Distance

Hervé Abdi[1]

1 Overview

The notion of distance is essential because many statistical techniques are equivalent to the analysis of a specific distance table. For example, principal component analysis and metric multidimensional scaling analyze Euclidean distances, correspondence analysis deals with a χ^2 distance matrix, and discriminant analysis is equivalent to using a *Mahalanobis* distance. To define a distance is equivalent to defining rules to assign positive numbers between *pairs* of objects. The most important distances for statistics are: Euclidean, generalized Euclidean (which include χ^2 and Mahalanobis), Minkowsky (which include the sorting and the symmetric difference distances), and the Hellinger distances.

2 Notation and definition

For convenience, we restrict our discussion to distance between vectors because they are the objects mostly used in statistics. Let **a**, **b**, and **c** be three vectors with J elements each, a distance is a function which associates to any pair of vectors a real positive number,

[1]In: Neil Salkind (Ed.) (2007). *Encyclopedia of Measurement and Statistics.* Thousand Oaks (CA): Sage.
Address correspondence to: Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utd.edu/~herve

denoted $d(\mathbf{a},\mathbf{b})$, which has the following properties:

$$d(\mathbf{a},\mathbf{a}) = 0 \tag{1}$$
$$d(\mathbf{a},\mathbf{b}) = d(\mathbf{b},\mathbf{a}) \quad \text{[symmetry]} \tag{2}$$
$$d(\mathbf{a},\mathbf{b}) \leq d(\mathbf{a},\mathbf{c}) + d(\mathbf{c},\mathbf{b}) \text{ [triangular inequality]} \tag{3}$$

2.1 An minimalist example: the sorting distance

The axioms defining a distance are very easily met. For example, suppose that we consider two objects and assign the number 1 if we find them different and 0 if we find them alike. This procedure defines a distance called the *sorting* distance because the number assigned to a pair of same objects will be equal to 0 (this satisfies Axiom 1). Axiom 2 is also satisfied because the order of the objects is irrelevant. For the third axiom, we need to consider two cases, if $d(\mathbf{a},\mathbf{b})$ is equal to 0, the sum $d(\mathbf{a},\mathbf{c})+d(\mathbf{c},\mathbf{b})$ can take only the values 0, 1 or 2 which will all satisfy Axiom 3. If $\mathbf{a}$ and $\mathbf{b}$ are different, $d(\mathbf{a},\mathbf{b})$ is equal to 1 and $\mathbf{c}$ cannot be identical to *both* of them, and therefore the sum $d(\mathbf{a},\mathbf{c}) + d(\mathbf{c},\mathbf{b})$ can take only the values 1, or 2 which will both satisfy Axiom 3.

With the same argument, we can see that if we ask a set of respondents to sort objects into piles, the number of participants who do not sort two objects together defines a distance between the sorted objects.

3 The Euclidean distance

The most well-known distance is the *Euclidean distance* which is defined as:

$$d(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\| = \sqrt{(\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{b})} = \sqrt{\sum_j (a_j - b_j)^2} \tag{4}$$

(with $\|\mathbf{a}\|$ being the norm of $\mathbf{a}$, and a_j and b_j being the j-th element of $\mathbf{a}$ and $\mathbf{b}$). Expressed as a squared distance (in an Euclidean world, it is always more practical to work with squared quantities

because of the Pythagorean theorem), it is computed as:

$$d^2(\mathbf{a},\mathbf{b}) = (\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{b}) \; . \tag{5}$$

For example, with:

$$\mathbf{a} = \begin{bmatrix} 2 \\ 5 \\ 10 \\ 20 \end{bmatrix} \text{ , and } \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \; , \tag{6}$$

the vector $\mathbf{a}-\mathbf{b}$ gives:

$$\mathbf{a}-\mathbf{b} = \begin{bmatrix} 2-1 \\ 5-2 \\ 10-3 \\ 20-4 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 7 \\ 16 \end{bmatrix} \tag{7}$$

and

$$\begin{aligned} d^2(\mathbf{a},\mathbf{b}) &= (\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{b}) \\ &= \sum_{j=1}^{4}(a_j - b_j)^2 \\ &= 1^2 + 3^2 + 7^2 + 16^2 \\ &= 315 \; . \end{aligned} \tag{8}$$

