-diagonal elements a_{ij} is also p.s.d. Similarly $\mathbf{d}$ with constant diagonal term v and off-diagonal elements d_{ij} is p.s.d.

The matrix of simple matching coefficients

From section 2.3, $S_{SM} = (a + d)/v$ and from (i) and (ii) above, $\mathbf{a}_0$ and $\mathbf{d}_0$ are p.s.d. with $a_{ii} + d_{ii} = v$. Hence $\mathbf{S}_{SM} = (\mathbf{a}_0 + \mathbf{d}_0)/v$ and by Theorem 1 is p.s.d.

The matrix formed from dichotomous variates

Provided $d < v$, $S_J = a/(v - d) = (a/v)(1 + d/v + d^2/v^2 + d^3/v^3 + \cdots)$. Writing $\mathbf{A} = \mathbf{a}/v$ and $\mathbf{D} = \mathbf{d}/v$ and forming quadratic forms in $\mathbf{x}$ we have

$$\mathbf{x}'\mathbf{S}_J\mathbf{x} = \mathbf{x}'[\mathbf{A} + \mathbf{A}^*(\mathbf{D} + \mathbf{D}^*\mathbf{D} + \mathbf{D}^*\mathbf{D}^*\mathbf{D} \cdots)]\mathbf{x}.$$

Since $\mathbf{A}$ and $\mathbf{D}$ are p.s.d., repeated application of Theorems 1 and 2 shows that so is every term on the right hand side of the series expansion of $\mathbf{S}_J$. That $\mathbf{x}'\mathbf{S}_J\mathbf{x}$ is the limit of the right hand side is elementary, and as every term on the right hand side is non-negative, so is $\mathbf{x}'\mathbf{S}_J\mathbf{x}$. This proves that $\mathbf{S}_J$ is p.s.d.

The matrix formed from qualitative variates

If there are A matches and B mis-matches amongst the v variates ($v = A + B$) then the similarity $S_Q = A/(A + B) = A/v$. To show that $\mathbf{A}$ is p.s.d., note that a qualitative variate with q levels could be scored as q different dichotomous variates by setting the qth variate + when the qth level is attained and — otherwise. There would then be A positive matches and by (i) above, $\mathbf{A}$ is p.s.d. The ith diagonal element of $\mathbf{A}$ is the number of + responses for the ith individual, and thus must be v because each qualitative variate is at some level. Therefore $\mathbf{S}_Q$ is p.s.d.

The matrix formed from quantitative variates

Suppose x_1 , x_2 , $\cdots$, x_n are the values taken by a single quantitative variate for each of a set of n objects. Then

$$S_{ij} = 1 - |x_i - x_j|/R,$$

where $R \geq \max |x_u - x_v|$; i.e., it is not less than the sample range. We shall first prove the theorem for $R = \max |x_u - x_v|$.

No generality is lost by assuming that $R = 1$, i.e., that the x_i are measured on a new scale in which R of the original units are one of the new units. We require to prove $S_{ij} = 1 - |x_i - x_j|$ is p.s.d. We can further assume that $1 = x_1 \geq x_2 \geq , \cdots , x_{n-1} \geq x_n = 0$ since S_{ij} can always be transformed into this form by permuting the rows of the data matrix and shifting the origin so that $x_n = 0$; thus

$$\mathbf{S}_N = \begin{bmatrix} 1 & 1-(x_1-x_2) & 1-(x_1-x_3) & \cdots & 1-(x_1-x_n) \\ 1-(x_1-x_2) & 1 & 1-(x_2-x_3) & \cdots & 1-(x_2-x_n) \\ 1-(x_1-x_3) & 1-(x_2-x_3) & 1 & \cdots & 1-(x_2-x_n) \\ \vdots & \vdots & \vdots & & \vdots \\ 1-(x_1-x_n) & 1-(x_2-x_n) & 1-(x_3-x_n) & \cdots & 1 \end{bmatrix}. \tag{A2}$$

To prove that $\mathbf{S}_N$ is p.s.d. requires the determinants of its principal leading minors. A series of elementary transformations gives Δ_{pp} , the principal $p \times p$ leading minor, as

$$\Delta_{pp} = 2^{p-1}[1 - \tfrac{1}{2}(x_1 - x_p)] \prod_{i=1}^{p-1} (x_i - x_{i+1}). \tag{A3}$$

Since $x_1 - x_p \leq x_1 - x_n = 1$ and $x_i - x_{i+1} \geq 0$ we have $\Delta_{pp} \geq 0$ and therefore, by Theorem 3, $\mathbf{S}_N$ is p.s.d.

