THE CHIRAL INDEX: APPLICATIONS TO MULTIVARIATE DISTRIBUTIONS AND TO 3D MOLECULAR GRAPHS

Michel Petitjean
MTi, INSERM UMR-S 973, University Paris 7
35 rue Hélène Brion, 75205 Paris Cedex 13, France.
petitjean.chiral@gmail.com
http://petitjeanmichel.free.fr/itoweb.petitjean.html

Abstract: We review the main properties of the chiral index. Its use as an asymmetry coefficient of multivariate probablity distributions is pointed out, and its application to measure the degree of chirality of rigid 3D molecular graphs is presented. Several extreme chirality sets are shown. Some open optimization problems are mentioned.

Keywords: chirality and symmetry measures, chiral index, asymmetry coefficient, colored mixture, colored Wasserstein distance, 3D molecular graphs.

1 INTRODUCTION

The historical definition of chirality is due to Lord Kelvin [5]: *I call any geometrical figure, or group of points, chiral, and say that it has chirality if its image in a plane mirror, ideally realized, cannot be brought to coincide with itself.* In other intuitive words, an object *identical* to one of its mirror images is achiral, i.e. not chiral: it has indirect symmetry. Despite what is believed since a long, the full mathematical definition of chirality does not relie on the existence of some oriented space. It is based on a general symmetry definition [17] and involves only basic group theory concepts [19]. In this paper we deal with a quantitative measure of the deviation from indirect symmetry. That problem goes back to the end of the 19^{th} century. It was of interest first for chemists and statisticians, but contributors from many fields are known (see [13] for a review). Although measuring the degree of asymmetry of the probability distribution of some random variable or vector is basically a geometric problem, the case of molecules is more complicated, even under assumption of a rigid model. To see this, we consider a simplified model of the molecule CHBrClF (bromochlorofluoromethane) with five ponctual atoms, four of them (H, F, Cl, Br) being the vertices of a regular tetrahedron with the fifth atom (C) at the center of the tetrahedron. Geometrically speaking we have an achiral object, but any chemist would say that this molecular object is chiral because a valid superposition of the molecule with any of its mirror image is expected to superpose an H atom with an H atom and so on with the four other atom types, and no valid superposition respecting these five constraints exists. The general situation for molecules is in fact more complicated because the labeling of the atoms does not depend only on their nature: it depends on the full molecular graph, where the punctual atoms are colored nodes, and the chemical bonds are colored edges. E.g., the graph of the water molecule H–O–H has three nodes and two edges. Such molecular graphs are of common use in chemistry [2, 6, 8]. The chiral index presented hereafter applies both to 3D molecular graphs and to multivariate distributions, discrete or continuous.

2 THE COLORED MIXTURE MODEL

A general process to define an indirect asymmetry coefficient of a multivariate distribution consists to consider a probability metric, and then to minimize the distance between the distribution and any of its indirect isometry image for all rotations and translations of that image.

The asymmetry coefficient is got via an adequate normalization of this minimized distance. Here, the L^2-Wasserstein distance D [3, 20] is considered: X_1 and X_2 being two random vectors in R^d, w being an element of the space W of their joint distributions and the quote denoting the transposition operator, then

$$D^2 = Inf_{\{w \in W\}} E[(X_1 - X_2)'(X_1 - X_2)] \tag{1}$$

In order to handle pairwise correspondences as required in chemistry, we first consider a probability space (C, A, P), where C ia a non empty set called the space of colors, A is a σ-algebra defined on C, and P is a probability measure. Then we define a mapping Φ from C on the space of probability distributions on (R^d, B), where B is the Borel σ-algebra of R^d. In other words, to each color $c \in C$ is associated a d-variate distribution $\tilde{P}_c = \Phi(c)$. The random variable (K, X) in the compound space $(C \times R^d, A \otimes B)$ is called a colored mixture [12, 14] because its distribution is viewed as a variant of the usual mixture distributions concept [4]. Then, considering a couple of random variables $(K_1, X_1), (K_2, X_2)$, the fundamental assumption of the colored mixture model is:

$$K_1 \overset{a.s.}{=} K_2 \tag{2}$$

It means that once a color is selected, we get two random vectors X_1 and X_2 which in general are not independant, and the set W_c of their joint distributions is a non empty subset of W introduced in eq. 1. The colored Wasserstein distance D_c is [12, 14]:

$$D_c^2 = Inf_{\{w \in W_c\}} E[(X_1 - X_2)'(X_1 - X_2)] \tag{3}$$

The case where C is of finite cardinality n is of interest. When $n = 1$, D_c and D coincide. For any n, when (a) the mixing distribution of K_1 (or K_2) is uniform, and (b) the mixed distributions are those of almost surely constant random vectors, D_c is the distance induced by the Frobenius norm, and this distance, minimized for some class of transformations of X_2 (e.g. linear, orthogonal, etc.), is the Procrustes distance [12]. This latter, with or without minimization for isometries of X_2, is called in the 3D case *RMS* or *RMSD* by many chemists and structural biologists. The colored mixture model is also a framework for defining shape complementarity and was used to define a geometric docking criterion when the expectation is replaced by a variance operator in the right member of eq. 3 [11, 14].

