Chapter 3.

Similarity functions*

Serge Joly†

Georges Le Calvé†

3.1. Introduction

The basic tool, in Statistics like in many branches of experimental sciences concerned with the study of information expressed in observations, is comparison analysis: in the field of statistical modelling, comparison to a theoretical model, in exploratory data analysis (EDA), comparison between data.

In EDA, these comparisons between data fall into two broad categories: analysis of similarities, when we measure how similar two objects look, analysis of dissimilarities, when we measure how different they are. These approaches are not contradictory, though each statistical technique is usually more specifically related to one or the other. For instance Principal Component Analysis (PCA) is related to the analysis of similarities (by means of covariances), Hierarchical cluster analysis (HCA) and additive trees to the analysis of dissimilarities (by means of distances) ; however, in both cases, we can associate with the index commonly used an index of the other category, in a natural way: with the covariance we can associate the Euclidean distance, while by taking the opposite of the ultrametric distance and adding a well chosen positive constant , we get an index of similarity.

Both examples show that these associations are basically worked out by means of decreasing functions ; it also appears that different models of functions are available: quadratic function (PCA), linear function (HCA).

We could of course use a linear link for the Euclidean geometry, or a quadratic link for HCA. Quite obviously these new indices would not be well adapted to the corresponding representation, and they would provide little if any lisibility.

This article is devoted to a number of functions that link indices of similarity and indices of dissimilarity. We call these links "similarity functions" (SF). We shall not make an analytical study of SF. We shall concern ourselves with their properties with respect to methods of representation.

* *In* Van Cutsem, B. (Ed.), (1994) *Classification and Dissimilarity Analysis*, Lecture Notes in Statistics, Springer-Verlag, New York.

† Université de Haute Bretagne, Rennes, France.

3.2. Definitions. Examples

3.2.1. Definitions

Let I be a finite set, $I = \{1, 2, ..., n\}$, and D and S be two $n \times n$-matrices.

Definition 3.2.1.

A $n \times n$-matrix D is a *dissimilarity on* I if and only if D is symmetric, with a null diagonal, and all the other terms are non-negative.

A $n \times n$-matrix D is a *semi-distance* if D is a dissimilarity and if the triangle inequality holds:

$$\forall (i, j, k) \in I^3,\ D_{ij} \leq D_{jk} + D_{kj}.$$

A $n \times n$-matrix D is *definite* (in the French literature "*propre*") if and only if

$$\forall (i, j) \in I^2,\ D_{ij} = 0 \iff i = j.$$

A $n \times n$-matrix D is *semi-definite* if and only if

$$D_{ij} = 0 \implies \left(\forall k \in I,\ D_{ik} = D_{jk}\right).$$

A semi-distance which is definite will be called a *distance*. □

Definition 3.2.2. A $n \times n$-matrix S is a *similarity on* I if and only if S is symmetric and

$$\forall (i, j) \in I^2,\ S_{ii} \geq S_{ij}$$

A similarity S is said to be *proper* if

$$\forall (i, j) \in I^2,\ S_{ii} > S_{ij}$$

A similarity S is said to be *normed* if

$$\forall i \in I,\ S_{ii} = 1.$$ □

Theorem 3.2.1. *Given f a decreasing real function with $f(0) = 1$, and D a dissimilarity, let us define S such that*

$$\forall (i, j) \in I^2,\ S_{ij} = f\left(D_{ij}\right).$$

Then S is a normed similarity. If f is strictly decreasing and D is definite, then S is proper.

Conversely, let g be a decreasing real function with $g(1) = 0$ and let S be a normed similarity. Then $D = g(S)$ is a dissimilarity.

If S is proper and g is strictly decreasing, then D is definite.

The proof is obvious. We will emphasize two points:

1 It is natural to consider using $g = f^{-1}$ and hence to take invertible functions. In that case any decreasing function is strictly decreasing.

2 Let g be a strictly decreasing function. In order for $g(S)$ to be a dissimilarity, S should have a constant diagonal. Then there is no point in looking for a dissimilarity $g(S)$ when S is not normed (to a multiplicative factor), and we restrict our choice of g's to invertible functions.

From now on, unless explicitly mentioned otherwise, we restrict ourselves to normed similarities and invertible functions.

3.2.2. Examples

We review here the main examples of application of Theorem 3.2.1.

3.2.2.1. Linear function

An example is $S = 1 - D$. Hence

$$\forall (i,j) \in I^2,\ s_{ij} = 1 - d_{ij} \qquad \text{and} \qquad d_{ij} = 1 - s_{ij}.$$

This is the most frequently used SF for qualitative "presence-absence" variables. (The reader will find a list of similarity indices for categorical variables in Appendix).

That kind of SF is well fitted to hierarchical representations (classifications, pyramids).

3.2.2.2. Homographic function

An example is

$$S = \frac{2}{1+D} - 1 \quad \Longleftrightarrow \quad S = \frac{1-D}{1+D}$$

$$D = \frac{2}{1+S} - 1 \quad \Longleftrightarrow \quad D = \frac{1-S}{1+S}$$

This SF is of interest mainly because of its analytical properties.

Many indices used for presence-absence variables are derived homographically from others. For example, Jaccard's distance is thus associated to the Sokal-Sneath-Anderberg similarity, and the Czenakowski-Dice distance to Jaccard's similarity. (See Appendix for definitions)

3.2.2.3. Quadratic function

An example is

$$S = 1 - \frac{1}{2} D^2 \qquad \Longleftrightarrow \qquad D = \sqrt{2\,(1 - S)}$$

This formula is well fitted to the Euclidean representation, and specially to variables which are representable on a sphere centered at the origin O.

