How Now Oblong Cow: the importance of anisotropy in benchmarking intrinsic dimension estimators

Anonymous authors
Paper under double-blind review

Abstract

The manifold hypothesis suggests that data lies on manifolds with smaller intrinsic dimension (ID) than their ambient dimension. However there is no empirical agreement on the estimates for ID from different estimators for realistic datasets. Thus it is important to test ID estimation methods with targeted stressors. In this work, we consider the role of anisotropy. To this end, we propose HomID, a collection of homogeneous spaces with anisotropic embedding, for benchmarking ID estimation methods. We observe that methods that perform well on standard benchmarks systematically degrade on HomID under identical resource allocation. We further observe that anisotropic distortion of such benchmarks also results in performance degradation. Finally, we analyze two particular IDE estimators and reveal that anisotropy induces such failure modes by causing shifts from the assumed distributions. [1]

1 Introduction

The success of machine learning (ML) algorithms on datasets with large representative dimensions is often attributed to the hypothesis that the data actually lies on a manifold with smaller dimension, but embedded in a larger space Gorban & Tyukin (2018b); Fefferman et al. (2013); Olah (2014); Cayton (2008). This is called the manifold hypothesis. An ML algorithm is able to generalize because it can infer this manifold (or sub-manifold) from the given training distribution. This notion is formalized by the concept of *intrinsic dimension* (ID).

The manifold hypothesis is quite intriguing and has been subject to many experimental investigations, whereby different methods have been proposed to estimate the ID of a particular data-cloud and applied to popular ML datasets such as MNIST Lecun et al. (1998); Deng (2012). However, a common theme is the disagreement in the estimated IDs of different methods on these real-world datasets Hein & Audibert (2005); Facco et al. (2017); Pope et al. (2021); Bahadur & Paffenroth (2019); Goldt et al. (2020); Candelori et al. (2025); Stanczuk et al. (2024); Tempczyk et al. (2022); Kamkari et al. (2024). This large variability between different methods suggests that either (i) the manifold hypothesis is incorrect, or (ii) these different methods have their inherent biases which may or may not be relevant to particular datasets.

Existing benchmarks primarily rely on isotropic or weakly anisotropic structures, which we show can systematically overestimate estimator robustness. In this work, we undertake a careful study of how local anisotropy in data can cause systematic biases which degrade IDE performance for several methods. However, existing methods to induce anisotropy necessarily induce inhomogeneities in the data. To this end, we propose **HomID** (**Hom**ogeneous Spaces for **I**ntrinsic **D**imension Estimation), a set of manifolds with homogeneous spaces and locally anisotropic embeddings known from classic differential geometry. Our proposed benchmark isolates the effect of local anisotropy, since the manifolds are otherwise homogeneous, have constant scalar curvature and have scalable embeddings. We summarize our core contributions below:

[1] The code for this project can be found with the accompanying Supplementary Material and will be later released as a public Github repo.

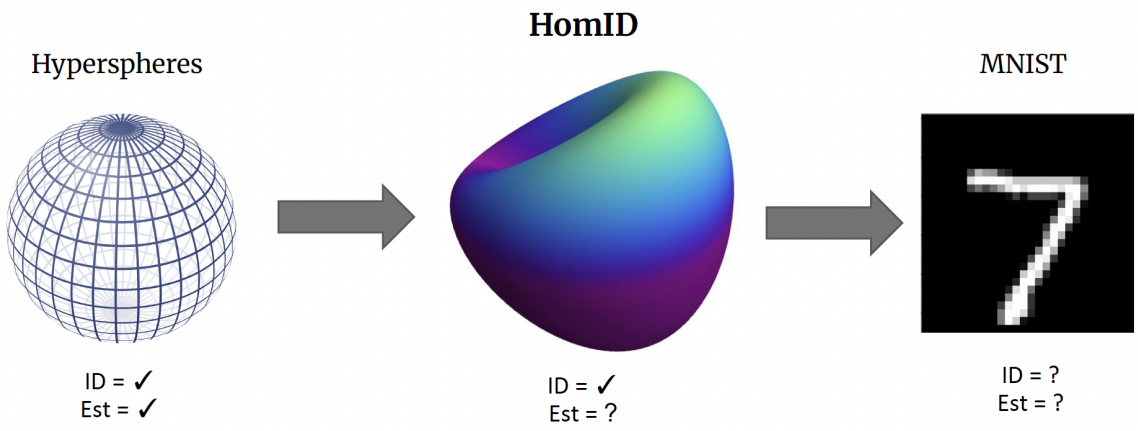


Figure 1: While most methods perform well when it comes to intrinsic dimension estimation (IDE) for simplistic manifolds like (hyper-)spheres, there's wide variability in their estimates for real-world datasets like MNIST. We propose benchmarking ID estimators with HomID—a family of manifolds known from classic differential geometry. HomID consists of homogeneous spaces with anisotropic embeddings, offering a clean way to understand the role of anisotropy in IDE evaluation.

1. We evaluate six ID estimators on HomID, a benchmark consisting of homogeneous spaces with anisotropic embeddings, derived using classic differential geometry results. We observe that all of these methods perform worse on HomID as compared to previous benchmarks except for a vector embedding of the Grassmanian manifold. HomID provides homogeneous, ground-truth manifolds whose embeddings exhibit strong local covariance anisotropy, and this stressor exposes failure modes not seen on comparable isotropic benchmarks.
2. We additionally stress-test intrinsic dimension estimators (IDEs) by controlled anisotropic distortions of otherwise isotropic manifolds such as spheres and linear nullspaces, which also cause performance degradation. Thus, we show that it is important to move beyond simple isotropic datasets for benchmarking IDE methods.
3. We analyze two particular ID estimators, namely lPCA and ABID, to test our hypothesis that anisotropy induce systematic biases in the underlying distributions, which lead to failure modes for the methods.

2 Background and Related Works

Manifolds in ML The manifold hypothesis states that real-world data lie on low-dimensional manifolds, even though expressed in high dimensions. (Gorban & Tyukin, 2018a; Cayton, 2008). This idea has been explored both theoretically and experimentally (Fefferman et al., 2013; Goldt et al., 2020). Topological properties of such data manifolds (Carlsson et al., 2008) and latent representations (Ansuini et al., 2019) have also been studied.

Intrinsic Dimension Intrinsic dimension (ID), as the name suggests, characterizes several inherent properties of the manifold (Narayanan & Niyogi, 2009; Pope et al., 2021). A plethora of estimators have been proposed for ID estimation (IDE), which we discuss in Appx. C. See (Campadelli et al., 2015) for a detailed survey.

Benchmarks for ID Estimation Practical benchmarks for IDE involve isotropic manifolds like hyperspheres, Gaussian distribution or non-scalable manifolds like the Möbius strip (Camastra, 2003; Hein & Audibert, 2005; Rozza et al., 2012; Campadelli et al., 2015; Bac et al., 2021). An interesting benchmark

proposed GAN-based data to lower bound ID (Pope et al., 2021). There also exist physics-inspired datasets (Weigend & Gershenfeld, 1993; Camastra & Filippone, 2009; Grassberger & Procaccia, 1983) which typically use non-linear dynamical models. These benchmarks, however, do not isolate the effect of different stressors in a dataset for evaluating ID estimation. In particular, previous benchmarks (Tempczyk et al., 2022; Kamkari et al., 2024) with anisotropic curvature typically have non-uniform density as well. We propose using homogeneous spaces as benchmarks to isolate the effect of anisotropy.

Manifold constructions Our methods are based on standard differential geometry, but the specific recipes we use originate from the quantum optics and quantum information literature. Namely, points in our manifolds correspond to Gilmore-Perelomov coherent states (Perelomov, 1977; 1986; Zhang et al., 1990) — the most "classical" states of a quantum system. Such states are typically parameterized by the sphere or complex plane, but coherent states on more general manifolds are relevant to quantum information, quantum metrology (Holevo, 2011) and, in the case of Grassmannians, the quantum Hall effect (Calixto & Pérez-Romero, 2014). As such, we sometimes refer to our construction as the "coherent-state method". We also describe these quantum-mechanical recipes in the language of the ML community in Appx. D.

A more comprehensive discussion of previous work may be found in Appx. C.

3 Preliminaries

Manifolds and Intrinsic Dimension A topological manifold M of dimension d_i is a topological space that locally "looks" like $\mathbb{R}^{d_i}$, in the sense that there exist local patches that are homeomorphic to $\mathbb{R}^{d_i}$ (Lee, 2012). For a given disjoint union of manifolds and a point p in this union, the local intrinsic dimension around p is the dimension of the submanifold to which it belongs (Kamkari et al., 2024).

In addition to this *intrinsic* viewpoint of a manifold, one can equivalently consider manifolds *extrinsically* by defining them as appropriate subsets of some ambient Euclidean space $\mathbb{R}^{d_a}$ (Lee, 2012). Given (samples from) a set $S \subset \mathbb{R}^{d_a}$ along with the promise that S is a manifold of dimension $d_i \leq d_a$ for some unknown ID d_i, one may naturally desire an algorithmic procedure to estimate d_i. Intuitively, the ID of a dataset can be thought of as the minimum number of parameters needed to represent the data with no loss of information (Ceruti et al., 2014). Importantly, though, a d_i-dimensional manifold can have nontrivial geometry (such as a direction-dependent, or anisotropic, curvature) and topology (such as being not simply connected) that makes it starkly different from real space or a sphere of the same dimension.

Our manifolds, by definition, exhibit a single ID at all points. In contrast, most methods return the local (scale-dependent) ID (LID) estimates at different points. Thus, we will refer to the *mean* of these LID estimates as *the* ID of the manifold as estimated by a given method. [2]

Homogeneous spaces All manifolds included in the HomID benchmark are parameterized by quotient spaces $\mathcal{G}/\mathcal{H}$ (a.k.a. homogeneous spaces), where $\mathcal{G}$ is a Lie group (i.e., a group that can also be considered as a manifold), and where $\mathcal{H}$ is one of its subgroups. For $\mathcal{G}$, we use either the orthogonal group $\mathrm{O}(n)$ or the unitary group $\mathrm{U}(n)$, which is the symmetry group of the n-dimensional real or complex sphere, respectively. We provide a pedagogical flavor of these spaces below, leaving technical details to Appx. D.

The two-dimensional real sphere, S^2, is a simple example of a homogeneous space. There exists a proper (i.e., orientation-preserving) rotation that can map the north pole to any other point on the sphere (whose action can be thought of as moving along the great circle connecting the north pole to the desired point). Therefore, we can label all points on the sphere by the rotations that take us there from the north pole, but we need to omit all rotations that rotate around the north pole since those do not take us anywhere.

Mathematically, this translates to the homogeneous space $\mathrm{SO}(3)/\mathrm{SO}(2) \cong S^2$, where $\mathrm{SO}(3)$ is the group of all proper three-dimensional rotations, and $\mathrm{SO}(2)$ is the subgroup of rotations around the north pole.

[2] We did not observe any significant differences with other statistical quantities such as using median, mode, mean of medians and median of means. This indicates that methods are not returning too much of variation for the IDE values.

The characteristics of the quotient space depend on the subgroup, which can be continuous or discrete. Spheres and more general homogeneous spaces whose $\mathcal{H}$ is a continuous group are constructed in the same spirit as this example. Their intrinsic dimension is $d_i = \dim \mathcal{G} - \dim \mathcal{H}$.

Local Anisotropy In this work, we use local anisotropy to refer to direction-dependent geometry within a small neighborhood of a point on the embedded manifold. Operationally, this can be detected through the spectrum of the local covariance matrix: if a neighborhood is locally isotropic, then its variance is distributed approximately evenly across the d_i tangent directions, whereas local anisotropy produces a non-flat tangent spectrum in which some tangent directions carry systematically more variance than others. This differs from global anisotropy, which refers to anisotropy of the entire data cloud, for example as measured by the spectrum of the global covariance matrix. A dataset can have weak or uninformative global covariance anisotropy while still exhibiting strong local anisotropy, and conversely global covariance anisotropy does not imply that every local neighborhood is anisotropic.

To operationalize local anisotropy, we compute the spectrum of the centered kNN covariance matrix around randomly chosen anchor points. For a locally isotropic d_i-dimensional reference manifold, such as a sphere, the first d_i normalized eigenvalues are approximately flat at small neighborhood scales, followed by a sharp drop. In contrast, several HomID embeddings exhibit a strongly sloped tangent spectrum, indicating direction-dependent local variance. This provides direct evidence that HomID embeddings exhibit local anisotropy under matched N, d_i, d_a conditions. We emphasize, however, that this is an operational finite-sample diagnostic rather than a complete differential-geometric invariant: it conflates, to some extent, curvature, embedding distortion, finite-neighborhood effects, and sampling effects. Precise intrinsic or extrinsic notions of anisotropy can be formulated through quantities such as the second fundamental form, directional curvature, or the pullback metric, but estimating these robustly and scalably from only high-dimensional samples is itself a difficult problem. Thus, we use local covariance anisotropy as a scalable proxy for the kind of direction-dependent local geometry that affects kNN-based IDE methods. This is discussed in more detail in Appx. F.1.

4 Proposed Datasets

Stiefel manifolds Stiefel manifolds are the closest relatives of spheres out of all HomID manifolds. Real Stiefel manifolds are parameterized by quotients of the form $\mathrm{O}(n)/\mathrm{O}(n-k)$ for $n \geq k \geq 1$ (James, 1977), while real spheres can also be understood as the quotients $\mathrm{O}(n)/\mathrm{O}(n-1)$.

A common alternative definition is the set of $n \times k$ real matrices X satisfying $X^T X = \mathbb{I}$, where T is the transpose map, and where $\mathbb{I}$ is the appropriately sized identity matrix. In this way, one can interpret the manifolds as all possible isometries of k-dimensional space into n dimensions. We note that the topology of general Stiefel manifolds is quite different from that of spheres (James, 1977).

The HomID benchmark includes two different embeddings of Stiefel manifolds. The first, called "St (Matrix)", is a simple vectorization of the matrices X. The second, called "St (Vec)", is a mixture of the coherent-state method and a vectorization of a matrix.

Grassmannians Grassmannians, or Grassmann manifolds, are defined as $\mathrm{Gr}(k, \mathbb{R}^n) \cong \mathrm{O}(n)/\mathrm{O}(n-k) \times \mathrm{O}(k)$ (Lee, 2012) and can be thought of as quotients of Stiefel manifolds by an extra $\mathrm{O}(k)$ subgroup. Grassmannians have a long history in ML (Zhang et al., 2018; Bendokat et al., 2020). A simple example of one is the real projective plane, $\mathbb{R}P^2 \cong \mathrm{Gr}(1, \mathbb{R}^3)$, which is the sphere S^2 with antipodal points identified.

The HomID benchmark includes two different embeddings of Grassmann manifolds. The first is based on the interpretation of the Grassmannian as a space of subspaces. The quotient of the Stiefel manifold by the extra $O(k)$ subgroup identifies two isometries related by a basis change as equivalent. This implies that Grassmannians parameterize all distinct k-dimensional subspaces of n-dimensional space. We represent a point on the Grassmannian by an n-dimensional projector onto a k-dimensional subspace, which yields our "Gr (Proj)" embedding when written as an n^2-dimensional vector.

