LIDL: Local Intrinsic Dimension Estimation Using Approximate Likelihood

Piotr Tempczyk [1 2 3] **Rafał Michaluk** [1 2] **Łukasz Garncarek** [2 4]
Przemysław Spurek [5] **Jacek Tabor** [5] **Adam Goliński** [6]

Abstract

Most of the existing methods for estimating the local intrinsic dimension of a data distribution do not scale well to high dimensional data. Many of them rely on a non-parametric nearest neighbours approach which suffers from the curse of dimensionality. We attempt to address that challenge by proposing a novel approach to the problem: Local Intrinsic Dimension estimation using approximate Likelihood (LIDL). Our method relies on an arbitrary density estimation method as its subroutine, and hence tries to sidestep the dimensionality challenge by making use of the recent progress in *parametric* neural methods for likelihood estimation. We carefully investigate the empirical properties of the proposed method, compare them with our theoretical predictions, show that LIDL yields competitive results on the standard benchmarks for this problem, and that it scales to thousands of dimensions. What is more, we anticipate this approach to improve further with the continuing advances in the density estimation literature.

1. Introduction

In this paper, we consider the problem of local intrinsic dimension (LID) estimation of a lower-dimensional data manifold embedded in a higher-dimensional ambient space.

Intrinsic dimension estimation is an established problem in data analysis and representation learning (Ansuini et al., 2019; Li et al., 2018; Rubenstein et al., 2018). It was studied in the context of dimensionality reduction, clustering, and classification problems (Vapnik, 2013; Kleindessner & Luxburg, 2015; Camastra & Staiano, 2016) and some prototype-based clustering algorithms (Claussen & Villmann, 2005; Struski et al., 2018), among others. It is also a powerful analytical tool to study the process of training and representation learning in deep neural networks (Li et al., 2018; Ansuini et al., 2019). Rubenstein et al. (2018) shows how the mismatch between the latent space dimensionality and the dataset's ID may hurt the performance of auto-encoder generative models like VAE (Kingma & Welling, 2014), WAE (Tolstikhin et al., 2018), or CWAE (Knop et al., 2020). Recent results of Pope et al. (2020) show that the global ID of the dataset impacts the training process of a machine learning model, sample efficiency, and its ability to generalize.

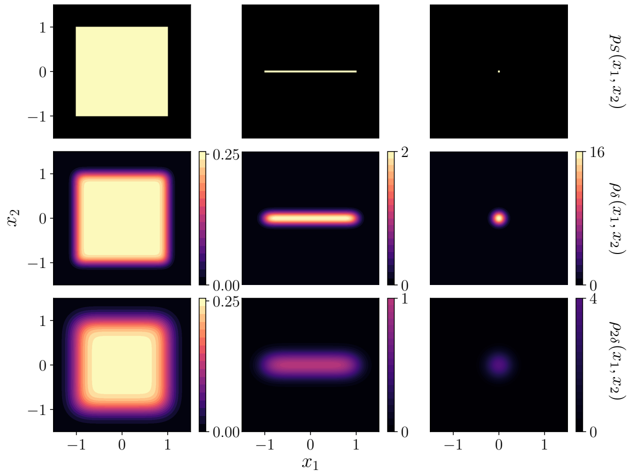

Figure 1: Illustration of LIDL's core insight. [Top] Three uniform distributions p_S supported respectively on a square, interval, and a point, with intrinsic dimensions 2, 1, 0. [Middle/bottom] Perturbed densities ρ_δ and $\rho_{2\delta}$ resulting from addition of Gaussian noise with different noise magnitudes: δ and 2δ. Our core insight is that the difference between the densities $\rho_\delta(x)$ and $\rho_{2\delta}(x)$ at any point x depends on the local intrinsic dimension (LID) at that point. Consider point $x = (0, 0)$. For the left column, that difference is zero; for the middle one, the density is halved; for the right one, it is quartered. We leverage this mechanism to estimate LID.

Recently, there has also been a rise of interest in methods for simultaneous manifold learning and density estimation

[1] Institute of Informatics, University of Warsaw [2] Polish National Institute for Machine Learning (www.opium.sh) [3] deeptale.ai [4] Applica [5] GMUM, Jagiellonian University [6] University of Oxford. Correspondence to: Piotr Tempczyk <piotr.tempczyk@mimuw.edu.pl>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

(Brehmer & Cranmer, 2020; Caterini et al., 2021; Ross & Cresswell, 2021). Most of these methods require setting the manifold dimension as a hyperparameter and the standard practice is to do this heuristically or by performing a hyperparamater sweep. Intrisic dimensionality estimation methods offer a principled alternative to this practice.

Outside of machine learning, an area where we hope reliable ID estimation might help is in the fields of science and engineering, where we nowadays collect enormous amounts of high-dimensional data. The first challenge such modern datasets present for LID estimation is scalability: we seek algorithms which yield unbiased estimates in high-dimensional spaces. Many of the existing methods for estimating ID rely on a non-parametric nearest neighbours-based approach which suffers from the curse of dimensionality. They estimate the LID by investigating the distribution of local distances and they run into the problem of "boundary effects, which distort the distribution and lead to a negative bias especially for high dimensions where any manifold with a boundary has almost all of its volume concentrated close to the boundary" (Johnsson et al., 2014), which is a symptom of the curse of dimensionality. ESS method (Johnsson et al., 2014) itself sidesteps this challenge by using angular rather than distance information, leading to its competitive performance in high dimensions, as highlighted in our evaluations. In contrast, the method we propose uses a *parametric* density estimation model which is a different mechanism to sidestep this challenge.

The second challenge is being able to deal with datasets that have highly non-isotropic structure (we refer to them as *multiscale*). A good example of this kind of dataset are images of a human face. In those datasets we have many latent factors of variation and some of them account for much less ambient-space variance than others. For instance, let us consider two of them: eye and head rotation. Each of them induces 2-dimensional manifold in the data space, but variance coming from the latter factor is much larger.

The third challenge is being able to deal with dataset dimensionality on different scales, which includes dealing with the ambient-space noise in the data. Most of the data we use is affected by measurement noise, and all the datasets are deformed further by being quantized and stored as finite precision numbers. To understand how it impacts LID estimation let us imagine a simple physical problem: we have a particle moving in an electromagnetic field, and our dataset is the set of 2-dimensional vectors describing its positions on a plane. Additionally, our measurements are quantized with some very small quantization step. When we magnify the dataset to the scale of quantization step, we see the dataset as a 0-dimensional manifold–it lies on a finite grid. When we zoom out to the scale of the measurement noise, we observe a 2-dimensional point cloud: one bigger dimension along the trajectory of the particle, and the other–smaller one–coming from the imperfect measurement. When we zoom out far enough, we observe only the 1-dimensional trajectory of that particle. It would be very convenient to be able to easily set an operating scale of an algorithm, e.g. have a possibility to ignore the dimensions that are artificially created by the measurement noise.

To address these problems, we propose a new method for Local Intrinsic Dimension estimation using approximate Likelihood (LIDL). Instead of using non-parametric methods based on local samples from the neighbourhood of a given point, it is based on parametric probability density estimation, which scales better to higher-dimensional setting. At the core of our method lies the observation that when we add Gaussian noise $\mathcal{N}(0, \delta^2 I)$ to the dataset X embedded in $\mathbb{R}^D$, the rate of change of the log-likelihood at $x \in X$ (at which LID equals d) is approximately linear in the logarithm of δ. Moreover, the proportionality constant is $\beta \approx d - D$ and we can estimate it using linear regression, thus estimating d. We may view δ as a scale parameter of our method, which may be of practical benefit beneficial when dealing with noisy datasets. Intuitive visualisation of this concept can be found in Fig. 1, and the formal derivation can be found in Sec. 2.

