J Math Chem (2013) 51:338–353
DOI 10.1007/s10910-012-0086-7

ORIGINAL PAPER

Collective Euclidian distances and quantum similarity

Ramon Carbó-Dorca

Received: 11 June 2012 / Accepted: 10 September 2012 / Published online: 22 September 2012

Abstract A simple algorithm to find out a collective distance between arbitrary assemblies of points in some vector space is defined at various levels of complexity. The most straightforward procedure is related to a sum of Euclidian distances, which can be easily obtained from any Gram matrix of the point vectors. Similar but more involved formulation can be obtained with the tensor products of an indefinite number of vectors. Simple and elaborate examples are provided to illustrate the procedures along the text. The use of collective distances in reference to quantum similarity is discussed as an important application issue and a numerical example given.

Keywords Distance · Euclidian distances · Gram matrices · Collective Euclidian distances · Quantum similarity · Carbó similarity index · Quantum dissimilarity indices

1 Introduction

1.1 Foreword

After the publication of several papers about the subject associated to quantum similarity (QS) theoretical background [1–4] and along the topic of Carbó index (CI) generalization as discussed in reference [3], it was also argued later on about the opportunity and difficulty to obtain expressions structuring collective Euclidian distances (ED), within an analysis connecting the Hodgkin-Richards similarity index (HRI) [5] with ED [6] in QS theoretical grounds.

R. Carbó-Dorca (✉)
Institut de Química Computacional, Universitat de Girona, 17071 Girona, Catalonia, Spain
e-mail: quantumqsar@hotmail.com

The present study will try to describe such a general distance description possibility employing several options, with the aim of trying to answer the simple question: can Euclidian-like distances involving N mathematical objects be defined?

In the plausible case of any affirmative response, then the application of the collective ED computational scheme will have an immediate application in QS, as a mean to consider the global dissimilarity definition of quantum object (QO) sets (QOS), see for more details reference [4].

1.2 Distance and dissimilarity

Thus, the main idea underlying the present discussion has to be referred to the axiomatic description of distance, consisting of three well-defined elements, according to reference [7], which defines a distance as:

Let S be a set. A function $D : \mathrm{S} \times \mathrm{S} \rightarrow \mathbf{R}^{+}$ is called a *distance* (or *dissimilarity*) on S if: $\forall x, y \in \mathrm{S}$, there holds:

a. $D(x, y) \geq 0$ (*non-negativity*).
b. $D(x, y) = D(y, x)$ (*symmetry*).
c. $D(x, x) = 0$ (*reflexivity*).

From this well-known previous axiomatic scheme, it will be tried here to extend the notion of dissimilarity into sets of mathematical objects contained in a pre-Hilbert vector space, where some kind of scalar product and an attached norm can be defined [8].

On the other hand, the interest in dissimilarities of the present author, do not pretends to find out just a solution for collective distances too intricate to be understood, nor difficult to be implemented. Because of this attitude, the mathematical framework where collective distances will be here defined will be just circumscribed to Euclidian vector spaces over the real field. Keeping in mind that it is not too hard to extend the definitions and simple mathematical structures, which will be here developed, to other kinds of vector spaces.

1.3 Quantum similarity and collective Euclidian distances

A justification of the interest of the present study, if the importance to extend the concept of dissimilarity measure to sets of points with cardinality greater than two needs justification, can be found in the domain of QS, where the concern for a distance or dissimilarity definition[1] has been present since the first paper [9] devoted to the classification of quantum objects (QO). See also reference [4] for more details on QO definitions, QO sets (QOS) and QS background definitions. The theoretical basis of QS developed up to 2010 can be found in detailed monographs [10,11].

[1] In fact, the initial intention of reference [9] was to answer the question: *how far is a molecule from another?*, and the draft title was posed accordingly. Some referee comment forced the authors to change the initial heading into: *how similar is a molecule to another?*. Such an occurrence has provided the literature with the field name of quantum similarity instead of the equally interesting and related quantum dissimilarity development.

Nowadays, in the mentioned reference [6], dedicated to discuss the geometrical origin of the HRI [5], there was discussed the difficulty and perhaps the impossibility to define a distance concept, involving not only two vectors or QO, but a well-defined collection of them. For example: it is interesting to try to find out a distance involving all the elements of a quantum point cloud or hut [4].

Such a mathematical reflection was originated due to the fact that many QO generalization of QS indices has been already discussed in some extent within the QS CI domain [1–3], providing a general expression involving an indeterminate number of vectors or QO and can be easily described. Originally CI [9] was trivially interpreted as the cosine of the angle subtended by two QO density functions (DF). In the same way, generalized CI's can be interpreted as the solid angle subtended by several vectors, or alternatively within the QS domain, by various quantum DF [3]. However, as far as the author has tried to find out, no such collective distance concepts have been so far defined. One can find out none, even within the previously mentioned exhaustive study of distances [7].

1.4 Structure of this study

Taking in consideration all the previous information, the structure of the present paper has been constructed as follows. First, in order to provide the reader with a simple mathematical background as a way to generate collective ED, the orthogonal decomposition of a square matrix will be discussed. In this way, the notation which will follow can be at the same time set up. In a second stage, the generation of ED between two mathematical objects defined within a vector space will be discussed, and a trivial extension to three objects, leading to N objects collective ED definition will be proposed and analyzed.