IDE Method	Gr (Proj)	Gr (Vec)	Flag (Vec)	St (Matrix)	St (Vec)	**Average (HomID)**
lPCA	1.5492	0.0700	3.5703	0.6210	1.4899	1.4601
MLE	0.1758	0.0474	0.3065	0.4613	0.7327	0.3447
CorrInt	1.0068	0.6832	4.0588	0.7366	0.7470	1.4465
TwoNN	0.0731	0.0713	0.1970	0.4536	0.4376	**0.2465**
ABID	0.3571	0.1858	0.4538	0.4588	0.6182	0.4147
DANCo	0.3090	0.1558	5.8647	0.7446	1.0756	1.6299
Average	0.5785	**0.2023**	2.4085	0.5793	0.8502	**0.9237**

Table 1: Mean relative error $|\delta|$ for various methods on HomID manifolds. A higher value indicates worse performance. We see that the vector embedding of Grassmannian consistently has low error for all methods, while TwoNN typically performs the best on all manifolds.

The second embedding, called "Gr (Vec)", is based on the coherent-state method and allows for an ambient dimension as low as $\binom{n}{k}$. It is also known as the Plücker embedding (Lee, 2012; Bendokat et al., 2020).

Flag manifolds Flag manifolds generalize Grassmannians to arbitrary *sets* of n-dimensional vectors. A t-flag manifold is defined as the manifold described by the quotient space $\mathrm{O}(n)/\mathrm{O}(k_1)\times\mathrm{O}(k_2)\times\cdots\times\mathrm{O}(k_{t+1})$, with the constraint $\sum_{i=1}^{t+1} k_i = n$. Our benchmark contains the "Flag (Vec)" embedding of the $t = 2$ case, which is based on the coherent-state method.

5 Results

5.1 Methods tested

We choose a few standard representative IDE methods of different flavors for testing on our benchmark. These are linear methods like linear subspace projection (lPCA (Cangelosi & Goriely, 2007; Fukunaga & Olsen, 1971; Fan et al., 2010)) as well as non-linear methods based on maximum likelihood estimation (MLE (Levina & Bickel, 2004; Haro et al., 2008)), fractal dimension estimation (CorrInt (Grassberger & Procaccia, 1983; Camastra & Vinciarelli, 2002)), distribution of distances (TwoNN (Facco et al., 2017)), angles (ABID (Thordsen & Schubert, 2022)) as well as concentration of measure (DANCo (Ceruti et al., 2014)). We emphasize that this list is not exhaustive, and our choice is based on diversity and relatively low computational requirements. We review these methods, including their implementation, in Appx. E.

Given a manifold embedded into a space with ambient dimension d_a, we define the *relative error* δ,

$$\boxed{\delta := \frac{\hat{d}_i}{d_i} - 1 \in \left[-1, \frac{d_a}{d_i} - 1\right]} \qquad \text{(relative error)} \tag{1}$$

Here, $\hat{d}_i$ is an estimated quantity, while d_i is the ground-truth ID. Note that $\delta < 0$ implies that the method underestimates the ID, while $\delta > 0$ indicates an overestimation for the ID up to the ambient dimension. Average performance of methods on HomID manifolds is summarised in Table 1, in which we list $|\delta|$ so as to compare over- and under-estimation on the same footing.

5.2 Comparison with other Benchmarks

We undertake a comprehensive evaluation of IDE methods on HomID. We perform hyper-parameter sweeps and, for each hyper-parameter combination, compare the performance of different IDE methods on HomID to other IDE benchmarks with isotropic manifolds. We notice that the chosen methods are almost always worse at estimating the ID for our manifold embeddings, with the notable exception being the embedding

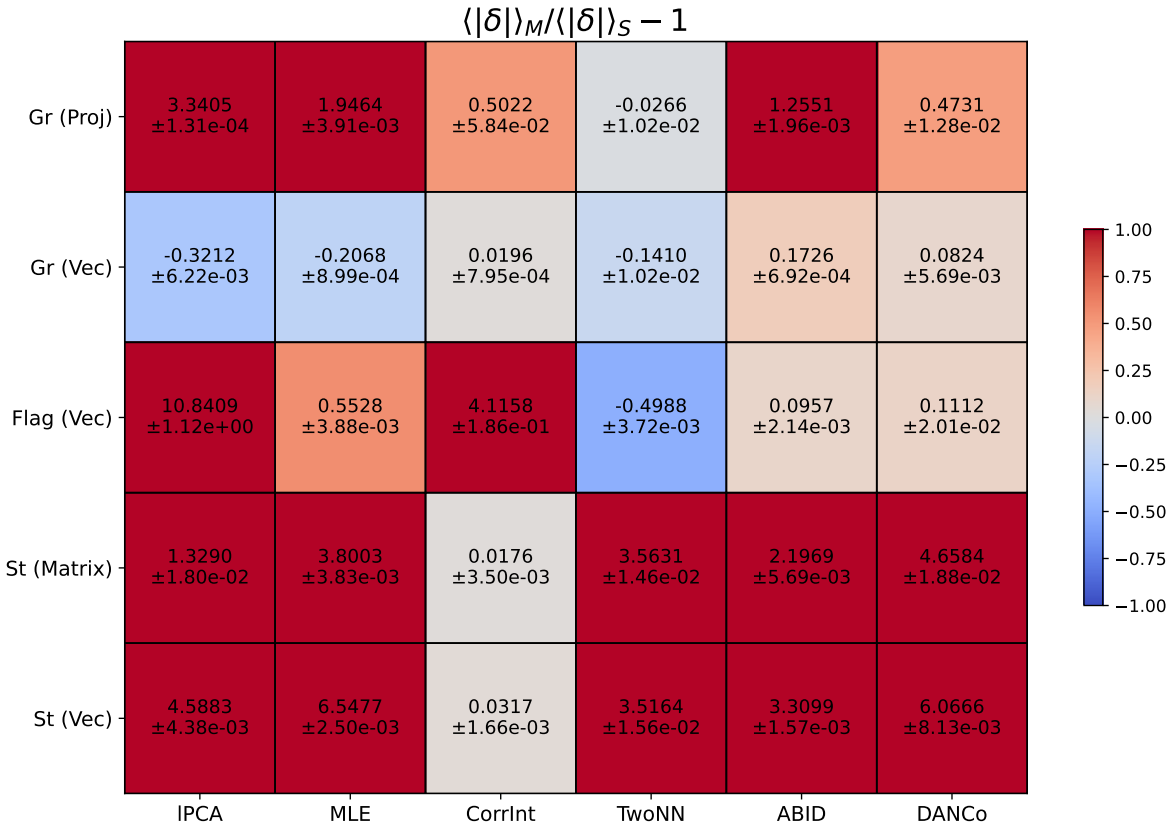


Figure 2: Comparison of the quantity $\langle|\delta|\rangle_M/\langle|\delta|\rangle_S - 1$, where the numerator is the average of the absolute value of the relative error $|\delta|$ over all instantiations of a given manifold family M, while the denominator is the corresponding average over all sphere embeddings with the same intrinsic and ambient dimensions. This relative comparison shows that tested methods tend to perform much *worse* against our manifolds than against spheres with the same dimensions. Interestingly, we observe a positive score with a change in embedding of the Grassmannian, hinting that method accuracy depends on the type of embedding. Due to high computational time, we choose a smaller range of hyper-parameter sweeps for DANCo. A 1-σ sampling error is reported here after the $\pm$ sign.

"Gr (Vec)". For brevity, we present the relative result on spheres here in the main text, cf. Fig 2. The reader is referred to Appx. F.2 for a comparison to other manifolds. We emphasize that the comparison is made across a range of scales, accessible by our computational resources, for manifolds with small ID. The details of the manifolds chosen are also discussed in Appx. F.2.

5.3 Stress-testing the role of anisotropy

The hypothesis that anisotropy limits robustness of IDE methods suggests that anisotropic distortion of isotropic manifolds should also cause degradation of IDE performance. We perform two sets of controlled experiments to obtain such distortions.

Random diagonal transformations We test IDE methods on anisotropic distorted versions of HomID manifolds, naturally obtained by amending the coherent-state method by linearly transforming the data with a random diagonal matrix. Such distorted manifolds have locally anisotropic curvature, but in addition they get non-uniform curvature and density. However, the ID is the same.

The strength of distortion is governed by a parameter ϵ, and we generate each diagonal entry by sampling uniformly from $[1 - \frac{\epsilon}{2}, 1 + \frac{\epsilon}{2}]$. The performance of methods is mostly unchanged upon distortion of the

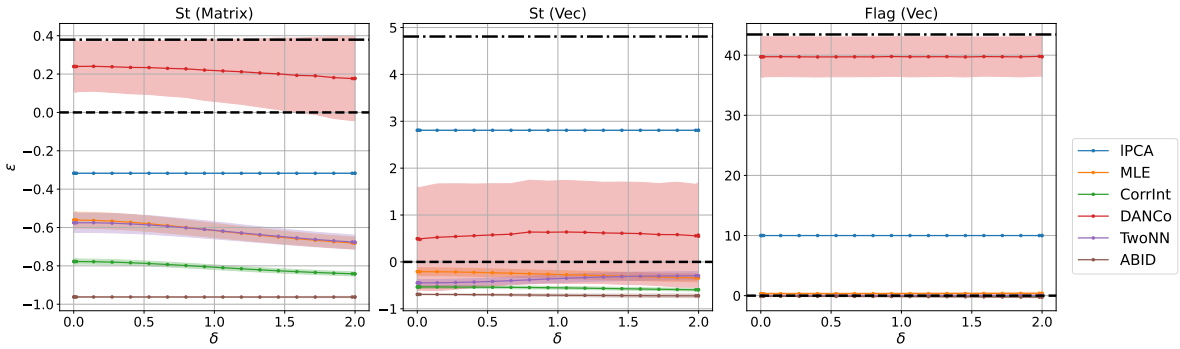


Figure 3: Effect of anisotropic distortion: We plot the relative error $\langle\delta\rangle$ as a function of the parameter ϵ, which is a direct measure of anisotropy. Except for St (Matrix), most methods show negligible change as ϵ is increased.

underlying manifolds. We demonstrate this for three typical manifolds in Fig 3, adding the rest to Appx G.1. Some methods show a slight degradation in performance, but the change is minimal and within the error margin. By contrast, the methods performed significantly worse on distorted spheres than on our distorted manifolds, cf. App F.4.

Additive noise An alternate way to understand the role of anisotropy is to perturb the original manifolds with additive noise, i.e. $\mathbf{x} \to \mathbf{x} + \epsilon$, where ϵ is sampled from a gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. The role of anisotropy can be clearly understood by choosing $\mathbf{\Sigma}$ to be either the identity matrix $\mathbf{I}_{d_a}$ (*isotropic*), a diagonal matrix Λ with elements chosen from $\mathcal{U}(0, \frac{2}{d_a})$ (*uncorrelated*), or a random positive definite matrix uu^T (*anisotropic*), where the elements of $u \in \mathbb{R}^{d_a \times d_a}$ are chosen from $\mathcal{N}(\mathbf{0}, 1)$. We then explicitly set $\text{Tr}\ \Sigma = 1$ by the transformation $\Sigma \to \Sigma' = \Sigma/(\text{Tr}\ \Sigma)$.

Most methods deviate in their estimations when $\sigma^2 \sim ||\mathbf{x}||_2^2 = \boldsymbol{O}(1)$. Rather curiously, we observe flat lines, indicating that, for certain methods, the data cloud is indistinguishable from pure noise. We also observe an improvement in IDE performance for certain manifold-method combinations, suggesting a certain regularizing effect emerging from the noise.

We observe that in the low to intermediate regime, there is no discernible change in the behavior of isotropic or anisotropic noise. However, in the high noise limit, we observe that anisotropic noise is always underestimated, further corroborating our conjecture.

The results are summarized in Fig. 4. We present three typical manifolds here, with plots for all the manifolds in Appx. G.1.

5.4 Relation to real-world data

As a sanity check, we performed the above anisotropic distortion experiments with MNIST. We show in Figure 5 how these distortions yield results that are qualitatively similar to that of HomID. This highlights the importance of benchmarking IDE methods with such controlled stressors which can be readily applied to realistic datasets. Note that this experiment is not validation of exact quantitative transfer, but rather that the anisotropy stressor is not artificial.

5.5 *Why* anisotropy hurts?

To leading order, the Euclidean and geodesic distances between two nearby points differ because of the curvature (Lang, 1995). For manifolds with anisotropic curvature, this difference depends on the points. Nearest-neighbor (NN) IDE methods use Euclidean distances to choose the kNNs of a point on the manifold, which can cause systematic bias in the choice of neighborhoods (assuming finite neighbors k and sample

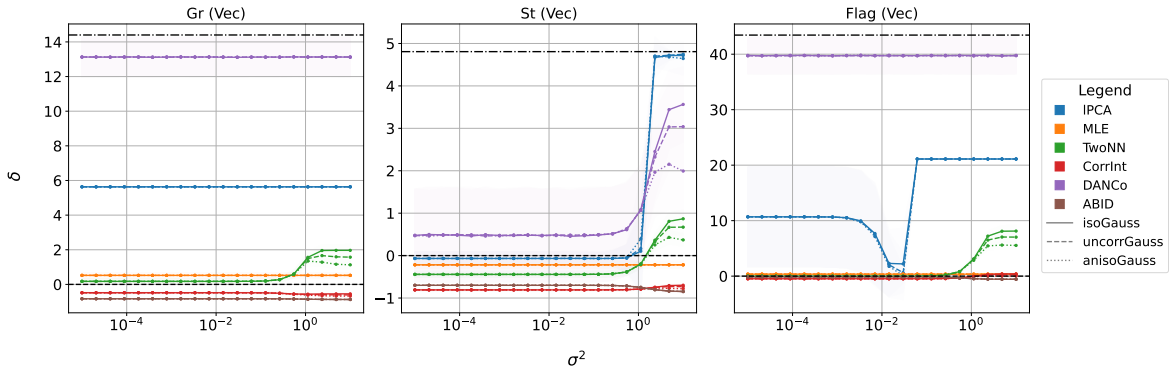


Figure 4: Effect of additive noise. We report IDE performance when the unperturbed data $\mathbf{x} \to \mathbf{x} + \epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\Sigma)$ with Σ chosen to be proportional to the identity (isotropic), diagonal (uncorrelated) or a random symmetric matrix (anisotropic). We then plot the relative error δ (on the true manifold) as a function of the noise scale σ^2. We observe all three types of behavior when noise is added, as discussed in Section 5.3. Notice that there is no discernible change in behavior between the different noise types, except in the high noise limit, where the anisotropic noises are consistently underestimated.A 1-σ sampling error is plotted.