To the best of our knowledge, LIDL is the first theoretically grounded method of LID estimation that uses global density estimation methods. In our method we relax the assumptions many algorithms make about uniformity of the density and the manifold flatness in the neighbourhood of x. We show theoretically and experimentally that we can deal with manifolds consisting of multiple multiscale connected components of different dimensions. We also compare our algorithm with a wide range of other LID estimation algorithms, and verify that only our algorithm can give unbiased estimates for high-dimensional datasets. The code to reproduce our results is available at `github.com/opium-sh/lidl`.

The empirical success of our method was made possible by using neural density estimators called *normalizing flows* (NF) (Rezende & Mohamed, 2015), which can estimate densities even in high-dimensional spaces as images. Although in this work we use NF as LIDL's density estimator, our method can be used with any density estimation method. Thus, we anticipate that its capabilities to grow further with continuing progress in the area of density estimation.

Contributions. Our main contribution is the introduction of a novel, accurate and scalable method of LID estimation. The method is backed up with solid mathematical foundations, and verified both theoretically and experimentally. The impact of LID itself on neural network performance is experimentally assessed. Finally, we identified some problems with existing approaches to LID estimation.

2. Method

In this section, we introduce the problem setting formally and we lay out the LIDL method theoretically, including its derivation and pointing out certain predictions about its behavior that we later verify empirically in Section 3.

It is often known that a particular dataset X is a subset of some data manifold M equipped with probability measure μ. However, the manifold M and the dataset X need not be directly observable. Instead, there may exist an embedding (or, more generally, an immersion satisfying some regularity conditions, see Appendix B) $j\colon M \to \mathbb{R}^D$ into a Euclidean space of larger dimension, through which we can view X.

The method we propose is based upon the observation, expressed in Theorem 2.1, that for a probability measure supported on an embedded submanifold of an Euclidean space, the dimension of its support can be recovered from its asymptotic behavior under small normally distributed perturbations (see Fig. 1 for an intuitive illustration).

2.1. The formal setting

We first define a class of measures we will restrict our considerations to, which we will call *smooth* measures, including all measures with continuous positive densities.

Definition 2.1. A positive measure ν on a manifold N will be called *smooth* if for any chart $\psi\colon U \to V \subset \mathbb{R}^n$ of N, the pushforward $\psi_*\nu$ is absolutely continuous with respect to the Lebesgue measure λ on V, and moreover, its density is locally bounded away from 0, i.e. any $x \in V$ admits a neighborhood on which $dj_*\nu/d\lambda > c$ for some $c > 0$.

Let $S \subset \mathbb{R}^D$ be a smooth connected d-dimensional embedded submanifold of a high-dimensional Euclidean space $\mathbb{R}^D$ (the more general case of a non-connected immersed manifold is dealt with in Appendix B). This is our observable data manifold, embedded in Euclidean space, i.e. $S = j(M)$. Furthermore, suppose we are given a smooth (according to Definition 2.1) probability measure p_S on S, representing the data probability distribution. In our notation, this is the pushforward of the probability μ on M, i.e. $p_S = j_*\mu$. We will implicitly treat p_S as a probability distribution on the whole ambient space $\mathbb{R}^D$.

The Gaussian function (i.e. the density of the standard normal distribution) on a Euclidean space V will be denoted by ϕ^V, or ϕ^n in the case where V is the standard $\mathbb{R}^n$ space. Also, for $\delta > 0$, let

$$\phi^V_\delta(x) = \delta^{-\dim V}\phi^V(x/\delta) \tag{1}$$

be the density of the normal distribution $\mathcal{N}(0, \delta^2 I)$ with covariance matrix $\delta^2 I$, where I is the identity matrix on V.

Under the above notation, if $X \sim p_S$ is a random vector representing the data, and $N_\delta \sim \mathcal{N}(0, \delta^2 I)$ a normally distributed random noise vector, the distribution of the perturbed random vector $X + N_\delta$ in $\mathbb{R}^D$ is given by the convolution $p_S * \mathcal{N}(0, \delta^2 I)$, and has density

$$\rho_\delta(x) = \int_S \phi^D_\delta(x - y)\, dp_S(y). \tag{2}$$

Finally, let us introduce a notation for uniform multiplicative estimates. We will write that $f(x, y) \asymp g(x, y)$ uniformly in x if for every y there exists $C > 0$ such that for all x

$$C^{-1}g(x, y) \le f(x, y) \le Cg(x, y). \tag{3}$$

This notation extends to any number of variables. We will use it to declutter the proofs from irrelevant constants.

2.2. The core estimate

At any $x \in S$ the tangent space of $\mathbb{R}^D$ admits a decomposition $T_x\mathbb{R}^D = T_xS \oplus N_xS$ into a direct sum of the tangent and normal spaces of S. Under the natural identification of $T_x\mathbb{R}^D$ with the underlying $\mathbb{R}^D$ (mapping the origin of T_xS to x), the tangent and normal spaces of S at x become two affine subspaces of $\mathbb{R}^D$ intersecting at x. Denote by $\pi_x\colon \mathbb{R}^D \to T_xS$ and $\pi_x^\perp\colon \mathbb{R}^D \to N_xS$ be the corresponding orthogonal projections. With this notation, the following decomposition of the Gaussian density holds

$$\phi^D_\delta(x - y) = \phi^{T_xS}_\delta(\pi_x(y))\phi^{N_xS}_\delta(\pi^\perp_x(y)). \tag{4}$$

By the Inverse Function Theorem applied to the restriction of π_x to S, in a small neighborhood of any $x \in S$, the manifold S can be represented as the graph of a smooth map $F_x\colon T_xS \to N_xS$. In particular, it follows that in this neighborhood one has $\pi^\perp_x = F_x \circ \pi_x$. Moreover, $F_x(0) = 0$, and since the graph of F_x is tangent to T_xS at the origin, the derivative of F_x at x vanishes. Hence, the Taylor expansion of F_x at 0 starts with the second-order term, and consequently, there exists $C > 0$ such that for small v

$$\|F_x(v)\|_{N_xS} \le C\|v\|^2_{T_xS}. \tag{5}$$

Denote by $B(x, r)$ the ball of radius r in $\mathbb{R}^D$, centered at x. Subsequent five lemmas are proven in Appendix A. Here we give only their statements, followed by the proof of our core estimate.

Lemma 2.1. *Let $x \in S$. For sufficiently small δ the projection $\pi_x(S \cap B(x, \delta^{1/2}))$ contains the ball $B(x, \delta) \cap T_xS$.*

Lemma 2.2. *For $x \in S$ and sufficiently small δ, the estimate*

$$\int_{S \cap B(x, \delta^{1/2})} \phi^{T_xS}_\delta(\pi_x(y))\, dp_S(y) \asymp 1 \tag{6}$$

holds uniformly in δ.

Lemma 2.3. *For sufficiently small δ and $y \in S \cap B(x, \delta^{1/2})$, where $x \in S$, the estimate $\phi_\delta^{N_xS}(\pi_x^\perp(y)) \asymp \delta^{d-D}$ holds uniformly in δ and y.*

Lemma 2.4. *For $x \in S$ and sufficiently small δ,*

$$\int_{S \cap B(x,\delta^{1/2})} \phi_\delta^D(x-y)\, dp_S(y) \asymp \delta^{d-D} \tag{7}$$

uniformly in δ.