Simple numerical examples to illustrate the collective ED functionality will be given at each stage and several equivalent algorithms described. One of them makes relevant the role of the point cloud centroid in the collective ED structure algorithm. Abstract application examples involving orthonormal vector sets and simplices will be given next. Afterwards, having set the structure of collective ED, some more general concept will be studied within tensorial double and v-tuple object structures. Then, collective ED between two mathematical object sets will be discussed in brief, and applied into another simple illustrative example. Finally, the application of collective ED within the QS environment made of quantum point clouds will be discussed.

2 Orthogonal decomposition of a square matrix

Known any arbitrary $(N \times N)$ real matrix $\mathbf{A}$, it can be trivially seen that such a matrix can be decomposed and then reconstructed by means of two orthogonal well-defined $(N \times N)$ matrix parts: a diagonal matrix $\mathbf{D} = Diag\,(\mathbf{A})$ and an off-diagonal matrix $\mathbf{F} = Offdiag\,(\mathbf{A})$ in the following manner:

$$\mathbf{A}_{\pm} = \mathbf{D} \pm \mathbf{F} \tag{1}$$

where $\mathbf{D}$ has the off-diagonal null, that is: $\forall I \neq J : D_{IJ} = 0$, and $\mathbf{F}$ the diagonal null, or: $\forall I : F_{II} = 0$. The double sign in Eq. (1) can be chosen at will and in fact it defines, in relationship to the initial information, two newly reconstructed matrices which can be linearly independent, except of course in the case one of the decomposition matrices is the null matrix.

To prove this interesting property, one can use the inward matrix product and the complete sum of a matrix symbols.[2] In this manner, resulting that it can be written: $\langle \mathbf{D} * \mathbf{F} \rangle = 0$, as some way to represent a null scalar product, evidencing that the two parts of the decomposition (1) are orthogonal. Thus, being both matrices in Eq. (1) a two-dimensional basis set, then one also has the following Euclidian norm and scalar product relations between the matrices appearing into definition (1):

$$\langle \mathbf{A}_+ * \mathbf{A}_+ \rangle = \langle \mathbf{D} * \mathbf{D} \rangle + \langle \mathbf{F} * \mathbf{F} \rangle = \langle \mathbf{A}_- * \mathbf{A}_- \rangle$$
$$\langle \mathbf{A}_+ * \mathbf{A}_- \rangle = \langle \mathbf{D} * \mathbf{D} \rangle - \langle \mathbf{F} * \mathbf{F} \rangle = \langle \mathbf{A}_- * \mathbf{A}_+ \rangle$$

Therefore, the determinant D of the Gram matrix constructed by these four scalar products can be written as:

$$D = (\langle \mathbf{D} * \mathbf{D} \rangle + \langle \mathbf{F} * \mathbf{F} \rangle)^2 - (\langle \mathbf{D} * \mathbf{D} \rangle - \langle \mathbf{F} * \mathbf{F} \rangle)^2 = 4 \langle \mathbf{D} * \mathbf{D} \rangle \langle \mathbf{F} * \mathbf{F} \rangle , \tag{2}$$

a result which can be supposed resulting as a trivial consequence of the orthogonal decomposition defined in Eq. (1).

Consequently, unless the matrices $\mathbf{D}$ or $\mathbf{F}$ are null, the Gram matrix determinant (2) is positive definite, as corresponds to the product of the two Euclidian norms of the matrices associated to the orthogonal decomposition.

Resuming: this result means the decomposition (1) might produce in general two orthogonal, linearly independent matrices.

3 Euclidian distance between two mathematical objects represented in a vector space

Suppose one knows two vectors of some pre-Hilbert space, which might represent any pair of coherently defined mathematical objects: $L_2 = \{|a\rangle ; |b\rangle\}$. One can construct the Gram matrix of such object pair, which may be generally written in a symmetrical way and can be orthogonally decomposed in the previously defined manner, as shown

[2] The result of an *inward matrix product* involving two matrices of the same dimension $(N \times M)$: $\{\mathbf{A}, \mathbf{B}\}$ is another $(N \times M)$ matrix: $\mathbf{C} = \mathbf{A} * \mathbf{B}$, defined as: $\forall I, J : C_{IJ} = A_{IJ} B_{IJ}$. The *complete sum* of a matrix is easily defined as: $\langle \mathbf{A} \rangle = \sum_I \sum_J A_{IJ}$. Both symbols can be employed to define the scalar product of two $(N \times M)$ matrices: $\langle \mathbf{A} | \mathbf{B} \rangle = \langle \mathbf{A} * \mathbf{B} \rangle = \sum_{I=1}^{N} \sum_{J=1}^{M} A_{IJ} B_{IJ}$. The Euclidian norm of a matrix can be thus defined as: $\langle \mathbf{A} | \mathbf{A} \rangle = \langle \mathbf{A} * \mathbf{A} \rangle = \sum_{I=1}^{N} \sum_{J=1}^{M} |A_{IJ}|^2$.

at Eq. (1):

$$\mathbf{S} = \begin{pmatrix} \langle a|a\rangle & \langle a|b\rangle \\ \langle a|b\rangle & \langle b|b\rangle \end{pmatrix} \rightarrow \mathbf{S}_{\pm} = \begin{pmatrix} \langle a|a\rangle & 0 \\ 0 & \langle b|b\rangle \end{pmatrix} \pm \begin{pmatrix} 0 & \langle a|b\rangle \\ \langle a|b\rangle & 0 \end{pmatrix} = \mathbf{D} \pm \mathbf{F}$$

Then, the squared Euclidian distance between the two objects can be easily written as the complete sum:

$$D_{ab}^2 = \langle \mathbf{S}_- \rangle = \langle \mathbf{D} \rangle - \langle \mathbf{F} \rangle = \langle a|a\rangle + \langle b|b\rangle - 2\,\langle a|b\rangle\,. \tag{3}$$

This unassuming definition sequence sets up how the Euclidian distance, involving two mathematical objects, can be constructed in terms of their Gram matrix orthogonal decomposition of type (1).

