# QUANTUM-INSPIRED BENCHMARK FOR INTRINSIC DIMENSION ESTIMATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Machine learning models can generalize well on real-world datasets. According to the manifold hypothesis, this is possible because datasets lie on a latent manifold with small intrinsic dimension (ID). There exist many methods for ID estimation (IDE), but their estimates vary substantially. This warrants benchmarking IDE methods on manifolds that are more complex than those in existing benchmarks. We propose a Quantum-Inspired Intrinsic-dimension Estimation (QuIIEst) benchmark consisting of infinite families of topologically non-trivial manifolds with known ID. Our benchmark stems from a quantum-optical method of embedding arbitrary homogeneous spaces while allowing for curvature modification and additive noise. The IDE methods tested were generally less accurate on QuIIEst manifolds than on existing benchmarks under identical resource allocation. We also observe minimal performance degradation with increasingly non-uniform curvature, underscoring the benchmark's inherent difficulty.

## 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 (40; 31; 64; 20). 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).

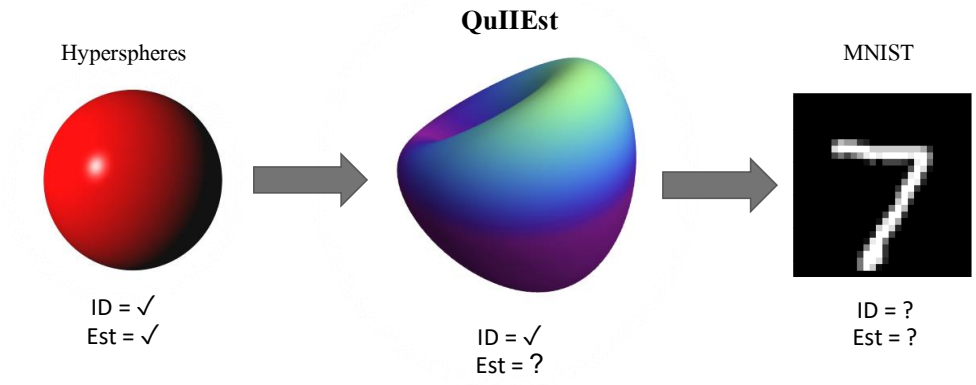
The manifold hypothesis is quite intriguing and has been subject to many experimental investigations, whereby different methods have been proposed to estimate the ID of a particular data-cloud, which have been then applied to popular ML datasets such as MNIST (56; 24). However, a common theme is the disagreement in the estimated IDs of different methods on these real-world datasets (44; 29; 69; 5; 36; 17; 73; 75; 53). 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.

Hence, it is important to benchmark existing methods against datasets consisting of more complicated manifolds whose ground-truth features and IDs are known. There have been a few proposed benchmarks (44; 53; 14; 4; 52) that include manifolds like spheres, hyper-cubes, Swiss rolls, Möbius strips, among others. However, each of these datasets comes with their own flaws — low dimensionality, presence of singular points, and, most importantly, lack of tractable yet non-trivial infinite families of manifolds with varying ID.

In this light, we propose the **QuIIEst** (**Qu**antum-**I**nspired **I**D **Est**imation) benchmark — a collection of synthetic datasets sampled from non-trivial manifolds constructed with tools from quantum information theory. We highlight our contributions in this regard below.

**Summary of Contributions:**

1. We propose a set of real and complex-valued *families of manifolds* with known ground-truth IDs to serve as a benchmark for current and future IDE techniques. By having an infinite *family* of manifolds, one can probe the effects of dimensionality on IDE while sampling from the same non-trivial distribution. The full comparison of advantages are summarized




| Property | Spheres | Gaussian vectors | Möbius strips | Nonlinear manifolds | Affine spaces | **QuIIEst** |
|---|---|---|---|---|---|---|
| Non-trivial topology | × | × | ✓ | × | × | ✓ |
| Scalable manifold families | ✓ | ✓ | × | ✓ | ✓ | ✓ |
| Multiple natural embeddings | × | × | × | × | × | ✓ |

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 QuIIEst- a family of topologically non-trivial manifolds to serve as an **intermediate confidence evaluation** for IDE. The QuIIEst dataset contains several different embeddings of infinite families of manifolds whose dimension is polynomial in their parameters, and which admit nontrivial geometry and topology.

in Figure 1. We also provide an easily generalizable framework to create embeddings of any manifolds that are also homogeneous spaces.

2. We demonstrate that the IDs estimated for QuIIEst manifolds deviate from the ground-truth. We show that our manifolds are more challenging relative to standard benchmarks, even at low dimensionality, under identical resource allocation.
3. We investigate IDE performance for distorted manifolds to understand aspects of manifolds not deriving purely from symmetry groups. We observe minimal degradation when we asymmetrically distort the manifold, indicating that QuIIEst is *already challenging enough*.
4. We leverage the scalability of QuIIEst manifolds to investigate scaling patterns with respect to ground-truth ID and sample size.
5. We observe that the performance of tested methods is only weakly correlated with the anisotropy and degree of correlation between different components of the data.

## 2 Background and Related Works

**Manifolds in ML** The manifold hypothesis states that real-world data lie on low-dimensional manifold, even though represented in high dimensions. (39; 20). This idea has been explored both theoretically and experimentally (31; 36). Topological properties of such data manifolds (19) and latent representations (2) have also been studied.

**Intrinsic Dimension** Intrinsic dimension (ID), as the name suggests, characterizes several inherent properties of the manifold (63; 69). A plethora of estimators have been proposed for ID estimation (IDE), which we discuss in Appendix C. See (16) for a detailed survey.

**Benchmarks for ID Estimation** Practical benchmarks for IDE involve topologically simple manifolds like hyperspheres or non-scalable manifolds like the Möbius strip (14; 44; 71; 16; 4). An interesting benchmark proposed GAN-based data to lower bound ID (69). Physics-inspired datasets (80; 15; 42) typically use non-linear dynamical models. However, the exact ground truth ID is unknown. QuIIEst, as we discuss, alleviates these concerns.

**Quantum and ML** Unlike quantum machine learning or quantum-inspired algorithms (21; 74), we take inspiration from quantum optics and quantum information theory to generate diverse, well-defined manifolds for testing ID estimators. Specifically, our benchmark uses Gilmore-Perelomov coherent states (65; 83) that correspond to the most "classical" states of a quantum system.

A more comprehensive discussion of previous work may be found in Appendix 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 $\mathbb{R}^{d_i}$ (58). 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 (53).

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 $\mathbb{R}^{d_a}$ (58). 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 many 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 (23). Importantly, though, a $d_i$-dimensional manifold can have nontrivial topology and geometry that makes it starkly different than $\mathbb{R}^{d_i}$.

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. We report the result of experiments with other statistical quantities, such as median and mode, in Appendix G.7.

**Homogeneous spaces** All manifolds included in the QuIIEst 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 Appendix 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.

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 $d_i = \dim \mathcal{G} - \dim \mathcal{H}$.

The remainder function, for which $\mathcal{G} = \mathbb{R}$ and $\mathcal{H} = \mathbb{Z}$, is an example of a quotient by a discrete, or finite, subgroup. The remainder is obtained from a real number $r$ by subtracting the closest integer less than or equal to $r$, and all remainders lie in the interval $[0, 1)$. This domain is periodic — the remainder cycles as $r$ increases past an integer — demonstrating that $\mathbb{R}/\mathbb{Z} \cong S^1$, the circle.

Homogeneous spaces with finite $\mathcal{H}$ can be thought of as subsets of points of the numerator with higher dimensional periodic identifications. Their intrinsic dimension is $d_i = \dim \mathcal{G}$ since finite groups are zero-dimensional.

| Manifold | Gr (Proj) | Gr (Vec) | St (Matrix) | St (Vec) | Flag (Vec) | Pauli |
|---|---|---|---|---|---|---|
| Int dim $d_i$ | 6-24 | 2-36 | 9-434 | 2-65 | 3-12 | 4-25 |
| Amb dim $d_a$ | 25-900 | 3-924 | 10-660 | 7-960 | 9-300 | 32-1250 |

Table 1: Table detailing the range of the datasets utilized.

## 4 Proposed Datasets

Coherent states are the states of a quantum system that are the closest, in both a technical and a heuristic sense, to the states the system can assume in the classical limit. Such states often parameterize a particular well-behaved manifold, such as the sphere or complex plane, but coherent states lying on more general manifolds are relevant to quantum information, quantum metrology (46) and, in the case of Grassmanians, the quantum Hall effect (12). Our framework uses several physically relevant manifolds as test beds for IDE.

Our embeddings are constructed using the Gilmore-Perelomov coherent-state method (65; 66; 83), vectorization of matrices, and combinations thereof. The coherent-state method should generalize to datasets with a natural notion of vectorization, e.g. images, embedded text tokens, etc.

Table 1 contains a summary of the advantages of our manifolds as compared to other benchmarks. We overview the manifold families below and present explicit constructions in Appendix D. Information about licensing, maintenance and dataset release can be found in Appendix A

**Stiefel manifolds** Stiefel manifolds are the closest relatives of spheres out of all QuIIEst manifolds. Real Stiefel manifolds are parameterized by quotients of the form $\mathrm{O}(n)/\mathrm{O}(n-k)$ for $n \geq k \geq 1$ (50), while real spheres are equivalent to $\mathrm{SO}(n)/\mathrm{SO}(n-1)$.

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 (50).

The QuIIEst 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 Gilmore-Perelomov coherent-state method and a vectorization of a matrix.

**Grassmanians** Grassmanians, or Grassman manifolds, are defined as $\mathrm{Gr}(k, \mathbb{R}^n) \cong \mathrm{O}(n)/\mathrm{O}(n-k) \times \mathrm{O}(k)$ (58) and can be thought of as quotients of Stiefel manifolds by an extra $\mathrm{O}(k)$ subgroup. Grassmanians have a long history in ML (82; 6). 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 QuIIEst benchmark includes two different embeddings of Grassman manifolds. The first is based on the interpretation of the Grassmanian 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 Grassmanians parameterize all distinct $k$-dimensional subspaces of $n$-dimensional space. We represent a point on the Grassmanian by an $n$-dimensional projector onto a $k$-dimensional subspace, which yields our "Gr (Proj)" embedding when written as an $n^2$-dimensional vector.

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 (58; 6).

**Flag manifolds** Flag manifolds generalize Grassmanians 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.

**Pauli quotients** This homogeneous space family of real dimension $2n^4$ consists of quotients by a discrete subgroup. It is of the form $\mathrm{U}(n)/\mathcal{P}_n^\star$, where $\mathcal{P}_n^\star$ is a finite subgroup of the unitary group that is defined in Appendix D and that is closely related to the Pauli (a.k.a. Heisenberg-Weyl) group. The

| IDE Method | Gr (Proj) | Gr (Vec) | Flag (Vec) | St (Matrix) | St (Vec) | Pauli | Average (QuIIEst) |
|---|---|---|---|---|---|---|---|
| lPCA | 1.5492 | 0.0700 | 3.5703 | 0.6210 | 1.4899 | 0.7491 | 1.3416 |
| MLE | 0.1758 | 0.0474 | 0.3065 | 0.4613 | 0.7327 | 0.3913 | 0.3525 |
| CorrInt | 1.0068 | 0.6832 | 4.0588 | 0.7366 | 0.7470 | 0.7570 | 1.3316 |
| TwoNN | 0.0731 | 0.0713 | 0.1970 | 0.4536 | 0.4376 | 0.4456 | **0.2797** |
| ABID | 0.3571 | 0.1858 | 0.4538 | 0.4588 | 0.6182 | 0.0956 | 0.3616 |
| DANCo | 0.3090 | 0.1558 | 5.8647 | 0.7446 | 1.0756 | 1.0692 | 1.5365 |
| **Average** | 0.5785 | **0.2023** | 2.4085 | 0.5793 | 0.8502 | 0.5846 | **0.8672** |

Table 2: Mean relative error $|\delta|$ for various methods on QuIIEst manifolds. A higher value indicates worse performance. We see that the vector embedding of Grassmanian consistently has low error for all methods, while TwoNN typically performs the best on all manifolds. Note however that the native `scikit-dimension` implementation of TwoNN often fails to return an estimate.

corresponding embedding, called "Pauli", is constructed using the coherent-state method with the help of recent results in quantum information theory (8).