Lemma 2.5. *For every $x \in S$*

$$\lim_{\delta \to 0^+} \int_{S \setminus B(x,\delta^{1/2})} \phi_\delta^D(x-y)\, dp_S(y) = 0. \tag{8}$$

Theorem 2.1 (The core estimate)**.** *Assume that $S \subset \mathbb{R}^D$ is a connected d-dimensional submanifold endowed with a smooth probability measure p_S. Let ρ_δ be the density of $p_S * \mathcal{N}(0, \delta^2 I)$ on $\mathbb{R}^D$. Then for $x \in S$ and sufficiently small δ, we have*

$$\log \rho_\delta(x) = (d-D)\log\delta + O(1). \tag{9}$$

Proof. Since $\delta^{d-D} \geq 1$ for $\delta \leq 1$, given sufficiently small δ, from Lemma 2.5 we get

$$\int_{S \setminus B(x,\delta^{1/2})} \phi_\delta^D(x-y)\, dp_S(y) < \delta^{d-D}. \tag{10}$$

By combining this with eq. (2) and Lemma 2.4, we obtain $\rho_\delta(x) \asymp \delta^{d-D}$, which yields the desired estimate after taking log. □

2.3. The LIDL algorithm

Now, let us consider how to use the core estimate derived above in practice. The core requirement of LIDL is access to the approximate densities $\rho_\delta(x)$, which we have to obtain by fitting a density estimator on the data points from the dataset perturbed with a normally distributed noise of an appropriate magnitude δ. Luckily, these days there exist density estimators which scale to data even as high-dimensional as images. For the purpose of empirical evaluation of our method, in this work we use three models from the family of NF, however we emphasize that our method could use absolutely any density estimation method. A viable alternative could be, for example, using diffusion models (Song et al., 2021), which are likely to lead to further improved accuracy of LIDL estimates.

Algorithm 1 LIDL algorithm

```
Require: $X \subset \mathbb{R}^D$; $x_1, \ldots, x_m \in \mathbb{R}^D$; $\delta_1, \ldots, \delta_n \in \mathbb{R}^+$;
  for $j = 1$ to $n$ do
    $X_j \leftarrow X$ perturbed with $\mathcal{N}(0, \delta_j^2 I_D)$
    Fit the density model $\hat{\rho}_j$ to $X_j$
  end for
  for $i = 1$ to $m$ do
    for $j = 1$ to $n$ do
      $\xi_j \leftarrow \log \delta_j$
      $\eta_j \leftarrow \log \hat{\rho}_j(x_i)$
    end for
    $\beta \leftarrow$ regression coefficient for a set of $n$ points $(\xi_j, \eta_j)$
    $\hat{d}_i \leftarrow D + \beta$
  end for
  return $(\hat{d}_1, \ldots, \hat{d}_m)$
```

Given a dataset $X \subset \mathbb{R}^D$, and a point $x \in \mathbb{R}^D$, at which we want to estimate LID (usually $x \in D$, as we want to take a point from the image of the data manifold in $\mathbb{R}^D$), we proceed as follows. First, we choose $n > 1$ values $\delta_1, \ldots, \delta_n$ of perturbation magnitude. We discuss how to choose δ in the following section. Then, we fit n probability densities $\hat{\rho}_i$, which will be our approximations of ρ_{δ_i}. Having estimated the densities $\hat{\rho}_i$, we consider the sequence of points of the form $(\log\delta_i, \log\hat{\rho}_i(x))$. Using linear regression to fit eq. (9), we get an estimate β for $d - D$, from which we obtain $\hat{d} = D + \beta$, an estimate for d. To estimate LID for multiple points, we can fit the densities once, and then loop over the points. Full algorithm is presented in Algorithm 1.

It is worth noting that our method fits nicely into the LID estimation framework presented in (Amsaleg et al., 2019). Roughly speaking, it is based on two observations. Firstly, the dimension of an Euclidean space can be recovered from the degree of the polynomial growth rate of its ball volume as a function of its radius. Secondly, this idea can be applied to discrete datasets by replacing the notion of ball volume with the likelihood function of finding a point of the dataset within a given distance from a fixed base point.

In our notation, this likelihood function is $r \mapsto p_S(B(x, r))$, where x is the base point. With reasonable assumptions on the measure p_S, it can be shown that for small r this function behaves like a polynomial of degree d, so the LID value we are estimating is the same as what is defined in (Amsaleg et al., 2019), which can be consulted for more details.

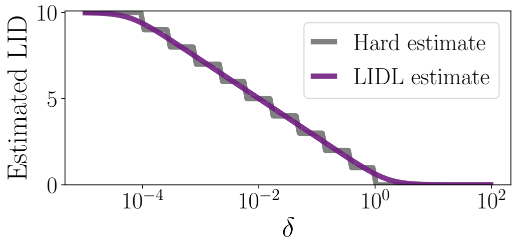

Figure 2: LIDL and hard estimates for different values of δ for 10D non-isotripic Gaussian. Notice how LIDL ignores dimensions smaller than δ, as predicted theoretically.

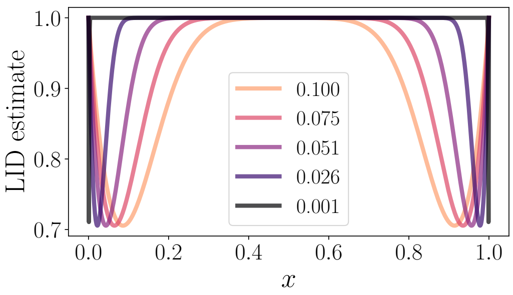

Figure 3: LIDL estimates for points from $\mathcal{U}(0, 1)$ for different values of δ marked with different colors, as explained in the legend of the plot.

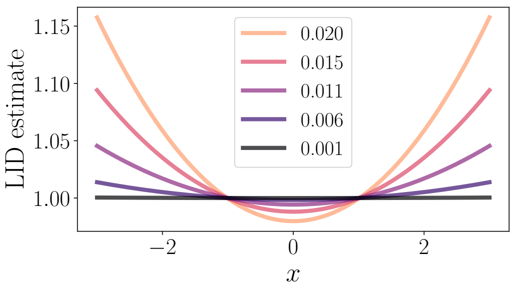

Figure 4: LIDL estimates for points from $\mathcal{N}(0, 1)$ for different values of δ marked with different colors, as explained in the legend of the plot.

2.4. Viewing δ as a scale parameter

In the previous section, we glossed over the fact that we have to choose the values of δ. From Sec. 2.2 we know that the core estimate is exact for infinitesimally small δ, so the first pressure is for the δ to be small, possibly as small as the numerical precision allows.

However, δ can also be viewed as a length-scale parameter that allows users to choose a certain minimum 'thickness' to be considered, such that dimensions 'thinner' than the threshold will be ignored. Consider an illustrative example: suppose that the probability distribution p_S is concentrated in a tubular neighborhood of another submanifold S' of dimension $d' < d$. In this case, it can be approximated by a probability distribution $p_{S'}$ supported on S'.

Now, if this approximation is 'good', in the sense that the thickness of the considered neighborhood is much smaller than the values of δ used, then intuitively the LIDL estimate should actually reflect the dimension d' of the submanifold S' instead of the true dimension d.

Continuing this example, we present the described behavior empirically. Let $S = \mathbb{R}^D$, and $p_S = \mathcal{N}(0, \Sigma)$, where Σ is a diagonal matrix with entries $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_D^2$. In Fig. 2, we plot the LIDL estimates for case $D = 10$ and for σ equally distributed on the logarithmic scale. We also plot the *hard estimate* which is simply counting the number of entries in σ that are larger than δ. The LIDL estimate follows the hard estimate, and this behavior is predicted theoretically in Appendix D. We further investigate the role of the scale parameter in the case of imperfect density estimates in Sec. 5.2.