In that case s_{ij} is equal to the inner product $\overrightarrow{Oi} \cdot \overrightarrow{Oj}$ (the multiplicative coefficient $\frac{1}{2}$ for D^2 is necessary in order to represent s_{ij} as the inner product $\overrightarrow{Oi} \cdot \overrightarrow{Oj}$).

3.2.2.4. Exponential function

An example is

$$\forall (i,j) \in I^2,\ s_{ij} = e^{-d_{ij}} \qquad \Longleftrightarrow \qquad d_{ij} = -\ln s_{ij}$$

or, more generally

$$s_{ij} = e^{-d_{ij}^p}$$

This kind of SF is seldom used in data analysis. It is well adapted to multiplicatively transformed variables (economic growth rate, for instance).

The exponential $e^{-d_{ij}^p}$ is well fitted to representation in L^p-spaces.

Given $p = 2$ and the Euclidean distance

$$D^2 = \sum_i (x_i - y_i)^2$$

we once again observe the strong link between the Euclidean geometry and the normal distribution.

3.2.2.5. Circular function

An example is

$$\forall (i,j) \in I^2,\ s_{ij} = \cos d_{ij} \qquad \Longleftrightarrow \qquad d_{ij} = \operatorname{Arccos} s_{ij}$$

In this case, d_{ij} is set as an angle.

Angular distances are well adapted to such notions as "apparent distances" in astronomy, and spherical representations.

3.2.2.6. Graphical representations

Let I be a set with n elements, and D a dissimilarity on I. Let E be a metric space, and Δ a dissimilarity on E.

We shall say that the n points $M_1, M_2, \ldots, M_n$ are a representation of (I, D) into (E, Δ) if and only if

$$\forall (i,j) \in I^2,\ \Delta(M_i, M_j) = D_{ij}$$

The analogous definition also holds for similarities.

The representation depends on Δ, which has to be specified.

- The representation most commonly used is the euclidean representation: set $E = \mathbb{R}^n$ and Δ the distance associated to the inner product norm.
- If $E = \mathbb{R}^n$, another choice for Δ could be the L_1-distance.

Some dissimilarities will permit a euclidean representation, others a L_1-representation, and some will permit neither of them.

In fact the set of possible representations is wide. We will make a survey of some cases with the objective of pointing out the type of SF underlying each model.

Let $M_1, M_2, \ldots, M_n$ be n points lying on a Euclidean sphere (representation often used for categorical variables). In that case E is the sphere (with radius one) of $\mathbb{R}^n$. There are numerous eligible choices for the dissimilarity Δ, because the relative position of any two points can be described in many ways. We could use for instance the Euclidean distance (length of the chord M_iM_j) or the geodetical distance (length of the shortest arc M_iM_j).

In fact any quantity tending to zero as M_i tends to M_j can be choosen as a dissimilarity between M_i and M_j. Any quantity increasing as M_i tends to M_j can be choosen as a similarity.

Phrased differently, the dissimilarity between two points M_i and M_j on the euclidean sphere can be expressed in a variety of mathematical notions, some of them we shall now review. Using θ, the central angle $\widehat{M_iM_j}$, we can define

$d_1 = 1 - \cos\theta$

$d_2 = \theta$

$d_3 =$ arc θ's sagitta

$d_4 =$ the length of the chord M_iM_j

$d_5 = \sin\theta$

$d_6 = \tan\frac{\theta}{2}$

$s_1 = cos\theta$

$s_2 = \frac{\pi}{2} - \theta$

$s_3 =$ arc θ's apothem

Here d_1 to d_6 may be viewed as measures of dissimilarities while s_1 to s_3 can be considered as measures of similarities. Furthermore, there is a linear link between s_1 and d_1, s_2 and d_2, s_3 and d_3, as between s_1^2 and d_5^2. Between s_1 and d_4, there is a quadratic link, an homographic one between s_1^2 and d_6^2.

It is worth remarking that, though it is of common use, with categorical variables, given a similarity S and a dissimilarity D, to produce a Euclidean representation on a sphere, this is absolutely not justified. On the one hand, it has been proved that most indices do not permit a Euclidean representation (Fichet, Le Calvé 1984). On the other hand, this representation makes use of s_1 and d_4, linked by a quadratic function, though a linear function links the given indices D and S.

It would be much more appropriate to use the couples (s_1, d_1), or (s_2, d_2), or else (s_3, d_3), more especially as by using them, most of the indices for categorical data can be exactly represented (Beninel, 1987).

3.3. The $W^M(D^p)$ forms

3.3.1. Definitions and properties

We will now concern ourselves with a very important family of SF, which we will call *W-forms*. To define a W-form, we first have to choose a point M of I that will act as origin for the form. Then a W-form will be a $n \times n$-matrix whose terms are linear combinations of D_{Mi}, D_{Mj}, D_{ij} hence the name "*W-form*".

