Geometric Data Science challenges and solutions

Vitaliy Kurlin
Department of Computer Science, University of Liverpool, Liverpool, United Kingdom

Abstract

This paper outlines the forthcoming book introducing the new area of *Geometric Data Science* (GDS), which develops polynomial-time algorithms for continuous metrics and geographic-style maps for spaces of equivalence classes of real data objects modulo practically important relations.

The major achievement of GDS is the *Crystal Isometry Principle* extending Mendeleev's table of elements to a continuous *Crystal Isometry Space* containing all known and not yet discovered periodic crystals considered up to isometry and canonically parameterized by complete invariants.

2012 ACM Subject Classification Theory of computation → Computational geometry

Keywords and phrases point cloud, periodic point set, isometry, invariant, metric, continuity

Funding *Vitaliy Kurlin*: EPSRC grant 'Application-driven Topological Data Analysis' (EP/R018472/1, 2018-2023, joint with the University of Oxford, UK), Royal Academy of Engineering Industrial Fellowship ' Data Science for Next Generation Engineering of Solid Crystalline Materials' (IF2122/186, 2021-2023), EPSRC grant 'Inverse design of periodic crystals' (EP/X018474/1, 2022-2024).

1 Computable complete invariants for finite clouds of unlabeled points

The essential examples of data and equivalence are a cloud of m points and an isometry in $\mathbb{R}^n$, which preserves inter-point distances, hence the rigidity of real structures from molecules to sculptures. The first step is to build complete invariants that should be also continuous under perturbation of data in a suitable metric. The continuity requirement is the key novelty needed for reliable comparisons and optimization, also motivated by noisy measurements.

If all m points have distinct labels, the classical distance matrix is a complete invariant uniquely determining a cloud of m labeled points up to isometry in $\mathbb{R}^n$. If the given points are unlabeled, all permutations of m points make comparisons of $m!$ distance matrices too costly. Hence the following problem for finite unlabeled clouds was highly non-trivial.

▶ **Problem 1.1** (mapping continuous spaces of isometry classes)**.** Find a complete isometry invariant I of finite clouds of unlabeled points in $\mathbb{R}^n$ with a continuous metric d. In detail,

(1.1a) *invariance* : if point clouds $A \cong B$ are *isometric* in $\mathbb{R}^n$ (meaning that $f(A) = B$ for an *isometry* $f : \mathbb{R}^n \to \mathbb{R}^n$ preserving all Euclidean distances $||f(p) - f(q)|| = ||p - q||$ for any points $p, q \in \mathbb{R}^n$), then $I(A) = I(B)$, so the invariant I has *no false negatives*;

(1.1b) *completeness* : if $I(A) = I(B)$, then $A \cong B$ are isometric, so I has *no false positives*;

(1.1c) a *metric* d on invariant values should satisfy all metric axioms below :
first axiom : $d(I(A), I(B)) = 0$ if and only if point clouds $A \cong B$ are isometric,
symmetry : $d(I(A), I(B)) = d(I(B), I(A))$,
triangle inequality : $d(I(A), I(C)) \le d(I(A), I(B)) + d(I(B), I(C))$ for all A, B, C;

(1.1d) *continuity* : for any point cloud $A \subset \mathbb{R}^n$ and $\varepsilon > 0$, there is $\delta > 0$ such that if B is obtained by perturbing every point of A within its δ-neighborhood, then $d(I(A), I(B)) < \varepsilon$.

(1.1e) *computability* : for a fixed dimension n, the invariant $I(A)$ and the metric $d(A, B)$ can be exactly computed in a polynomial time in the number m of points in $A, B \subset \mathbb{R}^n$. ■

One popular isometry invariant of finite clouds of unlabeled points is persistent homology in Topological Data Analysis. If we consider standard filtrations of Vietoris-Rips, Cech, Delaunay complexes, then persistent homology is invariant up to isometry of a cloud A, not up to more general deformations, and cannot distinguish generic families of inputs [5, 3, 13].

The simpler and generically complete isometry invariant is the Pair Distribution Function or the ordered list of distances between all points of A. Fig. 1 (left) shows one of the infinitely many counter-examples to completeness: a pair of non-isometric sets with the same list of pairwise distances. The stronger *local distributions of distances* [11] is similar to Pointwise Distance Distribution (PDD) with a continuous Earth Mover's Distance (EMD), see Definition 2.1 and [12]. Though PDD is conjectured to be complete for discrete sets up to isometry in $\mathbb{R}^2$, [15, Fig. 6] provides a counter-example to completeness in $\mathbb{R}^3$.