4 Trivial extension to three objects and generalization to *N*

Supposing the knowledge of a three object set: $L_3 = \{|a\rangle\,;\,|b\rangle\,;\,|c\rangle\}$ with an attached Gram matrix constructed and orthogonally decomposed similarly as in the previously defined two object case:

$$\mathbf{S} = \begin{pmatrix} \langle a|a\rangle & \langle a|b\rangle & \langle a|c\rangle \\ \langle a|b\rangle & \langle b|b\rangle & \langle b|c\rangle \\ \langle a|c\rangle & \langle b|c\rangle & \langle c|c\rangle \end{pmatrix}$$

$$\rightarrow \mathbf{S}_{\pm} = \begin{pmatrix} \langle a|a\rangle & 0 & 0 \\ 0 & \langle b|b\rangle & 0 \\ 0 & 0 & \langle c|c\rangle \end{pmatrix} \pm \begin{pmatrix} 0 & \langle a|b\rangle & \langle a|c\rangle \\ \langle a|b\rangle & 0 & \langle b|c\rangle \\ \langle a|c\rangle & \langle b|c\rangle & 0 \end{pmatrix} = \mathbf{D} \pm \mathbf{F};$$

then, one can describe the *triple* object Euclidian squared distance associated to the collection L_3 as:

$$\begin{aligned} D_{abc}^{(2)} &= \langle \mathbf{S}_- \rangle = 2\,\langle \mathbf{D} \rangle - \langle \mathbf{F} \rangle \\ &= 2\,(\langle a|a\rangle + \langle b|b\rangle + \langle c|c\rangle) - 2\,(\langle a|b\rangle + \langle a|c\rangle + \langle b|c\rangle)\,. \end{aligned}$$

Such a definition presents the following advantages:

(1) It possesses a structure of distance or dissimilarity, as the following axioms hold: (a) it is always real and positive (*non-negativity*), (b) any permutation in the canonical order of the involved objects leaves the collective distance invariant (*extended symmetry*) and (c) when all the objects taken into account coalesce and become a unique object it becomes null (*extended reflexivity*)

(2) The collective distance concept is easily generalizable when involving an indeterminate number N of mathematical objects. Provided that such N objects are defined by the vector set: $L_N = \{|I\rangle\,|I = 1, N\}$, belonging to some appropriate vector space and their $(N \times N)$ Gram matrix is written as: $\mathbf{S} = \{S_{IJ} = \langle I|J\rangle\}$. The orthogonal decomposition (1) of the Gram matrix can be defined in turn

by means of the two matrices: $\mathbf{D} = Diag(\mathbf{S}) \wedge \mathbf{F} = OffDiag(\mathbf{S})$. Then, in general one can write a collective dissimilarity involving the set L_N as:

$$_L D_N^{(2)} = (N-1)\langle \mathbf{D} \rangle - \langle \mathbf{F} \rangle = (N-1)\sum_{I=1}^{N} \langle I|I\rangle - 2\sum_{I}^{N-1}\sum_{J=I+1}^{N} \langle I|J\rangle \quad (4)$$

The factor $N-1$ in Eq. (4) is needed to take into account the repetition of the diagonal elements, in order to fill the computing needs of all the squared ED between every pair of elements, included within the collection L_N.

(3) It is multiplied by a unique factor when the objects are submitted to a homothetic transformation.

(4) It is invariant upon translation of the involved vector collection. This is due because expression (4) is the sum of all $\frac{1}{2}N(N-1)$ non-redundant ED between pairs of elements of the object set. Taken individually ED between pairs of objects are invariant upon translation.

4.1 Simple illustrative example

Defining a two dimensional square as:

$$Q = \{|a\rangle = (-1,-1)\,;\, |b\rangle = (1,-1)\,;\, |c\rangle = (-1,1)\,;\, |d\rangle = (1,1)\}, \quad (5)$$

one can construct the associated Gram matrix and the collective distance of the four vertices as:

$$\mathbf{S} = \begin{pmatrix} 2 & 0 & 0 & -2 \\ 0 & 2 & -2 & 0 \\ 0 & -2 & 2 & 0 \\ -2 & 0 & 0 & 2 \end{pmatrix} \rightarrow D_4^{(2)} = 24 - (-8) \rightarrow D_4^{(2)} = 32.$$

Multiplying by a scalar λ all the vectors of the square Q, providing the homothetic square:

$$Q_\lambda = \{|a\rangle = (-\lambda,-\lambda)\,;\, |b\rangle = (\lambda,-\lambda)\,;\, |c\rangle = (-\lambda,\lambda)\,;\, |d\rangle = (\lambda,\lambda)\},$$

yields the Gram matrix shown below and it is easy to see the collective distance appears multiplied by the square of the homothetic scale parameter:

$$\mathbf{S} = \begin{pmatrix} 2\lambda^2 & 0 & 0 & -2\lambda^2 \\ 0 & 2\lambda^2 & -2\lambda^2 & 0 \\ 0 & -2\lambda^2 & 2\lambda^2 & 0 \\ -2\lambda^2 & 0 & 0 & 2\lambda^2 \end{pmatrix} \rightarrow D_4^{(2)} = 32\lambda^2.$$