Definition 3.3.1. Let M be a point belonging to I and D a dissimilarity on I. We call *W-form of D evaluated at point M*, denoted $W^M(D)$, the $n \times n$-matrix whose elements are

$$\forall (i,j) \in I^2,\ W^M(D)_{ij} = \frac{1}{2}\left(D_{Mi} + D_{Mj} - D_{ij}\right).$$

For p belonging to $\mathbb{R}_+$ we will also consider the following form, called "*W-form of D^p evaluated at point M*"

$$\forall (i,j) \in I^2,\ W^M(D^p)_{ij} = \frac{1}{2}\left(D^p_{Mi} + D^p_{Mj} - D^p_{ij}\right). \qquad \square$$

It is a known property that if D is a dissimilarity, then, for any positive p, D^p is also a dissimilarity.

Some remarks concerning the W-forms can be of interest.

- The multiplicative coefficient -1 for D^p_{ij} is not sufficient in order for $W^M(D^p)$ to be a decreasing function in D, because of the positive terms D^p_{Mi} and D^p_{Mj}. Thus Theorem 3.2.1 cannot be applied to W^p-forms.

- It is easy to show that usually $W^M(D^p)$ is not a similarity, because of the condition $W_{ii} \geq W_{ij}$ failing to be true (this condition is equivalent to $D^p_{Mi} \geq D^p_{Mj} - D^p_{ij}$).

- On the other hand, whenever D^p is a distance, then, for any M, $W^M(D^p)$ is an index of similarity.

- However, when D_{Mi} is a constant with respect to i, $W^M(D^p)$, as a function of D, is decreasing. It happens, for instance, for many indices defined on "presence-absence" variables, where there exists a point O (the null variable) such that $D_{Oi} = 1$ and $0 \leq D_{ij} \leq 1$. Then, for any p the $W^O(D^p)$ are indices of similarity. In that peculiar but important case, $W^O(D^p)$ can be rewrited as

$$\forall (i,j) \in I^2,\ W^O\left(D^p\right)_{ij} = 1 - \frac{1}{2} D^p_{ij}.$$

This property is no longer true when we consider the $W^M(D^p)$-form at any point M different from O: the $W^M(D^p)$-forms generally are not similarities.

From the definition of $W^M(D^p)$ we derive

$$\begin{aligned} \forall (i,j) \in I^2,\ W^M\left(D^p\right)_{ii} &= D^p_{Mi} \\ D^p_{ij} &= W^M\left(D^p\right)_{ii} + W^M\left(D^p\right)_{jj} - 2W^M\left(D^p\right)_{ij} \end{aligned}$$

Finally, M and N being two points in I, we can explicit the relation between the W^M-form and the W^N-form:

$$\begin{aligned} \forall (i,j) \in I^2,\ W^M\left(D^p\right)_{ij} &= W^N\left(D^p\right)_{ij} + X_i + X_j \\ \text{with} \qquad X_i &= \frac{1}{2}\left(D^p_{Mi} - D^p_{Ni}\right) \end{aligned}$$

- A similarity S and a dissimilarity D are said to be "W-associated" if and only if

$$\forall (i,j) \in I^2,\ D_{ij} = S_{ii} + S_{jj} - 2\,S_{ij}$$

This is not a one-to-one relation: a dissimilarity D can be W-associated to several similarities. Let S and T be two such similarities, W-associated to D. Then S and T are related by:

$$\begin{aligned} \forall (i,j) \in I^2,\ S_{ij} &= T_{ij} + Z_i + Z_j \\ \text{with} \qquad Z_i &= \frac{1}{2}\left(S_{ii} - T_{ii}\right) \end{aligned}$$

Finally we shall note that $W^M(D)$ is a $n \times n$-matrix with a null-line and a null-column (because $\forall i \in I$, $W^M(D^p)_{Mi} = 0$). We will frequently use the restriction of $W^M(D)$, i.e. the $(n-1) \times (n-1)$-matrix obtained by cutting both null-line and null-column.

Both following cases $p = 1$ and $p = 2$ are of great importance:

- **p = 1**

$$\forall (i,j) \in I^2,\ W^M(D)_{ij} = \frac{1}{2}\left(D_{Mi} + D_{Mj} - D_{ij}\right)$$

$W^M(D_{ij})$ is a kind of measures how the three points M, i, and j deviate from the straight line.

- **p = 2**

$$\forall (i,j) \in I^2,\ W^M\left(D^2\right)_{ij} = \frac{1}{2}\left(D^2_{Mi} + D^2_{Mj} - D^2_{ij}\right)$$

$W^M(D^2_{ij})$ derives from the Cosine Law: in a triangle ABC, we have

$$BC^2 = AC^2 + AB^2 - 2\,AB\,AC\,\cos A$$

which leads to $\overrightarrow{AB} \cdot \overrightarrow{AC} = \frac{1}{2}(AB^2 + AC^2 - BC^2)$ and $W^M(D^2)_{ij}$ can be viewed, if D is Euclidean, as the inner product $\overrightarrow{Mi} \cdot \overrightarrow{Mj}$.

3.3.2. The $W^M(D^2)$ form

We investigate now the properties of the $W^M(D^2)$-forms.

The following theorem is generally known as Frechet's theorem (1935). With some alterations in the expressing, it can also be fastened to Gauss, Minkowsky, or Schoënberg.

Theorem 3.3.1. *A dissimilarity matrix D can be considered as a distance matrix between n points of a Euclidean space if and only if there exists a point M such that the matrix $W^M(D^2)$ is non-negative definite (NND).*

The dimension of the representative space is equal to the rank of the matrix. If there exists an M such that $W^M(D^2)$ is NND, then $W^M(D^2)$ is NND for any M.

