Geometric anomaly detection in data

Bernadette J. Stolz[a], Jared Tanner[a,b], Heather A. Harrington[a,b], and Vidit Nanda[a,b,1]

[a]Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom; and [b]The Alan Turing Institute, British Library, London NW1 2DB, United Kingdom

Edited by Tom C. Lubensky, University of Pennsylvania, Philadelphia, PA, and approved July 2, 2020 (received for review February 1, 2020)

The quest for low-dimensional models which approximate high-dimensional data is pervasive across the physical, natural, and social sciences. The dominant paradigm underlying most standard modeling techniques assumes that the data are concentrated near a single unknown manifold of relatively small intrinsic dimension. Here, we present a systematic framework for detecting interfaces and related anomalies in data which may fail to satisfy the manifold hypothesis. By computing the local topology of small regions around each data point, we are able to partition a given dataset into disjoint classes, each of which can be individually approximated by a single manifold. Since these manifolds may have different intrinsic dimensions, local topology discovers singular regions in data even when none of the points have been sampled precisely from the singularities. We showcase this method by identifying the intersection of two surfaces in the 24-dimensional space of cyclo-octane conformations and by locating all of the self-intersections of a Henneberg minimal surface immersed in 3-dimensional space. Due to the local nature of the topological computations, the algorithmic burden of performing such data stratification is readily distributable across several processors.

stratification inference | singularities | persistent cohomology

The *manifold hypothesis* (1) forms a cornerstone of modern data science; it assumes that the points in typical datasets tend to cluster near some unknown manifold of dimension substantially lower than the ambient dimension of the data. Manifold learning and dimensionality reduction techniques (2) rely on this assumption in order to infer faithful low-dimensional representations of high-dimensional data. Examples of such methods include a) classical *principal component analysis* (3, 4), where the approximating manifold is an affine subspace; b) *visual perception* (5), where continuous changes of configurations of an object yield smoothly varying changes along a curved manifold; c) *subspace clustering* (6), where data are clustered into disjoint sets that are well approximated by affine subspaces; and d) *generative adversarial networks*, which naturally produce data on pairs of manifolds (7). In sharp contrast to this profusion, one encounters a remarkable dearth of techniques designed for the analysis of data sampled from non-manifold, or singular, spaces. Among the simplest examples of singular spaces are unions of two manifolds along a common submanifold (as shown in Fig. 1); these arise organically when more than one class of data is present in the same set of observations. Recent techniques for the analysis of such heterogeneous data [see, for instance, *capsule networks* (8)] have focused primarily on coherently fusing together the multiple data classes.

The present work is motivated by an antipodal philosophy—singular regions of spaces that underlie modern datasets are inherently interesting, they will play an increasingly important role in the future of data analysis, and it is therefore of paramount importance to be able to detect these singularities directly from the data points. The task of fitting singular spaces to data is rendered difficult by the generic lack of observations which are located exactly on the geometric anomalies. For instance, most natural ways of sampling finitely many points from the space in Fig. 1 will not produce even a single point lying on the anomalous circle. Singular spaces occur quite naturally in several areas of data science—for instance, low-rank matrix approximation amounts to optimization over the determinantal variety of bounded-rank matrices, which is not a manifold (ref. 9, lecture 9). Moreover, even the simplest machine learning architectures, such as the multilayer perceptron, exhibit singularities in their parameter spaces (ref. 10, ch. 12.2). Here, we describe an algorithm that can detect singular regions from finitely many data points even when the points only lie near—rather than precisely on—the singularities. Our approach is based on *local cohomology* (11) and the theory of *stratifications* (12), which form particularly rich and fruitful enterprises in the study of singular spaces that arise in algebraic topology (13) and geometry (14). Recent computational advances in these fields (15, 16) have made it possible to bring this formidable theory to bear on the very concrete task of analyzing data that live on, or even near, spaces that are far more complicated than manifolds. Most of this existing machinery requires defining equations for a space in order to construct a stratification; in sharp contrast, here we only make use of a finite point sample.

Manifolds of dimension n are characterized by the requirement that a small neighborhood around each point should resemble the n-dimensional Euclidean disk (up to a standard equivalence relation called homeomorphism). While there can be no algorithmic procedure to determine whether two n-manifolds are homeomorphic or not (17) for $n > 4$, algebraic topology offers recourse to several rigorous descriptors for testing weaker forms of equivalence. Among the best known

Significance

The problem of fitting low-dimensional manifolds to high-dimensional data has been extensively studied from both theoretical and computational perspectives. As datasets get more heterogeneous and complicated, so must the spaces that are used to approximate them. Stratified spaces, built out of manifold pieces coherently glued together, form natural candidates for such geometric models. The key difficulty encountered when fitting stratified spaces to data is that none of the sampled points can be expected to lie exactly on the low-dimensional singular strata. The present work uses local cohomology to overcome this difficulty. Here, we describe an efficient and practical framework for singularity detection from finite samples and demonstrate its ability to detect interfaces in real and simulated data.