A translation by means of the vector: $|t\rangle = (1,1)$ transforms the initial square Q into the vertices of a translated equivalent square, which can be written as: $Q_t =$

$\{|a\rangle = (0,0); |b\rangle = (2,0); |c\rangle = (0,2); |d\rangle = (2,2)\}$. The resultant square provides a Gram matrix and an attached collective distance like:

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 4 \\ 0 & 0 & 4 & 4 \\ 0 & 4 & 4 & 8 \end{pmatrix} \rightarrow D_4^{(2)} = 48 - 16 = 32.$$

So, in this manner it is practically seen by means of a case example, how the sum of the collective distance definition becomes invariant upon translation.

5 An alternative equivalent definition

In order to observe how preserved the translation properties are in the present collective distances definition, one can perhaps also define the collective distance as the sum of distances between object pairs as:

$$\begin{aligned} D_N^{(2)} = \sum_{I=1}^{N-1} \sum_{J=I+1}^{N} D_{IJ}^2 &= \sum_{I=1}^{N-1} \sum_{J=I+1}^{N} \left(\langle I|I\rangle + \langle J|J\rangle - 2\langle I|J\rangle \right) \\ &= \sum_{I=1}^{N-1} \sum_{J=I+1}^{N} \left(\langle I|I\rangle + \langle J|J\rangle \right) - \langle \mathbf{F}\rangle = \frac{1}{2}\langle \mathbf{T}\rangle - \langle \mathbf{F}\rangle , \end{aligned} \quad (6)$$

where the matrix $\mathbf{T}$ can be easily defined as a matrix with the same off-diagonal structure as the matrix $\mathbf{F}$, but involving the Euclidian norms sums of the vectors, that is:

$$\mathbf{T} = \{T_{IJ} = \delta\,(I \neq J)\,(\langle I|I\rangle + \langle J|J\rangle)\}\,. \quad (7)$$

The matrix (7) above defined can be easily written, using again as a simple example the original two dimensional square Q coordinates setup as shown in Eq. (5), thus providing now the collective distance obtained as:

$$\mathbf{T} = \begin{pmatrix} 0 & 4 & 4 & 4 \\ 4 & 0 & 4 & 4 \\ 4 & 4 & 0 & 4 \\ 4 & 4 & 4 & 0 \end{pmatrix} \rightarrow \frac{1}{2}\langle \mathbf{T}\rangle = 24 \rightarrow D_4^{(2)} = 24 - (-8) = 32,$$

while in the translated square case Q_t one has the matrix $\mathbf{T}$ defined and the collective distance computed as:

$$\mathbf{T} = \begin{pmatrix} 0 & 4 & 4 & 8 \\ 4 & 0 & 8 & 12 \\ 4 & 8 & 0 & 12 \\ 8 & 12 & 12 & 0 \end{pmatrix} \rightarrow \frac{1}{2}\langle \mathbf{T}\rangle = 48 \rightarrow D_4^{(2)} = 48 - \langle \mathbf{F}\rangle = 48 - 16 = 32$$

which obviously yields an invariant collective distance value, as this formulation must be equivalent to the previous construction (4).

5.1 Average collective distance

The collective distance can be also transformed into some kind of average collective distance between N objects. Because taking into account the algorithm of Eq. (6) one can easily construct the average collective distance expression:

$$\overline{D_N^{(2)}} = \frac{2}{N\,(N-1)}\left(\frac{1}{2}\,\langle \mathbf{T}\rangle - \langle \mathbf{F}\rangle\right) = \frac{1}{N\,(N-1)}\left(\langle \mathbf{T}\rangle - 2\,\langle \mathbf{F}\rangle\right),$$

which in the previous square Q example will provide a value easily computed as:

$$\overline{D_4^{(2)}} = \frac{1}{12}\,(96-16) = \frac{20}{3} = 6.\widehat{6}\,.$$

6 The role of the point cloud centroid

Starting from the usual definition of the collective distance of a set of points as given in Eq. (6), forming a point cloud like: $\mathrm{L}_N = \{|I\rangle\,|I = 1, N\}$, then one can develop the following sequence of equivalent expressions:

$$\begin{aligned}
D_N^{(2)} &= \sum_{I=1}^{N}\sum_{J\neq I} D_{IJ}^2 = \sum_{I=1}^{N}\sum_{J\neq I}\left(\langle I|I\rangle + \langle J|J\rangle - 2\,\langle I|J\rangle\right)\\
&= \sum_{I=1}^{N}\sum_{J\neq I}\langle I|I\rangle + \sum_{I=1}^{N}\sum_{J\neq I}\langle J|J\rangle - 2\sum_{I=1}^{N}\sum_{J\neq I}\langle I|J\rangle\\
&= 2\sum_{I=1}^{N}\left[(N-1)\,\langle I|I\rangle - \sum_{J\neq I}\langle I|J\rangle\right] = 2\sum_{I=1}^{N}\left[N\,\langle I|I\rangle - \sum_{J}\langle I|J\rangle\right]\\
&= 2N\sum_{I=1}^{N}\left[\langle I|I\rangle - N^{-1}\sum_{J}\langle I|J\rangle\right] = 2N\sum_{I=1}^{N}\langle I|\left[|I\rangle - N^{-1}\sum_{J}|J\rangle\right]\\
&= 2N\sum_{I=1}^{N}\langle I|\left[|I\rangle - |\mathbf{c}\rangle\right] = 2N\sum_{I=1}^{N}\langle I|A_I\rangle \leftarrow \forall I : |A_I\rangle = |I\rangle - |\mathbf{c}\rangle
\end{aligned}$$

This final result corresponds to a scaled sum of the set of scalar products between the original point cloud vectors and the vectors resultant of translating the whole point cloud to a new origin, defined by the point cloud centroid: $|\mathbf{c}\rangle = N^{-1}\sum_{J}|J\rangle$.