Proof. Consider the spectral form of $W^M(D^2)$:

$$\forall (i,j) \in I^2,\ W^M\left(D^2\right)_{ij} = \sum_{k=1}^{r} \lambda_k\, X_i^k\, X_j^k.$$

Then

$$W^M\left(D^2\right)_{ii} = \sum_{k=1}^{r} \lambda_k \left(X_i^k\right)^2$$

and, as

$$\left(D^2\right)_{ij} = W^M\left(D^2\right)_{ii} + W^M\left(D^2\right)_{jj} - 2\,W^M\left(D^2\right)_{ij},$$

it follows that

$$\forall (i,j) \in I^2,\ (D^2)_{ij} = \sum_{k=1}^{r} \lambda_k \left(X_i^k\right)^2 + \sum_{k=1}^{r} \lambda_k \left(X_j^k\right)^2 - 2 \sum_{k=1}^{r} \lambda_k X_i^k X_j^k$$
$$= \sum_{k=1}^{r} \lambda_k \left(X_i^k - X_j^k\right)^2$$

and this is the square of a Euclidean distance if and only if $\lambda_k \geq 0$.

Conversely, assume there exists a representation such that

$$\forall (i,j) \in I^2,\ (D^2)_{ij} = \sum_{k=1}^{r} \mu_k \left(Y_i^k - Y_j^k\right)^2 \qquad \text{with} \quad \mu_k \geq 0.$$

Then, as

$$(D^2)_{Oi} = \sum_{k=1}^{r} \mu_k \left(Y_i^k\right)^2,$$

it follows that

$$W^O (D^2)_{ij} = \frac{1}{2} \sum_{k=1}^{r} \mu_k \left[\left(Y_i^k\right)^2 + \left(Y_i j^k\right)^2 - \left(Y_i^k - Y_j^k\right)^2 \right]$$
$$= \sum_{k=1}^{r} \mu_k Y_i^k Y_j^k$$

and since the μ_k are positive, this is a NND matrix.

If there exists a Euclidean representation, for any M the matrix of the inner products $\overrightarrow{Mi} \cdot \overrightarrow{Mj}$ is NND, and according to the Cosine Law this matrix is none other than $W^M(D^2)$. ■

Torgerson form

We can consider the value of W in any point, provided we know its relative distances to all others. To overrule the arbitrary choice of point M, Torgerson (1958) suggested taking the value at the average point, i.e. the gravity centre G. We then need to compute the values of D^2_{Gi}.

Using Koenig's theorem, and defining

$$\forall i \in I,\ D^2_{i\bullet} = \frac{1}{n} \sum_j D^2_{ij} \qquad \text{and} \qquad D^2_{\bullet\bullet} = \frac{1}{n} \sum_j D^2_{i\bullet}$$

we easily get

$$D^2_{Gi} = D^2_{i\bullet} - \frac{1}{2} D^2_{\bullet\bullet}$$

so that

$$\forall (i,j) \in I^2,\ W^G\left(D^2\right)_{ij} = \frac{1}{2}\left(D^2_{i\bullet} + D^2_{j\bullet} - D^2_{ii} - D^2_{\bullet\bullet}\right).$$

Now, in a Euclidean space, the centre of gravity belongs to that same space, so that $W^G(D^2)$ is NND if and only if $W^M(D^2)$ is NND, and we can apply the previous theorem to $W^G(D^2)$.

But we could prefer other choices. For example, instead of G, we could choose the average $W^\bullet(D^2)$ of the $W^M(D^2)$. That would mean defining a point H such that

$$\forall i \in I, D^2_{Hi} = Di_\bullet{}^2 \qquad \text{and} \qquad W^\bullet\left(D^2\right) = W^H\left(D^2\right).$$

There is no difficulty in proving that W^M, W^G, and W^H are simultaneously NND.

If D is not Euclidean, it is worth remarking that in that case G would be the very wrong choice:

On one hand, the positivity of D^2_{Gi} is no longer secured. Since D^2_{Gi} is defined as

$$D^2_{Gi} = D^2_{i\bullet} - \frac{1}{2}D^2_{\bullet\bullet}.$$

It is evident that this quantity can be negative.

But the main point leading to the rejection of G in that case is that the distances were calculated by means of Koenig's theorem ... which stands only if D is a Euclidean distance, and thus $W^G(D^2)$ has no meaning!

These remarks do not apply to $W^M(D^2)$ and $W^\bullet(D^2)$.

If D is Euclidean, the main interest of G comes from the fact that when G is the origin, the factorial plane corresponds to the maximum of the inertia. When the origin is an arbitrary M, the factorial plane corresponds to the maximum of the moment of order 2 about M.

When considering subsets of I in the analysis, the W^G-form is not easy to use: adding one point to the data leads to an $(n+1) \times (n+1)$-matrix, whose elements all have to be recalculated, since the mean point G has changed.

On the other hand, the new matrix W^M is obtained by adding one row and one column to the former W^M: the W^M-form on a subset of I is a submatrix of the W^M-form on I.

This last property does not hold for the Torgerson-form, which is thus unfitted for mathematical induction on n.

3.3.3. Transformations of D

What can be done when D is not a Euclidean distance or when it is not even a distance? A first approach could consist in looking for approximative representations: this is the field of *multidimensional scaling* (in its broadest sense). We will select another way of approach, consisting in transforming D so that we get the wanted property: this is the field of changes of metrics.

Among the possible transformations, our choice will be in favour of those which least modify the informations upon D. Thus we will select monotonous functions because they are order-preserving.