Author contributions: B.J.S., J.T., H.A.H., and V.N. designed research; B.J.S., J.T., H.A.H., and V.N. performed research; B.J.S. analyzed data; and B.J.S., J.T., H.A.H., and V.N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Database deposition: Both the cyclo-octane and the Henneberg datasets have been made available at GitHub (https://github.com/stolzbernadette/Geometric-Anomalies/tree/master/Data_Sets). Our Matlab implementation of the geometric anomaly detection algorithm is available at GitHub (https://github.com/stolzbernadette/Geometric-Anomalies).

[1]To whom correspondence may be addressed. Email: nanda@maths.ox.ac.uk.

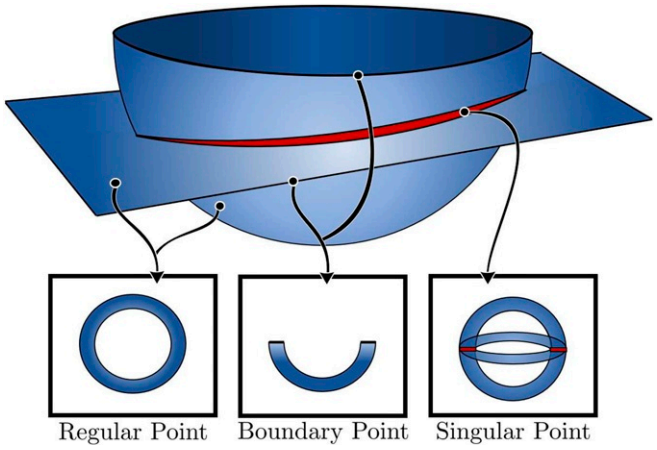

Fig. 1. Annular neighborhood classes A_x of several points x in union of a hemisphere with a plane (both blue) along a circle (red). *Regular points*, which lie away from this red circle and from the boundaries, have A_x, which looks like a standard annulus. All points lying in the boundary have A_x, which resembles a half or quarter annulus, as depicted in the central panel. All points x on the red circle itself have neighborhoods A_x, which resemble two annuli glued along two edges, as indicated in the right panel. The dimensions of $\mathbf{H}^1(A_x)$ count the number of independent loops in A_x, so from left to right, these are one, zero, and three, respectively.

computable homeomorphism invariants is *cohomology*, which assigns a sequence $\mathbf{H}^i(X)$ of vector spaces to a given topological space X. Although cohomology does not distinguish between Euclidean disks of different dimensions (all of these have the same cohomology as that of a point), it is an excellent tool for distinguishing n-dimensional spheres $\mathbb{S}^n$ from each other across different choices of n. Indeed, for all $n > 0$ and $i > 0$, we have

$$\dim \mathbf{H}^i(\mathbb{S}^n) = \begin{cases} 1 & \text{if } i = n, \\ 0 & \text{otherwise.} \end{cases}$$

Since the boundary of an n-dimensional disk is an $(n-1)$ -dimensional sphere, our strategy for detecting singular regions in a dataset P of points in Euclidean space $\mathbb{R}^n$ is as follows: we fix two real parameters $0 < r < s$, and around each point x of P, we examine the subset of *annular neighbors* A_x of x—this set consists of those points y in P whose Euclidean distance to x satisfies $r \leq \|x - y\| \leq s$, and it forms a discrete proxy for the boundary of a neighborhood around x. We then compute the cohomology of A_x at multiple scales (often called the *persistent cohomology* of A_x) and use this information to quantify whether or not A_x approximates a single sphere of some fixed dimension. If the answer is negative, then—provided we have made judicious choices of r and s—the point x lies near a singular region of X.

Results

Local persistent cohomology successfully identifies all of the non-manifold regions in two completely different datasets whose underlying spaces are known to admit singularities. The first of these is the *conformation space of the cyclo-octane molecule* C_8H_{16}. A single molecule consists of eight carbon atoms arranged in a ring, with each carbon atom being bound to two other carbon atoms and two hydrogen atoms. Under the influence of external chemical and physical forces, cyclo-octane assumes different forms, or *conformations*, in three-dimensional (3D) space. The locations of hydrogen atoms are completely determined by those of the carbon atoms, so each conformation may be represented by a point in $\mathbb{R}^{24}$ (i.e., three spatial coordinates for each of the eight carbon atoms). The space of all possible conformations forms the union of a Klein bottle and a sphere along two circles (18, 19). It is known that the conformations located on the sphere component are constrained by a specific type of symmetry, while the conformations on the Klein bottle are not (ref. 18, section IIIC). We consider points sampled from this conformation space and depict (a two-dimensional [2D] projection of) the partition of data points by local persistent cohomology in Fig. 2—points lying near the two singular circles are indeed distinct from all other points.