**Fractals** The definition of manifolds entails that the ID is *the same at every point*. In Appendix H, we discuss ID for fractals, which do not satisfy this definition and which are consequently not included directly in the QuIIEst benchmark. In particular, we discuss Hofstadter's butterfly (45) as an example of a fractal curve inspired from quantum physics.

These manifolds may appear exotic, but they are not as far-removed from real-world data as it seems. Grassmanians (81; 48) and Stiefel (78; 61) manifolds have been studied in ML before, with the former relevant to airfoil design (27). They are examples of manifolds derived from Lie group symmetries, which have approximate discretized counterparts for real data (38). We simulate loss of symmetry due to discretization by applying controlled distortion to QuIIEst manifolds in Secs. 5.3 and 5.4.

# 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 (18; 32; 30)] and non-linear methods such as maximum likelihood estimation [MLE (59; 43)], fractal dimension estimation [CorrInt (42; 13)], distribution of measure [TwoNN (29)], concentration of measure [DANCo (22)], and angle-based moments [ABID (76)]. We review these methods, including their implementation, in Appendix E.

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

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

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 QuIIEst manifolds is summarised in Table 2, 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 QuIIEst. We perform hyper-parameter sweeps and, for each hyper-parameter combination, compare the performance of different IDE methods on QuIIEst to other IDE benchmarks. 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 "Gr (Vec)".

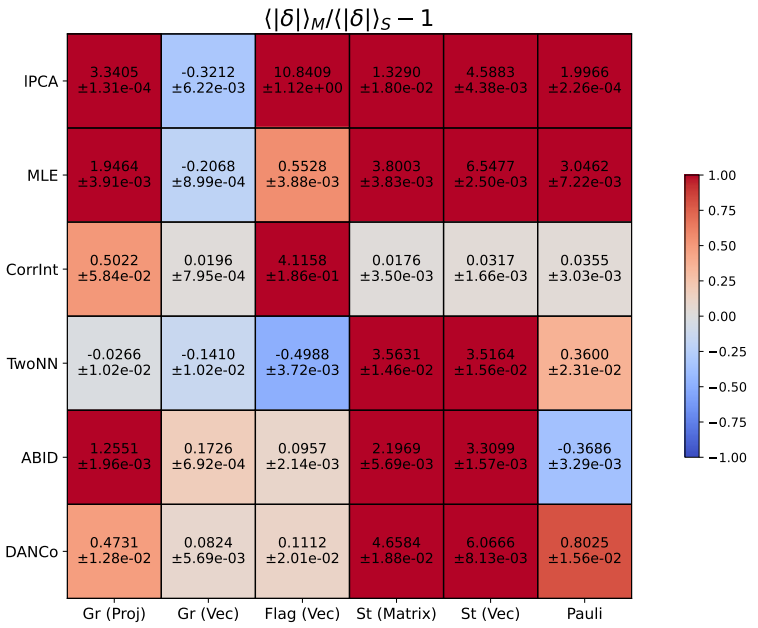



Figure 2: Comparison of the quantity $\langle|\delta|\rangle_M/\langle|\delta|\rangle_S - 1$, where the numerator is the average of the absolute value of the relative error $|\delta|$ over all instantiations of a given manifold family $M$, while the denominator is the corresponding average over all sphere 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 Grassmanian, 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-$\sigma$ sampling error is reported here after the $\pm$ sign.

For brevity, we present the relative result on spheres here in the main text, cf. Fig 2. The reader is referred to Appendix F 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 discussed in F.

### 5.3 Anisotropic distortions : "Squeezing"

We now test IDE methods on distorted versions of QuIIEst manifolds, naturally obtained by amending the coherent-state method with generalized “spin squeezing” effects from quantum optics.

Distorted versions of our manifolds are obtained by applying a fixed random diagonal matrix to the manifold vectors. The strength of distortion is governed by a parameter $\epsilon$, 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 underlying manifolds, cf. Fig 3. 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 G.2.

### 5.4 Distortion through additive noise

We perform experiments by perturbing our data from particular manifolds with additive noise, i.e. $\mathbf{x} \to \mathbf{x} + \epsilon$, where $\epsilon$ is sampled from a gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ either the identity matrix $\mathbf{I}_{d_a}$ (*isotropic*), a diagonal matrix $\Lambda$ 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 $\mathrm{Tr}\ \Sigma = 1$ by the transformation $\Sigma \to \Sigma' = \Sigma/(\mathrm{Tr}\ \Sigma)$. The results are summarized in Fig. 4.

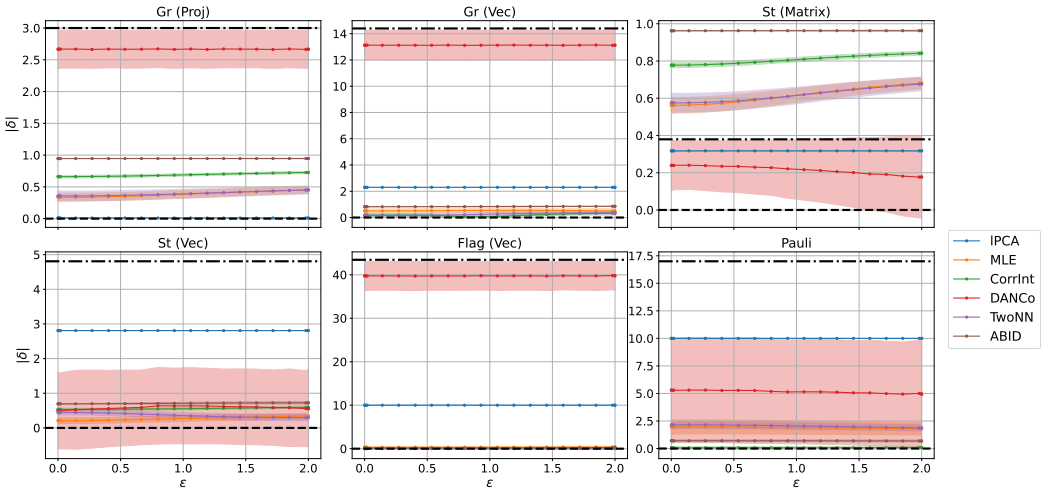



Figure 3: Effect of squeezing: We plot the relative error $\langle|\delta|\rangle$ as a function of the parameter $\epsilon$, which is a direct measure of anisotropy. Except for St (Matrix), most methods show negligible change as $\epsilon$ is increased.

Most methods deviate in their estimations when $\sigma^2 \sim ||\mathbf{x}||_2^2 = \mathcal{O}(1)$. 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 discernbile change in the behavior of isotropic or anisotropic noise. However, in the high noise limit, we observe that anisotropic noise is always underestimated.

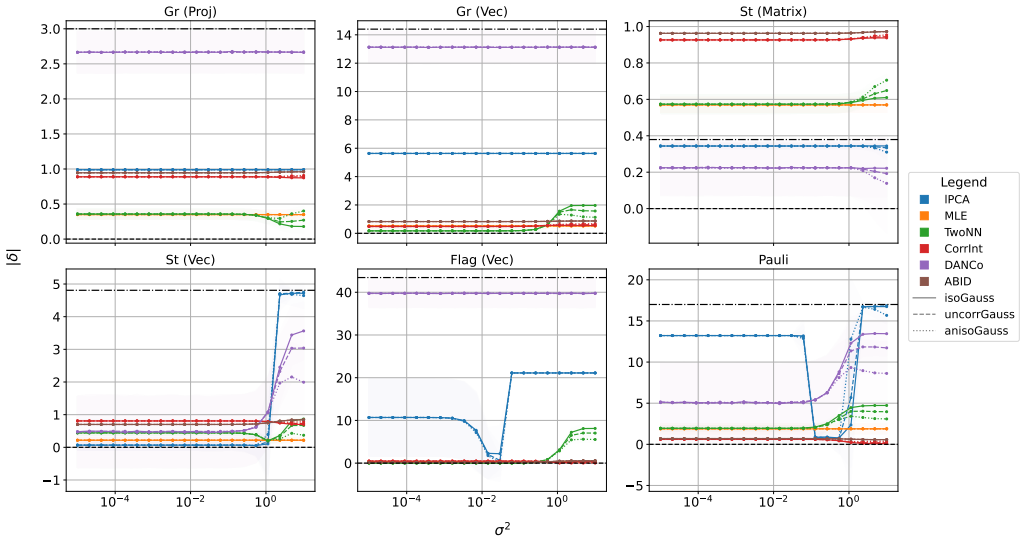



Figure 4: Effect of additive noise. We report IDE performance when the uncorrupted data $\mathbf{x} \to \mathbf{x} + \epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\Sigma)$ with $\Sigma$ chosen to be proportional to the identity (isotropic), diagonal (uncorrelated) or a complete random symmetric matrix (anisotropic). We then plot the relative error $\delta$ as a function of the noise scale $\sigma^2$. 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] A 1-$\sigma$ sampling error is plotted.

[a]The figures shown here plot the absolute value of $\delta$, but we numerically confirmed that $\delta$ is smaller.

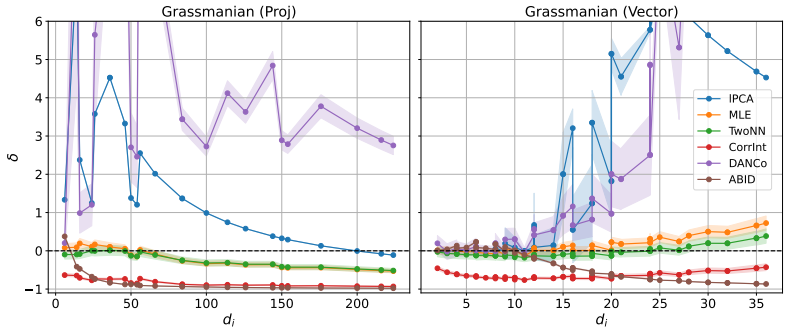



Figure 5: Effect of scaling with intrinsic dimension within the same family of manifolds. The relative error $\delta$ is plotted as a function of increasing $d_1$. Most manifolds show a transition from overestimation at small $d_i$ to underestimation at high $d_i$, corroborating earlier observations for other manifolds (59). The Gr (Vec) embedding shows some minor differences at the same range for $d_i$, but is overall consistent.

### 5.5 Scaling experiments

**Scaling with data dimensionality** One of the key advantages of QuIIEst manifolds is the independent tuning of the ID and ambient dimensions for the same family of manifolds. This allows us to probe the effect of data dimensionality while sampling from the *same* distribution. We observe that, for fixed sample size and hyper-parameters, the methods progressively become better at estimating the ID as we increase the true ID for most manifolds, with a transition from overestimation at low ID to underestimation at high ID. The notable exception is the vector embedding for the Grassmanian family, with all methods (except ABID) over-estimating the ID after a sufficiently large value, cf. Fig 5. Results for the other manifolds are provided in Appendix G.3.

**Scaling with sample size** Several arguments (44; 59; 77; 68; 28) exist to show that reliable ID estimates can be made with a number of samples exponential in the true $d_i$ of the manifold. Hence, the error $\delta$ should decrease with increase in sample size. We observe that this is generally true, but there is no universal convergence value. We detail the results of our investigations in Appendix G.4.

## 6 Analysis

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) (79) which is a direct measure of the anisotropy of the data and is given by $\frac{\text{Var}(\lambda)}{(\text{Mean}(\lambda))^2}$ ( $\lambda$ 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 that there is a slight negative correlation of performance with anisotropy, the effect being most prominent for the angle-based methods DANCo and ABID, cf. Fig 6. On the other hand, most methods show almost no correlation with the total variance, except for the angle-based methods. We also observe a slight positive correlation between performance and $\langle R^2 \rangle$. The plots for Tr $\mathbf{\Sigma}$ and $\langle R^2 \rangle$ are presented in Appendix G.5.