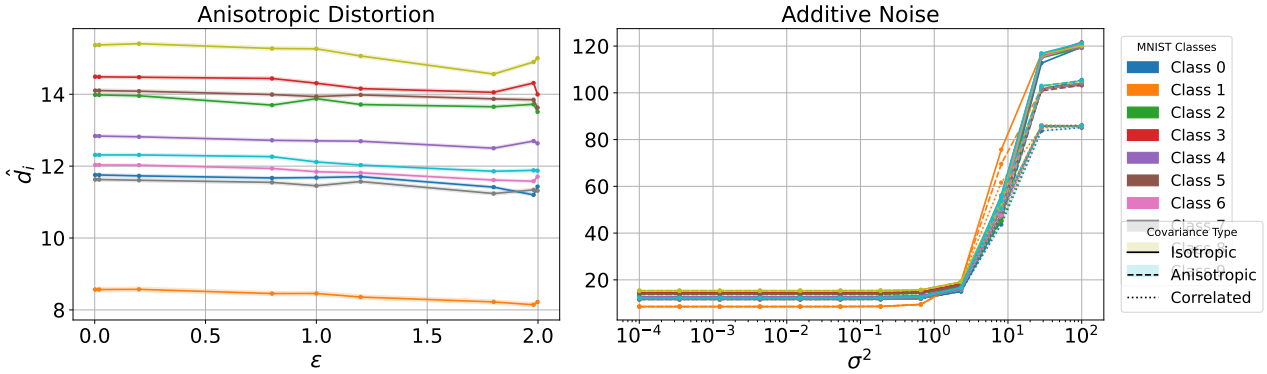


Figure 5: IDE performance for MLE on MNIST when subjected to anisotropic distortion and noise perturbation. Notice how additional anisotropy does not change the IDE by much, similar to HomID. The additive noise alters the prediction only at $\sigma^2 \sim \boldsymbol{O}(1)$. The other plots are presented in Appx. G.3.

size N). As such, kNN-based IDE methods should fail for manifolds with anisotropic curvature. This is confirmed by our experiments with the (anisotropic) manifolds of HomID.

5.6 Analysis of other statistical and geometric properties

We find that commonly used scalar quantities are insufficient to explain IDE performance differences. These include properties such as global variance, global anisotropy, correlations between different Euclidean components of the vectors, local density and local curvature. We report null results with all of these quantities in Appx. G.2.

6 Failure mode analysis for IPCA and ABID

We analyze the reasons for failure for two IDE methods in this section. Most methods usually rely on some assumptions about the underlying data distribution if they belong to a manifold. We show how anisotropy

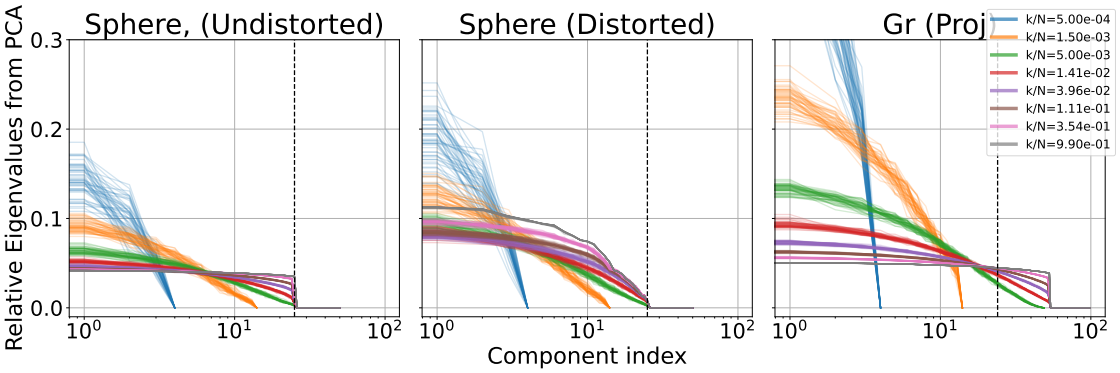


Figure 6: Analysis of relative eigenvalues for lPCA. Recall that the relative eigenvalues represent the fraction of total variance explained by top PCA components. We observe that the relative eigenvalues show a sharp decline at the correct ID value for undistorted spheres, the kink disappearing with distortion. However, for Grassmannian projectors, we observe that the sharp drop does happen but at a different ID. Dashed black lines represent the ID of the manifold.

causes deviations from such distribution for the lPCA and ABID methods. Note that an exhaustive quantitative treatment of this effect is beyond the scope of this project.

We conjecture that lPCA fails in the presence of anisotropic noise due to deviation from the expected distribution for manifolds. Recall that lPCA looks at the directions of maximal variance in a local neighborhood. For a manifold, the normal directions show zero variance. Thus we expect a sharp drop in the relative eigenvalues at the index corresponding to the ID. In Fig. 6, we observe that spheres show this behavior, but distorted spheres and HomID manifolds do not. In particular, this gap sharply decreases as we distort the sphere, while the Grassmanian exhibits a drop at a different index. This shows that the failure mode for lPCA arises from a decreasing signal-to-noise ratio for distorted spheres, while for HomID, the failure mode is because of an incorrect signal. This trend, however, does not occur exactly for all HomID manifolds, which we describe in detail in Appx. G.4.

We also analyze the failure mode for ABID. As implemented in (Thordsen & Schubert, 2022), the distribution of pairwise cosine similarities between independent random vectors on a d_i-dimensional hypersphere is expected to follow a Beta distribution. ABID calculates cosine similarity between vectors in a local neighborhood. Thus the distribution is slightly shifted, due to correlations, but still overall symmetric. This variance controls the dimension estimate, as outlined in Appx. E. With increasing anisotropy, we observe a gradual asymmetry in the distribution, which leads to an incorrect estimate. This is true for both distorted spheres as well as HomID datasets, cf. Fig. 7.

We believe that anisotropic curvature leads to similar systematic biases in distributions relevant for other methods. An in-depth investigation of all the methods tested is left for future work.

7 Conclusion, Limitations and Future Work

Conclusion In this work, we investigated the role of anisotropy as a stressor for intrinsic dimension estimation (IDE). We introduced a new benchmark family, HomID, consisting of homogeneous manifolds with anisotropic embeddings and known ground-truth intrinsic dimension. Across a range of commonly used estimators, we found that methods that perform well on standard synthetic benchmarks—such as hyperspheres—consistently degrade on HomID under matched sampling conditions. Furthermore, controlled anisotropic distortions of classical benchmarks reproduce similar degradation, suggesting that isotropy in existing benchmarks masks important failure modes. Our analysis of representative estimators (lPCA and ABID) indicates that anisotropy induces systematic biases in the distributions these methods rely on, leading to incorrect dimension estimates. While this does not fully characterize all failure mechanisms, it provides concrete evidence that anisotropic geometry is an important but underexplored factor in IDE evaluation.

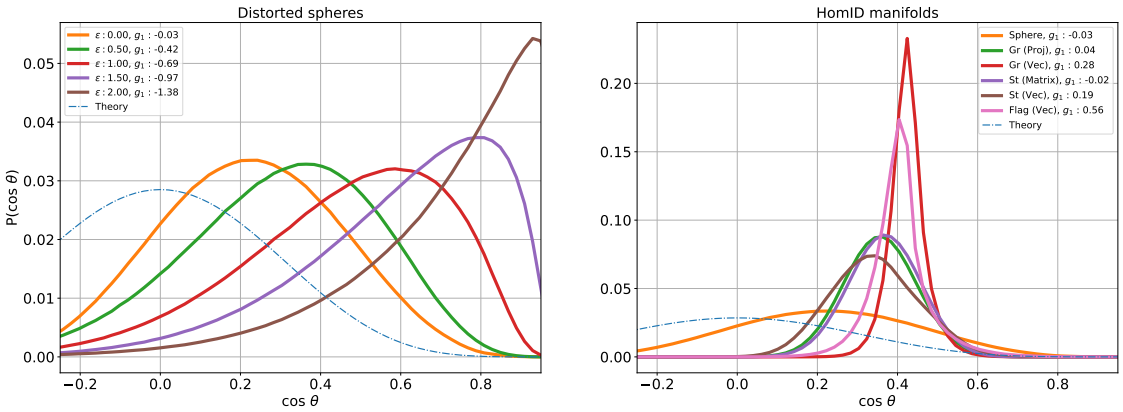


Figure 7: Cosine distribution for ABID: The distribution of cosine similarity between random unit vectors is expected to follow a Beta distribution centered at zero (dashed-dotted curve in blue). In practice due to correlations between the vectors chosen in a k-local neighborhood, there is a slight shift, but the overall symmetry is preserved. As we distort a sphere (left subpanel) by increasing ϵ, we notice a gradual change in asymmetry, which affects the ID estimate. This is reflected by the increased value of the skewness g_1 for the different samples. A similar behavior is seen for the HomID manifolds (right subpanel).

Overall, our results highlight a gap between standard synthetic benchmarks and more realistic geometric settings. HomID provides a controlled intermediate regime—retaining exact ground truth while exhibiting richer geometric structure—that can serve as a more informative testbed for future IDE methods.

Limitations While HomID introduces controlled anisotropy and nontrivial geometry, it remains a synthetic benchmark and does not fully capture the complexity of real-world data distributions. Second, our analysis focuses on a limited set of estimators and computational regimes; a broader evaluation across more specialized methods may reveal additional insight. Although we provide a scalable local covariance diagnostic for anisotropy, we do not claim that it is a complete differential-geometric invariant or that it fully explains all estimator failures. Establishing a quantitative causal model connecting anisotropy, curvature, sampling scale, and IDE error remains future work. Finally, our experiments are limited by computational resources, restricting the scale of datasets and hyperparameter sweeps considered. These constraints may affect the generality of some quantitative conclusions. Additionally, we do not provide a quantitative correlation of anisotropy due to the lack of a suitable scalable measure available, to the best of our knowledge.

Future Work A key direction for future work is to develop a more systematic understanding of geometric factors affecting intrinsic dimension estimation, beyond anisotropy alone. In particular, extending the benchmark to incorporate controlled variations in scale, noise structure, and "fuzzy" or non-manifold data could help bridge the gap to real-world datasets. Another promising direction is the design of geometry-aware IDE methods that explicitly account for anisotropy in local neighborhoods. Finally, expanding HomID to include a broader class of manifolds and integrating it with existing benchmarking libraries may facilitate more standardized and reproducible evaluation of intrinsic dimension estimators.

8 Impact Statement

This work contributes to the evaluation of intrinsic dimension estimation methods by identifying a previously underexplored failure mode—anisotropic geometry—and proposing a benchmark to study it in a controlled setting. Improved understanding of IDE behavior has implications for a range of applications in machine learning, including representation learning, data compression, and model complexity analysis. As our benchmark is fully synthetic and does not involve human data, it raises no direct ethical or privacy concerns. However, more reliable intrinsic dimension estimation may influence how practitioners interpret the structure and complexity of real-world datasets, which could impact downstream modeling decisions. We

hope that this work encourages more rigorous and realistic benchmarking practices in the study of intrinsic dimension and related geometric properties of data.

References

Victor V Albert, Jacob P Covey, and John Preskill. Robust encoding of a qubit in a molecule. *Physical Review X*, 10(3):031050, 2020.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks, 2019. URL https://arxiv.org/abs/1905.12784.

Jonathan Bac, Evgeny M. Mirkes, Alexander N. Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-dimension: A python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, October 2021. ISSN 1099-4300. doi: 10.3390/e23101368. URL http://dx.doi.org/10.3390/e23101368.

Nitish Bahadur and Randy Paffenroth. Dimension estimation using autoencoders, 2019. URL https://arxiv.org/abs/1909.10702.

Thomas Bendokat, Ralf Zimmermann, and P. A. Absil. A grassmann manifold handbook: Basic geometry and computational aspects, 11 2020.

Tolga Birdal, Aaron Lou, Leonidas Guibas, and Umut Şimşekli. Intrinsic dimension, persistent homology and generalization in neural networks, 2021. URL https://arxiv.org/abs/2111.13171.

Adam Block, Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. Intrinsic dimension estimation using wasserstein distances, 2022. URL https://arxiv.org/abs/2106.04018.

Bradley C. A. Brown, Jordan Juravsky, Anthony L. Caterini, and Gabriel Loaiza-Ganem. Relating regularization and generalization through the intrinsic dimension of activations, 2022. URL https://arxiv.org/abs/2211.13239.

Bradley C. A. Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data, 2023. URL https://arxiv.org/abs/2207.02862.

Manuel Calixto and Emilio Pérez-Romero. Coherent states on the grassmannian $u(4)/u(2)^2$: Oscillator realization and bilayer fractional quantum hall systems. *Journal of Physics A: Mathematical and Theoretical*, 47(11):115302, 2014. doi: 10.1088/1751-8113/47/11/115302.

F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(10):1404–1407, 2002. doi: 10.1109/TPAMI.2002.1039212.

Francesco Camastra. Data dimensionality estimation methods: a survey. *Pattern Recognition*, 36(12): 2945–2954, 2003. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(03)00176-6. URL https://www.sciencedirect.com/science/article/pii/S0031320303001766.

Francesco Camastra and Maurizio Filippone. A comparative evaluation of nonlinear dynamics methods for time series prediction. *Neural Computing and Applications*, 18(8):1021–1029, 2009. ISSN 0941-0643. doi: 10.1007/s00521-009-0269-1.

P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015(1):759567, 2015. doi: https://doi.org/10.1155/2015/759567. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2015/759567.

Luca Candelori, Alexander G. Abanov, Jeffrey Berger, Cameron J. Hogan, Vahagn Kirakosyan, Kharen Musaelian, Ryan Samson, James E. T. Smith, Dario Villani, Martin T. Wells, and Mengjia Xu. Robust estimation of the intrinsic dimension of data sets with quantum cognition machine learning. *Scientific Reports*, 15(1):6933, February 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-91676-8. URL https://doi.org/10.1038/s41598-025-91676-8.

Richard Cangelosi and Alain Goriely. Component retention in principal component analysis with application to cdna microarray data. *Biology Direct*, 2:2, 2007. doi: 10.1186/1745-6150-2-2. URL https://doi.org/10.1186/1745-6150-2-2.

Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76:1–12, 2008. doi: 10.1007/s11263-007-0056-x. URL https://doi.org/10.1007/s11263-007-0056-x.

Laurent Cayton. Algorithms for manifold learning. Technical report, University of California, San Diego, Department of Computer Science & Engineering, 2008. URL https://escholarship.org/uc/item/8969r8tc. Retrieved from eScholarship.

Claudio Ceruti, Simone Bassis, Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli. Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration. *Pattern Recognition*, 47(8):2569–2581, 2014. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2014.02.013. URL https://www.sciencedirect.com/science/article/pii/S003132031400065X.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

David L. Donoho and Carrie Grimes. Image manifolds which are isometric to euclidean space. *Journal of Mathematical Imaging and Vision*, 23:5–24, 2005. doi: 10.1007/s10851-005-4965-4. URL https://doi.org/10.1007/s10851-005-4965-4.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1), September 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-11873-y. URL http://dx.doi.org/10.1038/s41598-017-11873-y.

Mingyu Fan, Nannan Gu, Hong Qiao, and Bo Zhang. Intrinsic dimension estimation of data by principal component analysis, 2010. URL https://arxiv.org/abs/1002.2050.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis, 2013. URL https://arxiv.org/abs/1310.0425.

Paul Frost. Embedding real projective space in sphere. Mathematics Stack Exchange, 2021. URL https://math.stackexchange.com/q/4100495. URL:https://math.stackexchange.com/q/4100495 (version: 2021-04-13).

K. Fukunaga and D.R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, C-20(2):176–183, 1971. doi: 10.1109/T-C.1971.223208.

M. Gashler and T. Martinez. Tangent space guided intelligent neighbor finding. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN '11)*, pp. 2617–2624, August 2011. doi: 10.1109/IJCNN.2011.6033560. URL https://doi.org/10.1109/IJCNN.2011.6033560.

M. Gashler and T. Martinez. Robust manifold learning with cyclecut. *Connection Science*, 24(1):57–69, 2012. doi: 10.1080/09540091.2012.664122. URL https://doi.org/10.1080/09540091.2012.664122.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10:041044, Dec 2020. doi: 10.1103/PhysRevX.10.041044. URL https://link.aps.org/doi/10.1103/PhysRevX.10.041044.