Therefore, within the present definition of collective distance within a set of points, the result becomes equivalent to sum up $2N$ times the scalar products between the involved vectors with the same vectors submitted to a centroid origin shift.

7 Some abstract examples

In fact, collective distances can be computed for any known Gram matrix, including metric matrices, which are just Gram matrices build by linearly independent vector scalar products.

7.1 Collective distances of orthonormalized vector sets

For instance, if the involved vector collection is orthonormalized, then one will have: $\mathbf{D} = \mathbf{I} \wedge \mathbf{F} = \mathbf{0}$. Therefore, the collective distance of a given orthonormalized set of vectors can be easily written by using Eq. (4), yielding: ${}_O D_N^{(2)} = N(N-1)$. Therefore, for any other case studied with the same number of normalized vectors but not orthogonal, where the off diagonal matrix becomes non-null, Eq. (4) tells that if: $\langle \mathbf{F} \rangle \geq 0$, then: $D_N^{(2)} = N(N-1) - \langle \mathbf{F} \rangle < {}_O D_N^{(2)}$. Thus, whenever $\langle \mathbf{F} \rangle \geq 0$ holds, it can be easily stated that: sets of N normalized vectors possess their collective distance always inferior to the one, associated to the same number of orthonormalized vectors.

This is the same to say that, under the condition: $\langle \mathbf{F} \rangle \geq 0$, among all the vector sets made by an arbitrary but equal number of vectors, the orthonormalized vector sets provide the maximal collective distance.

7.2 Collective distances in simplices as a general geometric example

A simplex object, defined in any of the spaces which have been used here this far, can be considered as a set of N equidistant points constructed within a containing $N-1$ dimensional space. Thus, naming ℓ the edge length between any two vertices of the simplex, then it can be written in this case: $\forall I, J : D_{IJ}^2 = \ell^2$. The collective distance in any N vertex simplex can be obviously written as: ${}_S D_N^{(2)} = \frac{1}{2} N(N-1)\ell^2$. On the other hand, a simplex might be also considered as a polyhedron, which can be constructed from the centroid as the origin, by using N vectors starting at such origin and ending at the simplex vertices. As a simplex is a tetrahedron generalization to any N dimensional space, such vectors will have equal length λ and the subtended angles for any vector pair are all the same: θ. Then, the scalar products of these vectors can be written simply by the unique expression: $\langle P|Q \rangle = \lambda^2 \cos\theta = \lambda^2 \gamma$. Thus, the Gram matrix of this simplex centroid related vectors can be decomposed as: $\mathbf{D} = \lambda^2 \mathbf{I} \wedge \mathbf{F} = \gamma\lambda^2 (\mathbf{1} - \mathbf{I})$, which implies a collective distance which in turn might be written as:

$$\begin{aligned} {}_S D_N^{(2)} &= (N-1)\langle \mathbf{D} \rangle - \langle \mathbf{F} \rangle = N(N-1)\lambda^2 - \gamma\lambda^2 N(N-1) \\ &= N(N-1)\lambda^2 (1-\gamma). \end{aligned}$$

Therefore, one can deduce that: $\ell^2 = 2\lambda^2 (1 - \gamma) = 2\lambda^2 (1 - \cos\theta)$.

7.3 Orthonormalized sets and simplices

Using this result and considering again an orthonormalized set of vectors, one might also assume any orthonormalized set of N vectors as generating a $(N-1)$—dimensional simplex. On the other hand, an orthonormal vector set is isomorphic to the canonical basis set: $\mathrm{E}_N = \{|\mathbf{e}_I\rangle \,|I = 1, N\}$, which in turn can be associated to the columns or rows of the unit matrix $\mathbf{I}_N$. Each distance between any pair of vectors of E_N can be easily written as:

$$_E D^2_{IJ} = \langle I|I\rangle + \langle J|J\rangle - 2\langle I|J\rangle = 2 - 2\cdot 0 = 2.$$

This last result indicating that the set of orthonormalized vectors generates a $(N-1)$—dimensional simplex of Nvertices with edges measuring: $\sqrt{2}$, or $\ell^2 = 2$.

Such a situation is reproduced in the simplex description as given above, whenever using: $\lambda = 1 \wedge \theta = \frac{\pi}{2}$. This situation also produces a coherent collective distance: $_E D^{(2)}_{IJ} = {}_O D^{(2)}_{IJ}$.

Hence, one can state that: any orthonormalized set of Nvectors generates a $(N-1)$—dimensional simplex with $\ell^2 = 2$.

8 Euclidian distances (ED) between double object pairs

As customarily defined, ED's correspond to the usual distance between two objects. The collective ED definitions as discussed before can be associated to a composite distance between all the possible objects of a given set taking each non-redundant ED between object pairs one at a time.