Both following methods of transformations are often used:

- the additive constant: by adding a positive constant to every term of the matrix D (or to D^2) with the exception of the diagonal we obtain a distance or a euclidean distance.
- the D^α functions (generally with $0 \leq \alpha \leq 1$).

The case of the additive constant has already been largely considered. We shall only note that the constant often happens to be very large with respect to the values of d_{ij}, so that the distortion is important.

We will now consider the power functions D^α, and first define how to choose a in order for D^α to be a distance. Then we shall consider getting a euclidean distance.

Theorem 3.3.2.

a) *The set of all α's such that D^α is a distance is a closed set.*

b) *If D is a semi-definite dissimilarity, let us put*

$$k = \sup_{i,j} d_{ij} \qquad q = \inf_{i,j}\{d_{ij} : d_{ij} \neq 0\} \qquad \alpha = \frac{\ln 2}{\ln k - \ln q}$$

then α belongs to the set referred to in a)

c) *If D is not semi-definite, D^α is a distance if and only if $\alpha = 0$.*

Proof.

a) We need only prove that if D is a distance, D^α is a distance, for $\alpha \leq 1$. The triangle inequality holding for D also holds for D^α due to the inequality

$$\forall\,(a,b) \in \mathbb{R}_+,\ \forall \alpha \in [0,1]\,,\ a^\alpha + b^\alpha \geq (a+b)^\alpha\,.$$

b) This property follows from the fact that the triangle inequality holds for any dissimilarity Δ such that

$$\forall\,(i,j) \in I^2,\ i \neq j, \qquad \frac{1}{2} \leq \delta_{ij} \leq 1.$$

Let us put $k = \sup_{i,j} d_{ij}$. Let Δ be defined by $\delta_{ij} = \frac{1}{k}\, d_{ij}$. Then

$$\forall\,(i,j) \in I^2,\ 0 \leq \delta_{ij} \leq 1.$$

- If D is definite, then

$$\inf_{i,j} d_{ij} > 0.$$

Let us put

$$\alpha = \frac{\ln 2}{\ln(\sup d_{ij}) - \ln(\inf d_{ij})}$$

hence

$$\left(\frac{1}{k}\inf_{i,j} d_{ij}\right)^{\alpha} = \frac{1}{2}.$$

So Δ^{α}, with values between $\frac{1}{2}$ and 1, is a distance, and the same is true for D^{α}.

This value of α is not the supremum of all α's. The supremum is given by the formula

$$\alpha' = \sup_{\alpha} \{\inf_{i,j,k} (d_{ij}^{\alpha} + d_{jk}^{\alpha} - d_{ik}^{\alpha}) \geq 0\}.$$

- If D is semi-definite, let us define $I^0 = \{i : \exists j \text{ such that } d_{ij} = 0\}$ and $I^+ = I - I^0$. Then D restricted to I^+ is definite, and the precedent theorem holds. Furthermore, on I^0 D_{ij} is constant (and so is D_{ij}^{α}, and the proof of b) is achieved.

c) If D is not semi-proper, there exists i, j, k with $d_{ij} = 0$ and $d_{ik} \neq d_{jk}$. The triangle inequality applied to D^{α} holds only if $d_{ik}^{\alpha} = d_{jk}^{\alpha}$ and thus $\alpha = 0$. ∎

D^{α} and the Euclidean distances

It will be equivalent to prove that D is a Euclidean distance or that $W^M((f(D)^2)$ is NND. To that purpose the following property can be of great use:

Theorem 3.3.3. (generalized Schur's lemma)

Assume f to be a real function, such that

(C1) *The expansion of f into a serie about $t = 0$ exists, with radius of convergence r,*

$$f(t) = \sum_k c_k t^k, \qquad \text{with } c_k \geq 0, \text{ continuous in } t \text{ at the point } t = 1.$$

Assume $A = [a_{ij}]$ to be a real matrix such that A is symmetric, NND, and $a_{ij} \leq r$. Let us define $B = [b_{ij}] = [f(a_{ij})]$ and $A^{*n} = [a_{ij}^n]$. Then

1) the matrix B is NND,

2) the matrix B is not positive definite (PD) iff there exists an X such that,

$$\forall n \in \mathbb{N},\ X^t A^{*n} X = 0 \quad \text{(condition } \mathbf{C2}\text{)}.$$

Proof. (Joly, Le Calvé (1986)) Since condition (C1) refers to f and condition (C2) to A, it follows that

- for any f and any g both satisfying (C1)

$$\text{“}f(A) \text{ is PD”} \iff \text{“}g(A) \text{ is PD”},$$

- If $\lim_{n \longrightarrow +\infty} A^{*n} = I$, then (C2) holds for A.

- If S is a similarity matrix with positive terms, then (C2) holds for S. ∎

Corollary 3.3.1. *Assume S to be a NND similarity matrix. Then the matrices defined by their general term as follows are NND:*

$$\frac{1}{1-s_{ij}} \qquad 1-\sqrt{1-s_{ij}} \qquad \forall \alpha \in [0,1],\ 1-\left(1-s_{ij}^{\alpha}\right)$$

$$\forall \alpha \geq 1,\ \frac{1}{1-s_{ij}^{\alpha}} \qquad e^{s_{ij}} \qquad \ln \frac{1+s_{ij}}{1-s_{ij}} \qquad \operatorname{Arcsin} s_{ij}$$

For practical applications, it will be useful to remember that if D is a Euclidean distance, then Arccos D and D^{α} for $0 \leq a \leq 1$ are Euclidean too. This last characteristic of D^{α} implies that the set of α numbers such that D^{α} is Euclidean is a non empty set, since D^0 is Euclidean, and is a closed interval. Hence there exists a power of D that is Euclidean. Looking for the supremum of all these α's ($\alpha \leq 1$) seems an interesting method, competing with the additive constant technique. We don't know the infimum of this supremum, at least not for any general n.