As mentioned earlier, having an explicit length-scale parameter can be considered LIDL's feature as compared to other methods. Is allows the user to easily set an operating scale such that to ignore certain amplitude of noise in the original data, e.g. the observation noise if we are able to estimate its magnitude apriori. The empirically observed rule of thumb is to take at least $\delta \gtrsim 10\sigma$, where σ is standard deviation of the noise to be ignored.

Setting such operating scale characteristic can be difficult in many other non-parametric algorithms that calculate statistics based on nearest neighbors. In those approach, there is either of the two natural scale parameters: number of nearest neighbours k or radius r around the point where we search for neighbours.

When using k, our effective operating range depends on a combination of local density and the total number of samples used to run the algorithm. Using r allows to set an operating range. In this case, however, we expose ourselves to the risk of having not enough samples to estimate the local density. Most implementations of those methods set a default k.

3. Empirical Behavior of the Proposed Method

In this section we examine the behaviour of our method when confronted with certain isolated difficulties. Instead of relying on a computed approximation $\hat{\rho}_\delta$, we assume we are given the actual perturbed density ρ_δ explicitly or we compute it through numerical integration. This ensures that any error observed during this analysis is caused directly by our LIDL method and not the density estimator. However, it comes at a price of restricting us to relatively simple examples where we can efficiently compute ρ_δ.

3.1. Uniform density on an interval

We assume, that in the neighborhood of x the density is bounded from below by a positive constant. But for some real-world cases, this assumption is not fulfilled. To investigate how LIDL behaves in this case we ran it on $\mathcal{U}(0, 1)$. It can be seen as a distribution on the real line, whose density vanishes outside $[0, 1]$ interval, violating this assumption. Alternatively, in the vicinity of the interval endpoints, the size of the neighborhood admitting the parametrization required for the proof of the core estimate decreases to 0.

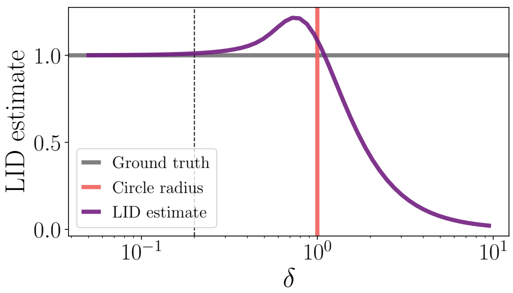

Figure 5: LIDL estimate as a function of δ for a uniform density on a unit circle. Vertical line at $\delta = 0.2$.

We analytically calculated the convolution of $\mathcal{U}(0, 1)$ with $\mathcal{N}(0, \delta^2)$ and used it to estimate LID at 1000 points between 0 and 1. We used just two points for linear regression, corresponding to $\delta_1 = \delta$ and $\delta_2 = 1.05\delta$. The estimates for different values of δ are plotted in Fig. 3. We can see that an error is introduced near the boundary as expected. In this case, its maximum value does not depend on the value of δ, and points affected by this problem lie in the part closer than $\sim 4\delta$ to the endpoints of the distribution support.

3.2. Normal distribution on a line

In this example, we study the LID estimates at different points of a line embedded in $\mathbb{R}^D$. In Fig. 4 we can see the estimates computed as per the previous example, for a few values of δ. At first glance, it is worrying that the error seems to explode with distance from the mean of the distribution. In Appendix D, we show that the error is quadratic in this distance, and, reassuringly, that its expected value over the whole distribution can be controlled. The reason for this behavior can be traced back to the proof of Lemma 2.2 (more specifically eq. (13) in the appendix), which depends on the positive constant locally bounding the density from below. In our example, the density decreases as $e^{-t^2/2}$, which produces the quadratic error term (the final error is bounded by $\sum_i |\log C_i|$, where C_i are the multiplicative estimate constants appearing in all the steps of the proof).

3.3. Uniform density on a curved manifold

The LIDL estimate is affected by the curvature of the manifold, which manifests in the constant C appearing in eq. (5), subsequently used in the proofs of Lemmas 2.1 and 2.3. To see empirically how the curvature influences the LIDL estimate, we numerically computed the convolution of the uniform density on the unit circle embedded in $\mathbb{R}^2$ with the noise distribution $\mathcal{N}(0, \delta^2 I)$ for 2 values of δ similarly as in the previous examples. We calculated LIDL for the range of $\delta \in (0.05, 10)$. We plot the estimate dependence on δ in Fig. 5.

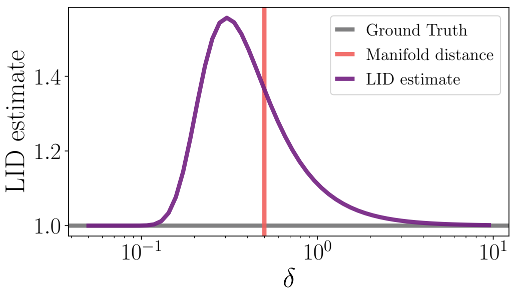

Figure 6: LIDL estimate as a function of δ for 2 long 1-dimensional manifolds parallel to each other.

We can see that for $\delta \lesssim 0.2$ the estimate error is relatively small. After the positive bias for $\delta < 1$ we can observe a monotonic drop in the estimate until it reaches nearly 0. This is by the effect described in Section 2.4 where LIDL was observed to ignore the directions in which the standard deviations were lower than δ.

3.4. Manifolds with neighboring components

In a real-world setting, it is possible for some connected components of the data manifold S to be close to each other in the observable data space $\mathbb{R}^D$, especially when some features in the dataset have discrete distribution (e.g. height and sex in a medical dataset). In those settings, for values of δ comparable to the distance between the components, additional bias may be introduced to the estimate. To investigate this we ran an experiment similar to the previous example, but with a uniform distribution supported on the union of two long parallel segments. We then calculated LIDL estimates for the midpoints of those segments, to minimize the error caused by proximity to the boundary. We present the results in Fig. 6. We can see positive bias in LIDL estimate appearing as δ is close to the distance between the segments, while for δ much larger than this distance, LIDL seems to view those two segments as a single line.

3.5. Impact of linear regression on LIDL estimate

Because our estimate depends on linear regression algorithm in order to estimate β, it may suffer from the same issues as any regression coefficient estimation algorithm (Li, 1985), so in the future, more robust algorithm for linear regression estimation may be considered. Because we estimate only the rate of change, and not the constant from linear regression equation, LIDL is prone to biased log-likelihood estimates, and noise added to log-likelihood estimate only affects the variance of the estimate.

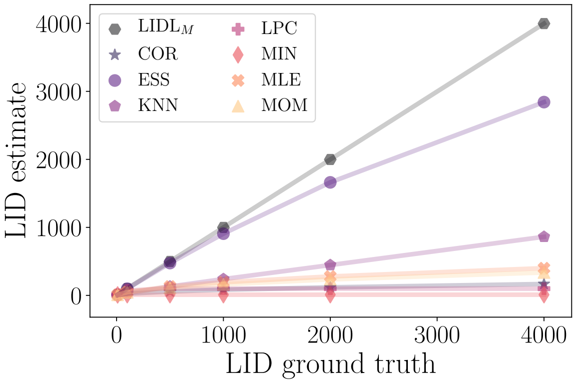

Figure 7: LID estimates for d dimensional uniform distribution on a hypercube. More results and abbreviation explanations can be found in Tables 1, 2 and 3 in Appendix F. The dimensionality d of the distribution is plotted on the horizontal axis and the estimates for different algorithms on the vertical axis.