Marina Gomtsyan, Nikita Mokrov, Maxim Panov, and Yury Yanovich. Geometry-aware maximum likelihood estimation of intrinsic dimension, 2019. URL https://arxiv.org/abs/1904.06151.

A. N. Gorban and I. Y. Tyukin. Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 376(2118):20170237, April 2018a. ISSN 1471-2962. doi: 10.1098/rsta.2017.0237. URL https://doi.org/10.1098/rsta.2017.0237.

Alexander N. Gorban and Ivan Y. Tyukin. Blessing of dimensionality: Mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2118):20170237, April 2018b. doi: 10.1098/rsta.2017.0237. URL https://doi.org/10.1098/rsta.2017.0237.

Daniele Granata and Vincenzo Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific Reports*, 6:31377, 2016. doi: 10.1038/srep31377. URL https://doi.org/10.1038/srep31377.

Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1):189–208, 1983. ISSN 0167-2789. doi: https://doi.org/10.1016/0167-2789(83)90298-1. URL https://www.sciencedirect.com/science/article/pii/0167278983902981.

Gloria Haro, Gregory Randall, and Guillermo Sapiro. Translated poisson mixture model for stratification learning. *International Journal of Computer Vision*, 80(3):358–374, December 2008. ISSN 1573-1405. doi: 10.1007/s11263-008-0144-6. URL https://doi.org/10.1007/s11263-008-0144-6.

Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 289–296, Bonn, Germany, August 7–11 2005. ACM. doi: 10.1145/1102351.1102383.

A. S. Holevo. *Probabilistic and Statistical Aspects of Quantum Theory*, volume 1 of *Quaderni del Consiglio Nazionale delle Ricerche*. Springer, Pisa, 2011. ISBN 978-88-7642-378-9.

I. M. James. *The Topology of Stiefel Manifolds*. London Mathematical Society Lecture Note Series. Cambridge University Press, 1977. doi: 10.1017/CBO9780511600753.

Kerstin Johnsson, Charlotte Soneson, and Magnus Fontes. Low bias local intrinsic dimension estimation from expected simplex skewness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37 (1):196–202, 2015. doi: 10.1109/TPAMI.2014.2343220.

Hamidreza Kamkari, Brendan Leigh Ross, Rasa Hosseinzadeh, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 38307–38354. Curran Associates, Inc., 2024. doi: 10.52202/079017-1211. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/43ab1646052dab79731f5d70bf40f6dc-Paper-Conference.pdf.

Bobak Kiani, Jason Wang, and Melanie Weber. Hardness of learning neural networks under the manifold hypothesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=dkkgKzMni7.

Kevin Knudson. Explicit formula for embedding projective spaces into euclidean space. Mathematics Stack Exchange, 2020. URL https://math.stackexchange.com/q/3505639. URL:https://math.stackexchange.com/q/3505639 (version: 2020-01-11).

Serge Lang. *Differential and Riemannian Manifolds*. Springer New York, 1995. doi: 10.1007/978-1-4612-4182-9.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

A.B. Lee, K.S. Pedersen, and D. Mumford. The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54:83–103, 2003. doi: 10.1023/A:1023705401078. URL https://doi.org/10.1023/A:1023705401078.

John M. Lee. *Introduction to Smooth Manifolds*. Springer New York, 2012. doi: 10.1007/978-1-4419-9982-5.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. In L. Saul, Y. Weiss, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/74934548253bcab8490ebd74afed7031-Paper.pdf.

German Magai and Anton Ayzenberg. Topology and geometry of data manifold in deep learning, 2022. URL https://arxiv.org/abs/2204.08624.

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/8a1e808b55fde9455cb3d8857ed88389-Paper.pdf.

Hariharan Narayanan and Partha Niyogi. On the sample complexity of learning smooth cuts on a manifold. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, Montreal, Canada, 2009.

Christopher Olah. Neural networks, manifolds, and topology. https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/, 2014. URL https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/. Blog post.

A M Perelomov. Generalized coherent states and some of their applications. *Soviet Physics Uspekhi*, 20(9): 703–720, September 1977. doi: 10.1070/pu1977v020n09abeh005459.

Askold Perelomov. *Generalized Coherent States and Their Applications*. Springer Berlin Heidelberg, 1986. doi: 10.1007/978-3-642-61629-7.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli. Novel high intrinsic dimensionality estimators. *Machine Learning*, 89(1):37–65, October 2012. ISSN 1573-0565. doi: 10.1007/s10994-012-5294-7. URL https://doi.org/10.1007/s10994-012-5294-7.

Santiago R Simanca. Canonical isometric embeddings of projective spaces into spheres, 2018. URL https://arxiv.org/abs/1812.10173.

David E Speyer. Does every irreducible representation of a compact group occur in tensor products of a faithful representation and its dual? MathOverflow, 2011. URL https://mathoverflow.net/q/58644. URL:https://mathoverflow.net/q/58644 (version: 2022-10-13).

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Piotr Tempczyk, Rafał Michaluk, Lukasz Garncarek, Przemysław Spurek, Jacek Tabor, and Adam Golinski. LIDL: Local intrinsic dimension estimation using approximate likelihood. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21205–21231. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tempczyk22a.html.

Erik Thordsen and Erich Schubert. Abid: Angle based intrinsic dimensionality — theory and analysis. *Information Systems*, 108:101989, 2022. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2022.101989. URL https://www.sciencedirect.com/science/article/pii/S0306437922000059.

Jing Wang, Zhenyue Zhang, and Hongyuan Zha. Adaptive manifold learning. In L. Saul, Y. Weiss, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/eb0ecdb070a1a0ac46de0cd733d39cf3-Paper.pdf.

Junya Watanabe. Statistics of eigenvalue dispersion indices: Quantifying the magnitude of phenotypic integration. *Evolution*, 76(1):4–28, January 2022. doi: 10.1111/evo.14382. URL https://doi.org/10.1111/evo.14382.

Andreas S. Weigend and Neil A. Gershenfeld (eds.). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Santa Fe Institute Studies in the Sciences of Complexity, Proceedings Volume XV. Addison-Wesley, Reading, MA, 1993. ISBN 9780201407852.

Jiayao Zhang, Guangxu Zhu, Robert W. Heath Jr., and Kaibin Huang. Grassmannian learning: Embedding geometry awareness in shallow and deep learning, 2018. URL `https://arxiv.org/abs/1808.02229`.

Wei-Min Zhang, Da Hsuan Feng, and Robert Gilmore. Coherent states: Theory and some applications. *Reviews of Modern Physics*, 62(4):867–927, October 1990. doi: 10.1103/revmodphys.62.867.

Yijia Zheng, Tong He, Yixuan Qiu, and David Wipf. Learning manifold dimensions with conditional variational autoencoders. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=Lvlxq_H96lI`.

Aneeq Zia, Ahmed Khamis, Joseph Nichols, Leonidas Guibas, Gunnar Carlsson, Remy Rampin, Brittany Fasy, Yusu Wang, Tamal Dey, P. Yian Lum, and Christopher Tralie. Topological deep learning: A review of an emerging paradigm. *Artificial Intelligence Review*, 57:77, 2024. doi: 10.1007/s10462-024-10710-9. URL `https://doi.org/10.1007/s10462-024-10710-9`.

A Licensing, Release, and Dataset Maintenance

The HomID dataset will be released under the **Creative Commons Attribution 4.0 International (CC BY 4.0)** license, allowing use, modification and redistribution with proper attribution. The dataset consists of embeddings for several quantum-inspired manifolds and contains no reference to any human data or sensitive information.

Release and Maintenance plan: The dataset will be released through a public Github repository. As presented in the Supplementary Material, scripts for generating embeddings of the different HomID manifolds will be uploaded, with clear instructions on the sampling process, including but not limited to relevant libraries, necessary computational budget, etc.

Due to memory constraints, we will not be releasing the actual data samples, especially very high-dimensional ones. Instead, we will release the scripts used to generate samples from a particular HomID manifold embedding.

The Github repo will be additionally tracked to allow users to report and fix bugs, additional manifold updates and so on. We will further invite the community to contribute to existing datasets through bug reports, suggestions for improvement and new dataset and feature suggestions. Each update will be accompanied by well-documented release notes, detailing changes and new requirements.

We anticipate integrating HomID with existing IDE libraries, providing additional avenues for long-term maintenance and community contributions. As such, since these are synthetic datasets, the original datasets can be maintained indefinitely. However, through the above practices, we hope to ensure that the dataset remains high-quality, well-maintained and easily accesible to the research community.

Intended Use: This dataset is intended primarily for research and benchmarking purposes. In particular, we envision these datasets to be useful and relevant for IDE performance evaluation and aspects of manifold hypothesis. Users are encouraged to cite the dataset and the accompanying paper when reporting results. While the dataset is released under an open license, any use must respect the intended research purpose and proper attribution requirements.

Ethics and Privacy Standards: No human subjects or sensitive information are involved in this dataset. The data is entirely synthetic, ensuring full compliance with privacy and ethical standards. Users are encouraged to adhere to best practices in computational research and reproducibility when using the dataset.

B Experimental details

We generated the data with the respective embeddings through custom scripts.

We tested 6 IDE methods — lPCA, MLE, CorrInt, TwoNN, DANCo and ABID. Among these, lPCA was implemented manually, while ABID was implemented via the methodology shared by the authors of (Thordsen & Schubert, 2022). All other methods were obtained directly from the `scikit-dimension` package. All methods involved computing k nearest neighbors (kNNs) which were pre-computed using the `Nearest Neighbors` module from sklearn.

For comparisons involving HomID and other benchmarks, we ran sweeps for the hyperparameter k and the sample size N for each of the methods. In particular, we performed three types of sweeps - sweeping N logarithmically from 100 to 10000, holding k fixed at 50, sweeping k logarithmically from 10 to 1000 while holding N fixed at 5000 and sweeping N from 100 to 10000 while holding k/N fixed at the values [0.08, 0.1, 0.15, 0.2, 0.5, 0.99]. However, as mentioned in the main text, in order to optimize computational resources, we had to make smaller sweeps for certain IDE-manifold combination. Any other hyper-parameters were held fixed at their default values, after small-scale experiments showed that the effect of these hyper-parameters were not as significant as compared to k and N. All runs were performed with 3 different random seeds.

For the scaling experiments, we held k fixed for all runs. The default values were $k = 100(200)$ for lPCA; $k = 100$ for MLE, ABID, TwoNN; $k = 10$ for DANCo; $k_1 = 10, k_2 = 20$ for CorrInt. In the case of $k >= N - 1$, in which case we chose $k = N - 2$; for CorrInt, this was modified as $k_2 = N - 2$ and $k_1 = k_2/2$. We used $k = 100$ for the experiment where we scaled the sample size N, while we used $k = 200$ for the case of scaling with noise. All other hyperparameters were kept at their default values.

Experiments (including data generation) were run on dual AMD 7763 64-core CPUs and took around 40,000 CPU hours with an approximate total of 1000 hours for data generation. Plots and inferences were then made locally with negligible overhead cost. [3]

There are two dominant sources of errors for our plots — spread in the local ID estimates for the entire sample, and the error due to using three different random seeds. The error reported is the mean error obtained for the local estimates, averaged over 3 random seeds.

C Detailed Background and Relevant Works

The manifold paradigm in ML The manifold hypothesis has been a well-known paradigm in the machine learning community (Gorban & Tyukin, 2018a; Olah, 2014; Cayton, 2008). This hypothesis has been subject to both theoretical and experimental investigations (Fefferman et al., 2013; Goldt et al., 2020; Narayanan & Mitter, 2010; Kiani et al., 2024; Brown et al., 2023; Donoho & Grimes, 2005; Lee et al., 2003). At the same time, there has been interest in understanding the topological structures of real-world datasets through the field of Topological Data Analysis (Carlsson et al., 2008; Zia et al., 2024). There is a separate notion involving manifolds in ML, which seeks to understand the learning process as modification of latent representations of the data manifold, as explored in (Magai & Ayzenberg, 2022; Ansuini et al., 2019).

Intrinsic Dimension The intrinsic dimension (ID) of a manifold is a very useful quantity for several reasons. It is a characteristic property of the manifold, and hence many properties or features of learning problems are dependent on the ID — dimensionality reduction (Wang et al., 2004; Gashler & Martinez, 2011; 2012), exponential scaling of samples with ID (Narayanan & Niyogi, 2009; Narayanan & Mitter, 2010), correlation of generalization capability with ID of data and internal representations (Pope et al., 2021; Ansuini et al., 2019; Birdal et al., 2021; Magai & Ayzenberg, 2022; Brown et al., 2022) as well as ID being a natural measure of local complexity of the data (Kamkari et al., 2024).

The importance of ID has thus led to a spate of research on coming up with estimators of ID. These method have different approaches to estimating LID — through various distance-based measures, such as those based on Euclidean distance measure (Camastra & Vinciarelli, 2002; Levina & Bickel, 2004; Gomtsyan et al., 2019;

[3] A typo in an earlier version reported the physical wall-clock time as the CPU hours.

Facco et al., 2017), non-Euclidean geodetic distances (Granata & Carnevale, 2016) and Wasserstein distances (Block et al., 2022); deviations of simplexes (Johnsson et al., 2015); angle-based measures (Ceruti et al., 2014; Thordsen & Schubert, 2022); measures based on generative models (Stanczuk et al., 2024; Tempczyk et al., 2022; Kamkari et al., 2024; Zheng et al., 2022) and quantum encoding-based algorithms (Candelori et al., 2025). See (Campadelli et al., 2015) for a more extensive survey of various ID estimators.

Benchmarks for IDE (Camastra, 2003) undertook one of the first surveys for ID estimators, outlining different datasets used for benchmarking ID estimators known at the time. (Hein & Audibert, 2005) constructed a series of manifolds to benchmark their ID estimator, including hyperspheres, isotropic Gaussian vectors, the Möbius strip, etc. However, the authors acknowledged the problem with evaluating ID estimators for MNIST, and attempted to construct datasets with variable ID by applying transformations such as translation, rotation, etc. on MNIST images. (Rozza et al., 2012) drew on these ideas and constructed several complicated manifolds, devising means of embedding data into higher dimensions without linear isometries. (Campadelli et al., 2015) took this further and proposed a benchmark with several such manifolds. A very interesting technique to benchmark ID estimators on distributions different from these manifolds was proposed in (Pope et al., 2021). In their paper, the authors propose using GANs with certain restriction on the dimension $\bar{d}$ of the latent noise vectors to generate images, whose ID was then bounded from above by $\bar{d}$. However, this method suffers from an obvious problem in that there is no exact ground-truth ID for the generated data. A few other examples with modern IDE methods include the use of "lollipop", Swiss rolls, double helix, mutliscale manifolds in (Tempczyk et al., 2022), synthetic image manifolds in (Stanczuk et al., 2024) and random affine transformations in (Kamkari et al., 2024). These benchmarks however are not simultaneously (i) scalable (ii) anisotropic or (iii) ID-preserving.