Definition 3.3.2. A dissimilarity is said to be "*quasi-hypermetric*" if its square root is a Euclidean distance.

Corollary 3.3.2. *The following dissimilarities are quasi-hypermetric with full rank: Jaccard, Sokal-Sneath-Anderberg, Czenakowski-Dice, Rogers-Tanimoto, Russel-Rao, Ochiaï.*

The proof, and the definitions of these indices (for categorical "presence-absence" variables), can be found in Fichet, Le Calvé (1984) or in Gower, Legendre (1986). (See also Appendix).

3.4. The $W^M(D)$ form

3.4.1. Geometrical interpretations and properties

In Section 3.3, the definition of the matrix $W^M(D)$ was introduced by the relation

$$\forall (i,j) \in I^2,\ i \neq j,\ W^M(D)_{ij} = \frac{1}{2}\left(D_{Mi} + D_{Mj} - D_{ij}\right).$$

Let us define the *metric segment* $[AB]$ as

$$[AB]_{\text{met}} = \{M : D_{AM} + D_{MB} = D_{AB}\}.$$

In an affine space, we define a *vector segment* by

$$[AB]_{\text{vec}} = \{M : \overrightarrow{AM} = k\,\overrightarrow{MB},\ 0 \leq k \leq 1\}.$$

In a Euclidean space, both definitions are equivalent. But in a normed space, they differ if the norm on the vector space is not the one inducing the metric. For example,

if, on $\mathbb{R}^2$, we define the L^1-norm, the metric segment $[AB]$ consists of all points within the rectangle with vertices A and B and whose sides are parallel to the axes.

If M belongs to the metric segment $[ij]_{\text{met}}$, $W^M(D)_{ij}$ is null and, in the general case, $W^M(D)_{ij}$ can thus measure how M deviates from the metric segment $[ij]$.

For another interpretation, let us suppose that D is a distance on a space E and consider the set A , the intersection of the metric segments $[Mi]$ and $[Mj]$:

$$A = [Mi] \cap [Mj]$$

The set A is non-empty since M belongs to A. Then,

$$\forall N \in A, \qquad W^M(D)_{ij} = D_{MN} + W^N(D)_{ij}\,.$$

It follows that

$$\forall N \in A, \qquad D_{MN} \leq W^M(D)_{ij}$$

Then $W^M(D)_{ij}$ can be considered as the length of the greatest metric segment included in both $[M_i]$ and $[M_j]$, and thus defines a kind of "metric similarity" between these two metric segments. Though $W^M(D)_{ij}$ is not an inner product, it sometimes plays a very analogous part.

Lastly, assume $X, Y, \ldots$ to be categorical variables. They can be viewed as characteristic functions (indicators) of some sets $X, Y, \ldots$. If we consider the Hamming metric $D_{X,Y} = |X \Delta Y|$, then

$$W^{\emptyset}(D)_{X,Y} = |X \cap Y|.$$

Theorem 3.4.1 lists some properties of $W^M(D)$.

Theorem 3.4.1. *Let D be a dissimilarity. Then*

- *D is a distance iff*
$$\forall M \in I, \forall (i,j) \in I^2,\ W^M(D)_{ij} \geq 0$$
- *if D is a chain, then for every triple i, j, k one and only one of the three quantities $W^i(D)_{jk}$, $W^j(D)_{ik}$, $W^k(D)_{ij}$ is null.*
- *D is a Hamming metric if and only if $\forall M \in I$, $W^M(D)_{ij}$ is an integer, and if $W^{\emptyset}(D)_{ij} = |X_i \cap X_j|$*
- *D is quasi-hypermetric if and only if $W^M(D)$ is NND*
- *If D is an ultrametric, then $W^i_{jk} = W^j_{jk} = \frac{1}{2} D_{ik}$*
- *If D is an additive tree metric, then for any arbitrary triple (i, j, k), there exists an M such that $W^M_{ij} = W^M_{jk} = W^M_{ik} = 0$.*

3.4.2. About metric projection

We defined earlier the notion of "metric segment". In the same way we can define the metric projection of a point onto a subset.

Definition 3.4.1. Given a subset A, we shall call *metric projection of i onto A* a point i^* such that

$$D_{ii^*} = \inf_{k,\ell \in A} W^i(D)_{kl} = \inf_{k,\ell \in A} \{\frac{1}{2}(D_{ik} + D_{i\ell} - D_{k\ell})\} \qquad \square$$

If there exists a point j such that $D_{ij} = D_{ii^*}$, j can be considered as i^*.

If there exists such a point j, and if D is a proper dissimilarity, then j is unique.

If there exists no j such that $D_{ij} = D_{ii^*}$, by adding to I the point i^* we define I^* by $I^* = I \cup \{i^*\}$, and by lining D with one column and one row, we get D^*.

The metric projection of i onto the total set I is called "*the foot of i*".