Now suppose $\mathbf{S}'_N$ is defined by $S'_{ij} = 1 - |x_i - x_j|/T$, where $T > \max |x_u - x_v|$. Then the above algebraic manipulation follows through, leading again to (A3). Now $x_1 - x_p \leq x_1 - x_n < 1$ and $x_i - x_{i+1} \geq 0$ so that again $\Delta_{pp} \geq 0$ and so $\mathbf{S}'_N$ is p.s.d.

The above proof is for just one variate but if we have k quantitative variates, $\mathbf{S}_N$ becomes the average of k p.s.d. matrices of the above type and by Theorem 1 remains p.s.d.

Equation (A3) is still true when $R < \max |x_u - x_v|$, but $1 - \frac{1}{2}(x_1 - x_p)$ is no longer necessarily positive. It need not be positive, for example, if R_k were taken to be the standard error of variate k. In this case Δ_{nn} is likely to be negative because the range $(x_1 - x_n)$ is almost certainly greater than two standard deviations. Cain and Harrison's [1958] mean character difference normalises by dividing by the maximum value x_n ; when standard errors are used the argument above suggests that the resulting coefficient is not suitable for the type of work briefly mentioned in section 3.

The general similarity matrix

The above has proved that the similarity matrix is p.s.d. when the variates are all qualitative, all quantitative, or all dichotomous. It is now shown that this remains true for any combination of these types.

The similarity matrix derived from a combination of quantitative and qualitative variates is merely a weighted mean of matrices of type S_N and S_Q and by Theorem 1 is also p.s.d. Suppose that $(1/V)u_{ij}$ is the general element of such a p.s.d. matrix based on V variates and let $t_{ij} = a_{ij}/(v - d_{ij})$ be the general term of a similarity matrix based on v dichotomous variates. The similarity matrix obtained by combining these two matrices has elements

$$S_{ij} = \frac{a_{ij} + u_{ij}}{v - d_{ij} + V} = \frac{(a_{ij} + u_{ij})}{(v + V)}\left(1 + \frac{d_{ij}}{v + V} + \frac{d_{ij}^2}{v + V} + \cdots\right).$$

Now the matrices with general terms given by a_{ij} , u_{ij} , and d_{ij} are all p.s.d., and it follows from repeated applications of Theorems 1 and 2 that the general similarity matrix must be p.s.d.

The effect of missing values

Missing values may cause the similarity matrix to loose its p.s.d. property as can be seen by considering the similarity matrix for three individuals derived from the following table.

Variate number	1	2	3	4
Individual 1	−	−	+	+
Individual 2	+	+	+	*
Individual 3	+	+	+	+

In this table * denotes a missing value and +/− may represent either presence/absence of dichotomous variates or alternative values of qualitative variates. In either case

$$\mathrm{S} = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{2} \\ \frac{1}{3} & 1 & 1 \\ \frac{1}{2} & 1 & 1 \end{bmatrix}.$$

The determinant is $-\frac{1}{36}$ and S is therefore not p.s.d. Note that if the * is replaced by + we have

$$\mathrm{S} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & 1 \\ \frac{1}{2} & 1 & 1 \end{bmatrix},$$

and if * is replaced by $-$ we have

$$\mathrm{S} = \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & 1 & \frac{3}{4} \\ \frac{1}{2} & \frac{3}{4} & 1 \end{bmatrix},$$

both of which are p.s.d.

Received March 1967, *Revised May* 1971