In addition, most ID estimators have a plethora of synthetic datasets on which they perform their zeroth-order evaluations - but most of them are low-dimensional, "simple" manifolds, whereas real-world data is more complicated and comes from vastly different distributions. There are some attempts to artificially induce complexity in these manifolds, examples of which are highlighted above. Scikit has compiled some of these manifolds into their python library (Bac et al., 2021). However, it contains only a few topologically non-trivial manifolds such as the Möbius strip, Swiss roll, helixes, etc. which do not admit natural embeddings for arbitrary intrinsic dimensions. Previous attempts to construct physics-inspired datasets for IDE evaluation involved nonlinear dynamical systems. These include the Santa Fe datasets (Weigend & Gershenfeld, 1993), DSVC1 dataset (Camastra & Filippone, 2009) and strange attractors (Grassberger & Procaccia, 1983). However, the true ID in these cases is not always known a priori, which makes it difficult to use as a benchmark. It is with this background that we propose HomID- an infinite family of manifolds which offer multiple natural notions of anisotropic embedding in higher-dimensional space. In addition, we propose a recipe to construct such manifolds through the concept of homogenous spaces and coherent states. The key novelty of this benchmark is that it is able to induce a particular complexity, namely anisotropy, while maintaining control over other aspects like density, curvature, etc.

D Details for construction of HomID manifolds

We present a brief overview of the manifolds included in HomID. The interested reader is referred to the references herein for a more in-depth discussion.

All manifolds included in the HomID benchmark are parameterized by quotients $\mathcal{G}/\mathcal{H}$, where $\mathcal{G}$ is a Lie group (i.e., a group that can also be considered as a manifold), and where $\mathcal{H}$ is one of its subgroups. Such spaces are called homogeneous spaces, and they are examples of quotient spaces. When $\mathcal{H}$ is continuous subgroup, the intrinsic dimension of such quotient spaces is $d_i = \dim \mathcal{G} - \dim \mathcal{H}$. When $\mathcal{H}$ is finite, $\dim \mathcal{H} = 0$, and therefore the intrinsic dimension is equal to the dimension of $\mathcal{G}$.

Our embeddings are constructed using the Gilmore-Perelomov method (Perelomov, 1977; 1986; Zhang et al., 1990), vectorization of matrices, and combinations thereof. We review Gilmore-Perelomov coherent-states in Section D.1 and use them to build the HomID manifolds in Section D.2.

D.1 Gilmore-Perelomov embeddings

Given the groups $\mathcal{H} < \mathcal{G}$ and some desired ambient dimension d_A in which we wish to embed the manifold $\mathcal{M} = \mathcal{G}/\mathcal{H}$, we must first construct an orthogonal representation Λ of $\mathcal{G}$[4]. Define $\Lambda|_{\mathcal{H}}$ to be the representation Λ of $\mathcal{G}$ restricted to $\mathcal{H}$. For compact Lie groups, all irreducible representations (irreps) appear in the isotypic decomposition of tensor-product representation (Speyer, 2011). Our starting point will therefore be to consider representations of the form $\Lambda(g) = g^{\otimes t}$, where g is in the defining irrep, but the recipe can be generalized to other representations.

The projector onto the trivial irreps in the isotypic decomposition of a representation Λ is $P(\Lambda) = \int_{\mathcal{G}} \Lambda(g)\, dg$, where dg denotes the unique unit-normalized Haar measure on $\mathcal{G}$. We suppose that A). $P(\Lambda|_{\mathcal{H}}) > P(\Lambda)$, meaning that there are nontrivial irreps of $\mathcal{H}$ contained in Λ, and B). that there are no subgroups $\mathcal{H} < \mathcal{K} < \mathcal{G}$ with $P(\Lambda|_{\mathcal{K}}) > P(\Lambda|_{\mathcal{H}})$.

Given the above assumptions, there exists at least one unit vector $v_{\mathcal{H}}$ that lies in the image of $P(\Lambda|_{\mathcal{H}})$ and does not lie in the image of $P(\Lambda|_{\mathcal{K}})$ for any $\mathcal{K} > \mathcal{H}$ (including $\mathcal{K} = \mathcal{G}$). By construction, $v_{\mathcal{H}}$ lies in a trivial irrep of $\Lambda|_{\mathcal{H}}$, and therefore $\Lambda(h)v_{\mathcal{H}} = v_{\mathcal{H}}$ for all $h \in \mathcal{H}$. Moreover, due to assumption (B), all other elements of $\mathcal{G}$ do not leave $v_{\mathcal{H}}$ invariant. Thus, because any $g \in \mathcal{G}$ can be uniquely written as $g = ah$ for a subgroup element $h \in \mathcal{H}$ and coset representative $a \in \mathcal{M} = \mathcal{G}/\mathcal{H}$, the mapping

$$\psi_{\mathcal{M}}(g) = \Lambda(g)v_{\mathcal{H}} = \Lambda(a)\Lambda(h)v_{\mathcal{H}} = \Lambda(a)v_{\mathcal{H}} \tag{2}$$

is a well-defined, injective mapping between $\mathcal{M}$ and $\mathbb{R}^{d_A}$ or $\mathbb{C}^{d_A}$ (see footnote 4). In fact, the embedding is injective into the real or complex sphere Ω_{d_A}.

This mapping is equivariant – $\Lambda(g_1)\psi_{\mathcal{M}}(g_2) = \psi_{\mathcal{M}}(g_1 g_2)$. It is, however, not in general isometric – this depends on the metric that is chosen on $\mathcal{M}$. In particular, if the metric g on $\mathcal{M}$ is induced from the Cartan-Killing metric on $\mathcal{G}$, then $\psi_{\mathcal{M}} \colon (\mathcal{M}, g) \to (\Omega_{d_A}, g_\Omega)$ will not be isometric, where g_Ω is the metric on the sphere. Of course, we could instead choose g to be the metric induced by the embedding; that is, the pullback $g = \psi^*_{\mathcal{M}} g_\Omega$. In this case, the embedding is isometric by construction.

Finally, $\psi_{\mathcal{M}}$ is an immersion simply because its derivative is everywhere injective. Specifically, suppose that A is a Lie algebra element so that $g(s) = g(0)e^{As}$ is a curve in $\mathcal{M}$. The embedded tangent space elements are then identified with $\frac{d}{ds}\psi_{\mathcal{M}}(g(s))|_{s=0}$.

D.1.1 Example of a Gilmore-Perelomov embedding

We use the method outlined above to parametrize points on the sphere:

$$\mathbb{S}^2 \cong SO(3)/SO(2)$$

Step 1. Pick a fixed vector: We choose the north pole which has the coordinates:

$$v_o = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

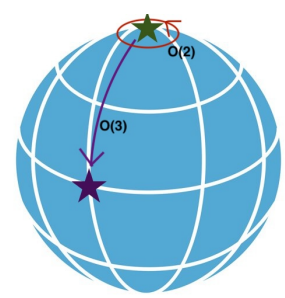


Figure 8: 2-sphere is O(3)/O(2).

Step 2. Representations of the groups:

- A general rotation $g \in SO(3)$ is parametrized by Euler angles:

$$g(\alpha, \beta, \gamma) = R_z(\alpha)R_y(\beta)R_z(\gamma)$$

where

$$R_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}$$

[4]When embedding into $\mathbb{R}^{d_A}$, the representation should be orthogonal; when embedding into $\mathbb{C}^{d_A}$, the representation should be unitary.

- The stabilizer subgroup $H = SO(2) \subset SO(3)$ consists of rotations about the z-axis:

$$h(\gamma) = R_z(\gamma)$$

Step 3. Invariance of the initial vector:

$$R_z(\gamma)v_0 = v_0 \qquad \Rightarrow \quad v_0 \text{ is invariant under } H$$

Incidentally these are the only rotations which preserve the north pole.

Step 4. Construction of the embedding:

$$v_g = g(\alpha, \beta, \gamma)v_0$$

Apply each part sequentially:

$$R_z(\gamma)v_0 = v_0, \quad R_y(\beta)v_0 = \begin{bmatrix} \sin\beta \\ 0 \\ \cos\beta \end{bmatrix}, \quad R_z(\alpha)\begin{bmatrix} \sin\beta \\ 0 \\ \cos\beta \end{bmatrix} = \begin{bmatrix} \sin\beta\cos\alpha \\ \sin\beta\sin\alpha \\ \cos\beta \end{bmatrix}$$

Final result:

$$\boxed{\psi_{\alpha,\beta} = \begin{bmatrix} \sin\beta\cos\alpha \\ \sin\beta\sin\alpha \\ \cos\beta \end{bmatrix}}$$

This parametrizes the full sphere $\mathbb{S}^2$ in the usual polar coordinate form.

D.1.2 Extension to natural data

The notion of group representations and fiducial vectors can be easily extended to their discrete counterparts. If there exists a natural vectorization of the data, one can make the correspondence $\mathbf{x} \to v_{\mathcal{H}}$ by setting $e_i, v_{\mathcal{H}} = x_i$ where e_i denotes some standard basis spanning a d-dimensional Hilbert space. One can then use these to generate an embedding for a homogenous space as outlined above.

As an example, consider the number '1' in MNIST. The image admits translation invariance, and so one can use any standard image representing '1' as the fiducial vector to represent a group $\mathcal{G}/\mathcal{T}$ where $\mathcal{T} < \mathcal{G}$ represents discrete translations.

D.2 Embedding manifolds into Euclidean space

Throughout this section, let $\{e_1, \ldots, e_n\}$ be the standard (unit-vector) basis of $\mathbb{R}^n$. Furthermore, for a matrix X, define $\vec{X}$ to be the vectorization of X – that is, the vector obtained by stacking the columns.

Stiefel manifolds and the "St (Matrix)" embedding

The Stiefel manifold is $\mathrm{St}(k, \mathbb{R}^n) = \left\{X \in \mathbb{R}^{n\times k} \mid X^T X = \mathbb{I}\right\}$ (James, 1977; Bendokat et al., 2020). This easily embeds into $\mathbb{R}^{nk}$ via the map $X \mapsto \vec{X}$. This yields the "St (Matrix)" embedding in the HomID dataset.

A more interesting embedding will be used as an intermediate step to derive the Grassmannian embedding used in HomID and is as follows. We will view $\mathrm{St}(k, \mathbb{R}^n)$ as $\mathrm{O}(n)/\mathrm{O}(n-k)$. Then a point $X \in \mathrm{St}(k, \mathbb{R}^n)$ is given by the first k columns of the corresponding orthogonal matrix. Using the notation from Section D.1, we let $\mathcal{G} = \mathrm{O}(n)$ and $\mathcal{H} = \mathrm{O}(n-k)$, and define the representation $\Lambda\colon \mathcal{O} \mapsto \mathcal{O}^{\otimes k}$ for $\mathcal{O} \in \mathrm{O}(n)$. An obvious

Family	**Symbol**	**Quotient space**	d_i	**HomID embedding**	$d_a^{\min}$
Stiefel manifold	$\mathrm{St}(k, \mathbb{R}^n)$	$\frac{\mathrm{O}(n)}{\mathrm{O}(n-k)}$	$nk - \frac{1}{2}k(k+1)$	St (Matrix)	nk
				St (Vec)	$\binom{n}{k} + k^2$
Grassmannian	$\mathrm{Gr}(k, \mathbb{R}^n)$	$\frac{\mathrm{O}(n)}{\mathrm{O}(n-k)\mathrm{O}(k)}$	$k(n-k)$	Gr (Proj)	n^2
	$\mathrm{Gr}^\star(k, \mathbb{R}^n)$	$\frac{\mathrm{O}(n)}{\mathrm{O}(n-k)\mathrm{SO}(k)}$	$k(n-k)$	Gr (Vec)	$\binom{n}{k}$
Flag manifold	$\mathrm{Flag}^\star(k_1, k_2, \mathbb{R}^n)$	$\frac{\mathrm{O}(n)}{\mathrm{SO}(k_1)\mathrm{SO}(k_2)\mathrm{O}(n-k_1-k_2)}$	$(k_1+k_2)n - k_1^2 - k_2^2 - k_1 k_2$	Flag (Vec)	$\binom{n}{k_1}\binom{n}{k_2}$

Table 2: Table listing manifold families used in the HomID benchmark, their mathematical symbols, their equivalent quotient/homogeneous spaces, and their intrinsic dimensions d_i. Here, $\mathrm{O}(n)$ denotes the orthogonal group in n dimensions, and $\mathrm{SO}(n)$ denotes the special orthogonal groups. The parameter $n \geq 1$. The last two columns list the five HomID embeddings and their lowest possible ambient dimensions, $d_a^{\min}$. Any embedding into a given ambient dimension can be further embedded into a space of larger ambient dimension via any isometry. The use of slightly different quotients for our two Grassmannian embeddings yields a lower possible ambient dimension for the latter embedding while maintaining the same intrinsic dimension. Note that the usual flag manifold can be obtained from $\mathrm{Flag}^\star(k_1, k_2, \mathbb{R}^n)$ by letting $\mathrm{SO} \to \mathrm{O}$. The symbol $\binom{a}{b}$ is the binomial coefficient.

choice for the state $v_{\mathcal{H}}$ is $e_1 \otimes \cdots \otimes e_k$. Given $\mathcal{O} \in \mathrm{O}(n)$, this yields the embedding

$$\begin{aligned}
\psi_{\mathrm{St}}(\mathcal{O}) &= \Lambda(\mathcal{O}) v_{\mathcal{H}} && (3)\\
&= \mathcal{O}e_1 \otimes \mathcal{O}e_2 \otimes \cdots \otimes \mathcal{O}e_k && (4)\\
&= \sum_{j_1,\ldots,j_k=1}^{n} \mathcal{O}_{j_1,1} \ldots \mathcal{O}_{j_k,k} e_{j_1} \otimes \cdots \otimes e_{j_k} && (5)\\
&= \mathcal{O}_{*,1} \otimes \cdots \otimes \mathcal{O}_{*,k}, && (6)
\end{aligned}$$

where $\mathcal{O}_{*,i}$ denotes the ith column of $\mathcal{O}$. From this we can easily see that $\psi_{\mathrm{St}}(\mathcal{O})$ depends only on the first k columns of $\mathcal{O}$, but will result in a different vector for different elements of the Stiefel manifold. It is therefore an injection of the Stiefel manifold into $\mathbb{R}^{n^k}$.

Grassmannians and the "Gr (Vec)" embedding

Consider the embedding $\psi_{\mathrm{St}}(\mathcal{O})$ of the Stiefel manifold from above. We need to slightly modify this to get an embedding of the Grassmannian,

$$\mathrm{Gr}^\star(k, \mathbb{R}^n) \cong \frac{\mathrm{St}(k, \mathbb{R}^n)}{\mathrm{SO}(k)} \cong \frac{\mathrm{O}(n)}{\mathrm{O}(n-k) \times \mathrm{SO}(k)}\,, \tag{7}$$

due to the extra quotient by $\mathrm{SO}(k)$. In particular, we define

$$v_{\mathcal{H}} = \frac{1}{\sqrt{k!}} \sum_{\sigma \in S_k} \mathrm{sgn}(\sigma) e_{\sigma(1)} \otimes \cdots \otimes e_{\sigma(k)}, \tag{8}$$

where $\mathrm{sgn}(\sigma)$ denotes the symmetric group on k elements.