When a set of mathematical objects is defined as belonging to some vector space: $\Omega^1_N = \{|I\rangle \,|I = 1, N\} \subset V$,then one can construct the tensor product of the $\frac{1}{2}N(N+1)$ non-redundant pairs:

$$\Omega^2_N = \{|I, J\rangle = |I\rangle \otimes |J\rangle \,|I = 1, N;\, J = I, N\} \subset V \otimes V.$$

With the tensor products as above defined, one might in turn construct the generalized scalar products: $\mathbf{S} = \{\langle I, J|K, L\rangle\}$,with the additional definition:

$$\langle I, J|K, L\rangle = \langle |I, J\rangle * |K, L\rangle\rangle\,.$$

Then, the Euclidian distances between the tensorial pairs can be obtained as usual:

$$D^2_{IJ;KL} = \langle |I, J\rangle * |I, J\rangle\rangle + \langle |K, L\rangle * |K, L\rangle\rangle - 2\langle |I, J\rangle * |K, L\rangle\rangle\,.$$

Therefore, a similar structure as in the classical case studied before can be described. For example, take a triangle in two dimensions represented by a set of three two-bit strings:

$$\mathrm{T}_3^1 = \{|a\rangle = (01); |b\rangle = (10); |c\rangle = (11)\};$$

the possible tensor pairs may be defined as the resulting bit strings:

$$\mathrm{T}_3^2 = \{|aa\rangle = (0001); |ab\rangle = (0010); |bb\rangle = (1000) \ldots |cc\rangle = (1111)\}.$$

Then, the scalar products between the elements of T_3^2 can be easily written:

$$\langle aa|aa\rangle = 1; \langle aa|ab\rangle = 0; \langle aa|bb\rangle = 0;$$
$$\langle ab|ab\rangle = 1; \langle ab|bb\rangle = 0; \langle bb|bb\rangle = 1; \ldots$$

and the Euclidian distances between some of the tensor doublets can be written in turn as:

$$D^2_{aa;ab} = \langle aa|aa\rangle + \langle ab|ab\rangle - 2\langle aa|ab\rangle = 1 + 1 - 2 \cdot 0 = 2$$
$$D^2_{aa;bb} = \langle aa|aa\rangle + \langle bb|bb\rangle - 2\langle aa|bb\rangle = 1 + 1 - 2 \cdot 0 = 2$$

and so on…

Then, a general collective distance, involving all the possible doublet distances: $\left\{D^2_{pq;rs}\right\}$, might be defined in the same way as in the already studied singleton case.

9 Euclidian distances between indeterminate object v-tuples

In the same fashion one could easily define the Euclidian distance, involving an indeterminate number of non-redundant tensor products involving v objects, using a scalar product which can be schematically written as:

$$|I_1, I_2, \ldots I_v\rangle = \bigotimes_{P=1}^{v} |I_P\rangle \rightarrow$$
$$\langle I_1, I_2, \ldots I_v|J_1, J_2, \ldots J_v\rangle = \langle |I_1, I_2, \ldots I_v\rangle * |J_1, J_2, \ldots J_v\rangle\rangle .$$

Thus, the ED between two of these v-tuples can be easily defined as:

$$D^2_{I(v);J(v)} = \langle |I_1, I_2, \ldots I_v\rangle * |I_1, I_2, \ldots I_v\rangle\rangle + \langle |J_1, J_2, \ldots J_v\rangle * |J_1, J_2, \ldots J_v\rangle\rangle$$
$$-2\langle |I_1, I_2, \ldots I_v\rangle * |J_1, J_2, \ldots J_v\rangle\rangle$$

In the same manner, such a definition is nothing else than the Euclidian norm of the difference between both v-tuples tensorial products. In order to ease the generalized scalar products symbols one can write: $|\mathbf{I}_v\rangle = |I_1, I_2, \ldots I_v\rangle$, then the ED can be expressed as:

$$D^2_{\mathbf{I}_v\mathbf{J}_v} = \left\langle (|\mathbf{I}_v\rangle - |\mathbf{J}_v\rangle)^{*2} \right\rangle = \langle (|\mathbf{I}_v\rangle - |\mathbf{J}_v\rangle) * (|\mathbf{I}_v\rangle - |\mathbf{J}_v\rangle) \rangle$$
$$= \langle \mathbf{I}_v|\mathbf{I}_v\rangle + \langle \mathbf{J}_v|\mathbf{J}_v\rangle - 2\langle \mathbf{I}_v|\mathbf{J}_v\rangle .$$

With the same idea in mind it might be also defined the generalized distance:

$$D^g_{\mathbf{I}_v\mathbf{J}_v} = \left\langle ||\mathbf{I}_v\rangle - |\mathbf{J}_v\rangle|^{*g} \right\rangle,$$

as in the case of the usual definition attached to one object cases. Care must be taken into defining the odd order g distances, as in this case one must take care of having positive definite Euclidean norms. The v-tuples Euclidian distances become in this way an extended formalism encompassing an indefinite v-tuple tensor rank and distance orders, thus including as a very particular case the single object algorithm, represented by Eq. (6).

Of course, the final collective distance of general order has to be written employing nested summation symbols [12–15]. Then, the generalized collective distance expression acquires a simple conceptual form, resembling the initial definition involving only singletons:

$$D^{(g)}_v = \sum (\mathbf{I}_v) \sum (\mathbf{J}_v)\, \delta\,(\mathbf{I}_v \neq \mathbf{J}_v)\, D^g_{\mathbf{I}_v\mathbf{J}_v}.$$

10 Collective distance between two sets of objects

The collective distances considering all the involved object elements within a set have been defined in several variants. It might be also interesting to define a composite collective ED between two object sets: {A;B}.