**Geometric properties** We measure the local curvature $H$, local density $\rho$ and a dimensionless parameter $\kappa \equiv \rho / H^{d_i}$. However, we fail to observe any significant dependence. We believe that this ties in with our observation that IDE performance depends both on the manifold *and* the method, and geometric properties capture only the former. We leave a more in-depth investigation of this effect to future work. We outline the details of this investigation in Appendix G.6

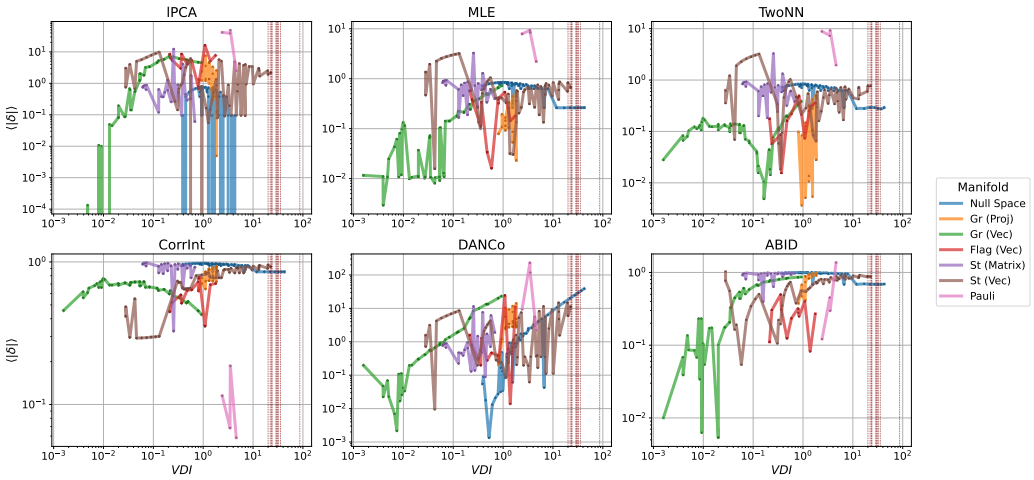



Figure 6: Effect of anisotropy on the performance of various IDE methods, for different manifolds. We observe that DANCo and ABID show a prominent negative correlation of performance with anisotropy, with ABID saturating at VDI $\sim 1.0$. For reference, the VDI values for various MNIST classes are plotted as light maroon vertical lines. We plot the average of the absolute values of relative error $\delta$.

## 7 Conclusion, Limitations and Future Works

We present QuIIEst— a set of manifold embeddings with complex topological and geometrical structure to be used as a benchmark for intrinsic dimension estimation (IDE). We believe this is an important step in using these IDE methods for the estimation of real-world datasets of unknown ID. Due to constraints on (compute) time and (human) effort, we restrict ourselves to only six IDE methods. Because of this, our analysis on correlation between data properties and IDE performance is limited: we observe weak correlations, but more samples are needed to relate our results to asymptotic estimates of method accuracy.

We do, however, notice that IDE methods perform differently on estimating the ID of a manifold embedded using different techniques. Investigating this further may yield more favorable embeddings of real-world data.

We generate embeddings of homogeneous spaces $\mathcal{G}/\mathcal{H}$ for group pairs $\mathcal{H} \subset \mathcal{G}$ using the Gilmore-Peremolov coherent-state method, which can be further extended to double coset spaces $\mathcal{K}\backslash\mathcal{G}/\mathcal{H}$ for subgroup $\mathcal{K}$ (1). Since these spaces need not be manifolds, this extension is a promising route to emulating real-world data not living on a manifold.

Going further, we believe it is possible to extend the coherent state method *directly* to data vectors. For example, given a relevant group of transformations, we can generate new data by applying group elements to a data vector. We hope this will yield concrete connections between the topological and geometrical features of our manifolds and real-world datasets. Since our manifolds have the same ID at every point, one can gauge how much these methods deviate in their IDE at different points. This can serve as a useful diagnostic to probe whether data actually lives on a *manifold*, confirming or refuting the manifold hypothesis. We also plan to integrate our benchmark with existing benchmarks for easier access by practitioners.

## Declaration of LLM usage

We declare that LLMs were used for minor polishing and formatting of the text and figures and correcting grammar.

## REFERENCES


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

[2] 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.

[3] Juan Miguel Arrazola, Alain Delgado, Bhaskar Roy Bardhan, and Seth Lloyd. Quantum-inspired algorithms in practice. *Quantum*, 4:307, August 2020. doi: 10.22331/q-2020-08-13-307.

[4] 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.

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

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

[7] 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.

[8] Lennart Bittel, Jens Eisert, Lorenzo Leone, Antonio A. Mele, and Salvatore F. E. Oliviero. A complete theory of the Clifford commutant, 4 2025.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Lawrence Cayton. Algorithms for manifold learning. Technical report, University of California, San Diego, June 2005. URL https://cs.ucsd.edu/~lcayton/papers/techreport.pdf. Technical Report.

[21] M. Cerezo, Guillaume Verdon, Hsin-Yuan Huang, Lukasz Cincio, and Patrick J. Coles. Challenges and opportunities in quantum machine learning. *Nature Computational Science*, 2(9): 567–576, September 2022. doi: 10.1038/s43588-022-00311-3.

[22] Claudio Ceruti, Simone Bassis, Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli. Danco: Dimensionality from angle and norm concentration, 2012. URL https://arxiv.org/abs/1206.3881.

[23] 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.

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

[25] Chen Ding, Tian-Yi Bao, and He-Liang Huang. Quantum-inspired support vector machine. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7210–7222, December 2022. doi: 10.1109/tnnls.2021.3084467.

[26] 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.

[27] Olga A Doronina, Zachary J Grey, and Andrew Glaws. Grassmannian shape representations for aerodynamic applications. *arXiv preprint arXiv:2201.04649*, 2022.

[28] Valerio Erba, Marco Gherardi, and Pietro Rotondo. Intrinsic dimension estimation for locally undersampled data. *Scientific Reports*, 9(1):17133, November 2019. doi: 10.1038/s41598-019-53549-9.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] Farhad Soleimanian Gharehchopogh. Quantum-inspired metaheuristic algorithms: comprehensive survey and classification. *Artificial Intelligence Review*, 56(6):5479–5543, November 2022. doi: 10.1007/s10462-022-10280-8.

[36] 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.

[37] 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.

[38] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[39] 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 2018. ISSN 1471-2962. doi: 10.1098/rsta.2017.0237. URL https://doi.org/10.1098/rsta.2017.0237.

[40] 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 2018. doi: 10.1098/rsta.2017.0237. URL https://doi.org/10.1098/rsta.2017.0237.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] Douglas R. Hofstadter. Energy levels and wave functions of bloch electrons in rational and irrational magnetic fields. *Phys. Rev. B*, 14:2239–2249, Sep 1976. doi: 10.1103/PhysRevB.14.2239. URL https://link.aps.org/doi/10.1103/PhysRevB.14.2239.

[46] 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.

[47] David E Speyer (https://mathoverflow.net/users/297/david-e speyer). Does every irreducible representation of a compact group occur in tensor products of a faithful representation and its dual? MathOverflow. URL https://mathoverflow.net/q/58644. URL:https://mathoverflow.net/q/58644 (version: 2022-10-13).

[48] Zhiwu Huang, Jiqing Wu, and Luc Van Gool. Building deep networks on grassmann manifolds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32. AAAI Press, 2018. doi: 10.1609/aaai.v32i1.11725. URL https://doi.org/10.1609/aaai.v32i1.11725.

[49] Larry Huynh, Jin Hong, Ajmal Mian, Hajime Suzuki, Yanqiu Wu, and Seyit Camtepe. Quantum-inspired machine learning: a survey, 2023.

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

[51] K. Johnsson, C. Soneson, and M. Fontes. Low bias local intrinsic dimension estimation from expected simplex skewness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):196–202, 2014. doi: 10.1109/TPAMI.2014.2343215. URL https://doi.org/10.1109/TPAMI.2014.2343215.

[52] Kerstin Johnsson. *Structures in High-Dimensional Data: Intrinsic Dimension and Cluster Analysis*. Doctoral thesis (compilation), Faculty of Engineering, LTH, August 2016.

[53] 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 *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024. URL https://openreview.net/forum?id=wc044k7QBj.

[54] Iordanis Kerenidis and Anupam Prakash. Quantum recommendation systems, 2016.

[55] 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.

[56] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[57] 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.

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

[59] 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.

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

[61] Estelle Massart and Vinayak Abrol. Coordinate descent on the stiefel manifold for deep neural network training. In *Proceedings of the 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium and online, October 4–6 2023. i6doc.com. ISBN 978-2-87587-088-9. URL https://www.esann.org/sites/default/files/proceedings/2023/ES2023-143.pdf.

[62] Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *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.

[63] 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.

[64] 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.

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

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

[67] permutation_matrix (https://math.stackexchange.com/users/913340/permutation matrix). Embedding real projective space in sphere. Mathematics Stack Exchange. URL https://math.stackexchange.com/q/4097673. URL:https://math.stackexchange.com/q/4097673 (version: 2021-04-11).

[68] Vladimir Pestov. An axiomatic approach to intrinsic dimension of a dataset. *Neural Networks*, 21 (2-3):204–213, March 2008. doi: 10.1016/j.neunet.2007.12.030. Erratum in: Neural Networks. 2008 May;21(4):698.

[69] 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.

[70] reuns (https://math.stackexchange.com/users/276986/reuns). explicit formula for embedding projective spaces into euclidean space. Mathematics Stack Exchange. URL https://math.stackexchange.com/q/3505749. URL:https://math.stackexchange.com/q/3505749 (version: 2020-01-16).

[71] 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.

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

[73] Jan Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Your diffusion model secretly knows the dimension of the data manifold, 2023. URL https://arxiv.org/abs/2212.12611.

[74] Ewin Tang. A quantum-inspired classical algorithm for recommendation systems. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC '19. ACM, June 2019.

[75] Piotr Tempczyk, Adam Golinski, Przemysław Spurek, and Jacek Tabor. LIDL: Local intrinsic dimension estimation using approximate likelihood. In *ICLR 2021 Workshop on Geometrical and Topological Representation Learning*, 2021. URL https://openreview.net/forum?id=ij1aPkDZBYV.

[76] 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.

[77] Nakul Verma. *Learning From Data With Low Intrinsic Dimension*. Phd dissertation, University of California, San Diego, San Diego, CA, USA, 2009.

[78] Bokun Wang, Shiqian Ma, and Lingzhou Xue. Riemannian stochastic proximal gradient methods for nonsmooth optimization over the stiefel manifold. *Journal of Machine Learning Research*, 23(106):1–33, 2022. URL http://jmlr.org/papers/v23/21-0314.html.

[79] 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.

[80] 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.

[81] Ryoma Yataka, Kazuki Hirashima, and Masashi Shiraishi. Grassmann manifold flows for stable shape generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=H2udtfMbl4.

[82] 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.

[83] 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.

[84] Zhenyue Zhang and Hongyuan Zha. Adaptive manifold learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 17, 2005.

[85] 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.

[86] 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 QuIIEst 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 QuIIEst 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 QuIIEst 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 QuIIEst 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 (76). 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 QuIIEst 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 32-core CPUs and took around 600 CPU hours with an approximate total of 10 hours for data generation. Plots and inferences were then made locally with negligible overhead cost.

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.

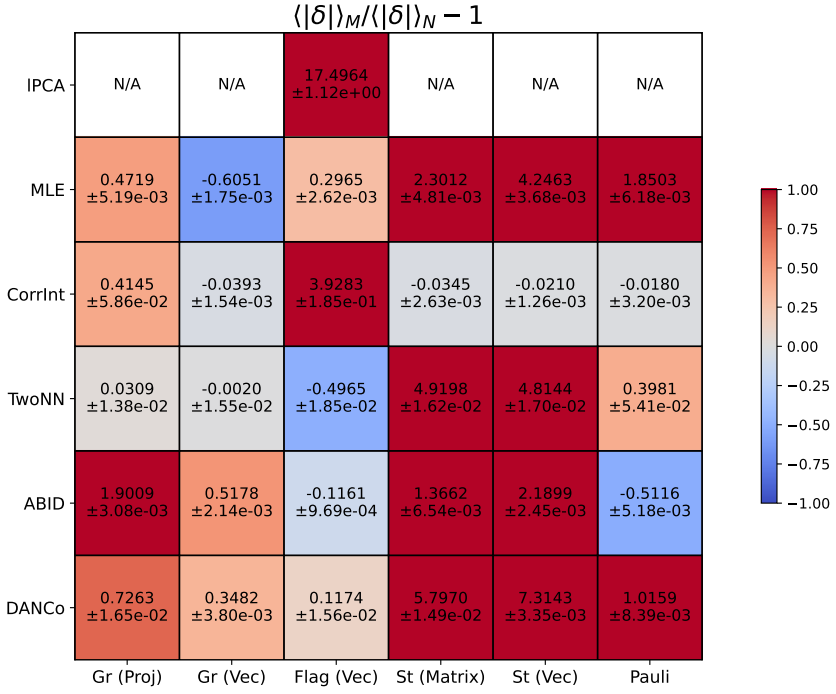



Figure 7: Relative performance of IDE methods on QuIIEst manifolds and the manifold of normal isotropic Gaussian vectors . N/A indicates that the method performs very well on $\mathcal{N}$.