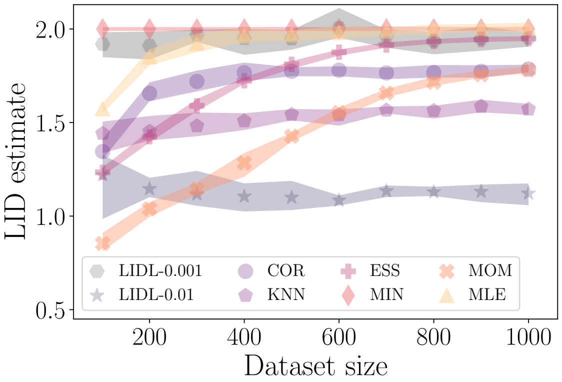

Figure 8: LID estimates for uniform distribution on a rectangle with edge lengths equal to 0.1 and 0.01. The size of the dataset is plotted on the horizontal axis and the estimate (with respective 95% confidence intervals) on the vertical axis. For most algorithms (except LIDL, KNN and MIN), we can see a disturbing phenomenon: the estimate depends on the sample size. LIDL-δ stands for LIDL with MAF density estimator and scale parameter δ.Other abbreviations are explained in appendix in Table 3.

3.6. Synthetic datasets

We ran evaluations of LIDL with density estimates computed using numerical integration on Swiss roll, uniform distribution on a helix, and Gaussians from 10 up to 4000 dimensions. We got almost exact estimates with mean absolute error (MAE) lower than 10^{-4} for every dataset. More information about these results is included in Appendix F.

4. Related Work

ID estimation methods can be divided into two broad categories: global and local (Camastra & Staiano, 2016). Global methods aim to give a single estimate of the dimensionality of the entire dataset, which however discards the nuanced manifold structure when the data lies on a union of different manifolds (which is often the case for real-world datasets).

On the contrary, the local methods (Carter et al., 2009; Kleindessner & Luxburg, 2015; Levina & Bickel, 2004; Hino et al., 2017; Camastra & Staiano, 2016; Rozza et al., 2012; Ceruti et al., 2014; Camastra & Vinciarelli, 2002) try to estimate the local ID of the data manifold at an arbitrary point. This approach gives more insight into the nature of the dataset, and provides more options to summarize the dimensionality of the manifold than the global perspective. A detailed overview of the methods used for global and local ID estimation is provided by Camastra & Staiano (2016), and for a good review of the local ID estimation methods we refer the reader to Johnsson et al. (2014). We list all the algorithms we compare to in Table 3 in the appendices.

5. Experiments

In this section, we compare LIDL with other algorithms, investigate its behavior with imperfect density estimators, and run it on real-world datasets. Details of training procedure can be found in Appendix F and in Sec. F.1 we describe how to reduce an error of our estimate.

5.1. Comparison on synthetic datasets

We collated LIDL with other LID estimation algorithms from `scikit-dimension` Python library (Bac et al., 2021), which covers all of the important algorithms for LID estimation, and compared them in three different aspects: *1*. Scalability, *2*. Multidimensional and curved manifolds, *3*. Multiscale manifolds.

We excluded FisherS and DANCo algorithms because they do not scale well to higher-dimensional settings. FisherS suffered from memory problems on medium datasets, and DANCo had unfeasibly long runtimes (multiple weeks) on the thousand-dimensional datasets. According to the convention in the field, we choose to make comparison only on synthetic datasets, because we have ground truth for them.

Scalability To test scalability we ran all algorithms on standard multidimensional uniform and normal distributions up to 4K dimensions. Detailed results of the comparison are gathered in Appendix F in Tables 1 and 2 (starting from the 7-th row). Each dataset consisted of 10K data points and each algorithm was run 5 times on different samples

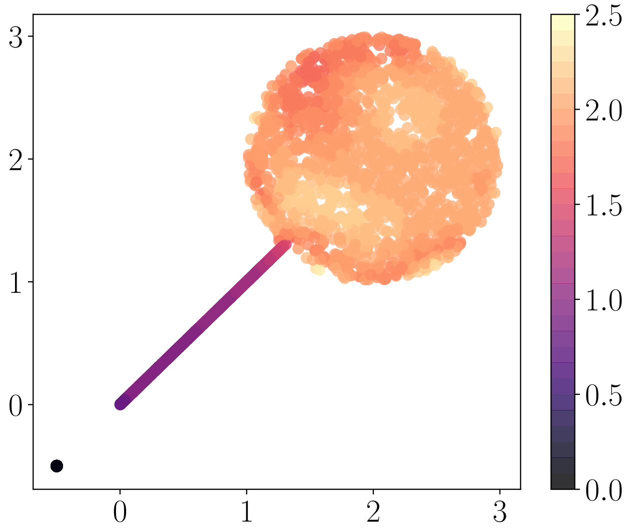

Figure 9: Points from the lollipop benchmark dataset and LIDL (with MAF) estimates for those points.

from the distribution. For each run, we calculated differences between true LID and estimate and averaged it over 5 runs. Then we divided the result by the average manifold dimensionality for each dataset, getting a relative bias of each algorithm. In subsequent tables, we report relative MAE and estimate standard deviation for the same procedure. From those tables, we can clearly see, that although in many cases LIDL does not have the lowest error and bias, for almost all datasets the results are in the $\pm 5\%$ range. Other algorithms fail to accurately estimate dimensions exceeding 100. One exception is ESS, which stands out from the rest but remains inferior to LIDL. We plot LID estimates for some of the algorithms (we omitted few for the sake of clarity) for multidimensional uniform distributions in Fig. 7. All the abbreviations used in the plot are explained in Appendix F.

Multiscale manifolds In the Introduction, we postulated that a useful LID estimation algorithm should operate properly on multiscale manifolds. In this section, we compare existing LID methods and LIDL on highly non-isotropic datasets. We observed that most of the algorithms with the same scale parameters (or those without such parameters, like ESS) give different results for different sizes of the dataset. We hypothesize that this may be caused by violating assumptions about the local uniformity of the distribution, but we did not investigate it further. Only LIDL, MiND-ML, DANCo, and KNN give stable estimates for different dataset sizes. We plot those results for selected algorithms in Fig. 8. For both scale parameter values, LIDL gives stable estimates for different dataset sizes. The rest of the unplotted algorithms also give unstable estimates, and we omitted them only to make the plot more readable.

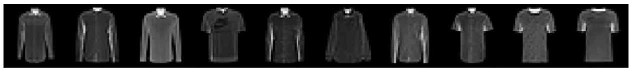

Figure 10: Images from the FMNIST dataset, for which the LID estimate is close to 0. This effect occurred when we used to high δs for this thin data manifold.

Curved manifolds and unions of manifolds We tested LIDL and other algorithms on some smaller but more complicated manifolds. We used three classical benchmarks from (Kleindessner & Luxburg, 2015): the Swiss roll dataset, uniform density on a helix, and uniform density on a sphere. These datasets lie on a curved manifolds (2-, 1- and 7-dimensional respectively) which may cause difficulties with fitting density estimators. Results of those experiments can be found in rows 4-7 of Tables 1 and 2. The results for LIDL are decent (Relative bias less than 0.05 and relative MAE less than 0.06), but lPCA and ESS gave estimates with relative bias and MAE less than 0.01. For perfect density estimates LIDL gives almost perfect estimates on those datasets, as presented in Sec. 3.6.

None of the above datasets however consisted of components of different dimensions, which may be the case for many real-world datasets. We used a *lollipop dataset*, which is composed of 0, 1, and 2-dimensional components. The dataset and its corresponding LIDL estimates are plotted in Fig. 9. On the 2- and 1-dimensional parts, many algorithms achieved good results, some even better than LIDL, but the 0-dimensional component, which consisted of replicas of the same point, caused most problems for other algorithms.