Suppose that the two object sets[3] are defined with the two sets of vectors: $\mathrm{A} = \{|a_I\rangle\,|I = 1, M\} \wedge \mathrm{B} = \{|b_J\rangle\,|J = 1, N\}$. The collective distance between them can be defined as:

$$D^{(2)}[\mathrm{A;B}] = {}_A D^{(2)}_M + {}_B D^{(2)}_N - \sum_{I\in A}\sum_{J\in B} D^2_{IJ}$$
$$= {}_A D^{(2)}_M + {}_B D^{(2)}_N - \sum_{I\in A}\sum_{J\in B} (\langle I; A|A; I\rangle + \langle J; B|B; J\rangle - 2\langle I; A|B; J\rangle)$$
$$= {}_A D^{(2)}_M + {}_B D^{(2)}_N$$
$$- \left(N\sum_{I\in A}\langle I; A|A; I\rangle + M\sum_{J\in B}\langle J; B|B; J\rangle - 2\sum_{I\in A}\sum_{J\in B}\langle I; A|B; J\rangle \right)$$
$$= {}_A D^{(2)}_M + {}_B D^{(2)}_N - (N\langle \mathbf{D}_A\rangle + M\langle \mathbf{D}_B\rangle - 2\langle \mathbf{P}_{AB}\rangle)$$

[3] The involved sets are free in their compositions: they might have common elements or can be disjoint. The furnished simple example below corresponds to a collective ED between two disjoint sets.

where the elements $\{\mathbf{D}_A; \mathbf{D}_B\}$ can be obtained in the ways previously discussed and the collective distance between both sets can be essentially associated to the complete sum of some matrix bearing the scalar products between the vectors of both sets, defined as:

$$\mathbf{P}_{AB} = \{P_{IJ} = \langle a_I | b_J \rangle \, | I = 1, M; \, J = 1, N\}.$$

Of course in case that: A = B, then it will occur that one can rewrite the definition of collective distance according to the two sets equality, for instance:

$$D^{(2)}\,[\mathrm{A};\mathrm{A}] = 2_A D_M^{(2)} - 2\,(M\,\langle \mathbf{D}_A \rangle - \langle \mathbf{P}_{AA} \rangle)$$

The matrix $\mathbf{P}_{AA}$ is just the Gram matrix of the set A, thus one can also write:

$$\langle \mathbf{P}_{AA} \rangle = \langle \mathbf{D}_A \rangle + \langle \mathbf{F}_A \rangle$$

And it is easy to see that this part of the collective distance becomes twice the one of the duplicated set, making zero the final expression:

$$\begin{aligned} D^{(2)}\,[\mathrm{A};\mathrm{A}] &= 2_A D_M^{(2)} - (2M\,\langle \mathbf{D}_A \rangle - 2\,(\langle \mathbf{D}_A \rangle + \langle \mathbf{F}_A \rangle)) \\ &= 2_A D_M^{(2)} - 2\,((M-1)\,\langle \mathbf{D}_A \rangle - \langle \mathbf{F}_A \rangle) = 0. \end{aligned}$$

Therefore it can be seen how the definition of the composite distance between two sets can be associated to the distance axioms of non-negativity, symmetry and reflexivity.

10.1 A simple case example of collective distance between two sets

A simple example can be sought seeking for the collective distance between a square made with two-bit vectors:

$$\mathrm{Q} = \{s_1 = (0,0)\,; s_2 = (1,0)\,, s_3 = (0,1)\,; s_4 = (1,1)\}$$

and a triangle made by three vectors like:

$$\mathrm{T} = \{t_1 = (-1,0)\,; t_2 = (0,-1)\,; t_1 = (-1,-1)\}.$$

For the square one can construct the scalar product matrix and compute the attached collective ED:

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix} \rightarrow {}_S D_4^{(2)} = 3\,\langle \mathbf{D}_\mathrm{Q} \rangle - \langle \mathbf{F}_\mathrm{Q} \rangle = 3 \cdot 4 - 4 = 8$$

and for the triangle it is also easily obtained:

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix} \rightarrow {}_T D_3^{(2)} = 2\langle \mathbf{D}_\mathrm{T} \rangle - \langle \mathbf{F}_\mathrm{T} \rangle = 2 \cdot 4 - 4 = 4.$$

Then, for the composite scalar product matrix between both sets it can be written the scalar product matrix and the complete sum:

$$\mathbf{K}_\mathrm{QT} = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & -1 \\ -1 & -1 & -2 \end{pmatrix} \rightarrow \langle \mathbf{K}_\mathrm{QT} \rangle = -8$$

Therefore, the collective distance between both sets can be written as:

$$D^{(2)}[\mathrm{Q;T}] = 3 \times 4 + 4 \times 4 - 2(-8) = 12 + 16 + 16 = 44.$$

11 Quantum similarity and collective distances between quantum point clouds

The concept of quantum point cloud (QPC) has been used since the initial times of QS development, see for example references [16,17], but as commented before a recent revision of terms and concepts around the QS definitions and procedures has been recently published [4] too.

Resuming all this previous work, QPC can be considered the geometrical image of QOS, as the DF tag set of any QOS can be considered as the vectors defining an infinite dimensional polyhedron possessing as many vertices as the cardinality N of the QOS. When using the DF forming the QPC, as a source of a QS matrix, it has been recently proven that this procedure constructs not only a discrete QOS (DQOS), but the columns of the QS matrix become a $N-$dimensional hologram of the original QPC [18].