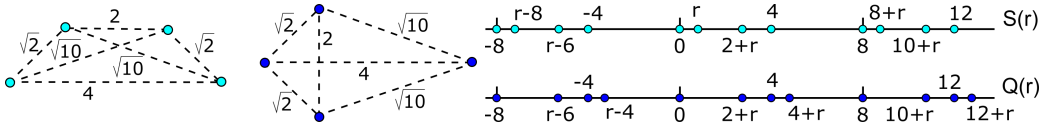

Figure 1 Left: point sets $K = \{(\pm 2, 0),\ (\pm 1, 1)\}$ and $T = \{(\pm 2, 0), (-1, \pm 1)\}$ can not be distinguished by their six pairwise distances $\sqrt{2}, \sqrt{2}, 2, \sqrt{10}, \sqrt{10}, 4$. **Right**: 1D periodic sets $S(r) = \{0, r, 2 + r, 4\} + 8\mathbb{Z}$ and $Q(r) = \{0, 2 + r, 4, 4 + r\} + 8\mathbb{Z}$ for $0 < r \le 1$ have the same Pair Distribution Function. All these pairs are distinguished by the PDD invariant in Definition 2.1.

Problem 1.1 was recently solved by the Principal Coordinates Invariant (PCI) in general position and Weighted Matrices Invariant (WMI) in all cases, see [8, Theorems 3.5 and 5.4]. For m-point clouds in $\mathbb{R}^n$, a continuous metric on PCIs is computed in time $O(m^{1.5}(\log^n m)2^n)$, see [8, Theorems 4.6 and 4.7]. The WMIs have a metric computable in time $O(m^{2n-0.5}\log^n m + m^{3n-3})$ for clouds of exactly m points. For $n = 2$, the time $O(m^{3.5}\log m)$ improves the time $O(m^5 \log m)$ of the *only exact algorithm* [4] for the Hausdorff distance under Euclidean motion. Any clouds of up to m points can be compared by EMD on WMIs in time $O((m^{2n+1} + nm^{3n-3})\log m)$, see [8, Theorems 6.3 and 6.6].

2 The new Crystal Isometry Principle for all real periodic crystals

The simplest periodic set is a *lattice* $\Lambda \subset \mathbb{R}^n$ consisting of all integer linear combinations $\sum_{i=1}^{n} c_i v_i$ of a basis $v_1, \dots, v_n$ whose vectors span the *unit cell* $U = \{\sum_{i=1}^{n} c_i v_i : 0 \le c_i < 1\}$.

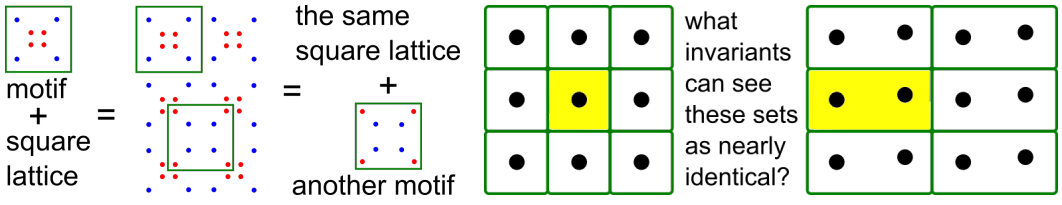

Figure 2 Left : even for a fixed cell of a lattice Λ, different motifs M can define isometric periodic sets $\Lambda + M$. **Right**: for almost any perturbation, the symmetries and (the minimum volume of) any reduced cell discontinuously change, which justifies continuity (1.1d) in Problem 1.1 for periodic sets.

Any *periodic point set* S is a sum $\Lambda + M = \{\vec{u} + \vec{v} : u \in \Lambda, v \in M\}$, where a *motif* M is a finite set of points in the basis of U. Basis vectors of U and atomic coordinates of motif points (atomic centers) in M form a conventional Crystallographic Information File (CIF). Fig. 2 (left) shows the ambiguity of the CIF pair (Λ, M) even if a basis of U is fixed.

Edelsbrunner et al [6] initiated continuous classifications of periodic point sets up to isometry (as in Problem 1.1 for finite clouds) motivated by the rigidity of crystal structures.