On I^* (I completed with all the feet), we define δ_{ij} "*distance*" between the feet, such that

$$\forall (i,j) \in I^2,\ D_{ij} = D_{ii^*} + \delta_{ij} + D_{jj^*}$$

Then we can establish the following theorem.

Theorem 3.4.2. *With the above notations,*

- *δ_{ij} is a distance.*
- *the binary relation defined on I^2 by: "$iRj \Leftrightarrow \delta_{ij} = 0$" is an equivalence relation.*
- $(iRj) \iff (\forall k \in I, \forall x \in I,\quad W^k_{ij} = W^k_{ix} = W^k_{jx})$.

We shall remark that δ is a distance, even when D fails to be one. Furthermore, the equivalence classes are all subsets on which $W^M(D)$ has a constant value (whatever D may be).

It should be noted that D is a star distance if and only if all points have the same foot. On an additive tree, the foot of a point is the node under which the point hangs. We deduce that D is "*representable*" by an additive tree if and only if, for any M, there exists a partition such that

$$\text{for any } i \text{ and } j \text{ belonging to distinct classes } W^M(D)_{ij} = D_{MM^*}.$$

(Le Calvé (1988)).

3.4.3. $W^M(D)$ and "$\mathcal{M}^1$-type" distance

Definition 3.4.2. A dissimilarity D is said to be an "$\mathcal{M}^1$-*type*" distance† iff the following equality holds:

$$D_{ij} = \sum_{k=1}^{r} |X_i^k - X_j^k|. \qquad \square$$

Such a distance is called "*city-block*" distance. It plays an important part in data analysis. Since rotating the axes causes the distances to change (rotations are not isometry), the axes have an intrinsic importance, which can be of interest in many problems. Furthermore the "$\mathcal{M}^1$-type" distance is the widest class of distances allowing easy to read representations, and permitting thus the best approximation (see Critchley, this volume). The absolute value makes the calculations difficult ; it explains why there are so few available results. In particular, we know of no result similar to:

"D is Euclidean if and only if $W^M(D^2)$ is NND"

We will now establish such a result.

Definition 3.4.3. A symmetric $n \times n$-matrix M is said to be *realisable* if and only if it can be written

$$\forall (i,j) \in I^2, M_{ij} = \sum_k a_k X_i^k X_j^k \qquad \text{with} \quad a_k \geq 0, \quad \text{and} \quad X_i^k \in \{0,1\}. \qquad \square$$

It may be noticed that this definition is very similar to the definition of a NND matrix. It was first used by Kelly (1972), with a simplificated form, owing to the fact that he was working with integers, and needed thus only to consider $a_k = 1$.

Theorem 3.4.3. *The dissimilarity D is of "$\mathcal{M}^1$-type" if and only if there exists an M such that $W^M(D)$ is realisable. If there exists an M such that $W^M(D)$ is realisable, then it is realisable for any M.*

Proof. Let us assume that $W^M(D)$ is realisable:

$$W^M(D)_{ij} = \sum_k a_k X_i^k X_j^k.$$

From the identity

$$D_{ij} = W^M(D)_{ii} + W^M(D)_{jj} - 2W^M(D)_{ij},$$

we deduce, as in the Euclidean case,

$$\begin{aligned} D_{ij} &= \sum_k a_k \left[\{X_i^k\}^2 + \{X_j^k\}^2 - 2 X_i^k X_j^k \right] \\ &= \sum_k a_k \left(X_i^k - X_j^k \right)^2 \\ &= \sum_k a_k |X_i^k - X_j^k| \end{aligned}$$

† We use the $\mathcal{M}^1$ notation, for Minkowski spaces, and L_1 for normed spaces.

Conversely, if

$$D_{ij} = \sum_k a_k |X_i^k - X_j^k| = \sum_k a_k \left(X_i^k - X_j^k\right)^2,$$

it follows from the definition of $W^M(D)_{ij}$ that

$$W^M(D)_{ij} = \sum_k a_k X_i^k X_j^k.$$

■

The following theorem strengthens the parallelism between NND matrices and realisable matrices.

Theorem 3.4.4. *Let f be a real function, whose expansion into a serie has positive coefficients and a convergence radius r, continuous at the point r. Then, if $A = [a_{ij}]$ is realisable, so is $B = [f(a_{ij}]$.*

Proofs of the Theorem 3.4.4 and of the below Corollary 3.4.1 can be found in Joly, Le Calvé (1992). The demonstration is analogous to that of Theorem 3.3.3.

Corollary 3.4.1. *The following indices of dissimilarity, defined for categorical variables, are "city-block" distances: Jaccard, Sokal-Sneath-Anderberg, Czenakowski-Dice, Rogers-Tanimoto, Russel-Rao, Ochiaï.*

This result completes the result of Corollary 3.3.2 and is a very strong incitation to use $\mathcal{M}^1$-type representation for categorical variables.

Appendix: Some indices of dissimilarity for categorical variables

Let I be a set of n individuals and J a set of p attributes. The $n \times p$-matrix X is a zero-one matrix defined by

$$X_{ik} = \begin{cases} 1 & \text{if the individual } i \text{ possesses the attribute } k, \\ 0 & \text{if the individual } i \text{ does not posses the attribute } k. \end{cases}$$

We define

n_{ij} to be the number of attributes common to i and j

$$n_{ij} = \sum_{k=1}^{p} X_{ik} X_{jk}$$

$n_{\bar{i}\bar{j}}$ to be the number of attributes missing both for i and j

$$n_{\bar{i}\bar{j}} = \sum_{k=1}^{p} (1 - X_{ik})(1 - X_{jk})$$

q_{ij} to be the number of disagreements between i and j

$$q_{ij} = \sum_{k=1}^{p} |X_{ik} - X_{jk}|$$

n_i to be the number of attributes the individual i possesses

$$n_i = \sum_{k=1}^{p} X_{ik}$$

We now list some of the most frequently used indices of similarity defined on categorical "presence-absence" variables.