Therefore, collective distances involving QPC can be of some interest when studying QOS either internal or external relationships. As QS matrices can be considered as Gram matrices of their DF elements, the procedures put forward in the previous paragraphs can be evidently used in a QS environment.

11.1 QS application example: some molecular sets QPC collective distances

A short program within the Molecular Quantum Similarity Program Suite (MQSPS) has been constructed in order to test the feasibility of the proposed collective distance algorithms. Five molecular sets already studied in a previous work on molecular superposition algorithms and QS [19] already used in MQSPS implementation, have been chosen as tests for computing Euclidian collective distances. The following résumé provides the figures obtained:

#	Mol. set ID	#Molecules	Collective distance obtained with QS Matrices via Eq. (6)	Arithmetic mean of collective distance	SQRT of collective distance
1	F and Cl Methanes	15	0.19492812E+06	1,856.4583	441.50664
2	Curious Molecules	10	0.78099801E+05	1,735.5511	279.46342
3	Cramer Steroids	21	0.28781891E+06	1,370.5662	536.48757
4	Assorted Flavonoids	12	0.94462718E+05	1,431.2533	307.34788
5	Varied Molecules	8	0.22072712E+05	788.31113	148.56888

The collective Euclidian distances produced seem depending on the number of molecules hold within the set. According to the magnitude of the collective distance one can order the five studied sets as: $3 \succ 1 \succ 4 \succ 2 \succ 5$. Using the arithmetic mean (considering the number of distances employed) it is obtained the ordering: $1 \succ 2 \succ 4 \succ 3 \succ 5$. The square root of the collective distance follows again the same trend as encountered with the bulk distance: $3 \succ 1 \succ 4 \succ 2 \succ 5$. Thus, one can conclude that the collective distances are somehow associated to the number of elements in the QPC, but their arithmetic means are probably related to the dominant magnitudes of the distances into the set. Whatever the case can be, collective distances constitute a simple way to obtain global information on QOS and QPC.

12 Conclusions

Collective distances over vector sets have been described within a simple formal collection of mathematical descriptions, leading to algorithms easily implemented with the philosophy of the MQSPS. Apparently this is the first time that collective Euclidian distances are described in the literature. They can constitute a global assessment, which can be associated to a given vector set of molecular descriptors. The collective distance calculations can be easily implemented in QS framework, providing extra information on QOS via the geometrical picture provided throughout QPC.

Acknowledgements This work was elaborated and finished during the author's stage at RIKEN, Kobe, Japan. The hospitality of Prof. K. Hirao is highly appreciated.

References

1. R. Carbó-Dorca, J. Math. Chem. **47**, 331–334 (2010)
2. L.D. Mercado, R. Carbó-Dorca, J. Math. Chem. **49**, 1558–1572 (2011)
3. R. Carbó-Dorca, J. Math. Chem. **49**, 2109–2115 (2011)
4. R. Carbó-Dorca, E. Besalú, J. Math. Chem. **50**, 210–219 (2012)
5. E.E. Hodgkin, W.G. Richards, Int. J. Quant. Chem. **32**(S14), 105–110 (1987)
6. R. Carbó-Dorca, J. Math. Chem. **50**, 734–740 (2012)
7. M.M. Deza, E. Deza, *Encyclopedia of Distances* (Springer, Berlin, 2009)
8. S.K. Berberian, *Introduction to Hilbert Space* (Oxford University Press, New York, 1961)
9. R. Carbó, L. Leyda, M. Arnau, Intl. J. Quant. Chem. **17**, 1185–1189 (1980)
10. P. Bultinck, X. Gironés, R. Carbó-Dorca, in *Molecular Quantum Similarity: Theory and Applications*; Rev. Comput. Chem. vol. 21, eds. K.B. Lipkowitz, R. Larter, T. Cundari (Wiley, Hoboken, USA, 2005) pp. 127–207

11. R. Carbó-Dorca, A. Gallegos, *Quantum Similarity and Quantum QSPR (QQSPR)*, in Entry: 176, Encyclopedia of Complexity and Systems Science, vol. 8, ed. R. Meyers (Springer, New York, 2009) pp. 7422–7480
12. R. Carbó, E. Besalú, J. Math. Chem. **13**, 331–342 (1993)
13. R. Carbó, E. Besalú, Comp. Chem. **18**, 117–126 (1994)
14. E. Besalú, R. Carbó, J. Math. Chem. **18**, 37–72 (1995)
15. E. Besalú, R. Carbó, *Applications of Nested Summation Symbols to Quantum Chemistry:Formalism and Programming Techniques*. In "Strategies and Applications in Quantum Chemistry: from Astrophysics to Molecular Engineering" An Hommage to Prof. eds. G. Berthier, M. Defranceschi, Y. Ellinger (Kluwer Academic Publisher, Amsterdam 1996) pp. 229–248
16. R. Carbó, B. Calabuig, Intl. J. Quan. Chem. **42**, 1681–1693 (1992)
17. R. Carbó, B. Calabuig, J. Chem. Inf. Comp. Sci. **32**, 600–606 (1992)
18. R. Carbó-Dorca, IQC Technical Report TR-2012-1, J. Math. Chem. (2012). doi:10.1007/s10910-012-0034-6
19. R. Carbó-Dorca, E. Besalú, L.D. Mercado, J. Comp. Chem. **32**, 582–599 (2011)