## 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 (39; 64; 20). This hypothesis has been subject to both theoretical and experimental investigations (31; 36; 62; 55; 11; 26; 57). At the same time, there has been interest in understanding the topological structures of real-world datasets through the field of Topological Data Analysis (19; 86). 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 (60; 2).

**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 (84; 33; 34), exponential scaling of samples with ID (63; 62), correlation of generalization capability with ID of data and internal representations (69; 2; 7; 60; 10) as well as ID being a natural measure of local complexity of the data (53).

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 (13; 59; 37; 29), non-Euclidean geodetic distances

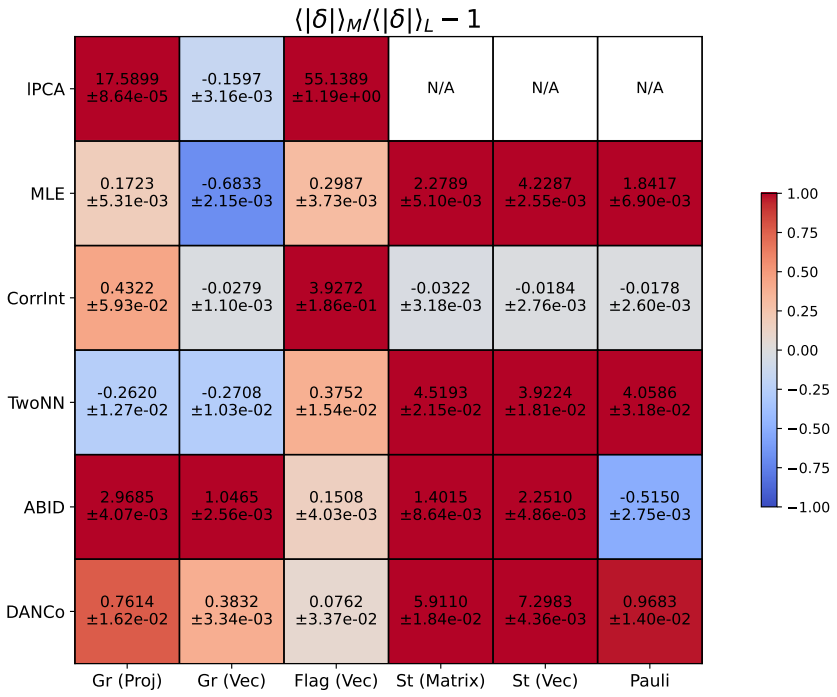



Figure 8: Relative performance of IDE methods on QuIIEst manifolds and the manifold of affine linear nullspace $\mathcal{L}$. N/A indicates that the method performs very well on $\mathcal{L}$.

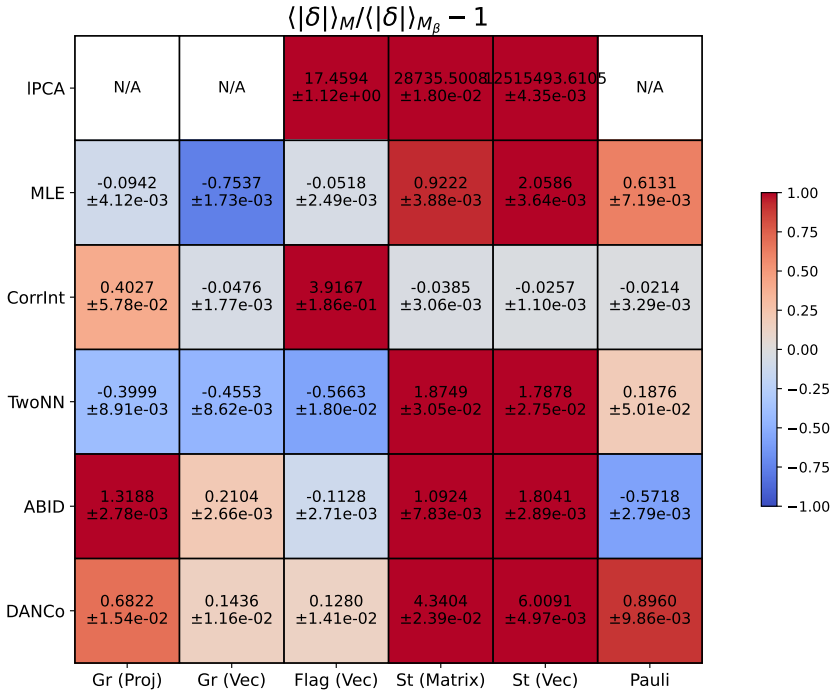



Figure 9: Relative performance of IDE methods on QuIIEst manifolds and the manifold $\mathcal{M}_\beta$. N/A indicates that the method performs very well on $\mathcal{M}_\beta$.

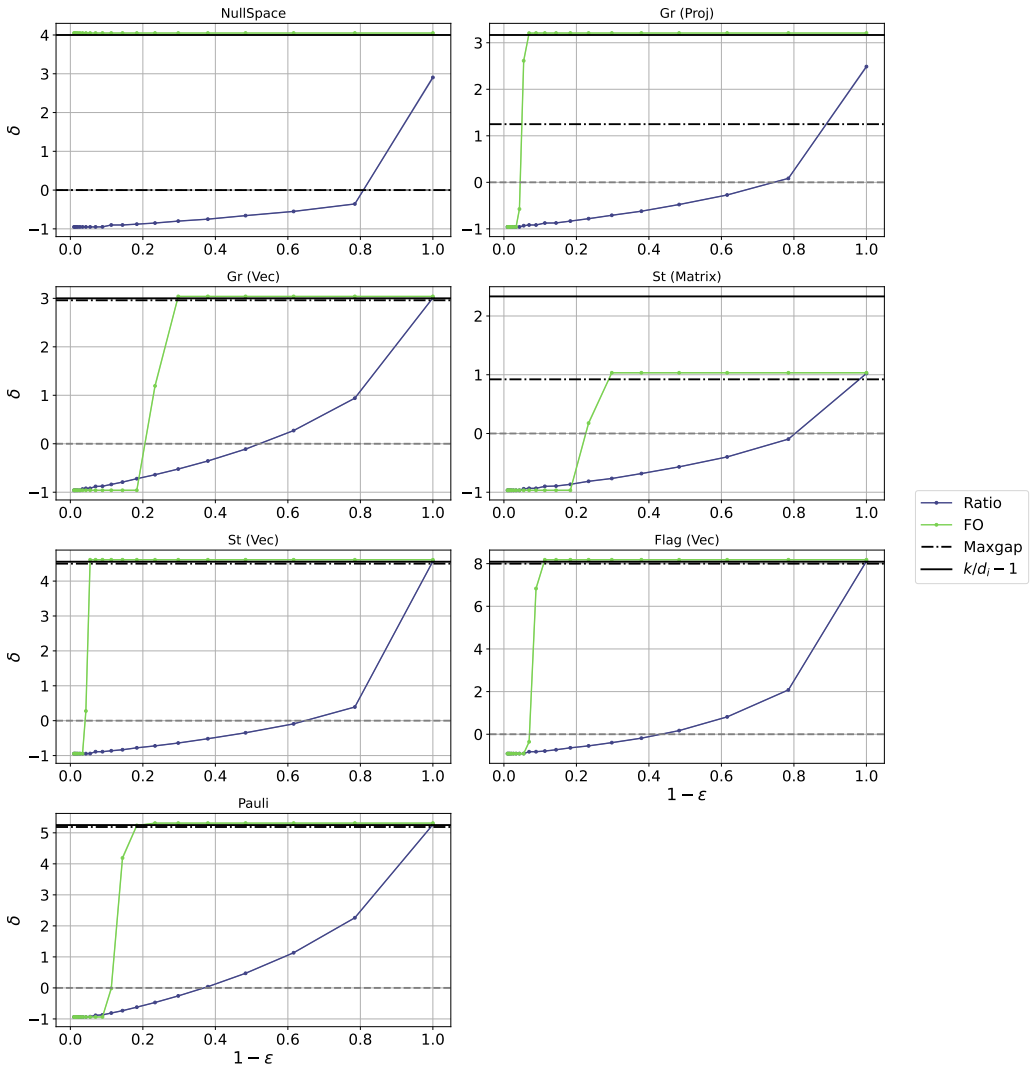



Figure 10: lPCA maxgap compared to lPCA ratio and FO versions. 'FO' usually shows a quicker convergence to lPCA maxgap value as compared to the 'ratio' version.

(41) and Wasserstein distances (9); deviations of simplexes (51); angle-based measures (23; 76); measures based on generative models (73; 75; 53; 85) and quantum encoding-based algorithms (17). See (16) for a more extensive survey of various ID estimators.

**Benchmarks for IDE** (14) undertook one of the first surveys for ID estimators, outlining different datasets used for benchmarking ID estimators known at the time. (44) 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. (71) drew on these ideas and constructed several complicated manifolds, devising means of embedding data into higher dimensions without linear isometries. (16) 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 (69). 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 ground-truth ID for the generated data.

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 (4). 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 (80), DSVC1 dataset (15) and strange attractors (42). 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 QuIIEst- an infinite family of manifolds which offer multiple natural notions of embedding in higher-dimensional space. In addition, we propose a recipe to construct such manifolds through the concept of homogenous spaces and coherent states.

**Quantum and ML** Any quantum algorithm can be simulated on a classical computer, but in a non-scalable way since the dimension of the underlying quantum state space increases exponentially in the number of quantum bits. Quantum-inspired algorithms (21; 3; 35) leverage the advantages of quantum algorithms on existing classical hardware in a scalable way. A well-known example is the quantum-inspired recommendation algorithm of Ewin Tang (74), which can be thought of as a "de-quantized" version of an earlier quantum recommendation algorithm (54). Various other instances have recently been developed in this nascent subfield, e.g., in Refs. (49; 25).

We also leverage tools from quantum mechanics, but to develop a benchmark instead of an algorithm. Specifically, we use Gilmore-Perelomov coherent states (65; 66; 83) to construct parameterized homogeneous spaces. Coherent states are the closest analogue of a classical state in a quantum system and have been used in high-energy physics, quantum optics, and, most recently, in quantum error correction. We further distort coherent-state spaces in a way akin to generalized spin squeezing, which is useful for measuring signals along certain directions in quantum metrology.

# D Detailed theory for QuIIEst Manifolds

All manifolds included in the QuIIEst 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 coherent-state method (65; 66; 83), vectorization of matrices, and combinations thereof. We review Gilmore-Perelomov coherent-states in Section D.1 and use them to build the QuIIEst 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 $\Pi$ of $\mathcal{G}$[1]. Define $\Pi|_{\mathcal{H}}$ to be the representation restricted to $\mathcal{H}$. For compact Lie groups, all irreducible representations (irreps) appear in the isotypic decomposition of tensor-product representation (47). Our starting point will therefore be to consider representations of the form $\Pi(g) = g^{\otimes t}$, but the recipe can be generalized to other representations.