RAO $S_1(i,j) = \dfrac{n_{ij}}{n}$

KULCYNSKI $S_2(i,j) = \dfrac{n_{ij}}{q_{ij}}$

JACCARD $S_3(i,j) = \dfrac{n_{ij}}{n_{ij} + q_{ij}}$

CZEKANOWSKI - DICE $S_4(i,j) = \dfrac{n_{ij}}{n_{ij} + \frac{1}{2} q_{ij}}$

ANDERBERG $S_5(i,j) = \dfrac{n_{ij}}{n_{ij} + 2\, q_{ij}}$

ROGERS - TANIMOTO $S_6(i,j) = (n_{ij} + n_{\bar{i}\bar{j}})(n_{ij} + 2\, q_{ij} + n_{\bar{i}\bar{j}})$

SOKAL - SNEATH $S_7(i,j) = (n_{ij} + n_{\bar{i}\bar{j}})(n_{ij} + \frac{1}{2}\, q_{ij} + n_{\bar{i}\bar{j}})$

Simple Matching $S_8(i,j) = \dfrac{n_{ij} + n_{\bar{i}\bar{j}}}{n}$

HAMMAN $S_9(i,j) = n_{ij} - q_{ij} + n_{\bar{i}\bar{j}}$

KULCYNSKI $S_{10}(i,j) = \frac{1}{2} \dfrac{n_{ij}}{n_i + n_{ij}\, n_j}$

ANDERBERG $S_{11}(i,j) = \frac{1}{4} \left(\dfrac{n_{ij}}{n_i} + \dfrac{n_{ij}}{n_j} + \dfrac{n_{\bar{i}\bar{j}}}{n_{\bar{i}}} + \dfrac{n_{\bar{i}\bar{j}}}{n_{\bar{j}}}\right)$

OCHIAI $S_{12}(i,j) = \dfrac{n_{ij}}{\sqrt{n_i\, n_j}}$

OCHIAI $S_{13}(i,j) = \dfrac{n_{ij}}{\sqrt{n_i\, n_j}} \dfrac{n_{\bar{i}\bar{j}}}{\sqrt{n_{\bar{i}}\, n_{\bar{j}}}}$

YULE $S_{14}(i,j) = \dfrac{n_{ij}\, n_{\bar{i}\bar{j}} - (n_i - n_{ij})\,(n_j - n_{ij})}{n_{ij}\, n_{\bar{i}\bar{j}} + (n_i - n_{ij})\,(n_j - n_{ij})}$

References

Al Ayoubi, B. (1991), Analyse des données de type M^1, Thèse, Université de Haute Bretagne, Rennes, France.

Beninel, F. (1987), Problèmes de représentations sphériques des tableaux de dissimilarité, Thèse de 3ème cycle, Université de Haute Bretagne, Rennes, France.

Fichet, B., Le Calvé, G. (1984), Structure géométrique des principaux indices de dissimilarité sur signes de présence-absence, *Statist. Anal. Données*, 9 (3), pp. 11–44.

Fréchet, M. (1935), Sur la définition axiomatique d'une classe d'espaces vectoriels distanciés applicables vectoriellement sur l'espace de Hilbert, *Ann. of Math.*, 36, pp. 705–718.

Gower, J.C., Legendre, P. (1986), Metric and Euclidean properties of dissimilarity coefficients, *J. Classification*, 3, pp. 5–48. [3].

Joly, S., Le Calvé, G. (1986), Metric and Euclidean properties of dissimilarity coefficients, *Statist. Anal. Données*, 11, pp. 30–50.

Joly, S., Le Calvé, G. (1992), Realisable 0-1 matrices and city-block distance, Rapport de recherche 92-1, Laboratoire Analyse des Données, Université de Haute Bretagne, Rennes, France.

Kelly, J.B. (1968), Products of zero-one matrices, *Canad. J. Math.*, 20, pp. 298–329.

Kelly, J.B. (1972), Hypermetric spaces and metric transforms, In: Shisha, O., ed, *Inequalities III*, Academic Press, New York, pp. 149–158.

Le Calvé, G. (1987), L_1-embeddings of a data structure (I, D), In: Dodge, Y., ed., *Statistical Data Analysis based on the L_1-norm and Related Methods*, North-Holland, Amsterdam, pp. 195–202.

Le Calvé, G. (1988), Similarities functions, In: Edwards, D., Raum, N.E., eds., *Proceedings of COMPSTAT 88*, Physica Verlag, Heidelberg, pp. 341–347.

Schoenberg, I.J. (1937), On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space, *Ann. of Math.*, 38, pp. 787–793.

Schoenberg, I.J. (1938), Metric spaces and positive definite functions, *Trans. Amer. Math. Soc.*, 44, pp. 522–536.

Schur, J. (1911), Bemerkungen zur Theorie der beschrankter Bilinearformen mit unendlich vielen Veränderlichen, *J. Reine Angew. Math.*, 140, pp. 1–28.

Torgerson, W.S. (1958), *Theory and methods of scaling.* Wiley, New York.