The Euclidean distance between two vectors can also be expressed *via* the notion of scalar product and cosine between vectors. By developing Equation 5 for the distance between vectors, we find that:

$$\begin{aligned} d^2(\mathbf{a},\mathbf{b}) &= (\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{b}) \\ &= \mathbf{a}^T\mathbf{a} + \mathbf{b}^T\mathbf{b} - 2\mathbf{a}^T\mathbf{b} \\ &= \|a\|^2 + \|b\|^2 - 2\|a\| \times \|b\| \times \cos(\mathbf{a},\mathbf{b}) \; . \end{aligned} \tag{9}$$

In the particular case of vectors with a unit norm, the distance between **a** and **b** simplifies to:

$$d^2(\mathbf{a},\mathbf{b}) = 2[1-\cos(\mathbf{a},\mathbf{b})] \; . \tag{10}$$

When two vectors are centered (*i.e.,* when their mean is equal to 0), their cosine is equal to the coefficient of correlation. This shows that we can define a (Euclidean) distance between two series of numbers as 1 minus their correlation.

4 Generalized Euclidean

The Euclidean distance can be generalized by taking into account constraints expressed by a matrix conformable with the vectors. Specifically, let **W** denote a $J \times J$ *positive definite* matrix, the generalized Euclidean distance between **a** and **b** becomes:

$$d^2_{\mathbf{W}}(\mathbf{a},\mathbf{b}) = (\mathbf{a}-\mathbf{b})^T\mathbf{W}(\mathbf{a}-\mathbf{b}) \; . \tag{11}$$

The most well-known generalized Euclidean distances are the χ^2 and the Mahalanobis distances.

4.1 χ^2 distance

The χ^2 distance is associated to correspondence analysis. It is a distance between profiles. Recall that a vector is called a *profile* when it is composed of numbers greater or equal to zero whose sum is equal to one (such a vector is sometimes called a *stochastic* vector). The χ^2 distance is defined for the rows (or the columns after transposition of the data table) of a contingency table such as the one shown, for example, in Table 1. The first step of the computation of the distance is to transform the rows into row profiles which is done by dividing each row by its total. There are I rows and J columns in a contingency table. The *mass* of each row is denoted r_i, and the mass vector **r**. The barycenter of the rows, denoted **c** is computed by transforming the total of the columns into a row profile. It can also be computed as the weighted average of the row profiles (with the weights being given by the mass vector

Table 1: Data for the computation of the χ^2, Mahalanobis, and Hellinger distances. The punctuation marks of six French writers (from Abdi & Valentin, 2006). The column labelled $N \times \mathbf{r}$ gives the total number of punctuation marks used by each author. The mass of each row is the proportion of punctuation marks used by this author. The row labelled $N \times \mathbf{c}^T$ gives the total of each column. This is the total number of times this punctuation mark was used. The centroid row (or barycenter, or center of gravity), gives the proportion of each punctuation mark in the sample. The weight of each column is the *inverse* of the centroid.

	Raw data					Row profiles		
Author's name	Period	Comma	Other	$N \times \mathbf{r}$	$\mathbf{r}$	Period	Comma	Other
Rousseau	7836	13112	6026	**26974**	**.0189**	.2905	.4861	.2234
Chateaubriand	53655	102383	42413	**198451**	**.1393**	.2704	.5159	.2137
Hugo	115615	184541	59226	**359382**	**.2522**	.3217	.5135	.1648
Zola	161926	340479	62754	**565159**	**.3966**	.2865	.6024	.1110
Proust	38177	105101	12670	**155948**	**.1094**	.2448	.6739	.0812
Giraudoux	46371	58367	14299	**119037**	**.0835**	.3896	.4903	.1201
$\sum N\mathbf{c}^T$	**423580**	**803983**	**197388**	**1424951**	1.0000			
$\mathbf{c}^T$	**.2973**	**.5642**	**.1385**					
$\mathbf{w}^T$	**3.3641**	**1.7724**	**7.2190**					

r). For the χ^2 distance, the **W** matrix is diagonal which is equivalent to assigning a weight to each column. This weight is equal to the inverse of the relative frequency of the column. This is expressed formally by expressing **W** as