The projector onto the trivial irreps in the isotypic decomposition of a representation $\Pi$ is $P(\Pi) = \int_{\mathcal{G}} \Pi(g)\, dg$, where $dg$ denotes the unique unit-normalized Haar measure on $\mathcal{G}$. Suppose that $P(\Pi|_{\mathcal{H}}) > P(\Pi)$, and there are no subgroups $\mathcal{H} < \mathcal{K} < \mathcal{G}$ with $P(\Pi|_{\mathcal{K}}) > P(\Pi|_{\mathcal{H}})$. Then there exists at least one unit vector $|\mathcal{H}\rangle$ that lies in the image of $P(\Pi|_{\mathcal{H}})$ and does not lie in the image of $P(\Pi|_{\mathcal{K}})$ for any $\mathcal{K} > \mathcal{H}$ (including $\mathcal{K} = \mathcal{G}$). By construction, $|\mathcal{H}\rangle$ lies in a trivial irrep of $\Pi|_{\mathcal{H}}$, and therefore $\Pi(h)\,|\mathcal{H}\rangle = |\mathcal{H}\rangle$ for all $h \in \mathcal{H}$. 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 $|\psi_{\mathcal{M}}(g)\rangle = \Pi(g)\,|\mathcal{H}\rangle$ is a well-defined, injective mapping between $\mathcal{M}$ and $\mathbb{R}^{d_A}$ or $\mathbb{C}^{d_A}$ (see footnote 1). In fact, the embedding is injective into the real or complex sphere $\Omega_{d_A}$.

This mapping is equivariant – $\Pi(g_1)\,|\psi_{\mathcal{M}}(g_2)\rangle = |\psi_{\mathcal{M}}(g_1 g_2)\rangle$. 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 $g_\Omega$ 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 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 $\mathbf{x} \to |\mathcal{H}\rangle$ by setting $\langle e_i|\mathcal{H}\rangle = x_i$ where $|e_i\rangle$ 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 $\{|1\rangle, \ldots, |n\rangle\}$ 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 $\mathrm{St}(k, \mathbb{R}^n) = \left\{X \in \mathbb{R}^{n\times k} \mid X^T X = \mathbb{I}\right\}$ (50; 6). This easily embeds into $\mathbb{R}^{nk}$ via the map $X \mapsto \vec{X}$. This yields the "St (Matrix)" embedding in the QuIIEst dataset.

A more interesting embedding will be used as an intermediate step to derive the Grassmanian embedding used in QuIIEst 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 $\Pi \colon \mathcal{O} \mapsto \mathcal{O}^{\otimes k}$ for $\mathcal{O} \in \mathrm{O}(n)$. An obvious choice for the state $|\mathcal{H}\rangle$ is $|1\rangle \otimes \cdots \otimes |k\rangle$. Given $\mathcal{O} \in \mathrm{O}(n)$,

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

| Family | Symbol | Quotient space | $d_i$ | QuIIEst 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) | $\binom{n}{k}+k^2$ |
| Grassmanian | $\mathrm{Gr}(k, \mathbb{R}^n)$ | $\frac{\mathrm{O}(n)}{\mathrm{O}(n-k)\mathrm{O}(k)}$ | $k(n-k)$ | 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 | $\mathrm{Flag}^\star(k_1, k_2, \mathbb{R}^n)$ | $\frac{\mathrm{O}(n)}{\mathrm{SO}(k_1)\mathrm{SO}(k_2)\mathrm{O}(n-k_1-k_2)}$ | $\substack{(k_1+k_2)n \\ -k_1^2-k_2^2-k_1k_2}$ | Flag (Vec) | $\binom{n}{k_1}\binom{n}{k_2}$ |
| Pauli quotient | $\frac{\mathrm{U}(n)}{\mathcal{P}_n^\star}$ | $\frac{\mathrm{U}(n)}{\mathcal{P}_n^\star}$ | $n^2$ | Pauli | $2n^4$ |

Table 3: Table listing manifold families used in the QuIIEst benchmark, their mathematical symbols, their equivalent quotient/homogeneous spaces, and their intrinsic dimensions $d_i$. Here, $\mathrm{U}(n)$ and $\mathrm{O}(n)$ denote the unitary and orthogonal groups in $n$ dimensions, respectively, $\mathrm{SO}(n)$ denotes the special unitary groups, and the group $\mathcal{P}_n^\star$ is defined in Sec. D. The parameter $n \geq 1$ for all rows except the last one, where it is assumed to be prime. The last two columns list the six QuIIEst 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 Grassmanian 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.

this yields the embedding

$$
\begin{aligned}
|\psi_{\mathrm{St}}(\mathcal{O})\rangle &= \Pi(\mathcal{O})\,|\mathcal{H}\rangle && (2)\\
&= \mathcal{O}\,|1\rangle \otimes \mathcal{O}\,|2\rangle \otimes \cdots \otimes \mathcal{O}\,|k\rangle && (3)\\
&= \sum_{j_1,\ldots,j_k=1}^{n} \mathcal{O}_{j_1,1} \ldots \mathcal{O}_{j_k,k}\,|j_1\rangle \otimes \cdots \otimes |j_k\rangle && (4)\\
&= |\mathcal{O}_{*,1}\rangle \otimes \cdots \otimes |\mathcal{O}_{*,k}\rangle\,, && (5)
\end{aligned}
$$

where $|\mathcal{O}_{*,i}\rangle$ denotes the $i^{\mathrm{th}}$ column of $\mathcal{O}$. From this we can easily see that $|\psi_{\mathrm{St}}(\mathcal{O})\rangle$ 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 $|\psi_{\mathrm{St}}(\mathcal{O})\rangle$ of the Stiefel manifold from above. We need to slightly modify this to get an embedding of the Grassmannian,

$$
\mathrm{Gr}^\star(k, \mathbb{R}^n) \cong \frac{\mathrm{St}(k, \mathbb{R}^n)}{\mathrm{SO}(k)} \cong \frac{\mathrm{O}(n)}{\mathrm{O}(n-k) \times \mathrm{SO}(k)}\,, \tag{6}
$$

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

$$
|\mathcal{H}\rangle = \frac{1}{\sqrt{k!}} \sum_{\sigma \in S_k} \mathrm{sgn}(\sigma)\,|\sigma(1)\rangle \otimes \cdots \otimes |\sigma(k)\rangle\,, \tag{7}
$$

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

We note that Eq. equation 6 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. equation 6, all have the same ID. In future work, we will add the standard and oriented Grassmannian to QuIIEst and further

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

Table 4: The first two homotopy groups of the three types of "Grassmannians" described below Eq. equation 6.

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 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. 4, 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

$$|\psi_{\mathrm{Gr}}(\mathcal{O})\rangle = \Pi(\mathcal{O})\,|\mathcal{H}\rangle \tag{8}$$

$$= \frac{1}{\sqrt{k!}}\sum_{\sigma\in S_k}\mathrm{sgn}(\sigma)\mathcal{O}\,|\sigma(1)\rangle\otimes\cdots\otimes\mathcal{O}\,|\sigma(k)\rangle \tag{9}$$

$$= \frac{1}{\sqrt{k!}}\sum_{\sigma\in S_k}\sum_{j_1,\ldots,j_k=1}^{n}\mathrm{sgn}(\sigma)\mathcal{O}_{j_1,\sigma(1)}\ldots\mathcal{O}_{j_k,\sigma(k)}\,|j_1\rangle\otimes\cdots\otimes|j_k\rangle \tag{10}$$

$$= \frac{1}{\sqrt{k!}}\sum_{j_1,\ldots,j_k=1}^{n}\det(\mathcal{O}_{(j_1,\ldots,j_k)})\,|j_1\rangle\otimes\cdots\otimes|j_k\rangle\,, \tag{11}$$

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

$$|\psi_{\mathrm{Gr}}(\mathcal{O})\rangle = \sum_{\substack{Q\subset\{1,\ldots,n\}\\|Q|=k}}\det\mathcal{O}_Q\frac{1}{\sqrt{k!}}\sum_{\sigma\in S_k}\mathrm{sgn}(\sigma)\,|\sigma(Q_1)\rangle\otimes\cdots\otimes|\sigma(Q_k)\rangle \tag{12}$$

$$= \sum_{\substack{Q\subset\{1,\ldots,n\}\\|Q|=k}}\det\mathcal{O}_Q\,|Q\rangle\,, \tag{13}$$

where we defined $Q_i$ to be the $i^{\mathrm{th}}$ 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

$$|Q\rangle = \frac{1}{\sqrt{k!}}\sum_{\sigma\in S_k}\mathrm{sgn}(\sigma)\,|\sigma(Q_1)\rangle\otimes\cdots\otimes|\sigma(Q_k)\rangle\ . \tag{14}$$

Notice that we are embedding into $\mathbb{R}^{n^k}$, but a full basis for the space is given by $|Q\rangle$ 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 3.

We note that this embedding is almost the Plücker embedding (6). The Plücker embedding is an embedding of the standard Grassmanian into real projective space, which is the real sphere with antipodal points identified. Above, we are embedding Eq. equation 6 into the real sphere. The reason that $\psi_{\mathrm{Gr}}$ is not an embedding of $\mathrm{O}(n)/\mathrm{O}(n-k)\times\mathrm{O}(k)$ is because the vector $|\mathcal{H}\rangle$ is not invariant under the action $\mathrm{O}(k)$. Instead, it is invariant under the action of $\mathrm{SO}(k)$, and yields a $\pm1$ 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 (72; 67; 70). 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

$$|\tilde{\psi}_{\mathrm{St}}\rangle = |\psi_{\mathrm{Gr}}(\mathcal{O})\rangle \oplus \vec{\mathcal{V}}\,, \tag{15}$$

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 $\psi_{\mathrm{St}}$, 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 QuIIEst in future work.

We begin with the case $t = 2$, as presented in Table 3. Again, in the notation of Section D.1, we work with the representation $\Pi\colon \mathcal{O} \mapsto \mathcal{O}^k$, and we use the fiducial vector

$$|\mathcal{H}\rangle = \left(\frac{1}{\sqrt{k_1!}}\sum_{\sigma\in S_{k_1}} |\sigma(1)\rangle\otimes\cdots\otimes|\sigma(k_1)\rangle\right)\otimes\left(\frac{1}{\sqrt{k_2!}}\sum_{\sigma\in S_{k_2}} |\sigma(k_1+1)\rangle\otimes\cdots\otimes|\sigma(k_1+k_2)\rangle\right). \tag{16}$$

This yields the embedding

$$\begin{aligned}
|\psi_F(\mathcal{O})\rangle &= \Pi(\mathcal{O})\,|\mathcal{H}\rangle && (17)\\
&= \tfrac{1}{\sqrt{k_1!k_2!}}\sum_{j_1,\ldots,j_k=1}^{n} \det\!\big(\mathcal{O}_{(j_1,\ldots,j_{k_1}),(1,\ldots,k_1)}\big)\det\!\big(\mathcal{O}_{(j_{k_1+1},\ldots,j_k),(k_1+1,\ldots,k)}\big)|j_1\rangle\otimes\cdots\otimes|j_k\rangle && (18)\\
&= \tfrac{1}{\sqrt{k_1!k_2!}}\sum_{\substack{Q\subset\{1,\ldots,n\}\\|Q|=k_1}}\sum_{\substack{P\subset\{1,\ldots,n\}\\|P|=k_2}}\sum_{\sigma\in S_{k_1}}\sum_{\pi\in S_{k_2}}\times && (19)\\
&\quad\times\mathrm{sgn}(\sigma)\mathrm{sgn}(\pi)\det\!\big(\mathcal{O}_{Q,(1,\ldots,k_1)}\big)\det\!\big(\mathcal{O}_{P,(k_1+1,\ldots,k)}\big)|\sigma(Q_1)\rangle\otimes\cdots\otimes|\sigma(Q_{k_1})\rangle\otimes|\pi(P_1)\rangle\otimes\cdots\otimes|\pi(P_{k_2})\rangle\\
&= \sum_{\substack{Q\subset\{1,\ldots,n\}\\|Q|=k_1}}\sum_{\substack{P\subset\{1,\ldots,n\}\\|P|=k_2}} \det\big(\mathcal{O}_{Q,\{1,\ldots,k_1\}}\big)\det\big(\mathcal{O}_{P,\{k_1+1,\ldots,k\}}\big)\,|Q\rangle\otimes|P\rangle\,. && (20)
\end{aligned}$$

As is, this embedding is into $\mathbb{R}^{n^k}$, but because an orthonormal basis is given by tensor products of $|Q\rangle$, 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 $d \geq \prod_{i=1}^{t}\binom{n}{k_i}$.