▶ **Definition 2.1** (Pointwise Distance Distribution PDD as introduced in [15]). For a motif of points $M = \{p_1, \dots, p_m\}$ in a unit cell U of a lattice Λ, let $S \subset \mathbb{R}^n$ be a finite set coinciding with M or a periodic set $S = \Lambda + M$. For an integer $k \geq 1$, consider the $m \times k$ matrix $D(S;k)$, whose i-th row consists of the ordered distances $d_{i1} \leq \dots \leq d_{ik}$ measured from p_i to its first k nearest neighbors in the full set S. The rows of $D(S;k)$ are *lexicographically* ordered as follows. A row $(d_{i1}, \dots, d_{ik})$ is *smaller* than $(d_{j1}, \dots, d_{jk})$ if a few first distances coincide: $d_{i1} = d_{j1}, \dots, d_{il} = d_{jl}$ for $l \in \{1, \dots, k-1\}$ and the next $(l+1)$-st distances satisfy $d_{i,l+1} < d_{j,l+1}$. If w rows are identical to each other, any such group is collapsed to one row with the *weight* w/m. For each row, put this weight in the first column. The final $m \times (k+1)$-matrix is the *Pointwise Distance Distribution* $\mathrm{PDD}(S;k)$. ■

The matrix $D(T;3)$ in Table 1 has two pairs of identical rows, so $\mathrm{PDD}(T;3)$ consists of two rows of weight $\frac{1}{2}$ below. The matrix $D(K;3)$ in Table 1 has only one pair of identical rows, so $\mathrm{PDD}(K;3)$ has three rows of weights $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{4}$. Then $\mathrm{PDD}(T;3) \neq \mathrm{PDD}(K;3)$

Table 1 Each point in $T, K \subset \mathbb{R}^2$ from Figure 1 has ordered distances to three other points.

T points	neighb.1	neighb.2	neighb.3
$(-2,0)$	$\sqrt{2}$	$\sqrt{10}$	4
$(+2,0)$	$\sqrt{2}$	$\sqrt{10}$	4
$(-1,1)$	$\sqrt{2}$	2	$\sqrt{10}$
$(+1,1)$	$\sqrt{2}$	2	$\sqrt{10}$

K points	neighb.1	neighb.2	neighb.3
$(-2,0)$	$\sqrt{2}$	$\sqrt{2}$	4
$(+2,0)$	$\sqrt{10}$	$\sqrt{10}$	4
$(-1,-1)$	$\sqrt{2}$	2	$\sqrt{10}$
$(-1,+1)$	$\sqrt{2}$	2	$\sqrt{10}$

$$\mathrm{PDD}(T;3) = \left(\begin{array}{c|ccc} 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/2 & \sqrt{2} & \sqrt{10} & 4 \end{array}\right) \neq \mathrm{PDD}(K;3) = \left(\begin{array}{c|ccc} 1/4 & \sqrt{2} & \sqrt{2} & 4 \\ 1/2 & \sqrt{2} & 2 & \sqrt{10} \\ 1/4 & \sqrt{10} & \sqrt{10} & 4 \end{array}\right).$$

[15, Theorem 3.2]: $\mathrm{PDD}(S;k)$ is an isometry invariant of a periodic point set $S \subset \mathbb{R}^n$;

continuity in [15, Theorem 4.3]: for any $k \geq 1$, if each point of S is perturbed within its ε-neighborhood, $\mathrm{PDD}(S;k)$ changes up to 2ε in the Earth Mover's Distance (EMD).

generic completeness in [15, Theorem 4.4]: any periodic point set $S \subset \mathbb{R}^n$ in general position can be uniquely reconstructed up to isometry from a lattice of S (or lattice invariants from [10, 9] for $n = 2, 3$), the size m of a motif, and $\mathrm{PDD}(S;k)$, where k is big enough so that all distances in the last column of $\mathrm{PDD}(S;k)$ are larger than the double covering radius of S.

[15, Theorem 5.1]: for a fixed dimension n, using fast nearest neighbors [7], $\mathrm{PDD}(S;k)$ can be computed in time $O(km(5\nu)^n V_n \log(m) \log^2(k))$, where m is the size of a motif of $S \subset \mathbb{R}^n$, V_n is the unit ball volume in $\mathbb{R}^n$, d and $\nu = \frac{d}{\sqrt[n]{\mathrm{Vol}[U]}}$ are the diameter and *skewness* of U.

The near-linear time algorithm above completed 200+ billion pairwise comparisons of all 660K+ real periodic crystals in the Cambridge Structural Database (CSD), the world's largest dataset of known materials. This huge experiment took only two days on a modest desktop and detected five pairs of identical 'needles in a haystack', see [14, section 7], [15]. In each pair, the crystals are truly isometric to the last decimal place but one atom is replaced with a different atom, for example, Cd with Mn for the CSD entries HIFCAB vs JEPLIA.

Since this replacement is physically impossible without perturbing atomic neighbors, five journals are investigating the integrity of the underlying articles. Traditional comparisons would take 22+ thousand years on the same machine. Even if some singular structures are not distinguished by PDD, the cubic time isoset invariant [1] is provably complete.

Much more importantly, the above experiment has justified the *Crystal Isometry Principle* saying that the map {periodic crystal} → {periodic point sets} is injective modulo isometry.