Our second dataset is obtained by uniformly sampling points from the nonorientable *Henneberg minimal surface*, which is an immersion of the punctured 2D projective space in standard 3D space. The results are depicted in Fig. 3: again, the points that lie near the four self-intersections are manifestly separated from manifold-like points and boundary points. Additional details involving both datasets, including an explicit parameterization of the Henneberg minimal surface, can be found in *Materials and Methods*.

Discussion

The two singular circles in the cyclo-octane conformation space were originally discovered using a local version of principal component analysis (PCA). This method uses spectral techniques to fit affine subspaces to local neighborhoods of data points (18, 19). While such methods work remarkably well for detecting intersections of flat manifolds (i.e., manifolds with curvature almost zero), they tend to require extremely dense samples and very small local neighborhood sizes in the presence of high curvature. Our approach differs substantially from such affine embedding techniques because cohomology, being a purely topological invariant, remains largely agnostic to the vagaries of local geometry such as curvature. Thus, it identifies dimensionally anomalous regions correctly even in highly curved regimes. As another pleasant side effect of the relative coarseness of cohomology, one obtains a far greater degree of robustness to the choice of neighborhood size than for local PCA. This latter phenomenon is illustrated for the cyclo-octane dataset in Fig. 4.

The enormous quantities of heterogeneous data being generated by modern experimental tools demand a concordant increase in the variety and sophistication of available geometric models. The procedure described here enables us to transcend the ubiquitous manifold hypothesis by allowing us to fit singular spaces to datasets. Aside from the data-dependent choice of radius parameters r and s, which determine the sizes of annular neighborhoods A_x, the method described here is entirely unsupervised. Moreover, it enjoys three remarkably convenient properties for our purposes. First, it can be iterated to discover more refined singularities of lower dimension: for instance, had the red points from Fig. 2 formed a singular space of their own (such as a figure eight rather than disjoint circles), we could have repeated our cohomological clustering operation on the subset of red points to isolate the points lying near the lower singularities. Second, the local cohomology computations that form the backbone of this procedure are easily distributed across a host of processors: the persistent cohomology of annular neighborhoods A_x and A_y for distinct points x and y in a dataset can—and should—be computed in parallel. Third, since persistent cohomology is stable with respect to bounded noise (20), the clustering produced by this method inherits a degree of robustness to perturbations of the original dataset. Similarly, local cohomology may be useful for the detection of bifurcations and phase transitions in certain high-dimensional dynamical systems.

Materials and Methods

Detailed accounts of the first three topics described below may be found in the textbooks of Hatcher (21), Oudot (22), and Kirwan and Woolf (12), respectively.

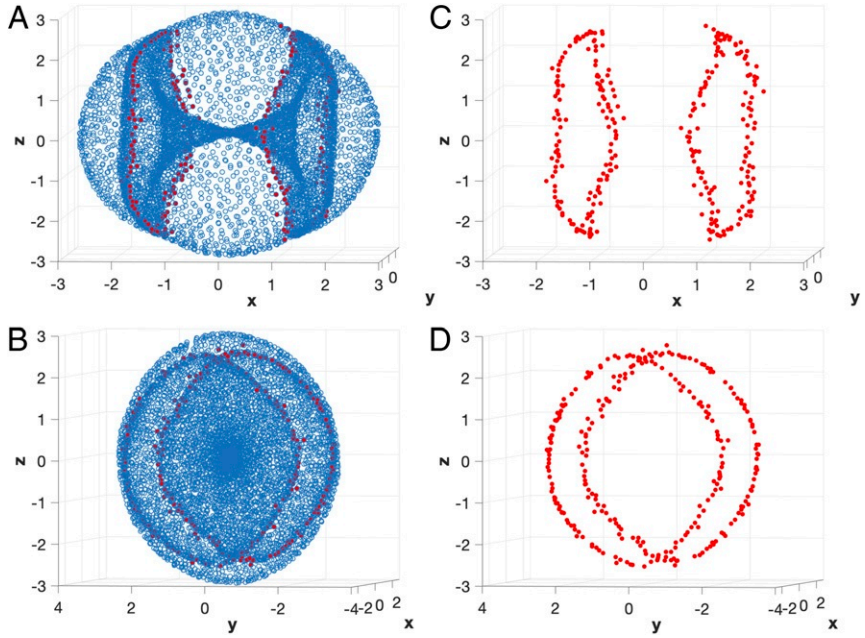

Fig. 2. Two-dimensional depiction of the 3D IsoMAP projection (4) of points sampled from the 24-dimensional conformation space of cyclo-octane. Points x for which $\dim \mathbf{H}^1(A_x) > 1$ have been colored red, and these clearly appear to cluster near the two embedded circles where the two surfaces intersect. We show the full set of points in *A* and *B* and additionally highlight the intersection points identified by our method separately in *C* and *D*, using the same perspectives as *A* and *B*, respectively. The perspective in *B* and *D* corresponds to a counterclockwise rotation around the z axis ($< 90^\circ$) and a counterclockwise rotation around the x axis ($< 45^\circ$) of the perspective in *A* and *C*.