THE "PAULI" EMBEDDING

We would like to construct an embedding for the quotient space $\mathcal{G}/\mathcal{H} = \mathrm{U}(n)/\mathcal{P}_n$, where $\mathcal{P}_n = \langle e^{i\frac{2\pi}{n}}, X, Z\rangle$ is the Pauli (a.k.a. Heisenberg-Weyl) group of prime dimension $n$, and where $Z$ and the real-valued $X$ are the standard $n$-dimensional qudit Pauli matrices (8). To construct the quotient space using the Gilmore-Perelomov prescription in Section D.1, we require a vector $|\mathcal{H}\rangle$ in some representation of $\mathrm{U}(n)$ that is invariant under $\mathcal{P}_n$ and not invariant under any $\mathrm{U}(n)$-subgroup that contains $\mathcal{P}_n$.

We pick the $n^4$-dimensional four-fold tensor-product unitary representation $\Pi\colon U \mapsto U\otimes\overline{U}\otimes U\otimes\overline{U}$ for $\mathrm{U}(n)$. The corresponding Pauli representation is then $Z\otimes\overline{Z}\otimes Z\otimes\overline{Z}$ and $X\otimes X\otimes X\otimes X$, where we recall that $X$ is real. There is an $n^2$-dimensional subspace $S$ that is invariant under this Pauli representation. It is spanned by the vectors

$$|\overline{a,b}\rangle = \frac{1}{\sqrt{n}}\sum_{c\in\mathbb{Z}_n}|c, c+a, c+b, c+a+b\rangle\ , \tag{21}$$

where $a, b \in \mathbb{Z}_n$, and where addition inside the kets is done modulo $\mathbb{Z}_n$. Our "Pauli" embedding is constructed by defining $|\mathcal{H}\rangle$ to be a random unit vector in $S$. Then, as in Section D.1, we apply a the unitary rotation in the four-fold tensor-product representation, and embed the resulting $n^4$-dimensional complex vector into $\mathbb{R}^{2n^4}$.

We now narrow down the quotient space that is spanned by our "Pauli" embedding. It has also been shown Ref. (8) that $S$ is not invariant under the larger Clifford group $\mathcal{C}_n \supseteq \mathcal{P}_n$, defined as the normalizer of the Pauli group inside the unitary group. Leaving open the possibility that there exists some "in-between" group $\mathcal{P}_n^\star$ satisfying

$$\mathcal{P}_n \subseteq \mathcal{P}_n^\star \subset \mathcal{C}_n\ , \tag{22}$$

we conclude that the quotient space of the "Pauli" embedding is $\mathrm{U}(n)/\mathcal{P}_n^\star$. Since the Clifford group is finite, the intrinsic dimensions of both $\mathrm{U}(n)/\mathcal{P}_n^\star$ and $\mathrm{U}(n)/\mathcal{P}_n$ are equal to the dimension $n^2$ of the unitary group.

# E OVERVIEW OF IDE METHODS TESTED

**IPCA :** Since a manifold is locally isomorphic to $\mathbb{R}^{d_i}$, 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 is 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 $\epsilon$ 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, but we also report results of small-scale experiments with the 'ratio' and 'fo' versions in G.1 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.

(43) 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 $\sigma$, 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 $\rho$ 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 $\alpha$ 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 $\mu$ represents the ratio $\frac{r_2}{r_1}$ for the $k$NNs of the particular point.

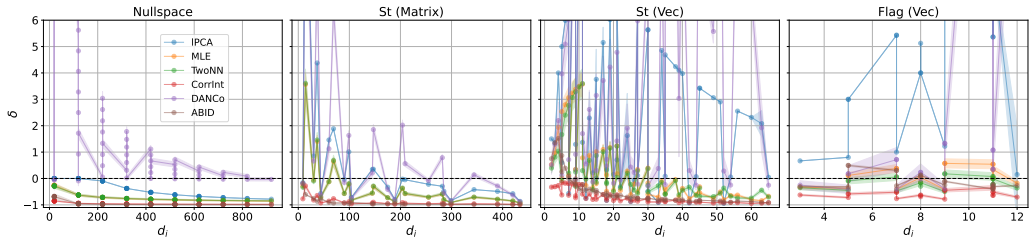



Figure 11: Scaling of relative error with ID for the other manifolds. We observe the same sort of trend as discussed in the main text, except for flag, which has limited resolution.

**DANCo :** For a manifold $\mathcal{M} \subseteq \mathbb{R}^{d_i}$, consider an embedding $\phi : \mathbb{R}^{d_i} \to \mathbb{R}^D$ 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

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

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 $\tau$ 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 (76), 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 COMPARISON WITH OTHER BENCHMARKS

Here we present results of comparing our manifolds against other standard manifolds. In Appendix F, 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 (44), we compare our manifold against isotropic Gaussian vectors $\in \mathbb{R}^{d_i}$ linearly embedded into $\mathbb{R}^{d_a}$, cf. Fig‘7

**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 $A \in \mathbb{R}^{d_a \times d_a}$ 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‘8

**Nonlinear manifolds** Based on (71; 16), we use a generalized version of the manifolds denoted by $\mathcal{M}_\beta$ — where we uniformly sample from $X : [0, 1)^{d_i}$ and then construct $Y = \sin(\cos(2\pi X))$, and finally linearly embed it into $\mathbb{R}^{d_a}$. This differs from the original formulation in that they append a $\tilde{Y} = \cos(\sin(2\pi X)$ and finally duplicate this to get $d_a = 4d_i$, cf. Fig‘9

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

# G ADDITIONAL EXPERIMENTS AND PLOTS

## G.1 OTHER VERSIONS OF LOCAL PCA

We observe that the $\hat{d}_i$ from lPCA vary significantly on the hyperparameter $\epsilon$; in particular this seems to suggest that there is no clear spectral gap in the singular values of the data covariance matrix $XX^T$. However, we do note that there always exist some value of $\epsilon^*$ for which $\delta(\epsilon^*) = 0$, but there is no clear pattern or consistent value for the choice of $\epsilon^*$ even for distinct versions for the same manifold, atleast within the scope of the experiments performed. We thus relegate this interesting investigation to a future project. The results are plotted in Fig. 10.

## G.2 SQUEEZING : COMPARISON WITH SPHERES

We present in Fig 3 the effect of squeezing for spheres, as compared to QuIIEst 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 QuIIEst manifolds are already challenging enough for the IDE methods tested, and enhances it applicability as a more robust performance evaluator for IDE.

## G.3 RELATIVE ERROR AS A FUNCTION OF ID

We include here some other plots on scaling of the relative error for different manifolds in Fig 11. We once again notice the trend mentioned in the main text, where an initial over-estimation gives way to an under-estimation. This trend is not immediately obvious for the Flag manifolds due to lack of sufficient points. The Pauli manifold is excluded since we could only test 4 different IDs.

## G.4 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 (59)

$$\frac{k}{N} = \Omega_{d_i} \rho(x) T_k^{d_i} \tag{34}$$

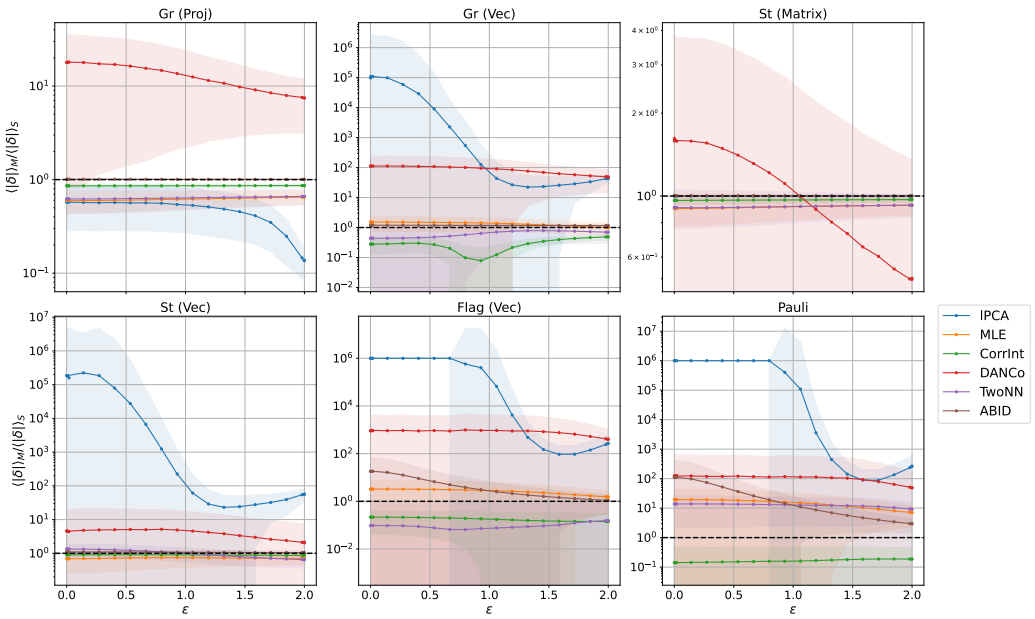



Figure 12: Comparison of errors when distorting QuIIEst manifolds versus spheres. The large change indicates that spheres become drastically more difficult to do IDE on with increasing distortion, as opposed to QuIIEst manifolds.

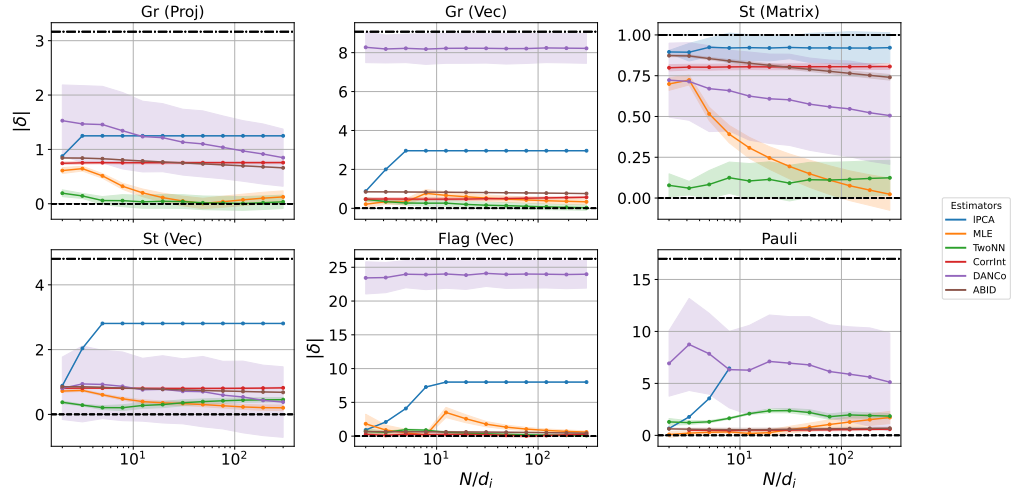
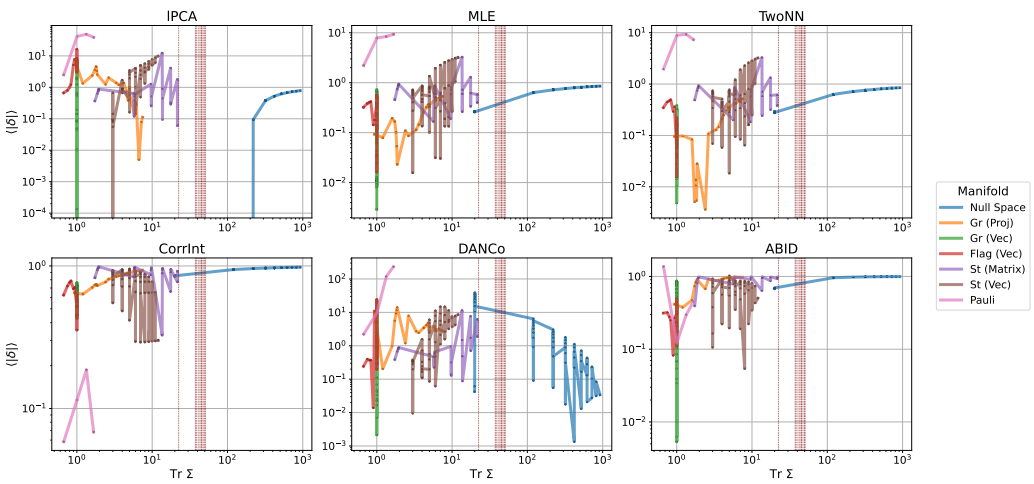



Figure 13: Performance of IDE methods as a function of $N/d_i$. We observe a gradual convergence with increasing sample size. The shaded region shows the 1-$\sigma$ error in the local ID estimates, averaged over three seeds.