$$\mathbf{W} = (\text{diag}\{\mathbf{c}\})^{-1}. \tag{12}$$

With this coding schema, variables which are used often contribute less to the distance between rows than variables which are used rarely. For example, from Table 1, we find that the weight matrix is equal to

$$\mathbf{W} = \mathbf{D}_{\mathbf{w}} = \text{diag}\{\mathbf{w}\} = \begin{bmatrix} .2973^{-1} & 0 & 0 \\ 0 & .5642^{-1} & 0 \\ 0 & 0 & .1385^{-1} \end{bmatrix} = \begin{bmatrix} 3.3641 & 0 & 0 \\ 0 & 1.7724 & 0 \\ 0 & 0 & 7.2190 \end{bmatrix}. \tag{13}$$

For example, the χ^2 distance between Rousseau and Chateaubriand is equal to

$$\begin{aligned} &d^2(\text{Rousseau}, \text{Chateaubriand}) \\ &= 3.364 \times (.291 - .270)^2 + 1.772 \times (.486 - .526)^2 + 7.219 \times (.223 - .214)^2 \\ &= .0036\,. \end{aligned} \tag{14}$$

This distance is called the χ^2 distance because the sum of the weighted distances from the rows to their barycenter is proportional to the χ^2 computed to test the independence of the rows and the columns of the Table. Formally, if we denoted by N the grand total of the contingency table, by $d^2(i,g)$ the distance from row i to the barycenter of the table, and by $d^2(i,i')$ the distance from row i to row i', we obtain the following equality:

$$\sum_{i}^{I} r_i d^2(i,g) = \sum_{i>i'} r_i r_{i'} d^2(i,i') = \frac{1}{N}\chi^2\,. \tag{15}$$

The metric multidimensional scaling analysis of a χ^2 distance matrix (with masses given by **r**) is equivalent to correspondence analysis.

4.2 Mahalanobis distance

The Mahalanobis distance is defined between rows of a Table. The weight matrix is obtained as the inverse of the columns variance / covariance matrix. Formally, if we denoted by $\mathbf{S}$ the variance / covariance matrix between the columns of a data table, the weight matrix of the Mahalanobis distance is defined as $\mathbf{W} = \mathbf{S}^{-1}$.

Using, again, the data from Table 1, we obtain:

$$\mathbf{S} = 10^{10} \times \begin{bmatrix} 0.325 & 0.641 & 0.131 \\ 0.641 & 1.347 & 0.249 \\ 0.131 & 0.249 & 0.063 \end{bmatrix} \tag{16}$$

and

$$\mathbf{S}^{-1} = 10^{-7} \times \begin{bmatrix} 0.927 & -0.314 & -0.683 \\ -0.314 & 0.134 & 0.124 \\ -0.683 & 0.124 & 1.087 \end{bmatrix} . \tag{17}$$

With these values, we find that the Mahalanobis distance between Rousseau and Chateaubriand is equal to

$$\begin{aligned} & d^2(\text{Rousseau}, \text{Chateaubriand}) \\ & = 10^{-7} \times \left(\begin{bmatrix} -45819 \\ -89271 \\ -36387 \end{bmatrix}^{\mathsf{T}} \times \begin{bmatrix} 0.927 & -0.314 & -0.683 \\ -0.314 & 0.134 & 0.124 \\ -0.683 & 0.124 & 1.087 \end{bmatrix} \times \begin{bmatrix} -45819 \\ -89271 \\ -36387 \end{bmatrix} \right) \\ & \approx 4.0878 . \end{aligned} \tag{18}$$

The Mahalanobis distance can be seen as a multivariate equivalent of the Z-score transformation. The metric multidimensional scaling analysis of a Mahalanobis distance matrix is equivalent to discriminant analysis.