The Cohomology of Simplicial Complexes. A *simplicial complex* K is a collection of subsets of a finite set V (usually called the set of vertices) satisfying the following condition: if $\sigma \subset V$ is in K and $\tau \subset \sigma$, then τ is also in K. The dimension of a simplex σ is one less than its cardinality, and the set of all i-dimensional simplices in K is denoted $K(i)$. The most familiar simplicial complexes are graphs, where $K(0)$ and $K(1)$ correspond to vertices and edges, respectively. For each i-dimensional simplex σ, denote by $1_\sigma : K(i) \to \mathbb{R}$ the characteristic function, which evaluates to 1 on σ and 0 on all other simplices. The vector space obtained by treating all such characteristic functions as an orthonormal basis is written $\mathbf{C}^i(K)$ and called the space of *i-cochains*. It is possible to construct a sequence of *coboundary operators* $\delta^i : \mathbf{C}^i(K) \to \mathbf{C}^{i+1}(K)$ with the following matrix representation in our chosen basis: the entry in 1_σ's column and 1_τ's row equals ± 1 if $\sigma \subset \tau$ and is 0 otherwise. It is always possible to choose signs of the nonzero entries consistently so that the kernel of δ^i contains the image of δ^{i+1}, and the i-th *cohomology* of K is the quotient vector space $\mathbf{H}^i(K) = \ker \delta^i / \operatorname{img} \delta^{i-1}$.

Cohomology is an extremely well-studied (21) descriptor of simplicial complexes and related spaces; it enjoys many wonderful properties, but only two of them are relevant to our purposes here. First, it is a *homeomorphism invariant*, meaning that any two different triangulations of the same space X will produce identical cohomologies even though the cochain spaces and coboundary operators might be wildly different. For instance, the cohomology vector spaces of an n sphere depend neither on geometric intricacies (such as its radius or its embedding in Euclidean space) nor on the combinatorics of a particular choice of simplicial decomposition. Second, cohomology is *functorial* with respect to the subcomplex relation among simplicial complexes. A subset L of simplices in K is called a subcomplex if it happens to be a simplicial complex in its own right. Whenever L is a subcomplex of K, there are well-defined linear maps $\mathbf{H}^i(K) \to \mathbf{H}^i(L)$ induced on the associated cohomology vector spaces.

The Persistent Cohomology of Data. Given a finite dataset P embedded in Euclidean space $\mathbb{R}^n$ and a scale parameter $t \geq 0$, the *Vietoris–Rips* simplicial complex $\underline{\mathrm{VR}}_t(P)$ contains as its i-dimensional simplices all subsets $\{p_0, \ldots, p_i\}$ of P whose pairwise Euclidean distances $\|p_j - p_k\|$ are no larger than t. It follows that $\underline{\mathrm{VR}}_t(P)$ is a subcomplex of $\underline{\mathrm{VR}}_u(P)$ whenever $t \leq u$. By the functoriality of cohomology, in each dimension $i \geq 0$ we obtain not only a one-parameter family of cohomology vector spaces

$$V(t) = \mathbf{H}^i\left(\underline{\mathrm{VR}}_t(P)\right)$$

but also, a compatible family of induced linear maps $V(u) \to V(t)$ for all pairs of real numbers $t \leq u$. Such collections of vector spaces and linear maps indexed by the positive real numbers are called *persistence modules*, and their systematic study—which forms the theoretical core of topological data analysis—has been greatly facilitated by three miraculous properties.

The first property is algebraic—although persistence modules appear to involve an infinite amount of information prima facie, any V arising from the Vietoris–Rips cohomology of a finite dataset $P \subset \mathbb{R}^n$ is completely determined by a finite collection $\mathrm{Bar}(P)$ comprising certain half-open subintervals of $\mathbb{R}$, called the *barcode* of P. The second property is computational; barcodes can be extracted via elementary matrix algebra, and there are several software packages dedicated to their efficient computation (23). The third crucial property of persistence modules is geometric and takes the form of a *stability theorem* (20). Roughly, this result asserts that if the points of P are perturbed by an amount $\epsilon > 0$, then the intervals in $\mathrm{Bar}(P)$ also have their end points shifted by no more than ϵ. As a consequence, one can conclude that Vietoris–Rips persistent cohomology barcodes are robust to the presence of bounded noise in the original dataset.