When algorithms tried to estimate LID for this 0-dimensional part, only lPCA and LIDL were able to estimate its dimensionality properly, and almost all other algorithms failed to converge. When we jittered those points a little with $\mathcal{N}(0, 10^{-6})$, almost all of the algorithms converged but all of them yielded estimates close to 2. Thanks to the possibility of setting operating scale in LIDL, we could estimate the latter dimension correctly, regardless of noise in the data. Results for each component of the manifold treated separately can be found in the first 3 rows of Tables 1 and 2.

5.2. Operating range

As stated in Sec. 2.4, δ can be seen as a scale parameter. We introduced some numerical and theoretical results to support this hypothesis, and in this section, we are going to present some experiments investigating this topic. In Fig. 13 we present a similar experiment to that from Fig. 2, but this time with 4-dimensional uniform density. Results seem quite similar to previous theoretical results. For similar Gaussian distribution, we get an almost identical relation between dimension variance, LIDL estimate and δ.

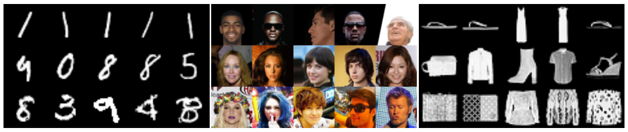

Figure 11: Samples from different image datasets (MNIST, Celab-A, FMNIST from left to right) presented according to their LIDL estimates (top to bottom). Those results are highly correlated with the complexity of an image.

We also tuned a δ range on image MNIST and FMNIST to reduce dequantization noise influence on the LIDL estimate. More on those experiments can be found in Appendix F. Although this scale parameter has to be used with care. In one experiment on FMNIST (normalized to have values between -0.5 and 0.5) for values of $\delta > 0.1$ we observed that the whole cluster of darker clothes had been estimated as being 0-dimensional. We present some samples from this cluster in Fig. 10.

5.3. Experiments on image datasets

We ran LIDL on MNIST, FMNIST, and Celeb-A (D = 1K, 1K, 12K respectively) datasets using Glow as a density estimator. We sorted those datasets according to LIDL estimates and observed that visually more complex examples have higher LID. Some small, medium, and high dimensional images from those datasets are shown in Fig. 11 and Fig. 20, 21, 22 from Appendix F. In the aforementioned section we plot a distribution of LIDL estimates for different classes from MNIST and FMNIST, and show how the estimate is affected by the dequantization used in NF training.

In Appendix F.3 we used LIDL to show, that LID negatively correlates with local (per image) accuracy of the classification model for images and that LID is positively correlated with image reconstruction error of VAE.

6. Conclusions

We identified three challenges in LID estimation and explained how the existing methods do satisfy those desiderata. To overcome those limitations we introduced an algorithm for LID estimation which relies on powerful neural parametric density estimators, and provided solid theoretical justification for the method. Our experiments showed that it can scale to datasets of thousands of dimensions and give accurate estimates on complicated manifolds. We investigated its strengths and limitations and showed that LID is connected with local model performance, especially in unsupervised learning and classification settings.

There is a number of future research directions stemming from this work. The first one is a more theoretically grounded and experimentally tested procedure for choosing δ in the presence of observation noise, which might be important for practitioners. Another one is further investigating the connection of LID estimates and classifier performance: LID estimates could be used in active learning, semi supervised learning or curriculum learning.

7. Acknowledgements

Some experiments were performed using the Entropy cluster at the Institute of Informatics, University of Warsaw, funded by NVIDIA, Intel, the Polish National Science Center grant UMO2017/26/E/ST6/00622 and ERC Starting Grant TOTAL. Some experiments were performed using the GUŚLARZ 9000 workstation at the Polish National Institute for Machine Learning.

Jacek Tabor is supported by Foundation for Polish Science (grant no POIR.04.04.00-00-14DE/18-00) carried out within the Team-Net program co-financed by the European Union under the European Regional Development Fund. Przemysław Spurek is supported by the National Centre of Science (Poland) Grant No. 2019/33/B/ST6/00894. Adam Goliński is supported by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems.

We wish to thank various people for their contribution to this project: Marek Cygan for overall support during this project and valuable and constructive suggestions on the manuscript; Maciej Dziubiński, Piotr Kozakowski, Dominik Filipak, and Maciej Śliwowski for their valuable and constructive suggestions on the manuscript.

8. CRediT Author Statement

	PT	RM	ŁG	PS	JT	AG
Conceptualization	■				■	■
Methodology	■		■		■	■
Software	■	■	■			
Validation	■					
Formal analysis	■		■			
Investigation	■	■				
Writing - Original Draft	■		■	■		■
Writing - Review & Editing	■		■	■	■	■
Visualization	■					■
Supervision	■				■	■
Project administration	■					
Data Curation		■				

References

Albergante, L., Bac, J., and Zinovyev, A. Estimating the effective dimension of large biological datasets using fisher separability analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Amsaleg, L., Chelly, O., Furon, T., Girard, S., Houle, M. E., Kawarabayashi, K.-i., and Nett, M. Extreme-value-theoretic estimation of local intrinsic dimensionality. *Data Mining and Knowledge Discovery*, 32(6): 1768–1805, 2018.

Amsaleg, L., Chelly, O., Houle, M. E., Kawarabayashi, K.-I., Radovanović, M., and Treeratanajaru, W. Intrinsic dimensionality estimation within tight localities. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 181–189. SIAM, 2019.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 6111–6122, 2019.

Bac, J., Mirkes, E. M., Gorban, A. N., Tyukin, I., and Zinovyev, A. Scikit-dimension: A python package for intrinsic dimension estimation. *Entropy*, 23(10), 2021. ISSN 1099-4300.

Brehmer, J. and Cranmer, K. Flows for simultaneous manifold learning and density estimation. 2020.

Camastra, F. and Staiano, A. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328: 26–41, 2016.

Camastra, F. and Vinciarelli, A. Vinciarelli, a.: Estimating the intrinsic dimension of data with a fractal-based method. ieee trans. pami 24, 1404-1407. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 05 2002. doi: 10.1109/TPAMI.2002.1039212.

Cangelosi, R. and Goriely, A. Component retention in principal component analysis with application to cdna microarray data. *Biology direct*, 2(1):1–21, 2007.

Carter, K. M., Raich, R., and Hero III, A. O. On local intrinsic dimension estimation and its applications. *IEEE Transactions on Signal Processing*, 58(2):650–663, 2009.

Caterini, A. L., Loaiza-Ganem, G., Pleiss, G., and Cunningham, J. P. Rectangular Flows for Manifold Learning. *NeurIPS*, 2021.

Cavallari, G. B., Ribeiro, L. S., and Ponti, M. A. Unsupervised representation learning using convolutional and stacked auto-encoders: a domain and cross-domain feature space analysis. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 440–446. IEEE, 2018.

Ceruti, C., Bassis, S., Rozza, A., Lombardi, G., Casiraghi, E., and Campadelli, P. Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration. *Pattern Recognition*, 47(8):2569–2581, 2014.

Claussen, J. C. and Villmann, T. Magnification control in winner relaxing neural gas. *Neurocomputing*, 63:125–137, 2005.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Facco, E., d'Errico, M., Rodriguez, A., and Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017.

Farahmand, A. M., Szepesvári, C., and Audibert, J.-Y. Manifold-adaptive dimension estimation. In *Proceedings of the 24th international conference on Machine learning*, pp. 265–272, 2007.

Gassberger, P. and Procaccia, I. Measuring the strangeness of the strange attractor. *Physica D*, 189, 1983.

Hino, H., Fujiki, J., Akaho, S., and Murata, N. Local intrinsic dimension estimation by generalized linear modeling. *Neural Computation*, 29(7):1838–1878, 2017.

Johnsson, K., Soneson, C., and Fontes, M. Low bias local intrinsic dimension estimation from expected simplex skewness. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):196–202, 2014.