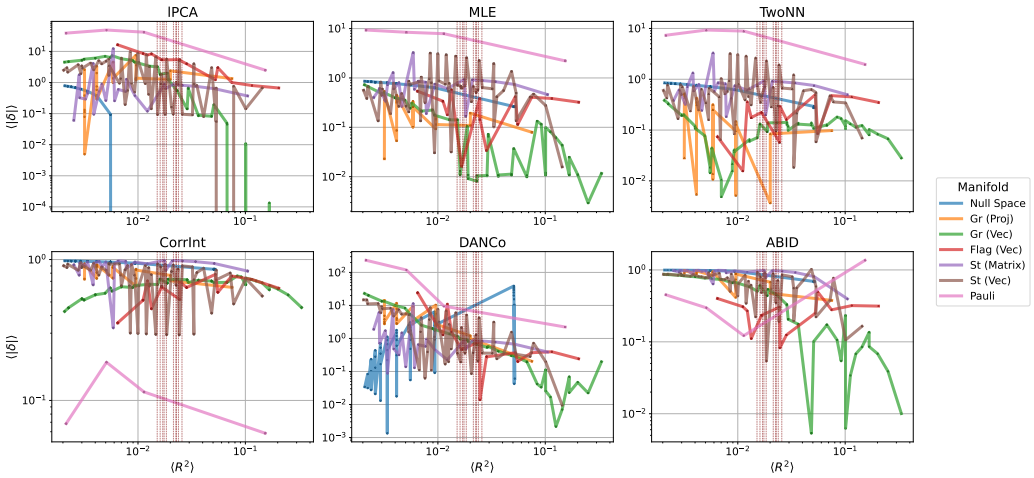
Thus in order to get a uniform and locally dense sampling of points, we need to hold $k/(T_k^{d_i} N)$ fixed, which translates to the condition that

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

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. The results are summarized in Fig 13




(a) Tr $\Sigma$




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

Figure 14: IDE performance seems to be almost independent for Tr $\Sigma$ with slight dependence observed for the angle-based methods. On the other hand, most methods except lPCA, seem to show weak positive correlation with performance and $\langle R^2 \rangle$. The light vertical maroon lines represent the corresponding quantities for different classes in MNIST.

### G.5 Correlation of IDE performance with Tr $\boldsymbol{\Sigma}$ and $\langle R^2 \rangle$

We present the results for the correlation of IDE performance with Tr $\boldsymbol{\Sigma}$ and $\langle R^2 \rangle$ here. Most methods seem to show no dependence on the total variance, with few exceptions being DANCo for linear nullspaces. On the other hand, there is a weak positive correlation between performance and inter-component correlation. The latter makes *a posteriori* sense since this implies that the manifold embeddings show structrual similarites, which makes it easier for the methods to discover the latent dimension.

### G.6 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-

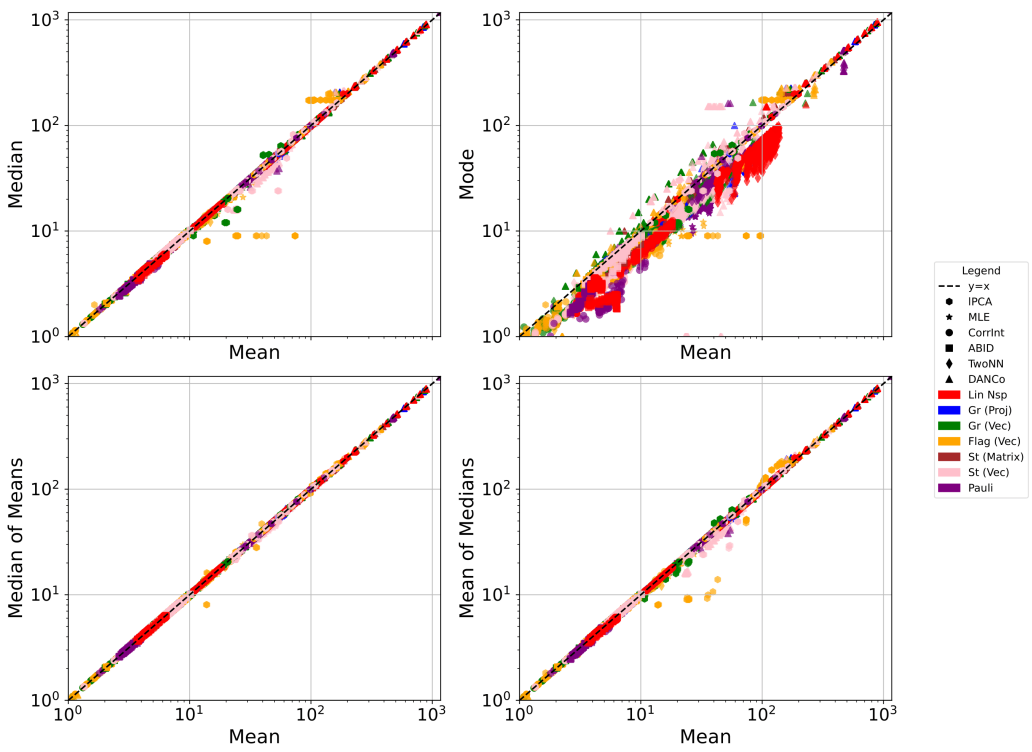



Figure 15: Results of using different statistical quantities. Most of them are concentrated around the $y = x$ line, indicating that they are quite consistent with each other.

spaced $k$ values and choose the median value as the desired geometric property. The tabulated results are then shown in Table 5

| Quantity | Gr (Vec) | Gr (Proj) | St (Matrix) | Pauli | St (Vec) | Flag (Vec) |
|---|---|---|---|---|---|---|
| $\langle H \rangle$ | 1.0313 | 1.1417 | 0.6702 | 0.4667 | 0.5348 | 1.5032 |
| $\langle \rho \rangle$ | 0.8527 | 0.5933 | 0.1766 | 0.0373 | 0.2846 | 2.9473 |
| $\langle \kappa \equiv \rho / H^{d_i} \rangle$ | 0.7758 | 0.3736 | 0.9097 | 0.8430 | 2.1033(*) | 0.5772 |
| $\langle \lvert\delta\rvert \rangle$ | **0.2023** | 0.5785 | 0.5793 | 0.5846 | 0.8502 | 2.4085 |

Table 5: Average absolute values by manifold (manifolds sorted by increasing $\langle \lvert\delta\rvert \rangle$).
(*) 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.7 Experiments with other statistical averages

We hereby report the result of replacing the mean as the GID estimate with four other statistical quantities — the median, the mode, the median of means and the mean of medians. The results are summarized in Fig.15. The relative squared errors between mean and the other quantities are respectively are $1.9 \times 10^{-3}, 9.0 \times 10^{-2}, 5.2 \times 10^{-5}, 7.9 \times 10^{-4}$, suggesting that the different statistical estimates are consistent with each other.

### G.8 Effect of hyper-parameters

We present a representative run for hyper-parameter sweep for the MLE method for Grassmanian manifolds. We notice that the absolute error decreases as a function of $N$, but then it switches from underestimation to overestimation. We also see that the error decreases mostly with increasing $k$ when $N$ is held fixed. Holding $k/N$ fixed on the other hand, results in convergence, but **not** necessarily to $\delta \approx 0$.

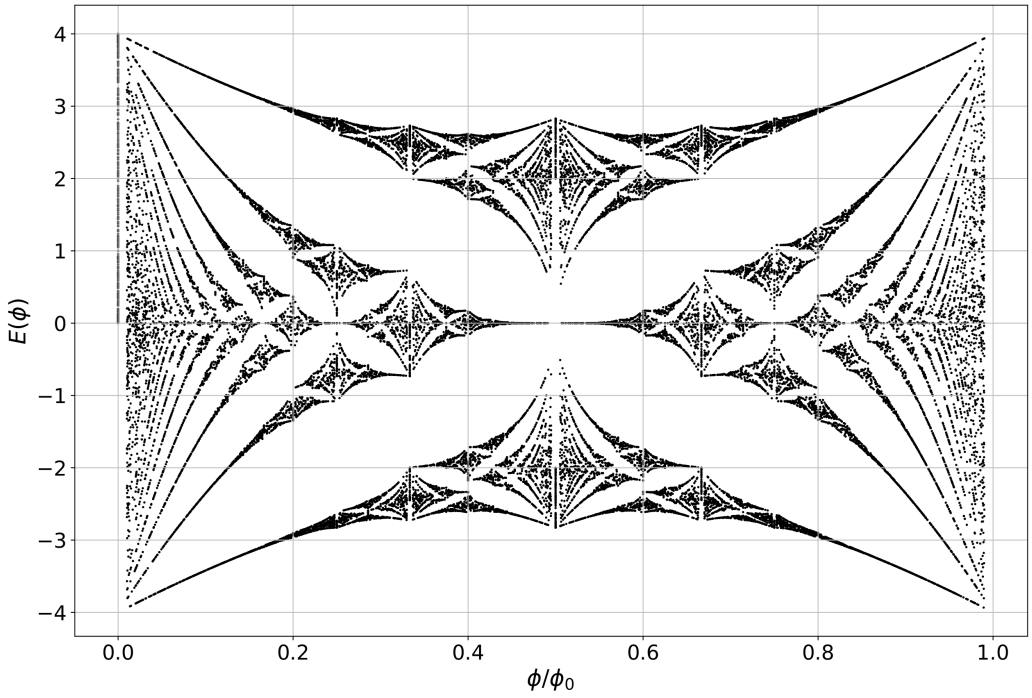



Figure 16: The Hofstadter butterfly. Numerical simulations indicate a fractal dimension of $d_i = 1.445$.

## H FRACTAL CURVES

Fractal curves present an interesting class of objects since they usually have *fractional dimension*. This is because the dimension for fractal curves is determined by the box-counting method, where if the entire box is partitioned into hypercubes of side $\epsilon$, the number of boxes with non-zero points depends on $\epsilon$ through a power-law decay with dimension

$$N(\epsilon) = N_0(\epsilon/\epsilon_0)^{-d}$$

However, fractal curves are *not* manifolds and are not locally isomorphic to hyperplanes, partly due to the discontinuous nature of the fractal sets.

### H.1 HOFSTADTER'S BUTTERFLY

Hofstadter's butterfly (45) is an example of a fractal curve obtained from quantum physics, by solving the system of electrons with nearest-neighbor hopping on a 2D lattice, while being subjected to a perpendicular constant magnetic field. The butterfly emerges when one plots the gapped energy spectra as a function of the magnetic flux through a plauette, as shown in Fig. 16

We include fractals in the appendix because while they are not really manifolds, they can serve as useful tools to test the manifold hypothesis. This is because fractals do have non-uniform local ID, and as such a method to estimate LID *should* give different answers at different points, which can be probed by (1) looking at the standard deviation and (2) visually for the butterfly which lives on a 2D plane. We report the results of these investigations on the Hofstadter butterfly only with methods which can give fractional ID estimates, in particular, we test MLE, ABID, CorrInt. [2]

By far, ABID seems to perform the best in estimating the ID, with the standard deviation decreasing as $k$ increases, indicating that more and more points are estimated to have the ambient dimension due to the increasing neighborhood size, see Fig. 19 Also the errors fall within the reasonable value of

[2]We investigated TwoNN but the native implementation could not handle `NaN` values properly, hence we omit the results here.