Stratified Spaces. Singular spaces, such as algebraic varieties and quotients of group actions on manifolds, are often analyzed via their *stratifications*. We remark that most stratifications are derived from algebraic or analytic equations, rather than data. Each stratification $Y_\bullet$ of an n-dimensional space Y is an ascending sequence of closed subspaces

$$\varnothing = Y_{-1} \subset Y_0 \subset Y_1 \subset \cdots \subset Y_{n-1} \subset Y_n = Y,$$

where the connected components of successive differences $Y_i - Y_{i-1}$, called the *i-strata*, are open i-dimensional submanifolds of Y. Every simplicial complex, for instance, admits a natural stratification whose i strata are precisely the *i-simplices*. It is customary to impose two additional constraints on the strata in order to render the study of stratified spaces tractable. The first requirement, called the *frontier axiom*, ensures that the set of all strata is partially ordered by the boundary relation $\sigma \leq \tau$ whenever the closure of τ intersects σ (this mirrors the ordering on simplices given by the containment relation $\sigma \subset \tau$). The second requirement, called *equisingularity* or *normal triviality*, imposes severe topological constraints on intersections of small neighborhoods in Y around various points of a single i-stratum with the higher strata Y_j for $j \geq i$.

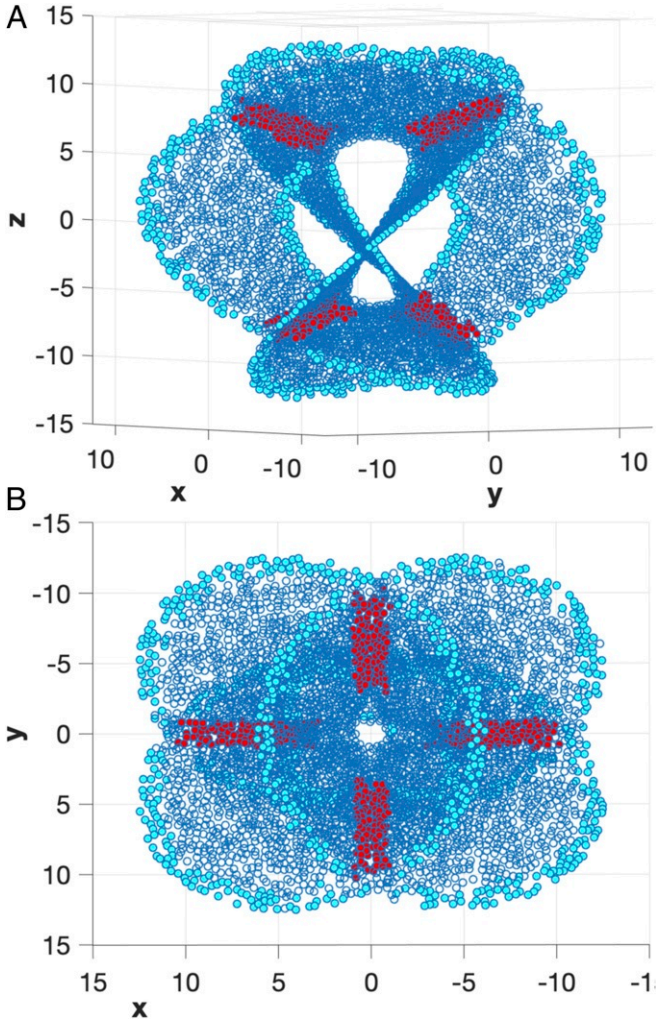

Fig. 3. Two-dimensional projections of points sampled from Henneberg's minimal surface immersed in 3D space. Points x for which $\dim \mathbf{H}^1(A_x) > 1$ are shown in red, and these lie along the four self-intersections. Similarly, points x for which $\dim \mathbf{H}^1(A_x) = 0$ have been colored cyan and appear near the boundary. The perspective in B corresponds to a counterclockwise rotation around the z axis ($< 90^\circ$) and a counterclockwise rotation around the x axis (90°) of the perspective in A; we indicate the x, y, and z axes to facilitate comparison.

As a consequence of equisingularity, to each i-stratum σ one can assign a single $(n-i-1)$ -dimensional stratified space $L_\bullet$, called the *link* of σ, so that the following property holds. For each point y in σ and all choices of small neighborhoods $U_y \subset Y$ of y, the intersection of U_y with higher strata Y_j admits a tangent $\times$ normal decomposition of the form

$$U_y \cap Y_j = \mathbb{R}^i \times \mathrm{Cone}(L_{j-i-1}),$$

where $\mathrm{Cone}(L_\bullet)$ is the quotient of $L_\bullet \times [0,1)$ obtained by identifying all pairs of the form $(\ell, 0)$ with a single point. When $j = i$, we have $U_y \cap \sigma = \mathbb{R}^i$, thus guaranteeing that σ is an i-dimensional manifold. Additionally, for $j = n$, we have $U_y = \mathbb{R}^i \times \mathrm{Cone}(L)$, so it follows that the homeomorphism type—and hence, the cohomology—of the boundary ∂U_y is independent of the choice of y in σ. This is the key property of stratified spaces, which is used in our algorithm to identify singular regions within datasets. In this discrete setting, we have no direct access to ∂U_y for a given data point y; however, we are able to approximate its cohomology via the persistent cohomology of all of the data points lying within an annular neighborhood A_y of y.