Kingma, D. and Welling, M. Auto-encoding variational Bayes. *arXiv:1312.6114*, 2014.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Kleindessner, M. and Luxburg, U. Dimensionality estimation without distances. In *Artificial Intelligence and Statistics*, pp. 471–479, 2015.

Knop, S., Spurek, P., Tabor, J., Podolak, I., Mazur, M., and Jastrzebski, S. Cramer-wold auto-encoder. *Journal of Machine Learning Research*, 21, 2020.

Levina, E. and Bickel, P. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17:777–784, 2004.

Li, C., Farkhoor, H., Liu, R., and Yosinski, J. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

Li, G. Robust regression. *Exploring data tables, trends, and shapes*, 281:U340, 1985.

Opitz, D. and Maclin, R. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2020.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538. PMLR, 2015.

Ross, B. L. and Cresswell, J. C. Tractable Density Estimation on Learned Manifolds with Conformal Embedding Flows. *NeurIPS*, 2021.

Rozza, A., Lombardi, G., Ceruti, C., Casiraghi, E., and Campadelli, P. Novel high intrinsic dimensionality estimators. *Machine learning*, 89(1):37–65, 2012.

Rubenstein, P. K., Schoelkopf, B., and Tolstikhin, I. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*, 2018.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Struski, Ł., Tabor, J., and Spurek, P. Lossy compression approach to subspace clustering. *Information Sciences*, 435:161–183, 2018.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 2013.

Supplementary material

A. Proofs

Proof of Lemma 2.1. Assume that δ is sufficiently small for F_x to be defined on $B(x,\delta)\cap T_xS$. Let $v\in B(x,\delta)\cap T_xS$. Under our identifications, $y=(v,F_x(v))\in T_xS\oplus N_xS$ is a point of S such that $v=\pi_x(y)$. Moreover, by eq. (5), for sufficiently small δ

$$\|v\|^2_{T_xS}+\|F_x(v)\|^2_{N_xS}\leq\delta^2(1+C\delta^2)<\delta, \tag{11}$$

so $y\in B(x,\delta^{1/2})$, and $v\in\pi_x(S\cap B(x,\delta^{1/2}))$. □

Proof of Lemma 2.2. Denote $B=S\cap B(x,\delta^{1/2})$. Integrating by substitution, we obtain

$$\int_B\phi_\delta^{T_xS}(\pi_x(y))\,dp_S=\int_{\pi_x(B)}\phi_\delta^{T_xS}(v)\,d(\pi_x)_*p_S, \tag{12}$$

where the pushforward $(\pi_x)_*p_S$ is a smooth measure on T_xS. Hence, for sufficiently small δ

$$\int_{\pi_x(B)}\phi_\delta^{T_xS}(v)\,d(\pi_x)_*p_S(v)\asymp\int_{\pi_x(B)}\phi_\delta^{T_xS}(v)\,dv \tag{13}$$

uniformly in δ. The integral on the right is at most 1, and simultaneously, by Lemma 2.1 we have

$$\int_{\pi_x(B)}\phi_\delta^{T_xS}(v)\,dv\geq\int_{B(x,\delta)\cap T_xS}\phi_\delta^{T_xS}(v)\,dv. \tag{14}$$

The last integral is the probability that a normal random variable falls within one standard deviation from the mean, which is a constant independent of δ. □

Proof of Lemma 2.3. By eq. (1),

$$\phi_\delta^{N_xS}(\pi_x^\perp(y))=\delta^{d-D}\phi^{N_xS}(\delta^{-1}\pi_x^\perp(y)). \tag{15}$$

Since π_x is a contraction, we have $\|\pi_x(y)\|\leq\delta^{1/2}$. It follows from eq. (5), that for sufficiently small δ

$$\|\pi_x^\perp(y)\|=\|F_x(\pi_x(y))\|\leq C\delta \tag{16}$$

for some $C>0$. Therefore $\delta^{-1}\pi_x^\perp(y)$ lies inside a fixed ball independent of δ, and in consequence

$$\phi^{N_xS}(\delta^{-1}\pi_x^\perp(y))\asymp 1 \tag{17}$$

uniformly in δ, concluding the proof. □

Proof of Lemma 2.4. Denote $B=S\cap B(x,\delta^{1/2})$. By eq. (4) and Lemma 2.3, for sufficiently small δ and $y\in B$, we have

$$\phi_\delta^D(x-y)\asymp\delta^{d-D}\phi_\delta^{T_xS}(\pi_x(y)) \tag{18}$$

uniformly in δ. It follows that the original integral can be estimated as

$$\int_B\phi_\delta^D(x-y)\,dp_S\asymp\delta^{d-D}\int_B\phi_\delta^{T_xS}(\pi_x(y))\,dp_S. \tag{19}$$

The proof concludes by applying Lemma 2.2 to the last integral. □

Proof of Lemma 2.5. Observe that if $\|v\|\geq\delta^{1/2}$, we have

$$\phi_\delta^D(v)\asymp\delta^{-D}\exp\left(-\frac{\|v\|^2}{2\delta^2}\right)\leq\delta^{-D}\exp\left(-\frac{1}{2\delta}\right) \tag{20}$$

uniformly in v. This bound on the integrand converges to 0 as $\delta\to 0$, and the measure p_S is finite, proving the convergence of the considered integral. □

B. Non-connected data manifolds and intersections

Earlier we assumed that the data comes from a connected manifold M, whose local dimension is constant. Moreover, it was embedded in $\mathbb{R}^D$, precluding self-intersections. These restrictions can be relaxed as follows. Firstly, we may allow M to contain multiple connected components. Secondly, instead of an embedding, we may consider a *good* immersion $j\colon M \to \mathbb{R}^D$, satisfying the following finiteness condition.

Definition B.1. We will call an immersion $j\colon M \to N$ *good*, if M admits an open cover $\mathcal{C}$ such that for every $U \in \mathcal{C}$ the restriction of j to U is an embedding, and moreover, every $x \in N$ has an open neighborhood whose preimage intersects only finitely many sets in $\mathcal{C}$.

In the non-connected case, the dimension is no longer constant on the manifold, but can differ between its components. We will denote by $\dim_x M$ the dimension of M at a point $x \in M$.

Before we proceed, we will prove a simple technical lemma.

Lemma B.1. *Let $j\colon M \to N$ be a good immersion. Then every $x \in N$ has a neighborhood whose preimage intersects only finitely many connected components of M.*

Proof. Let $\mathcal{C}$ be an open cover of M satisfying conditions of Definition B.1. Take $x \in N$, and let $V \subset N$ be a neighborhood of x such that $j^{-1}(V)$ intersects only finitely many sets $U_1, \ldots, U_n \in \mathcal{C}$. On each U_i the restriction of j is an embedding, so there exists a neighborhood $V_i \subset V$ of x whose preimage is contained in a single connected component of U_i, and hence in a single connected component M_i of M.

The intersection $\bigcap_i V_i$ is the required neighborhood of x, as its preimage is contained in the finite union of connected components $\bigcup_i M_i$. □

This more general case reduces to the one studied in Section 2.2, as the following reasoning shows.

