Invariants of molecules, including proteins

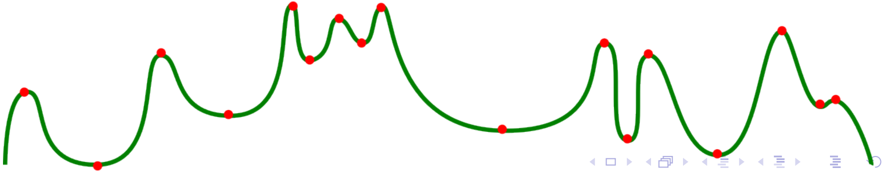
open-table discussion with new students
and colleagues in applied areas

THE ROYAL SOCIETY

Sorites paradox (of a heap of sand)

If a heap is reduced by a single grain at a time, when does it cease to be considered the [same] heap?

Any *discrete classification* of *continuous values* creates discontinuities similar to splitting our planet into countries by artificial boundaries, which can be avoided via *continuous invariants* playing the role of geographic-style coordinates.



Descriptors vs invariants

Real objects are often described by ambiguous *descriptors*, e.g. lists of x, y, z coordinates, that easily change under important equivalences. An **invariant** I is a function (property) whose values are preserved under a given equivalence.

If molecules $S \cong Q$ are exactly matched under rigid motion, then $I(S) = I(Q)$. Equivalently, if $I(S) \neq I(Q)$, then $S \not\cong Q$ are rigidly different.

The number m of atoms is invariant under rigid motion. A photo is a descriptor, not an invariant.

Invariants distinguish objects

An *invariant* is a function $I : \{\text{equivalence classes of objects}\} \to$ a simpler space such that if $A \sim B$ then $I(A) = I(B)$ or, equivalently,

if $I(A) \neq I(B)$ then $A \not\sim B$ meaning that I has

no false negatives = no pairs of equivalent inputs (representations) $A \sim B$ with $I(A) \neq I(B)$.

The size of a cloud is an isometry invariant. The center of mass is not invariant under translation. Any other invariants of unordered point clouds?

(In)complete invariants

An invariant can be weak, such as the number of atoms, distinguishing some (not all) objects.

An invariant I is called *complete* if I guarantees equivalence: if $I(A) = I(B)$, then $A \sim B$, so

I distinguishes all non-equivalent objects or has *no false positives*, which means no pairs of different $A \not\sim B$ with equal values $I(A) = I(B)$.

What is a complete invariant of a set of 2 points under rigid motion in the plane? For 3 points?

Equivariance vs invariance

Let a group, say $G = E(n)$ of isometries in Euclidean space $\mathbb{R}^n$, act on (finite) point sets.

A function $h : \{\text{point sets}\} \to$ a simpler space is G-**equivariant** if $h(g(C)) = T_g(h(C))$, where T_g is a map depending on g, e.g. T_g is g acting on the mid-point $h(C)$ between closest neighbours.

The *stronger* (restrictive) concept is **invariance** when T_g is the identity. For a point set $S \subset \mathbb{R}^n$, its centre of mass is equivariant, not invariant.

Do invariants suffice? Yes!

Non-constant invariants distinguish some objects. All non-invariant descriptors don't.

non-invariant descriptors

non-invariant **equivariants**

$\neq$

non-constant isometry **invariants** of periodic point sets

generically **complete, polynomial-time** computable

continuous under noise **reconstructable**

Equivariants are used to predict forces (vectors at atoms) that move one point set to another.

Any such sequence of (rigid classes of) sets $C_t \subset \mathbb{R}^n$ depending on a time t can be studied in terms of *only invariants* $I(C_t)$ without vectors.

Easy case: ordered points

If points $p_1, \dots, p_m \in \mathbb{R}^n$ are **ordered**, they can be *reconstructed* from all distances $|p_i - p_j|$ or scalar products $p_i \cdot p_j$, uniquely under isometry.

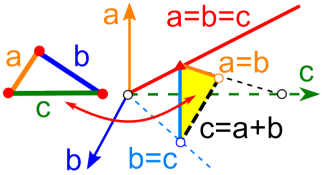
A faster invariant of protein backbones exposed thousands of duplicate chains in the PDB, see Anosova et al. MATCH v.94(1), p.97-134, 2025.

In practice, many point clouds are **unordered**.

The brute-force way to compare clouds of m unordered points by $m!$ distance matrices is *unrealistic* because of the exponential time.

Euclid's ideal solution for triangles

SSS theorem for $m = 3$ points in any $\mathbb{R}^n$. Two triangles are congruent (isometric) *if and only if* they have the same triple of sides a, b, c (under all 6 permutations). For rigid motion (without reflections), allow only 3 cyclic permutations.



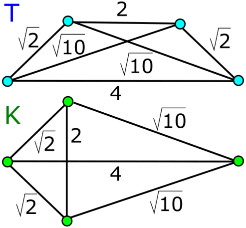
A complete isometry invariant of all triangles lives in the cone $\{0 < a \le b \le c \le a + b\} \subset \mathbb{R}^3$ bi-continuously parametrised by inter-point distances a, b, c.

Generically complete invariants

Is the problem *open for quadrilaterals* in $\mathbb{R}^2$?

One can train neural networks to experimentally output isometry invariants, but it can be hard to prove completeness and continuity under noise.

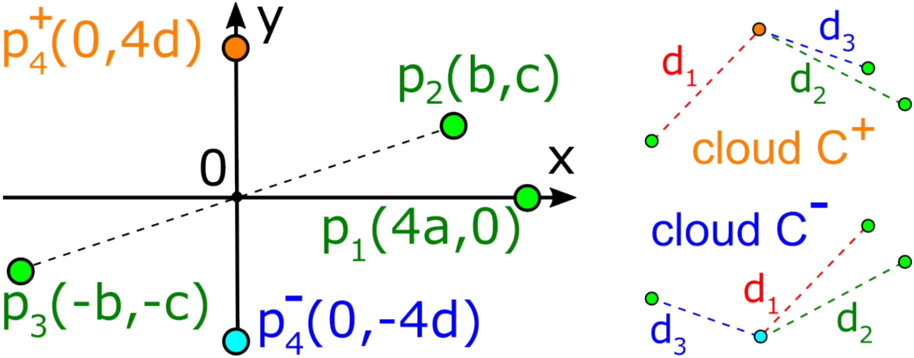
Boutin, Kemper, 2004: the vector of all sorted pairwise distances is *generically complete* in $\mathbb{R}^n$



distinguishing almost all clouds of unordered points except singular examples. These non-isometric clouds have the same 6 pairwise distances.

Pairs of singular quadrilaterals

The 5-dimensional space of 4-point clouds has non-isometric $\{p_1, p_2, p_3, p_4^{\pm}\}$ with the same 6 pairwise distances depending on 4 parameters.



Are there *stronger invariants* of point clouds?

Summary: equivalences, invariants

Any well-defined classification requires an **equivalence relation** (choices are possible).

Manual labels apply only to a labeled dataset.

Objects described by real numbers should be distinguished under all perturbations to avoid trivial classifications by the transitivity axiom.

An **invariant** is a property preserved under a given equivalence and has *no false negatives*.

A **complete invariant** has *no false positives*.