3 THE CHIRAL INDEX

The chiral index χ was introduced for finite sets in 1997 [7]. Then it was extended to weighted sets [10] before receiving its more general definition in 2002 for a colored mixture of finite inertia T [12], this inertia being referred to the marginal in R^d. The squared colored Wasserstein distance D_c^2 between a colored mixture and its image through any indirect isometry applied to its marginal in R^d (e.g. a mirror reflection), is minimized for all translations t and rotations R of the image, and then a normalization factor is applied so that $\chi \in [0; 1]$:

$$\chi = d \cdot [Inf_{\{R,t\}} D_c^2]/4T \tag{4}$$

The chiral index depends only on the distribution of the colored mixture and it is insensitive to isometries and scaling. It is null if and only if the distribution is indirect symmetric. The optimal translation is null for a centered distribution, and the optimal rotation is analytically known for $d = 2$ and $d = 3$ [12]. A direct symmetry index was defined for finite sets of points [9], but it cannot work for continuous distribution (see the discussion at the end of ref. [13]).

3.1 An asymmetry measure of multivariate distributions

When C is of cardinality 1, there is only one color and χ is an asymmetry coefficient of the distribution of the random vector associated to this unique color. In the unidimensional case, the chiral index of a distribution is expressable from the lower bound r_m of the correlation coefficient between two random variables following that distribution, taken over the space of their joint distributions:

$$\chi = (1 + r_m)/2 \tag{5}$$

Because r_m cannot be positive, in eq. 5 we have $\chi \in [0; 1/2]$. The chiral index should be compared with the skewness M_3, i.e. the reduced third order centered moment of the distribution. This latter is often presented as an asymmetry coefficient, and is such that $M_3^2 \leq M_4 - 1$, M_4 being the reduced fourth order centered moment [21, 23]. That inequality is itself a trivial consequence of equation A10 in [14] for a random vector G of null expectation:

$$Var(G'G) \geq E(GG'G) \cdot [E(GG')]^{-1} \cdot E(GG'G) \tag{6}$$

Unfortunately, the skewness can be null even for indirect symmetric distributions (see section 4.2 in [13]), although χ is null if and only if the distribution is achiral. Remark: an univariate *symmetric* distribution should be called achiral, because it has a mirror symmetry. An other advantage of χ over the skewness and its multivariate analogs is that χ is defined even when the third order moments do not exist.

From the convergence theorem section IV in [12], the sample chiral index is a consistent estimator of the chiral index of the parent distribution. Then, a class of open problems is to find simple asymptotic expressions of the distribution of the sample chiral index under hypothesis of interest for the experimentalist about the parent population, such as normality, uniformity, or else, in order to build symmetry tests.

In the case of a sample of n reals, r_m is got via correlating the ordered sample sorted in increasing order with the one sorted in decreasing order, and χ in eq. 5 is very easy to compute with a pocket calculator. Furthermore, χ offers simple expressions of the squared midranges or of the squared range lengths of the ordered sample (see section 2.9 in [13]).

Setting $d = 1$ and $n = 3$, and denoting by α the ratio of the lengths of the two adjacent segments defined by the three points, the chiral index is:

$$\chi = (1 - \alpha)^2/4(1 + \alpha + \alpha^2) \tag{7}$$

For this set, the chiral index satisfies to five properties:

1. χ is function of only the unique parameter of the set
2. χ is a continuous function of α
3. $\chi(1) = 0$
4. $\chi = 0 \Rightarrow \alpha = 1$
5. $\chi(\alpha) = \chi(1/\alpha)$ (invariance for scaling)

It has been emphasized in [16] that any safe chirality measure should first satisfy to the five properties above for this set, which is the simplest possible non trivial test set. By far it is not the case of many ones encoutered in the literature [13].

3.2 Colored sets and chemical graphs

The mechanism provided in section 2 permits to handle the constraints on pairwise correspondences (i.e. selecting permutations) between two sets of n points. When this constraint is relaxed, we are left to compute the Wasserstein distance between two uniform discrete distributions of n points, which needs to minimize the expectation in eq. 1 over the $n!$ pairwise correspondences. In the general case (e.g. continuous distributions), it is recalled that the constraints apply to a set of joint distributions. For molecules, the most used model is an undirected simple graph, where the nodes are colored by the Mendeleiev nature of the atoms and the edges are colored by the nature of the chemical bonds [6]. Molecular graphs are realized in R^3, and are assumed to be connected and rigid in the present framework.