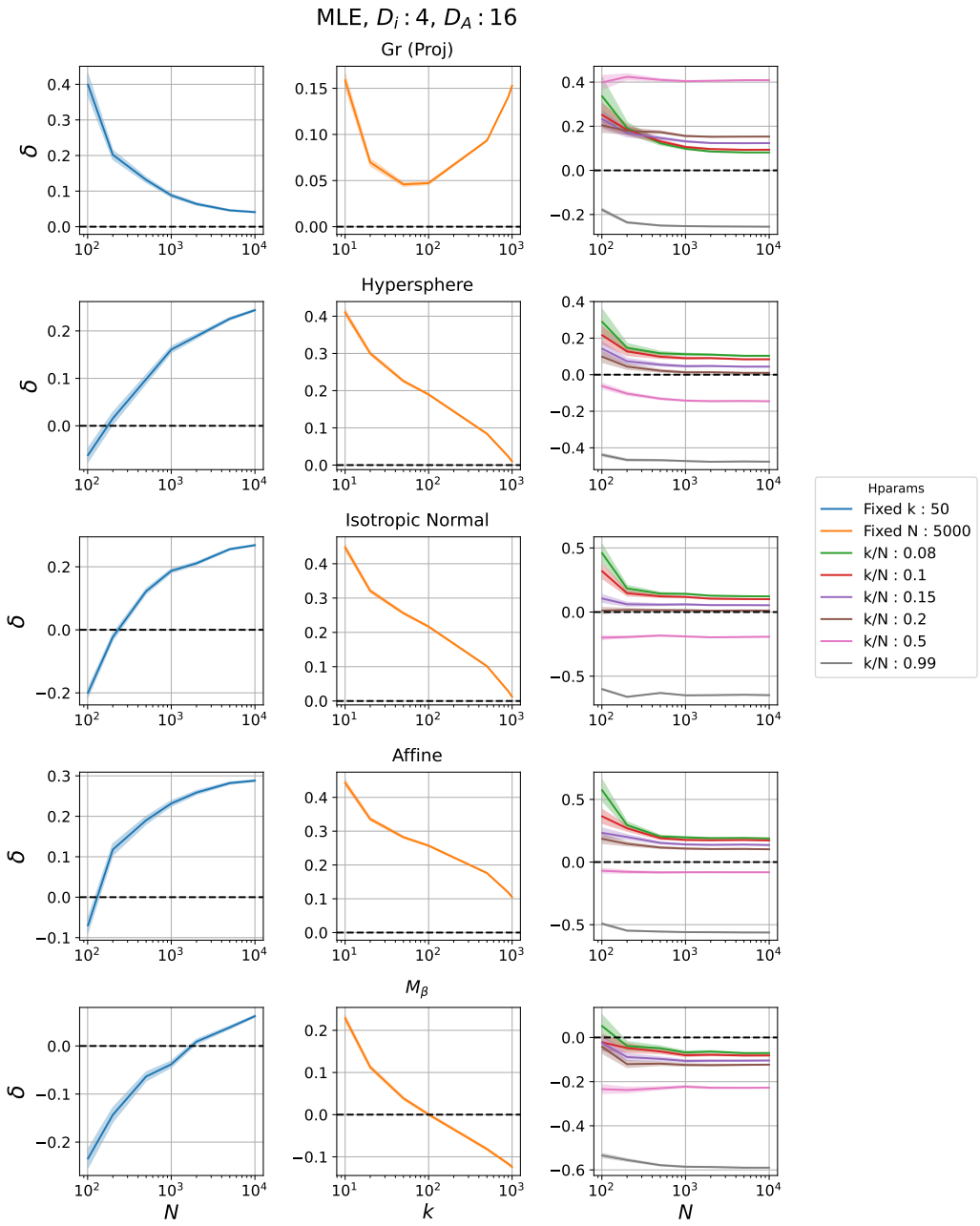



Figure 17: Results of extensive hyper-parameter run with MLE on grassmaian projector representation.

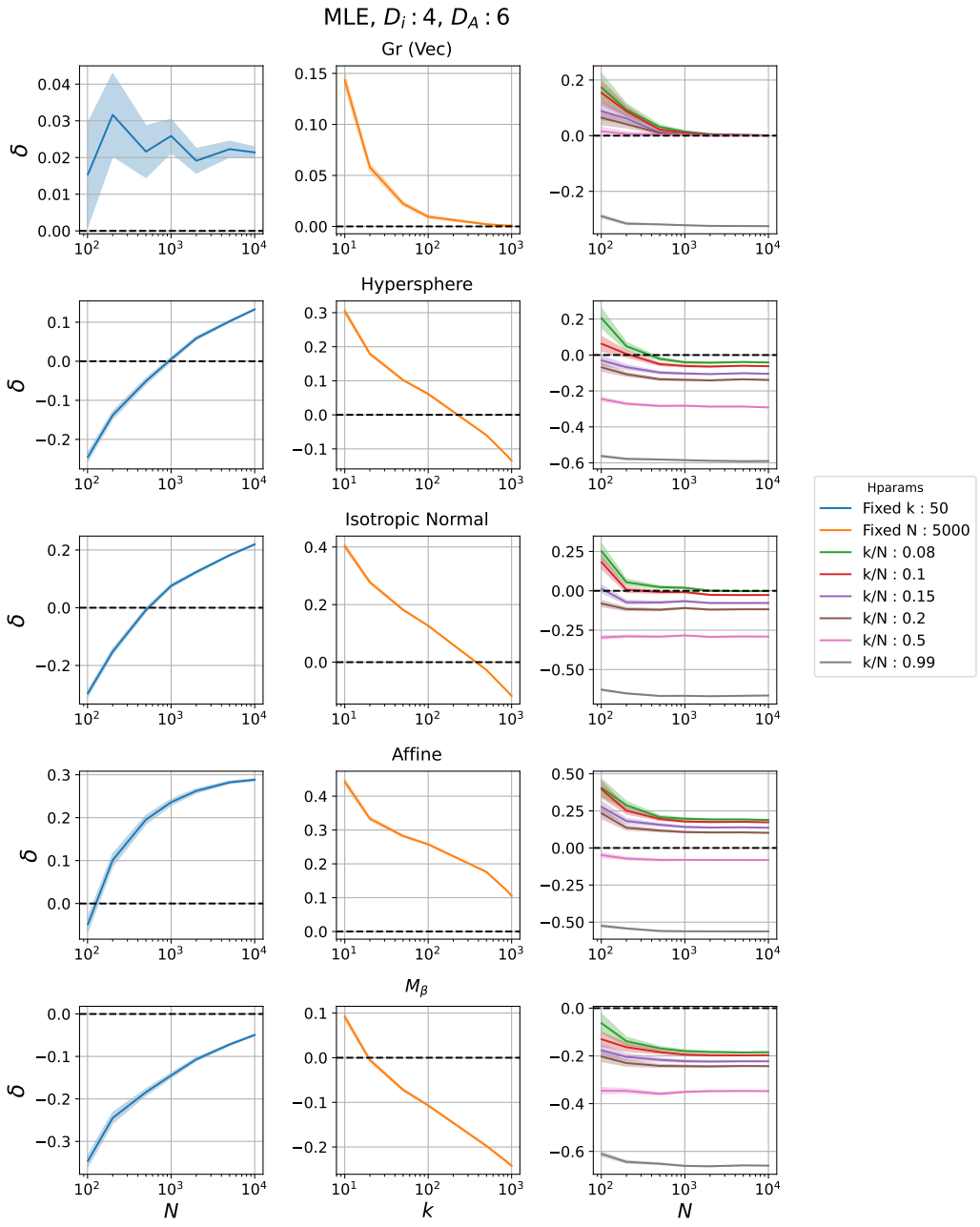



Figure 18: Results of extensive hyper-parameter run with MLE on grassmaian vector embedding.

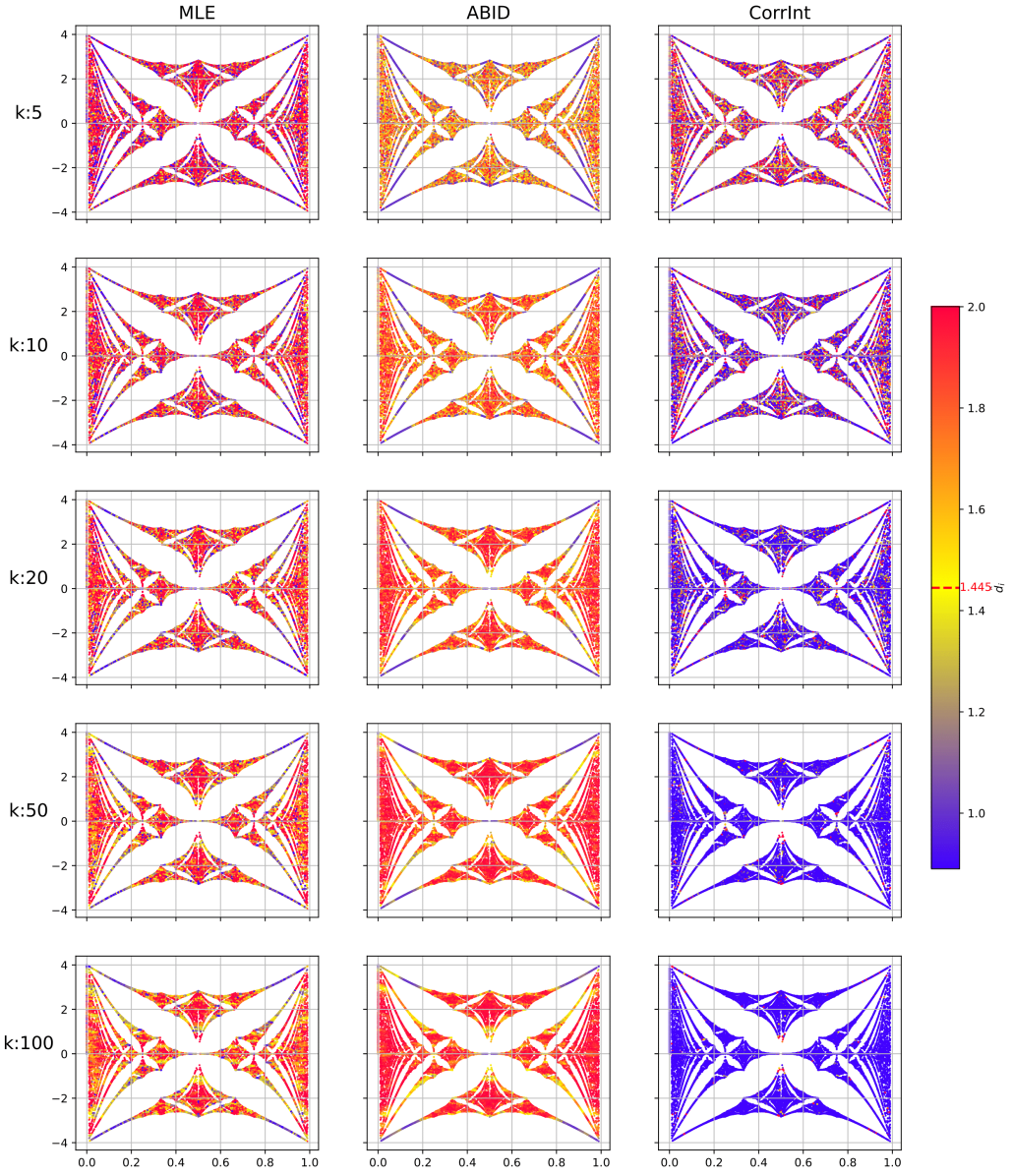



Figure 19: Red dashed line indicates the actual fractal dimension. Row indicates same $k$, column refers to same method. Note how ABID produces ID estimates for small $k$, while most other methods return LID estimates very different from the fractal dimension. $N = 1000$.

the ambient dimension, thereby showing that ABID can be trusted as a method to estimate if data lies on a *manifold* or not.

| $k$ | MLE | ABID | CorrInt |
|---|---|---|---|
| 5 | $2.285 \pm 2.083$ | $1.485 \pm 0.323$ | $1.858 \pm 1.581$ |
| 10 | $1.808 \pm 1.142$ | $1.620 \pm 0.303$ | $1.397 \pm 1.540$ |
| 20 | $1.663 \pm 0.781$ | $1.690 \pm 0.291$ | $0.885 \pm 0.933$ |
| 50 | $1.625 \pm 0.576$ | $1.742 \pm 0.274$ | $0.598 \pm 0.620$ |
| 100 | $1.631 \pm 0.502$ | $1.765 \pm 0.262$ | $0.486 \pm 0.442$ |

Table 6: Intrinsic dimension estimates (mean $\pm$ std) for different methods and values of $k$. Recall from Figure 19 that the true ID of the butterfly is $d_i = 1.445$.