Datasets. The cyclo-octane dataset, which was introduced by Martin et al. (18), consists of 6,040 points in $\mathbb{R}^{24}$ subsampled from a far larger dataset containing over a million cyclo-octane conformations. This dataset is publicly available as part of the JAVAPLEX software package (24). The Henneberg surface dataset was provided by Martin and Watson (19); it consists of 5,456 points sampled from the Henneberg surface using the following parametrization:

$$x = \frac{2(\beta^2-1)\cos(\phi)}{\beta} - \frac{2(\beta^6-1)\cos(3\phi)}{3\beta^3},$$
$$y = -\frac{6\beta^2(\beta^2-1)\sin(\phi) + 2(\beta^6-1)\sin(3\phi)}{3\beta^3},$$
$$z = \frac{2(\beta^4+1)\cos(2\phi)}{\beta^2},$$

where $\beta \in [0.4, 0.6]$ and $\phi \in [0, 2\pi]$. In this range of β values, the surface does not have triple intersections.

Algorithm and Implementation. The procedure *Geometric Anomaly Detection* discovers intersections of dimension $(k-1)$ from points sampled on k-dimensional submanifolds of $\mathbb{R}^n$ for $n > k$. The key step, as indicated previously, is the calculation of persistent cohomology of annular neighborhoods around data points and testing whether the number of sufficiently long intervals in the barcode for dimension $(k-1)$ is 0, 1, or larger. The partition produced by *Geometric Anomaly Detection* decomposes the original dataset P into the k-manifold points P_{man}, the boundary points P_{bnd}, and the desired intersection points P_{int}. We have implemented this procedure in MATLAB for surfaces (i.e., for the case $k = 2$) using the inbuilt function RANGESEARCH to compute the annuli A_y. The persistent cohomology calculations are performed using the RIPSER software package (25). The annulus parameters (r, s) equal (0.4, 0.25) for the cyclo-octane data and (2, 1.5) for the Henneberg surface data. The projections of Fig. 2 were obtained by initializing IsoMAP (4) with five nearest neighbors.

***Algorithm*: *Geometric Anomaly Detection*.**
In: Finite point set $P \subset \mathbb{R}^n$, real parameters $0 < r < s$.
Out: A partition of P into subsets P_{man}, P_{bnd}, and P_{int}.

```
initialize P_man, P_bnd, and P_int to ∅
for all y ∈ P
  find A_y = {x ∈ P satisfying r ≤ ||x − y|| ≤ s}
  compute Bar_{k−1}(A_y), the (k − 1) -dim barcode of A_y
  calculate N_y = #{[a, b) ∈ Bar_{k−1}(A_y) with (b − a) >(s − r)}
  if N_y is 0
    add y to P_bnd
  else if N_y is 1
    add y to P_man
  else
    add y to P_int
  end if
end for
return
```

This algorithm as written will also cast points lying in certain nongeneric intersections of k manifolds into P_{int}. For instance, consider the union of paraboloids $\pm z = x^2 + y^2$, which is singular at $z = 0$. The nonsingular subsets $\{z > 0\}$ and $\{z < 0\}$ share a limiting tangent plane at the origin and hence, do not intersect transversely. Given a sufficiently dense point sample lying on this paraboloid (say for x and y constrained within the square $[-1, 1]^2$), points near the origin will admit annular neighborhoods that resemble two disjoint circles, one in each nonsingular subset. Therefore, such points will have at least two prominent bars in their one-dimensional local persistent cohomology barcode and hence, will be correctly flagged as lying in P_{int}. In practice, one does not expect to encounter such nongeneric singularities precisely because they are not robust to perturbation. Small fluctuations in the defining equation of this paraboloid may, for example, either turn the singular set into a circle or make it disappear entirely.

More relevant for practical applications is the fact that this algorithm can be suitably iterated in order to also detect certain lower-dimensional singularities as follows. Since the subset of points $P_{\mathrm{man}} \subset P$ is expected to lie on a union of k-dimensional manifolds, removing them produces a subset of points lying on a union of manifolds, all of which have dimension strictly smaller than k. Thus, we may apply *Geometric Anomaly Detection* a second time, with P being replaced by $(P - P_{\mathrm{man}})$ and k by $(k-1)$ throughout. This allows us to discover singular strata of dimension $\leq (k-2)$ and so forth.