We note that Eq. 7 is not the usual definition of the Grassmannian. Indeed, Grassmannian is typically $\mathrm{Gr}(k, \mathbb{R}^n) = \frac{\mathrm{O}(n)}{\mathrm{O}(n-k) \times \mathrm{O}(k)}$, and the oriented Grassmannian is $\widetilde{\mathrm{Gr}}(k, \mathbb{R}^n) = \frac{\mathrm{SO}(n)}{\mathrm{SO}(n-k) \times \mathrm{SO}(k)}$. These two spaces, along with the space in Eq. 7, all have the same ID. In future work, we will add the standard and oriented Grassmannian to HomID and further compare the performance of standard ID estimation methods on these three very similar manifolds. In particular, the geometry and topology of these three manifolds are slightly

$\mathcal{M}$	$=$	$\mathcal{G}/\mathcal{H}$	$\pi_1(\mathcal{M})$	$\pi_2(\mathcal{M})$
$\mathrm{Gr}(k,\mathbb{R}^n)$	$=$	$\mathrm{O}(n)/\mathrm{O}(n-k)\times\mathrm{O}(k)$	$\mathbb{Z}_2$	$\mathbb{Z}_2$
$\widetilde{\mathrm{Gr}}(k,\mathbb{R}^n)$	$=$	$\mathrm{SO}(n)/\mathrm{SO}(n-k)\times\mathrm{SO}(k)$	0	$\mathbb{Z}_2$
$\mathrm{Gr}^\star(k,\mathbb{R}^n)$	$=$	$\mathrm{O}(n)/\mathrm{O}(n-k)\times\mathrm{SO}(k)$	$\mathbb{Z}_2$	$\mathbb{Z}_2$

Table 3: The first two homotopy groups of the three types of "Grassmannians" described below Eq. 7.

different despite the manifolds themselves being morally very similar, thus yielding an interesting testing ground. For example, we list the first two homotopy groups of these manifolds in Table. 3, which we calulate using the long exact sequence of homotopy groups induced from the fiber bundle $\mathcal{H}\to\mathcal{G}\to\mathcal{G}/\mathcal{H}$.

Given $\mathcal{H}$, we achieve the embedding

$$\psi_{\mathrm{Gr}}(\mathcal{O}) = \Lambda(\mathcal{O})v_{\mathcal{H}} \tag{9}$$

$$= \frac{1}{\sqrt{k!}}\sum_{\sigma\in S_k}\mathrm{sgn}(\sigma)\mathcal{O}e_{\sigma(1)}\otimes\cdots\otimes\mathcal{O}e_{\sigma(k)} \tag{10}$$

$$= \frac{1}{\sqrt{k!}}\sum_{\sigma\in S_k}\sum_{j_1,\ldots,j_k=1}^{n}\mathrm{sgn}(\sigma)\mathcal{O}_{j_1,\sigma(1)}\ldots\mathcal{O}_{j_k,\sigma(k)}e_{j_1}\otimes\cdots\otimes e_{j_k} \tag{11}$$

$$= \frac{1}{\sqrt{k!}}\sum_{j_1,\ldots,j_k=1}^{n}\det(\mathcal{O}_{(j_1,\ldots,j_k)})e_{j_1}\otimes\cdots\otimes e_{j_k}, \tag{12}$$

where $\mathcal{O}_{(j_1,\ldots,j_k)}$ denotes the $k\times k$ matrix obtained from $\mathcal{O}$ by taking the first k columns and taking the rows $j_1,\ldots,j_k$, and similarly $\mathcal{O}_{\{j_1,\ldots,j_k\}}=\mathcal{O}_{\mathrm{sorted}(j_1,\ldots,j_k)}$.

Let $\sigma\in S_k$ be the permutation that sorts $(j_1,\ldots,j_k)$. Then $\det\mathcal{O}_{(j_1,\ldots,j_k)}=\mathrm{sgn}(\sigma)\det\mathcal{O}_{\{j_1,\ldots,j_k\}}$. Therefore, we have that

$$\psi_{\mathrm{Gr}}(\mathcal{O}) = \sum_{\substack{Q\subset\{1,\ldots,n\}\\|Q|=k}}\det\mathcal{O}_Q\frac{1}{\sqrt{k!}}\sum_{\sigma\in S_k}\mathrm{sgn}(\sigma)e_{\sigma(Q_1)}\otimes\cdots\otimes e_{\sigma(Q_k)} \tag{13}$$

$$= \sum_{\substack{Q\subset\{1,\ldots,n\}\\|Q|=k}}\det\mathcal{O}_Q u_Q, \tag{14}$$

where we defined Q_i to be the ith element of the set Q (of course this does not make sense generally, but because we are summing over all permutations, it is fine to pick some arbitrary ordering of the set), and we defined the state

$$u_Q = \frac{1}{\sqrt{k!}}\sum_{\sigma\in S_k}\mathrm{sgn}(\sigma)e_{\sigma(Q_1)}\otimes\cdots\otimes e_{\sigma(Q_k)}\ . \tag{15}$$

Notice that we are embedding into $\mathbb{R}^{n^k}$, but a full basis for the space is given by u_Q for all $Q\subset\{1,\ldots,n\}$ with $|Q|=k$. Thus, this embedding in general gives us an embedding into $\mathbb{R}^{d_A}$ for any $d_A\geq\binom{n}{k}$. We denote this by the shorthand "Gr (Vec)" in Table 2.

We note that this embedding is almost the Plücker embedding (Bendokat et al., 2020). The Plücker embedding is an embedding of the standard Grassmannian into real projective space, which is the real sphere with antipodal points identified. Above, we are embedding Eq. 7 into the real sphere. The reason that ψ_{Gr} is not an embedding of $\mathrm{O}(n)/\mathrm{O}(n-k)\times\mathrm{O}(k)$ is because the vector $v_{\mathcal{H}}$ is not invariant under the action $\mathrm{O}(k)$. Instead, it is invariant under the action of $\mathrm{SO}(k)$, and yields a ±1 phase factor under the action of $\mathrm{O}(k)$. This yields a well-defined embedding of $\mathrm{O}(n)/\mathrm{O}(n-k)\times\mathrm{O}(k)$ into projective space, but further embedding projective space into the sphere via the standard maps (e.g., $(x_i)\mapsto(x_ix_j)_{i\leq j}$ for projective-space vectors $(\cdots x_j\cdots)$) would come at a price of quadratically increasing the lowest possible ambient dimension (Simanca, 2018; Frost, 2021; Knudson, 2020). We thus stick with our original mapping in order to be able to run smaller-scale numerical experiments.

Stiefel manifold revisited: the "St (Vec)" embedding

A point in the Stiefel manifold can be represented by a point on the Grassmannian and a matrix $\mathcal{V} \in \mathrm{SO}(k)$. In other words, any element of $\mathrm{O}(n)/\mathrm{O}(n-k)$ can be expressed as an element of $\mathrm{O}(n)/\mathrm{O}(n-k) \times \mathrm{SO}(k)$ and an element of $\mathrm{SO}(k)$. Therefore, given an element $(\mathcal{O}, \mathcal{V})$ on $\mathrm{St}(k, \mathbb{R}^n)$, where $\mathcal{O} \in \mathrm{O}(n)$ represents a point on $\mathrm{Gr}^\star(k, \mathbb{R}^n)$ and $\mathcal{V} \in \mathrm{SO}(k)$, we can define an embedding

$$\tilde{\psi}_{\mathrm{St}} = \psi_{\mathrm{Gr}}(\mathcal{O}) \oplus \vec{\mathcal{V}} \ , \tag{16}$$

recalling that $\vec{\mathcal{V}}$ is the vectorized matrix $\mathcal{V}$. This embedding has dimension $d + k^2$, where $d \geq \binom{n}{k}$, and we denote this by "St (Vec)". This is much less than the dimension of the embedding ψ_{St}, which is n^k.

To generate random points on the Stiefel manifold with this embedding, we can just generate a random $\mathcal{O} \in \mathrm{O}(n)$ and a random $\mathcal{V} \in \mathrm{SO}(k)$ and then construct the embedding.

The "Gr (Proj)" embedding

Recall that $\mathrm{Gr}(k, \mathbb{R}^n) = \mathrm{O}(n)/\mathrm{O}(n-k) \times \mathrm{O}(k)$ is the manifold of k-dimensional subspaces of $\mathbb{R}^n$. Thus, we can uniquely represent a point on this manifold by a projector that projects onto this corresponding subspace. In particular, given an $n \times k$ orthogonal matrix $\mathcal{O}, \mathcal{O}^T\mathcal{O} = \mathbb{I}_{k\times k}$ representing a point on $\mathrm{St}(k, \mathbb{R}^n)$, we can create the projector $P_{\mathcal{O}} = \mathcal{O}\mathcal{O}^T$ that projects onto the span of the columns of $\mathcal{O}$. From this projector, we define the "Gr (Proj)" embedding as its vectorization $\vec{P}_{\mathcal{O}}$.

As is, the Gr (Vec) and Gr (Proj) embed different spaces — $\mathrm{O}(n)/\mathrm{O}(n-k) \times \mathrm{SO}(k)$ versus $\mathrm{O}(n)/\mathrm{O}(n-k) \times \mathrm{O}(k)$ — but the intrinsic dimension remains the same.

Flag manifolds and the "Flag (Vec)" embedding

In this section, we consider a general flag manifold $\frac{\mathrm{O}(n)}{\mathrm{SO}(k_1)\times\cdots\times\mathrm{SO}(k_t)\times\mathrm{O}(n-k)}$ where $\sum_{i=1}^t k_i = k$. We note, as with the Grassmannians, that our definition of the Flag manifolds also slightly differs from the standard. Namely, the typical Flag manifold is $\frac{\mathrm{O}(n)}{\mathrm{O}(k_1)\times\cdots\times\mathrm{O}(k_t)\times\mathrm{O}(n-k)}$. We will add these to HomID in future work.

We begin with the case $t = 2$, as presented in Table 2. Again, in the notation of Section D.1, we work with the representation $\Lambda\colon \mathcal{O} \mapsto \mathcal{O}^k$, and we use the fiducial vector

$$v_{\mathcal{H}} = \left(\frac{1}{\sqrt{k_1!}} \sum_{\sigma \in S_{k_1}} e_{\sigma(1)} \otimes \cdots \otimes e_{\sigma(k_1)} \right) \otimes \left(\frac{1}{\sqrt{k_2!}} \sum_{\sigma \in S_{k_2}} e_{\sigma(k_1+1)} \otimes \cdots \otimes e_{\sigma(k_1+k_2)} \right). \tag{17}$$

This yields the embedding

$$\begin{aligned}
\psi_F(\mathcal{O}) &= \Lambda(\mathcal{O}) v_{\mathcal{H}} && (18)\\
&= \tfrac{1}{\sqrt{k_1!k_2!}} \sum_{j_1,\ldots,j_k=1}^{n} \det\!\big(\mathcal{O}_{(j_1,\ldots,j_{k_1}),(1,\ldots,k_1)}\big) \det\!\big(\mathcal{O}_{(j_{k_1+1},\ldots,j_k),(k_1+1,\ldots,k)}\big) e_{j_1} \otimes \cdots \otimes e_{j_k} && (19)\\
&= \tfrac{1}{\sqrt{k_1!k_2!}} \sum_{\substack{Q\subset\{1,\ldots,n\}\\|Q|=k_1}} \sum_{\substack{P\subset\{1,\ldots,n\}\\|P|=k_2}} \sum_{\sigma\in S_{k_1}} \sum_{\pi\in S_{k_2}} \times && (20)\\
&\quad \times \mathrm{sgn}(\sigma)\mathrm{sgn}(\pi) \det\!\big(\mathcal{O}_{Q,(1,\ldots,k_1)}\big) \det\!\big(\mathcal{O}_{P,(k_1+1,\ldots,k)}\big) e_{\sigma(Q_1)} \otimes \cdots \otimes e_{\sigma(Q_{k_1})} \otimes e_{\pi(P_1)} \otimes \cdots \otimes e_{\pi(P_{k_2})} \\
&= \sum_{\substack{Q\subset\{1,\ldots,n\}\\|Q|=k_1}} \sum_{\substack{P\subset\{1,\ldots,n\}\\|P|=k_2}} \det\!\big(\mathcal{O}_{Q,\{1,\ldots,k_1\}}\big) \det\!\big(\mathcal{O}_{P,\{k_1+1,\ldots,k\}}\big)\, u_{QP}. && (21)
\end{aligned}$$

As is, this embedding is into $\mathbb{R}^{n^k}$, but because an orthonormal basis is given by tensor products of u_Q, we see that this embedding works into $\mathbb{R}^d$ for any $d \geq \binom{n}{k_1}\binom{n}{k_2}$. We call this the "Flag (Vec)" embedding.

A straightforward extension to general t yields an embedding of $\frac{\mathrm{O}(n)}{\mathrm{SO}(k_1)\times\cdots\times\mathrm{SO}(k_t)\times\mathrm{O}(n-k)}$ into $\mathbb{R}^d$ for any ambient dimension $d \geq \prod_{i=1}^t \binom{n}{k_i}$.

D.3 Extension to double coset spaces

The Gilmore-Perelomov embedding, introduced above, produces vectors $\psi_{\mathcal{G}/\mathcal{H}}(a)$ valued in a homogeneous space $\mathcal{M} = \mathcal{G}/\mathcal{H}$ for coset representatives a. Such vectors transform under elements $g \in \mathcal{G}$ according to the group's multiplication rules, $|\psi_{\mathcal{G}/\mathcal{H}}(a)\rangle \to |\psi_{\mathcal{G}/\mathcal{H}}(ga)\rangle$.

A double coset space $\mathcal{K}\backslash\mathcal{G}/\mathcal{H}$ can be constructed from these vectors by taking their orbits under the subgroup $\mathcal{K}$,

$$v_{\mathcal{K}a\mathcal{H}} = \frac{1}{\sqrt{|\mathcal{K}|}} \sum_{k\in\mathcal{K}} \psi_{\mathcal{G}/\mathcal{H}}(ka) \tag{22}$$

for all coset representatives a. Above, we use a sum to denote the linear combination of vectors, but this can substituted for an integral over the Haar measure in the case of a continuous group.

As an example, let us take the case of the sphere, $\mathcal{M} = \mathrm{SO}(3)/\mathrm{SO}(2)$, and consider $\mathcal{K} = \mathrm{SO}(2)$ to be the group of proper rotations around the **z**-axis. The orbit of a generic point on the sphere under this subgroup is the latitude on which the point is located. Therefore, double cosets are labeled by points in a fixed longitude connecting the north and south poles, which is no longer a manifold but an orbifold (Albert et al., 2020). The reason is because the poles do not move under $\mathcal{K}$-rotations and are singular points whose neighborhood does not resemble flat space.

Double-coset spaces for finite subgroups of SO(2) come with enlarged versions of a longitudinal line that are called spherical lunes. In such cases, the poles remain the two singular points of these orbifolds. In general, orbifolds of other discrete subgroups of SO(3) correspond to fundamental domains — patches of the sphere whose translates under $\mathcal{K}$ tile the entire sphere.

E Overview of IDE methods tested

lPCA : Since a manifold is locally isomorphic to $\boldsymbol{R}^{d_i}$, the directions normal to the hyperplanes have zero variance. Given k nearest-neighbor (NN) of a point sampled from $\mathbb{R}^D$, singular value decomposition (SVD) of the $k \times D$ matrix is performed to determine the principal components. The intrinsic dimension of the manifold can be then estimated by different means:

- 'maxgap' : the component showing the biggest spectral gap is returned as the intrinsic dimension, i.e.
$$\hat{d}_i(x;k) = \mathrm{argmax}_{j\in 1,\ldots,\min(D,k)-1} \frac{e_{j-1}}{e_j} \tag{23}$$
- 'ratio' : the minimum number of components needed to explain $1-\epsilon$ of the total variance.
$$\hat{d}_i(x;k,\epsilon) = \min\{j : R_j \geq 1-\epsilon\} \tag{24}$$
where $R_j := \frac{\sum_{i=1}^{j} \sigma_{[i]}^2}{\sum_{i=1}^{N} \sigma_{[i]}^2}$ is the cumulative variance ratio ($(.)_{[i]}$ denotes that the quantity is sorted from largest to smallest (algebriacally).
- 'fo' : the index j for which the (sorted) eigenvalues cross $(1-\epsilon)$ of the largest eigenvalue.
$$\hat{d}_i(x;k,\epsilon) = \min\{j : \sigma_{[i]}^2 \geq (1-\epsilon)\sigma_{[k]}^2\} \tag{25}$$

It is easy to check that these definitions are equivalent with suitable choices of the hyperparameter ϵ for manifolds with a *clear* spectral gap. In order to preserve computational resources, we therefore report the results of the 'maxgap' technique in the main section. We also checked results with the 'ratio' and 'fo' versions in several small-scale experiments. We noticed that in order to match the results, careful hyper-parameter tuning needs to be done for the other methods. As a result, we omit discussing these results here.