So any real periodic crystal is uniquely determined by the geometry of its atomic centers without chemical data. Hence all periodic crystals have unique locations in a common *Crystal Isometry Space* continuously parameterized by complete invariants. Fig. 3 shows maps [10] with geographic-style coordinates for 2.6+ million 2D lattices in real CSD crystals [2].

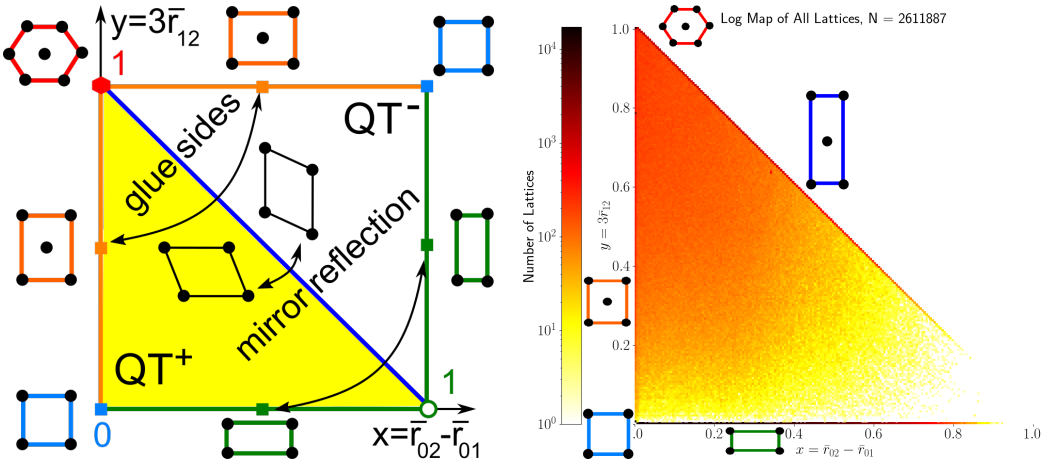

Figure 3 Left: any 2D lattice up to rigid motion and uniform scaling has a unique obtuse superbase with $v_0 + v_1 + v_2 = 0$, *root products* $r_{ij} = \sqrt{-v_i \cdot v_j}$, and complete invariant (x, y) in the unit square. **Right**: the log-scale heat map of all 2D lattices extracted from CSD crystals.

References

1 O Anosova and V Kurlin. An isometry classification of periodic point sets. In *Proceedings of Discrete Geometry and Mathematical Morphology*, 2021.

2 M Bright, A Cooper, and V Kurlin. Geographic-style maps for 2D lattices. *arxiv:2109.10885*. URL: `http://kurlin.org/projects/periodic-geometry-topology/lattices2Dmap.pdf`.

3 M Catanzaro, J Curry, B Fasy, J Lazovskis, G Malen, H Riess, B Wang, and M Zabka. Moduli spaces of morse functions for persistence. *J Applied Comp. Topology*, 4(3):353–385, 2020.

4 P Chew, M Goodrich, D Huttenlocher, K Kedem, J Kleinberg, and D Kravets. Geometric pattern matching under Euclidean motion. *Computational Geometry*, 7:113–124, 1997.

5 Justin Curry. The fiber of the persistence map for functions on the interval. *Journal of Applied and Computational Topology*, 2(3):301–321, 2018.

6 H Edelsbrunner, T Heiss, V Kurlin, P Smith, and M Wintraecken. The density fingerprint of a periodic point set. In *Proceedings of SoCG*, pages 32:1–32:16, 2021.

7 Yury Elkin. New compressed cover tree for k-nearest neighbors (PhD). *arxiv:2205.10194*.

8 V Kurlin. Computable complete invariants for finite unlabeled point clouds. *arXiv:2207.08502*.

9 V Kurlin. A complete isometry classification of 3-dimensional lattices. *arxiv:2201.10543*, 2022.

10 V Kurlin. Mathematics of 2-dimensional lattices. *arxiv:2201.05150*, 2022.

11 Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

12 Y Rubner, C Tomasi, and L Guibas. The Earth Mover's Distance as a metric for image retrieval. *Int. J Computer Vision*, 40:99–121, 2000.

13 Philip Smith and Vitaliy Kurlin. Families of point sets with identical 1D persistence. *arxiv:2202.00577*, 2022.

14 D Widdowson, M Mosca, A Pulido, V Kurlin, and A Cooper. Average minimum distances of periodic point sets. *MATCH Communications in Math. Comp. Chemistry*, 87:529–559, 2022.

15 Daniel Widdowson and Vitaliy Kurlin. Resolving the data ambiguity for periodic crystals. In *NeirIPS: Neural Information Processing Systems (arXiv:2108.04798)*, 2022.