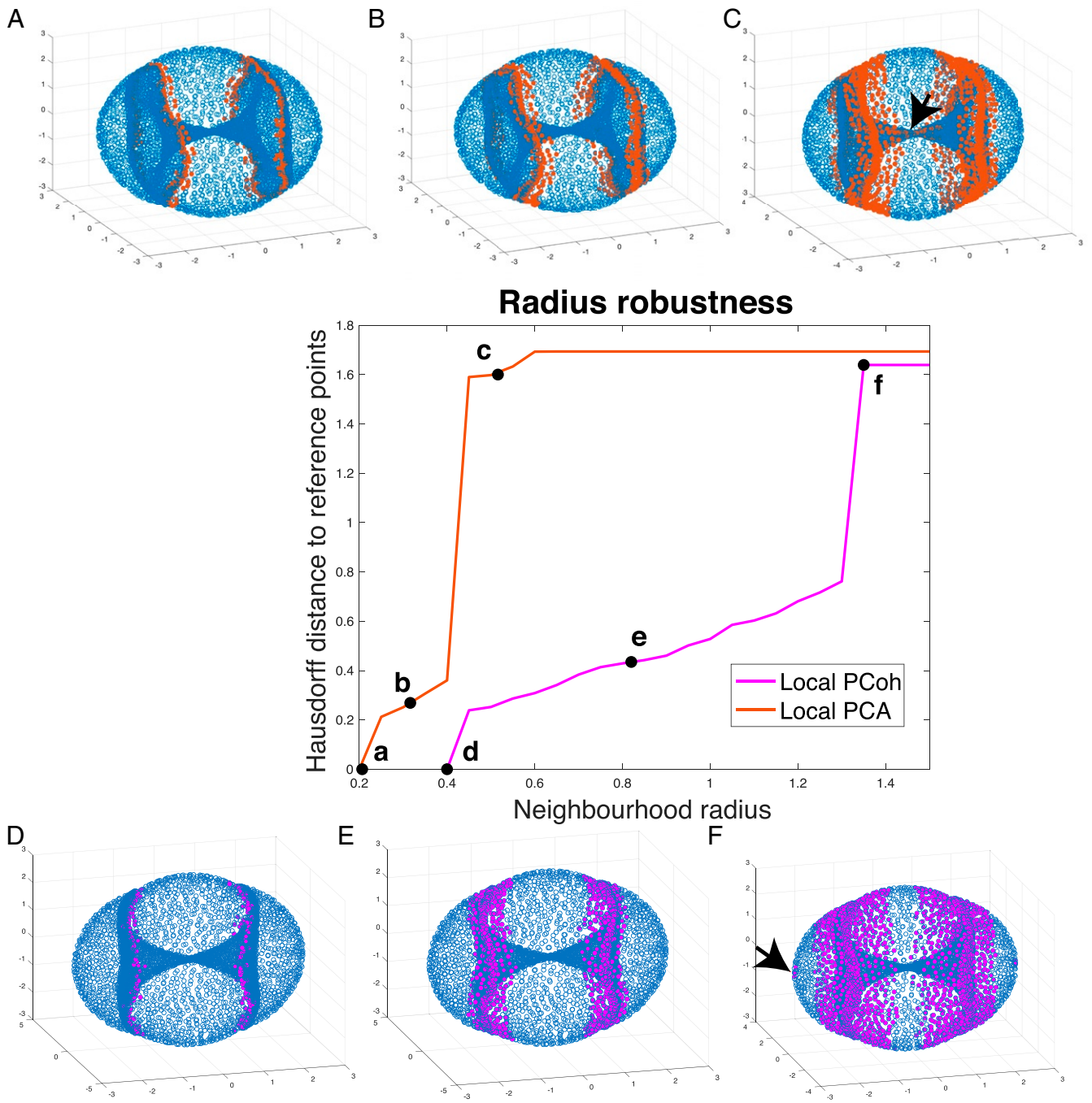

Fig. 4. Robustness with respect to the choice of local neighborhood size for the cyclo-octane dataset using local persistent cohomology (PCoh; purple line) and local PCA (orange). The horizontal axis represents local neighborhood size, while the vertical axis corresponds to the *Hausdorff distance* between the intersection points S_r selected by each method with neighborhood radius r on one hand and a set of ideal reference points $R \subset S_r$ on the other. This distance is defined to be the smallest $\epsilon > 0$ so that the union of radius ϵ balls around points in R contains all of the points in S_r. *A–F* illustrate the singularities detected at neighborhood radii corresponding to points a–f, respectively. The extreme points responsible for the steep increase in the Hausdorff distance for each method are indicated with black arrows.

The partition of P obtained in this manner may not produce points lying on a genuine stratified space since the algorithm does not check for normal triviality. To obtain such a stratification, more sophisticated methods (11) are needed.