We created a custom function to compute the $\hat{d}_i$ according to Eq. 23 above, closely drawing from the implementation in scikit-dimension.

The following methods, namely MLE, DANCo, CorrInt and TwoNN were implemented by accessing the implementations directly from the `scikit-dimension` package.

MLE : Given a set of samples $X_1, X_2, ...X_n$, related to a sample in lower-dimensional space m and equipped with a smooth density f, one considers the Possion process $\lambda(t)$ of number of points in a small sphere around the point with radius t. The log-likelihood of this process (assuming a constant density $f(x)$ in the sphere of radius $R \geq t$) yields

$$\hat{d}_i(x; R) = [\frac{1}{N_R(x)} \sum_{j=1}^{N_R(x)} \log \frac{R}{T_j(x)}]^{-1} \tag{26}$$

where $T_j(x)$ is the Euclidean distance between x and its j-th NN. Some results on the statistics of these log distances were considered too.

(Haro et al., 2008) goes one step further and considers a noisy translation of the observed distances, which leads to a non-linear recursive equation, which they can solve self-consistently. In particular for isotropic Gaussian noise with scale σ, their ID estimate reads

$$\hat{d}_i(x; R) = [\frac{1}{N_R(x)} \sum_{j=1}^{N_R(x)} \frac{\int_{r=0}^{R} \exp(-\frac{(T_i - r)^2}{2\sigma^2}) \log(\frac{T_k}{r}) dr}{\int_{r=0}^{R} \exp(-\frac{(T_i - r)^2}{2\sigma^2}) dr}]^{-1} \tag{27}$$

CorrInt : For a set $X_1, ..., X_n$ of i.i.d. samples with a smooth density $f(x)$ in $\mathbb{R}^{d_i}$, the Euclidean distance between a point x and its k-th NN $T_k(x)$ satisfies

$$k/n \approx f(x) V_{d_i}(T_k(x)) \tag{28}$$

where $V_{d_i}(R)$ is the volume of a d_i-dimensional sphere of radius R. Since $V_{d_i}(R) \propto R^{d_i}$, the number of points k in a ball of radius R grows exponentially with d_i. This is the motivation behind the fractal dimension definition.

One computes the correlation integtral (sum) as

$$C(r) = \frac{1}{N(N-1)} \sum_{i>j} \mathbf{1}_{||\mathbf{x}_i - \mathbf{x}_j|| \leq r} \tag{29}$$

One can then estimate the dimension as

$$\hat{d}_i(x; k_1, k_2) = \frac{\log(C(r_2)/C(r_1))}{\log(r_2/r_1)} \tag{30}$$

where the hyper-parameters k_1 and k_2 are used to find the median distances r_1 and r_2

TwoNN : For a constant density ρ around a point x, the volume of the hyper-spherical shell between i and $i+1$-th NN is drawn from an exponential distribution in the volume $\Delta\nu_l = \omega_{d_i}(r_l^{d_i} - r_{l-1}^{d_i})$. Define $R = \frac{\Delta\nu_2}{\Delta\nu_1}$ and then it follows that $f(R) = (1+R)^{-2}$. It then follows that $f(\mu) = d_i \mu^{-d_i - 1}$ where $\mu = \frac{r_2}{r_1} \in [1, \inf)$. (Basically note that $f(\mu) = f(R)\frac{dR}{d\mu}$ and $R = \mu^{d_i} - 1$) To make matters less prone to computational errors, discard α of the largest $\frac{r_2}{r_1}$ values. Thus the ID estimate is given by

$$\hat{d}_i(x; k, \alpha) = -\frac{\log(1 - F(\mu))}{\log \mu} \tag{31}$$

where μ represents the ratio $\frac{r_2}{r_1}$ for the kNNs of the particular point.

DANCo : For a manifold $M \subseteq \boldsymbol{R}^{d_i}$, consider an embedding $\phi : \boldsymbol{R}^{d_i} \to \boldsymbol{R}^D$ which is locally isometric, smooth and possibly non-linear. Then the points in a local neighborhood are drawn uniformly from the

hyperspheres. The distribution for distances for the unit hypersphere normalized by the distance of the k-th NN follows the distribution

$$g(r; k, d_i) = kd_i r^{d_i - 1}(1 - r^{d_i})^{k-1}$$

while the mutual angles follow the von Mises-Fisher (VMF) distribution

$$q(\mathbf{x}; \nu, \tau) = \mathbf{C_{d_i}}(\tau) \exp(\tau \nu^{\mathbf{T}} \mathbf{x})$$

where $C_{d_i}(\tau)$ is a normalization constant. It should be noted that the parameter τ in the VMF distribution denotes the concentration of angles around the mean — the parameter $\tau = 0$ reducing this distribution to the uniform distribution on the sphere. The joint distribution of the normalized distance and mutual angles factorizes into the product of marginals for the unit hypersphere. The ID is then estimated by minimizing the KL-divergence between the theoretical and experimental joint distribution of normalized distance and mutual angles.

$$\hat{d}_i(x; k) = \text{argmin}_{d=1,\ldots,d_a} \int_{-\pi}^{\pi} d\theta \int_0^1 dr h_d(r, \theta) \log(\frac{\hat{h}_d(r, \theta)}{h_d(r, \theta)}) \tag{32}$$

where $\hat{h}$ refers to the experimental joint distribution, and $h(r, \theta) = g(r) \cdot q(\theta)$ is the theoretical joint distribution.

ABID : As discussed in (Thordsen & Schubert, 2022), the distribution of pairwise cosines between two points drawn randomly and uniformly from a d-ball (excluding the origin; also holds for any such spherical distribution) follows a Beta distribution on the interval [-1,1].

$$P(\cos\theta) = \frac{1}{2} B(\frac{1 + \cos\theta}{2}; \frac{d_i - 1}{2}, \frac{d_i - 1}{2})$$

from which it follows that

$$\mathbb{E}[\cos^2\theta] = d_i^{-1}$$

This motivates the following definition for the ABID ID estimator

$$\hat{d}_i(x; k) = (\mathbb{E}_{\mathbf{x}_i, \mathbf{x}_j \sim B_k(x)}[\cos^2(\mathbf{x}_i, \mathbf{x}_j)])^{-1} \tag{33}$$

where $B_k(x)$ denotes the ball containing the kNN of x, i.e. a ball of radius $T_k(x)$

F Additional Experiments

F.1 Local anisotropy diagnostics

We use a simple sample-based diagnostic for local anisotropy. For each anchor point x_i, we compute its k nearest neighbors in the ambient Euclidean metric, center the resulting neighborhood, and form the local covariance matrix

$$C_i^{(k)} = \frac{1}{m-1} \left(Y_i^{(k)}\right)^\top Y_i^{(k)}, \tag{34}$$

where $Y_i^{(k)}$ is the centered neighborhood matrix. Equivalently, if $Y_i^{(k)} = U\Sigma V^\top$, then the covariance eigenvalues are $\lambda_j = \sigma_j^2/(m-1)$. We normalize these eigenvalues by their sum and average the resulting spectra over anchor points.

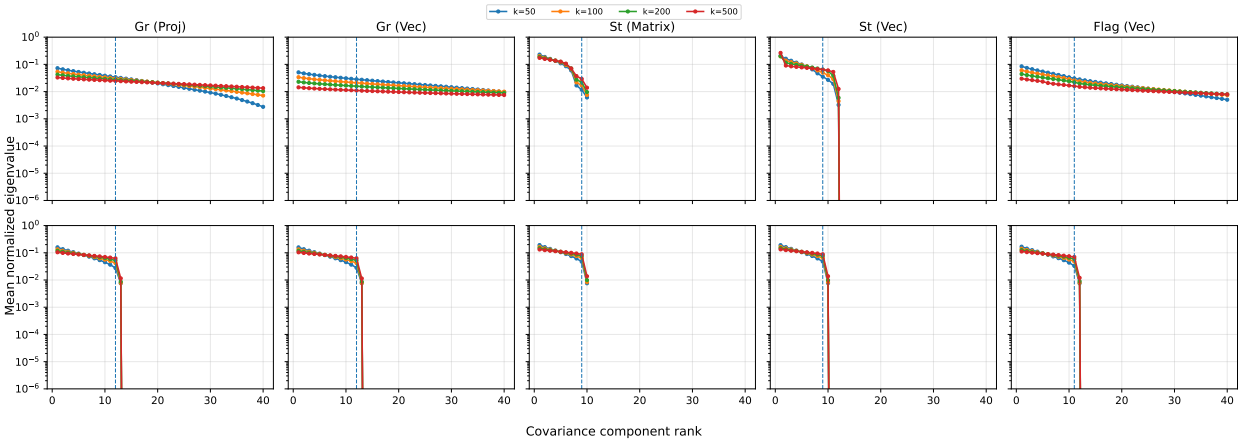


Figure 9: Local covariance spectra from centered kNN neighborhoods. Each HomID manifold is compared to a matched sphere with the same N, d_i, and d_a. Matched spheres show an approximately flat tangent spectrum up to d_i, followed by a sharp drop. Several HomID embeddings, especially the Grassmannian and flag embeddings, instead show a sloped tangent spectrum, indicating local covariance anisotropy.

For a locally isotropic d_i-dimensional reference manifold, such as a matched sphere, the first d_i normalized eigenvalues should be approximately flat at sufficiently small neighborhood scales, followed by a sharp drop. A sloped leading spectrum instead indicates that different tangent directions carry different amounts of local variance, which we interpret as local covariance anisotropy.

To control for trivial effects of sample size and dimension, we compare each HomID manifold to a matched sphere with the same N, d_i, and d_a. Figure 9 shows that several HomID embeddings, especially the Grassmannian and flag embeddings, have substantially more sloped local covariance spectra than their matched spheres. The Stiefel embeddings show weaker differences, suggesting that the degree of local covariance anisotropy is embedding-dependent (recall that Stiefel manifolds are the closest relatives of spheres as discussed in the main text).

This diagnostic should be interpreted as an operational proxy, not as a complete differential-geometric invariant. It depends on k, N, the embedding, and the ambient metric, and it can mix tangent anisotropy, curvature, finite-sample effects, and density fluctuations. More precise notions could be based on the pullback metric, the second fundamental form, principal curvatures, or direction-dependent geodesic-to-Euclidean distortion, but estimating these robustly and scalably from high-dimensional samples is itself difficult. We therefore use local covariance anisotropy as a practical diagnostic aligned with the local kNN structure used by the IDE methods studied in this work.

F.2 Comparison with other benchmarks

Here we present results of comparing our manifolds against other standard manifolds. In Appx. F.2, we present evidence that our manifolds are also adversarial as compared to other standard benchmarks. Note that one of our main considerations, was to compare these manifolds at fixed parameters and resources, hence we only include manifolds for which we can tune the d_i and d_a. This is the reason why we exclude some manifolds like the Moebius strip, torus, Swiss rolls, etc.

Isotropic Gaussian vectors Based on (Hein & Audibert, 2005), we compare our manifold against isotropic Gaussian vectors $\in \boldsymbol{R}^{d_i}$ linearly embedded into $\boldsymbol{R}^{d_a}$, cf. Fig. 10

Affine spaces Affine spaces are isomorphic to the linear nullspaces we considered in our main text. Given d_i, d_a, the nullspace of matrix $A \in \boldsymbol{R}^{d_a \times d_a}$ with rank $d_a - d_i$ consists of a hyperplane of intrinsic dimension d_i. In order to preserve the "smoothness" of our manifold, we sample the coefficients of the basis vectors of the nullspace from the standard unit normal, cf. Fig‘10

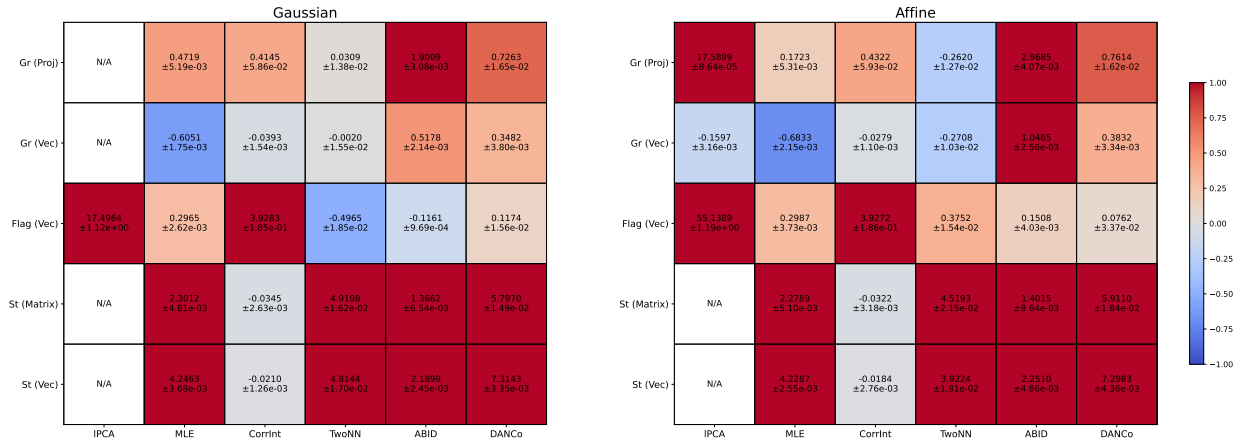


Figure 10: Relative performance of IDE methods on HomID manifolds compared against (left) normal isotropic Gaussian vectors $\mathcal{N}$ and (right) affine linear nullspace $\mathcal{L}$. N/A indicates that the reference error is zero or numerically negligible, making the relative ratio unstable.

For the purposes of this experiment, we run extensive scaling experiments by choosing small-dimensional manifolds. In particular we choose Grassmannians with ID of 2,3,4,5; Stiefels with ID of 3,5; Flags with ID of 4,12.