In a molecular graph, a node x_2 is equivalent to a node x_1 when x_2 is the image of x_1 through a graph automorphism. The equivalence of all n nodes in a molecular graph does not mean that there are $n!$ automorphisms: e.g. consider a ring of 6 carbons with 6 single bonds such as in the cyclohexane squeleton, there are only 12 automorphisms, not 6!. For a general molecular graph, computing the chiral index needs to enumerate the permutations P associated to the graph automorphisms and to find the optimal rotation R for each permutation [7]. Let Y be the the array of n lines and d columns containing the coordinates of the n points, assumed to be centered, i.e. the mean of the n points is null. Q being an arbitrary negative determinant orthogonal matrix, the chiral index is:

$$\chi = \frac{d}{4Tr(Y'Y)} Min_{\{P,R\}}[Tr(Y - PYQ'R')'(Y - PYQ'R')] \tag{8}$$

For a molecular graph $d = 3$, and the optimal rotation R is known analytically [9].

4 SOME EXTREME CHIRALITY DISTRIBUTIONS

In eq. 4, a necessary condition to reach the upper bound $\chi = 1$ is to have the covariance matrix V proportional to the identity [12], i.e., σ being some positive real:

$$V = \sigma^2 I \tag{9}$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ be the eigenvalues of V and let us consider n equiprobable points with not two having the same color. The chiral index is [9]:

$$\chi = d\lambda_d / Tr(V) \tag{10}$$

In this situation, $\chi = 0$ iff the set is subdimensional and $\chi = 1$ iff eq. 9 is satisfied, which is the case for the regular simplex, the d-cube, etc. The most chiral triangles (i.e. sets of $n = 3$ points in the plane) have been computed [7]. When the 3 points have 3 different colors, it is equilateral. When 2 points have the same color and the last one has an other color, the squared side lengths ratios of the optimal triangle are $1 : 1 - \sqrt{6}/4 : 1 + \sqrt{6}/4$, and $\chi = 1 - \sqrt{2}/2$. When the 3 points have the same color, these ratios are $1 : 4 + \sqrt{15} : (5 + \sqrt{15})/2$ and $\chi = 1 - 2\sqrt{5}/5$. These three triangles are shown fig. 1. It can be checked from their cartesian coordinates given in [7] that they satisfy to the following property: each squared side length is proportional to three times a squared distance vertex-barycenter. That property appears also for the two triangles maximizing the direct symmetry index defined in [9]. It is symmetrical for all permutations of the 3 vertices only in the case of the equilateral triangle.

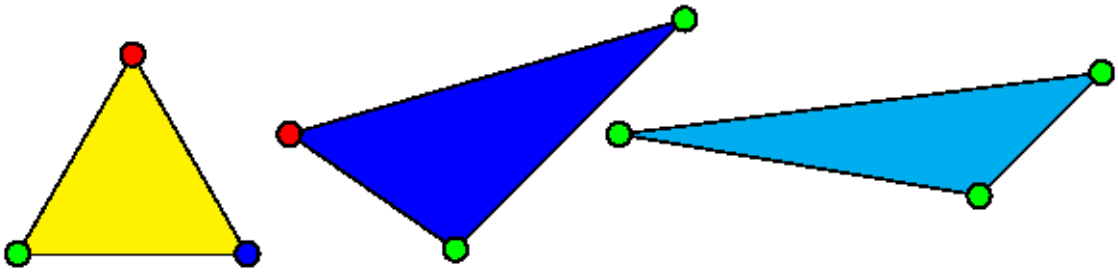

Figure 1: The maximal chirality triangles. From left to right, three different colors on vertices, two vertices with the same color, and three vertices with the same color.

We look now for the upper bound $\chi^*(d)$ of the chiral index in the case there is only one color, i.e. in the case of d-variate distributions (in fact, no need of color here). We get the following results for χ_1^*, χ_2^* and χ_d^* ($d \geq 1$) respectively from refs. [12], [1] and [18]:

$$\chi_1^* = 1/2 \tag{11}$$
$$\chi_2^* \in [1-1/\pi; 1-1/2\pi] \tag{12}$$
$$\chi_d^* \in [1/2; 1] \tag{13}$$

The Bernoulli distribution with parameter tending to 0 or to 1 has a chiral index tending to χ_1^* [12]. For $d \geq 2$, finding χ_d^* is an open problem. As mentioned in sect. 3.1, the sample chiral index is a consistent estimator of the parent population chiral index, so that χ_d^* can be seeked among samples of increasing size n. The case $d = 2$ is of interest. Defining $z \in C^n$, $z = x + \mathbf{i}y$, where x and y are the vectors in R^n of the marginals of the bidimensional sample, and P being the permutation matrix associated to their joint distribution matrix P/n, it is known that ithe optimal P is symmetric and the chiral index takes a simple expression [1]:

$$\chi = 1 - [Max_{\{P\}}|z'Pz|]/\|z\|^2 \tag{14}$$

Let Y be the matrix $[x|y]$, and μ_1 and μ_2 be the eigenvalues of $Y'PY$ ($\mu_1 \geq \mu_2$). Eq. 14 can be rewritten:

$$\chi = 1 - [Max_{\{P\}}(\mu_1 - \mu_2)]/Tr(Y'Y) \tag{15}$$

It was conjectured in [1] that $\chi_2^* = 1 - 1/\pi$ and a family of distributions in which the chiral index can be arbitrarily close to $1 - 1/\pi$ was exhibited.