Data Availability. Both the cyclo-octane and the Henneberg datasets have been made available at https://github.com/stolzbernadette/Geometric-Anomalies/tree/master/Data_Sets. Our MATLAB implementation of the *Geometric Anomaly Detection* algorithm is available at https://github.com/stolzbernadette/Geometric-Anomalies.

ACKNOWLEDGMENTS. We thank Barbara Mahler for performing the isomap projection of the cyclo-octane data to $\mathbb{R}^3$ and S. Martin for answering our queries about these data. We also thank André Henriques, Frances Kirwan, and Ulrike Tillmann for valuable feedback on an early version of this manuscript and the two anonymous referees for their insightful suggestions. B.J.S., H.A.H., and V.N. are members of the Center for Topological Data Analysis funded by Engineering and Physical Sciences Research Council (EPSRC) Grant EP/R018472/1.The doctoral studies of B.J.S. were funded by the EPSRC and Medical Research Council Grant EP/G037280/1 as well as F. Hoffmann-La Roche AG. H.A.H. acknowledges funding from EPSRC Fellowship EP/K041096/1 and a Royal Society University Research Fellowship.

1. C. Fefferman, S. Mitter, H. Narayanan, Testing the manifold hypothesis. *J. Am. Math. Soc.* **29**, 983–1049 (2016).
2. J. A. Lee, M. Verleysen, *Nonlinear Dimensionality Reduction* (Springer-Verlag, 2008).
3. Markus. Ringner, What is principal component analysis?. *Nat. Biotechnol.* **26**, 303–304 (2008).
4. J. B. Tenenbaum, V. De Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
5. H. Sebastian Seung, D. D. Lee, The manifold ways of perception. *Science* **290**, 2268–2269 (2000).
6. R. Vidal, Subspace clustering. *IEEE Signal Process. Mag.* **28**, 52–68 (2011).
7. T. Che, Y. Li, A. Paul Jacob, Y. Bengio, W. Li, "Mode regularized generative adversarial networks" in *5th International Conference on Learning Representations, ICLR 2017* (2017). https://openreview.net/forum?id=HJKkY35le. Accessed 28 July 2020.
8. S. Sabour, N. Frosst, G. E. Hinton, "Dynamic routing between capsules" in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds. (Curran Associates, Inc., 2017), vol. 30, pp. 3856–3866.
9. J. Harris, *Algebraic Geometry: A First Course* (Springer-Verlag, 1995).
10. S.-i. Amari, *Information Geometry and Its Applications* (Springer, 2016), vol. 194.
11. V. Nanda. Local cohomology and stratification. *Found. Comput. Math.*, **20**, 195–222 (2020).
12. F. Kirwan, J. Woolf, *An Introduction to Intersection Homology Theory* (Chapman and Hall/CRC, 2006).
13. M. Goresky, R. MacPherson, Intersection homology II. *Invent Math.* **71**, 77–129 (1983).
14. W. Fulton, *Intersection Theory* (Springer-Verlag, 1998).
15. K. Mischaikow, V. Nanda, Morse theory for filtrations and efficient computation of persistent homology. *Discrete Comput. Geom.* **50**, 330–353 (2013).
16. G. Henselman, R. Ghrist, Matroid filtrations and computational persistent homology. arXiv:1606.00199 (17 October 2017).
17. A. A. Markov, O konstruktivnykh funkciyakh. *Trudy Mat. Instituta im. Steklova* **52**, 315–348 (1958).
18. S. Martin, A. Thompson, E. A. Coutsias, J.-P. Watson, Topology of cyclo-octane energy landscape. *J. Chem. Phys.* **132**, 234115 (2010).
19. S. Martin, J.-P. Watson, Non-manifold surface reconstruction from high-dimensional point cloud data. *Comput. Geom.* **44**, 427–441 (2011).
20. D. Cohen-Steiner, H. Edelsbrunner, J. Harer, Stability of persistence diagrams. *Discrete Comput. Geom.* **37**, 107–120 (2007).
21. A. Hatcher, *Algebriac Topology* (Cambridge University Press, 2002).
22. S. Oudot, *Persistence Theory: From Quiver Representations to Data Analysis* (American Mathematical Society, Providence, RI, 2015), vol. 209.
23. N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, H. A. Harrington, A roadmap for the computation of persistent homology. *EPJ Data Sci.* **6**, 17 (2017).
24. A. Tausz, M. Vejdemo-Johansson, H. Adams, "JavaPlex: A research software package for persistent (co)homology" in *Proceedings of ICMS 2014, Lecture Notes in Computer Science*, H. Hong, C. Yap, Eds. (Springer, Berlin, Germany, 2014), vol. 8592, pp. 129–136.
25. U. Bauer, Data from "Ripser: A lean C++ code for the computation of Vietoris–Rips persistence barcodes." GitHub. https://github.com/Ripser/ripser. Accessed 1 May 2017.