5 Minskowki's distance

The Euclidean distance is a particular case of the more general family of Minkowski's distances. The p-distance (or a Minkowski's distance of degree p), between 2 vectors is defined as

$$\mathbf{a} = [a_1, \ldots, a_j, \ldots, a_J]^T \quad \text{and} \quad \mathbf{b} = [b_1, \ldots, b_j, \ldots, b_J]^T \tag{19}$$

as

$$d_p(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\|_p = \left[\sum_j^J |a_j - b_j|^p\right]^{\frac{1}{p}} . \tag{20}$$

The most frequently used Minkowski's distances are the distances of degree 1, 2, and ∞. A distance of degree 1 is also called the *city-block* or *taxi-cab* distance. When the vectors are binary numbers (*e.g.,* 1 and 0), the elements of the vector code for membership to a set (*i.e.,* 1 means the elements belongs to the set, 0 means it does not). In this case, the degree 1 distance is commonly referred to as the *Hamming distance* or the *symmetric difference distance* (the symmetric difference distance is a set operation which associates to two sets a new set made of the elements of these sets that belong to only one of them—*i.e.,* elements that belong to both sets are excluded. The symmetric difference distance gives the number of the elements of the symmetric difference set).

When p is equal to 2, we obtain the usual Euclidean distance. With $p = \infty$, we take the largest absolute value of the difference between the vectors as defining the distance between vectors.

For example, with the vectors:

$$\mathbf{a} = \begin{bmatrix} 2 \\ 5 \\ 10 \\ 20 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} , \tag{21}$$

the Minkowski distance of degree 1 is:

$$d_1(\mathbf{a},\mathbf{b}) = \sum_{j=1}^{4} |a_j - b_j| = 1+3+7+16 = 27 , \tag{22}$$

and the Minkowski distance of degree ∞ is:

$$d_\infty(\mathbf{a},\mathbf{b}) = \max_j |a_j - b_j| = \max\{1,3,7,16\} = 16 . \tag{23}$$

6 Hellinger

The Hellinger distance is defined between vectors having only positive or zero elements. In general (like the χ^2 distance), it is used

for row profiles. The Hellinger distance between vectors **a** and **b** is defined as

$$d(\mathbf{a},\mathbf{b}) = \left[\sum_{j}^{J}\left(\sqrt{a}_j - \sqrt{b}_j\right)^2\right]^{\frac{1}{2}} \tag{24}$$

(*cf.* Escofier, 1978, Domenges & Volle, 1979; Rao, 1995). Because the Hellinger distance is not sensitive to discrepancies between columns, it is sometimes used as an alternative to the χ^2 distance. An interesting property of the Hellinger distance when applied to row profiles is that the vectors representing these profiles can be represented as points on a sphere (or hypersphere when the number of elements of the vector is larger than 3).

For our example, we find that the Hellinger distance between the row profiles of Rousseau and Chateaubriand is equal to

$$\begin{aligned} &d(\text{Rousseau},\text{Chateaubriand}) \\ &= \left[\sqrt{2905} - \sqrt{.2704})^2 + (\sqrt{.4861} - \sqrt{.5259})^2 + (\sqrt{.2234} - \sqrt{.2137})^2\right]^{\frac{1}{2}} \\ &= .0302\ . \end{aligned} \tag{25}$$

7 How to analyze distance matrices?

Distance matrices are often computed as the first step of data analysis. In general, distance matrices are analyzed by finding a convenient graphic representation for their elements. These representations approximate the original distance by another distance such as 1) a low dimensionality Euclidean distance (*e.g.,* multidimensional scaling, DISTATIS) or 2) a graph (*e.g.,* cluster analysis, additive tree representations).

References

[1] Abdi, H. (1990). Additive-tree representations. *Lecture Notes in Biomathematics*, **84**, 43–59.

[2] Abdi, H. (2003). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks: Sage.

[3] Abdi, H., Valentin, D. (2006). *Mathématiques pour les sciences cognitives (Mathematics for Cognitive Sciences)*. Grenoble: PUG.
[4] Domenges, D., Volle M. (1979). Analyse factorielle sphérique: une exploration. *Annales de l'INSEE*, **35**, 3–83.
[5] Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statistiques Appliquées*, **26**, 29–37.
[6] Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
[7] Rao, C.R. (1995) Use of Hellinger distance in graphical displays. In E.-M. Tiit, T. Kollo, & H. Niemi (Ed.): *Multivariate statistics and matrices in statistics*. Leiden (Netherland): Brill Academic Publisher. pp. 143–161.