The 3D molecular graph of an hydrocarbon designed by A. Schwartz [22] has, among several remarkable properties, a chiral index of 0.9824 and its carbon skeleton has $\chi = 1.0000$.

An attempt to define the closest achiral distribution to a given chiral one was done[15], but no satisfactory general approach to that problem is known.

References

[1] Coppersmith, D., Petitjean, M., 2005. *About the Optimal Density Associated to the Chiral Index of a Sample from a Bivariate Distribution*. Compt. Rend. Acad. Sci. Paris, Série I, **340**[8],599–604.

[2] Diudea, M.V., Petitjean, M., 2008. *Symmetry in Multi-Tori.*. Symmetry Cult. Sci. **19**[4], 285–305.

[3] Dobrushin, R.L., 1970. *Prescribing a system of random variables by conditional distributions*. Theor. Probab. Appl. **15**[3], 458–486.

[4] Everitt, B.S., Hand, D.J., 1981. *Finite Mixture Distributions*. Chap. 1, Chapman and Hall, London.

[5] Lord Kelvin, 1904. *Baltimore Lectures on Molecular Dynamics and the Wave Theory of Light*, Appendix H., chap. 22, footnote p. 619. C.J. Clay and Sons, Cambridge University Press Warehouse, London.

[6] Petitjean, M., 1992. *Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds*. J. Chem. Inf. Comput. Sci. **32**[4],331–337.

[7] Petitjean, M., 1997. *About Second Kind Continuous Chirality Measures. 1. Planar Sets*. J. Math. Chem. **22**[2-4],185–201.

[8] Petitjean, M., 1999. *Calcul de chiralité quantitative par la méthode des moindres carrés*. Compt. Rend. Acad. Sci. Paris, Série IIc, **2**[1],25–28.

[9] Petitjean, M., 1999. *On the Root Mean Square Quantitative Chirality and Quantitative Symmetry Measures*. J. Math. Phys. **40**[9],4587–4595.

[10] Petitjean, M., 2001. *Chiralité quantitative: le modèle des moindres carrés pondérés*. Compt. Rend. Acad. Sci. Paris, Série IIc, **4**[5],331–333.

[11] Petitjean, M., 2002. *Solving the Geometric Docking Problem for Planar and Spatial Sets*. Internet Electron. J. Mol. Des. **1**[4],185–192.

[12] Petitjean, M., 2002. *Chiral mixtures*. J. Math. Phys. **43**[8],4147–4157.

[13] Petitjean, M., 2003. *Chirality and Symmetry Measures: A Transdisciplinary Review*. Entropy **5**[3],271–312.

[14] Petitjean, M., 2004. *From Shape Similarity to Shape Complementarity: toward a Docking Theory*. J. Math. Chem. **35**[3],147–158.

[15] Petitjean, M., 2006. *À propos de la référence achirale*. Compt. Rend. Chim. **9**[10],1249–1251.

[16] Petitjean, M., 2006. *Minimal Symmetry, Random and Disorder*. Symmetry Cult. Sci. **17**[1-2], 197–205.

[17] Petitjean, M., 2007. *A Definition of Symmetry*. Symmetry Cult. Sci. **18**[2-3], 99–119.

[18] Petitjean, M., 2008. *About the Upper Bound of the Chiral Index of Multivariate Distributions*. AIP Conf. Proc. **1073**, 61–66.

[19] Petitjean, M., 2010. *Chirality in Metric Spaces*. Symmetry Cult. Sci. **21**[1-3], 27–36.

[20] Rachev, S.T., 1991. *Probability Metrics and the Stability of Stochastic Models*. Chap. 6, Wiley, New York.

[21] Rohatgi, V.K., Székely, G.J., 1989. Sharp Inequalities between Skewness and Kurtosis. *Stat. Prob. Lett.*, **8**, 297–299.

[22] Schwartz, A., Petitjean, M., 2008. *[6.6]Chiralane: A Remarkably Symmetric Chiral Molecule*. Symmetry Cult. Sci. **19**[4], 307–316.

[23] Wilkins, J.E. (1944). A Note on Skewness and Kurtosis. *Ann. Math. Stat.* **15**, 333–335.