F.3 Scaling with sample size

The dependence of the intrinsic dimension estimate $\hat{d}_i$ comes from the fact that the fraction of neighbors in a small neighborhood for manifolds are given by (Levina & Bickel, 2004)

$$\frac{k}{N} = \Omega_{d_i} \rho(x) T_k^{d_i} \tag{35}$$

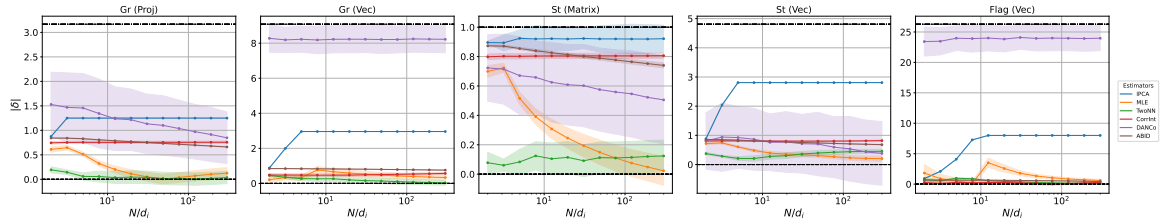
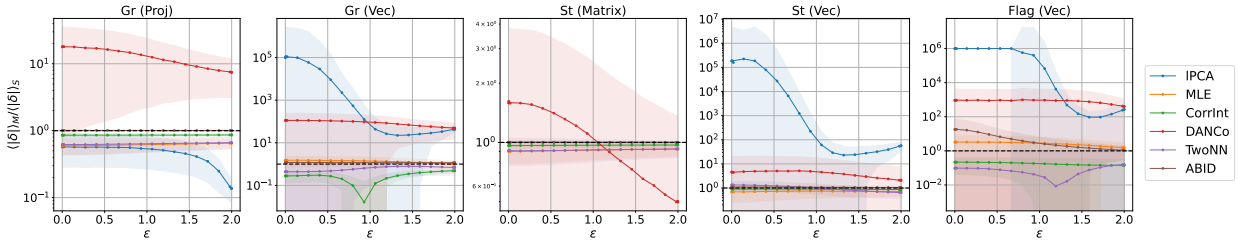


Figure 11: Performance of IDE methods as a function of N/d_i. We observe a gradual convergence with increasing sample size. The shaded region shows the 1-σ error in the local ID estimates, averaged over three seeds.

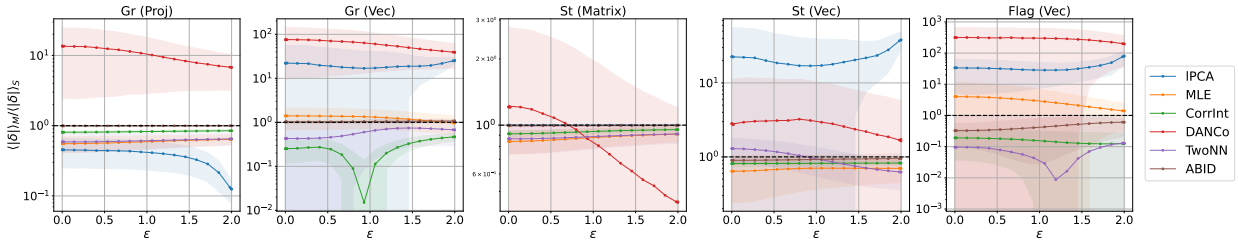
Thus in order to get a uniform and locally dense sampling of points, we need to hold $k/(T_k^{d_i} N)$ fixed, which translates to the condition that

$$N \propto \exp(d_i)$$

We investigate the effect of changing N while holding hyper-parameters fixed. We logarithmically sample N so that N/d_i goes from 2 to 300. As Fig 11 suggests, the ID estimates saturates with exponentially many samples, thus we always choose the maximum value of N possible for our experiments.



(a) Hyperspheres



(b) Linear nullspaces sampled from isotropic Gaussian

Figure 12: Comparison of errors when distorting HomID manifolds versus isotropic manifolds. The large change indicates that these manifolds become drastically more difficult to perform IDE on with increasing distortion, as opposed to HomID manifolds.

F.4 Squeezing : comparison with isotropic manifolds

We present in Fig. 12 the effect of distortion for spheres, as compared to HomID manifolds. In particular, we observe that spheres exhibit a very large increase in the error rate as anisotropy is increased. In conjunction with Section 5.3, this demonstrates that HomID manifolds are already challenging enough for the IDE methods tested, and enhances it applicability as a more robust performance evaluator for IDE.

G Additional Plots

G.1 Additional plots for anisotropic distortion and noise perturbation

We present here the additional plots for anisotropic distortion and additive noise perturbation for completeness. The plots presented in the main text are a subset of these plots.

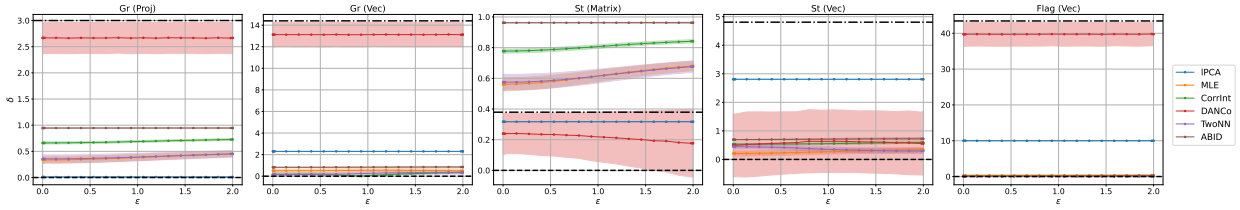
G.2 Null results on correlation of IDE performance

We analyze IDE performance in terms of different statistical and geometric features of the data.

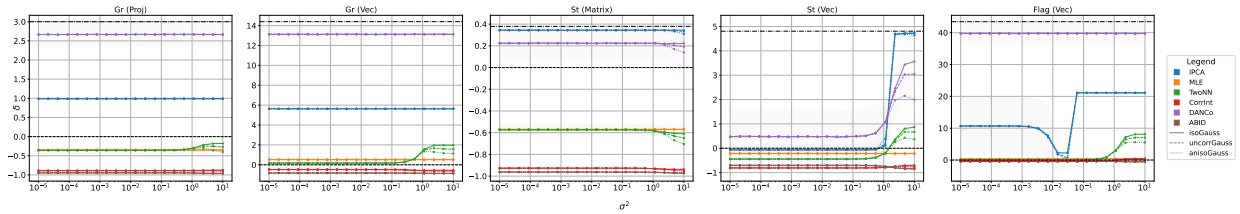
Statistical properties We look at three important features of the data covariance matrix $\mathbf{\Sigma}$: (1) the total variance given by Tr $\mathbf{\Sigma}$, (2) the variance dispersion index (VDI) (Watanabe, 2022) which is a direct measure of the global anisotropy of the data and is given by $\frac{\text{Var}(\lambda)}{(\text{Mean}(\lambda))^2}$ (λ refers to the eigenvalues of $\mathbf{\Sigma}$), and (3) the $\langle R \rangle^2$ value defined as $\frac{1}{d_a^2} \sum_{i,j} \frac{\mathbf{\Sigma}_{ij}^2}{\mathbf{\Sigma}_{ii}\mathbf{\Sigma}_{jj}}$, which captures the inter-component correlation. Note that anisotropy does not imply that components are correlated, but correlated components necessarily imply anisotropy.

We observe very little statistically significant correlation of IDE performance with the total variance and $\langle R^2 \rangle$ for these methods. There is a slight negative correlation of performance with VDI, the effect being most prominent for the angle-based methods DANCo and ABID, but nothing statistically significant.

The plots for these results are presented in Fig. 14.

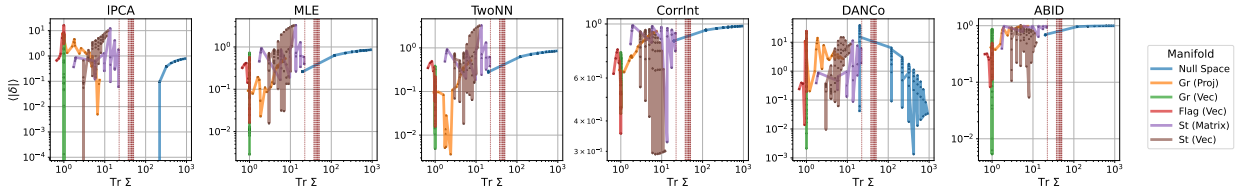


(a) Anisotropic distortion experiments (Sec. 5.3).

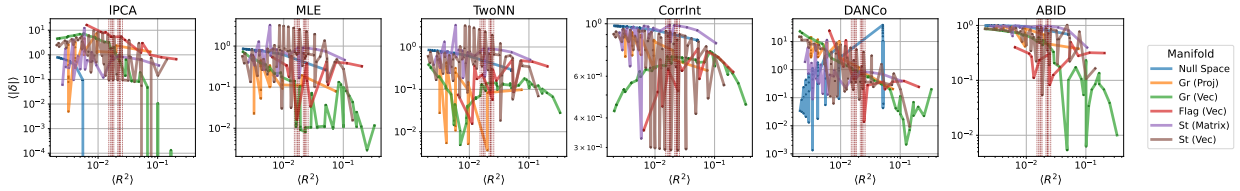


(b) Additive perturbation experiments (Sec. 5.3).

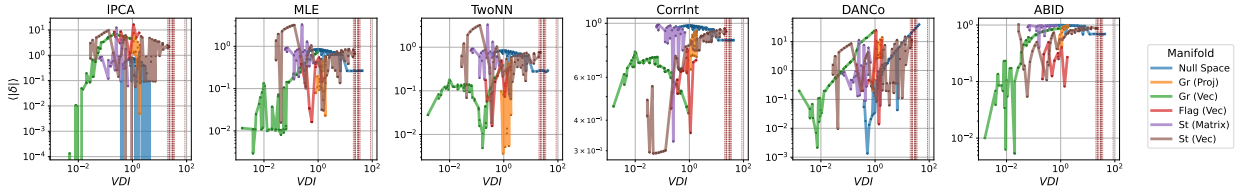
Figure 13: Full experimental results on all HomID manifolds. Top: anisotropic distortion (squeeze) experiments. Bottom: additive perturbation (noise) experiments.



(a) Tr Σ



(b) $\langle R^2 \rangle$



(c) VDI

Figure 14: IDE performance seems to be almost independent for these various quantities of interest. The light vertical maroon lines represent the corresponding quantities for different classes in MNIST.

Geometric Properties We measure the local mean curvature $H(x)$ and the local density $\rho(x)$. From these two measurements, we calculate a dimensionless parameter $\kappa(x) = \rho/H^{d_i}$. In order to find the curvature, we fit a quadratic surface in a local neighborhood and use standard results due to Gauss. We sample a kNN-neighborhood and estimate the density based on the ratio $k/\Omega_{d_i} T_k^{d_i}$. We sample 5 logarithmically-spaced k values and choose the median value as the desired geometric property. However, we fail to observe any

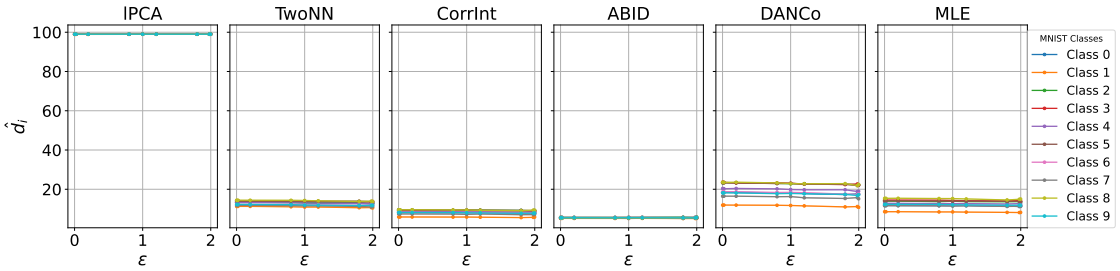
definite correlation of IDE performance with these geometrical quantities as well. The tabulated results are shown in Table 4

Quantity	Gr (Vec)	Gr (Proj)	St (Matrix)	St (Vec)	Flag (Vec)
$\langle H \rangle$	1.0313	1.1417	0.6702	0.5348	1.5032
$\langle \rho \rangle$	0.8527	0.5933	0.1766	0.2846	2.9473
$\langle \kappa \equiv \rho / H^{d_i} \rangle$	0.7758	0.3736	0.9097	2.1033(*)	0.5772
$\langle \lvert\delta\rvert \rangle$	**0.2023**	0.5785	0.5793	0.8502	2.4085

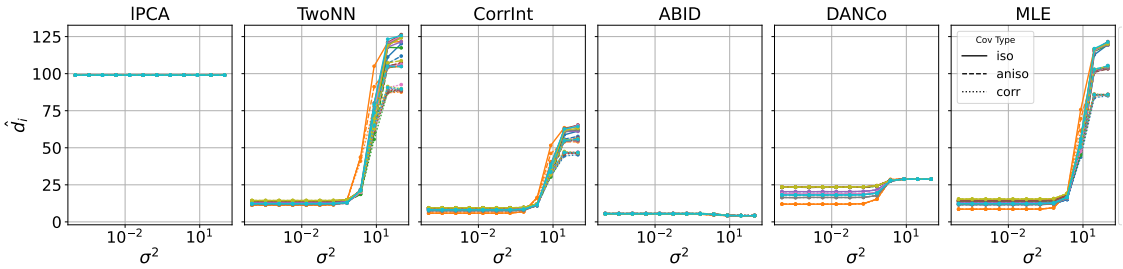
Table 4: Average absolute values by manifold (manifolds sorted by increasing $\langle |\delta| \rangle$).
(*) This number was skewed by a manifold with intrinsic dimension $d_i = 12$ which we do not include in the table. Including that we get 4080.9902.

G.3 Additional plots on MNIST

We provide additional plots here regarding our distortion experiments in MNIST. As emphasized in the main text, our observations are similar qualitatively to the analogous experiments for HomID manifolds. See Fig. 15



(a) Squeezing perturbation



(b) Additive noise perturbation

Figure 15: Effect of (a) distortion and (b) additive noise perturbation on MNIST dataset. We observe minimal degradation in (a) and degradation when $\sigma^2 \sim \boldsymbol{O}(1)$ in (b), both observed for HomID manifolds as well.

G.4 Additional lPCA interp plots

We present additional plots on lPCA here in the appendix. As alluded to in the main text, our main conjecture is that the kNN-based IDE methods fail due to deviations of their expected distributions. In the case of lPCA, this is true for distorted spheres and HomID manifolds, cf. Fig. 16.

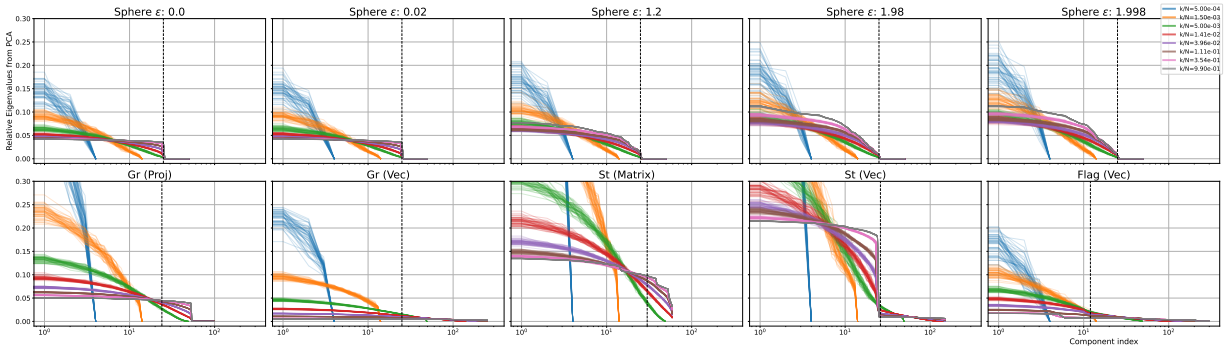


Figure 16: Extra plots for lPCA various manifolds. ϵ denotes the degree of anisotropy — higher ϵ indicates a higher anisotropy. Notice that only the undistorted sphere (top left) shows the expected behavior of a sharp drop at the ID.