Proposition B.1. *Suppose $j\colon M \to N$ is an immersion of manifolds. Moreover, let μ be a smooth measure on M. Then there exists a manifold $\tilde{M}$ endowed with a measure $\tilde{\mu}$ and a local diffeomorphism $f\colon \tilde{M} \to M$, such that*

1. *the measure $\tilde{\mu}$ is smooth*
2. *the pushforward $f_*\tilde{\mu}$ equals μ;*
3. *$\tilde{\jmath} = j \circ f\colon \tilde{M} \to N$ restricted to every connected component of $\tilde{M}$ is an embedding.*
4. *if j is good, then so is $\tilde{\jmath}$;*

Proof. Since j is an immersion, there exists an open cover $\mathcal{C}$ of M such that on every $U \in \mathcal{C}$ the restriction of j is an embedding. Let $\{\psi_U : U \in \mathcal{C}\}$ be a partition of unity subordinate to $\mathcal{C}$. Denote $M_U = \{x \in M : \phi_U(x) > 0\}$, and let $f_U\colon M_U \to M$ be the corresponding inclusion map. Finally, let $\tilde{M}$ be the disjoint union of $\{M_U : U \in \mathcal{C}\}$, and define $f\colon \tilde{M} \to M$ by gluing together the inclusions f_U.

The measure $\tilde{\mu}$ can be defined as

$$\tilde{\mu} = \sum_{U \in \mathcal{C}} (f_U^{-1})_*(\psi_U \mu), \tag{21}$$

i.e. for every $U \in \mathcal{C}$ we multiply μ by density function ψ_U, restrict it to M_U and pull it to $\tilde{M}$ through f_U. Since by definition ψ_U is continuous and positive on M_U, the measure $\tilde{\mu}$ is smooth. Moreover, by construction we have $f_*\tilde{\mu} = \mu$, and the restriction of $\tilde{\jmath}$ to every M_U (and therefore every to every connected component) is an embedding.

To show the last assertion, assume that j is good. In this case, the cover $\mathcal{C}$ defined above can be chosen in such a way that for every $x \in N$ there exists a neighborhood $V \subset N$ whose preimage $j^{-1}(V)$ intersects only finitely many sets in $\mathcal{C}$. It is then easy to see that the cover $\{M_U : U \in \mathcal{C}\}$ of $\tilde{M}$ satisfies the conditions of Definition B.1, so $\tilde{\jmath}$ is good. □

Now, suppose that in Theorem 2.1, instead of an embedded submanifold S, we are dealing with the image of a proper immersion $j\colon M \to \mathbb{R}^D$, and that p_S is the pushforward of a probability measure μ on M. Thanks to Proposition B.1, this reduces to the situation where j restricted to every connected component of M is an embedding.

Proposition B.2. *Suppose $j\colon M \to \mathbb{R}^D$ is a good immersion, and its restriction to every connected component of M is an embedding. Let μ be a smooth probability measure on M, and $p_S = j_*\mu$. For $x \in S = j(M)$ and sufficiently small δ we have*

$$\log \rho_\delta(x) = (d - D)\log\delta + O(1), \tag{22}$$

where

$$d = \min_{j(y)=x} \dim_y M. \tag{23}$$

Proof. By Lemma B.1, for sufficiently small r the preimage $j^{-1}(B)$ of the ball $B = B(x, r)$ centered at x intersects only finitely many connected components of M. Denote them by $M_1, \dots, M_k$, and let M_0 be the union of the remaining components. The measure μ can be decomposed as

$$\mu = \sum_{i=0}^{k} \mu(M_i)\mu_i, \tag{24}$$

where μ_i is the restriction of μ to M_i, normalized to a probability measure. If we put $p_i = j_*\mu_i$, a similar decomposition holds for p_S.

If we apply Theorem 2.1 to $j(M_i)$ endowed with the measure p_i, for $i > 0$, the corresponding perturbed density ρ^i_δ satisfies

$$\rho^i_\delta(x) \asymp \delta^{\dim M_i - D} \tag{25}$$

for sufficiently small δ. Moreover, for $\delta < r^2$, we have $j(M_0) = j(M_0) \setminus B(x, \delta^{1/2})$, so by Lemma 2.5

$$\lim_{\delta\to 0^+} \rho^0_\delta(x) = 0. \tag{26}$$

Consequently, for small $\delta < 1$

$$\rho_\delta(x) = \sum_{i=0}^{k} \mu(M_i)\rho^i_\delta(x) \asymp \sum_{i=1}^{k} \delta^{\dim M_i - D}, \tag{27}$$

and the term with the lowest exponent dominates. □

C. Examples with explicit derivations

C.1. A motivating example

Consider the standard embedding $\mathbb{R}^d \subset \mathbb{R}^D$. Take for S a bounded open subset of $\mathbb{R}^d$, endowed with the uniform probability measure p_S with constant density $\rho \equiv \mathrm{vol}(S)^{-1}$ on S. If we denote by x_1 and x_2 the components of a vector $x \in \mathbb{R}^D$ corresponding to the standard decomposition $\mathbb{R}^D = \mathbb{R}^d \times \mathbb{R}^{D-d}$, it follows from (2) and properties of the Gaussian function, that

$$\rho_\delta(x) = \frac{\phi^{D-d}_\delta(x_2)}{\mathrm{vol}(S)} \int_S \phi^d_\delta(x_1 - y_1)\, dy_1. \tag{28}$$

Now, if x is an interior point of S, then $x_2 = 0$. Moreover, for sufficiently small δ, the integral above is arbitrarily close to 1, as most of the mass of the integrand falls into a small neighborhood of x_1, which is contained in S. Therefore, for sufficiently small δ

$$\rho_\delta(x) \asymp \phi^{D-d}_\delta(0) = \delta^{d-D}\phi^{D-d}(0) \asymp \delta^{d-D} \tag{29}$$

uniformly in δ.

It follows that

$$\log \rho_\delta(x) = (d - D)\log\delta + O(1), \tag{30}$$

and hence

$$d - D = \lim_{\delta\to 0} \frac{\log \rho_\delta(x)}{\log\delta}. \tag{31}$$

In practice, $d - D$, and in consequence d, can be estimated by considering $\rho_\delta(x)$ for multiple small values of δ, and using linear regression.

C.2. Normal distribution in $\mathbb{R}^D$

Suppose that $S = \mathbb{R}^D$, and $p_S = \mathcal{N}(0, \Sigma)$, where Σ is a diagonal matrix with entries $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_D^2$. In this case, the perturbation with $\mathcal{N}(0, \delta^2 I)$ yields another normal distribution $\mathcal{N}(0, \Sigma + \delta^2 I)$, whose density at 0 is

$$\rho_\delta(0) = (2\pi)^{-D/2} \prod_{k=1}^{D} (\sigma_k^2 + \delta^2)^{-1/2}. \tag{32}$$

Proposition C.1. *Let $1 \leq d < D$, and denote $\tau = (\sigma_d \sigma_{d+1})^{1/2}$. For $\lambda \geq 1$ and $\delta \in [\lambda^{-1}\tau, \lambda\tau]$ we have*

$$\log \rho_\delta(0) = (d - D) \log \delta + M - C_\lambda, \tag{33}$$

where M is independent of δ, and $0 \leq C_\lambda \leq \frac{D\sigma_{d+1}}{2\sigma_d}\lambda^2$.

In other words, the above proposition states that for δ between two consecutive deviations σ_d and σ_{d+1}, our LID estimate is approximately the number d of dimensions in which the Gaussian distribution is 'thicker' than δ, and the approximation error decreases with the growth of the ratio σ_d/σ_{d+1} and the distance of δ from σ_d and σ_{d+1}.

Proof. Let us denote $\eta = \lambda(\sigma_{d+1}/\sigma_d)^{1/2}$. The, for $k \leq d$ we may compute $\lambda\tau = \eta\sigma_d \leq \eta\sigma_k$, which leads to

$$\sigma_k^2 + \delta^2 \leq (1 + \eta^2)\sigma_k^2. \tag{34}$$

On the other hand, for $k \geq d+1$, we have $\lambda^{-1}\tau = \eta^{-1}\sigma_{d+1} \geq \eta^{-1}\sigma_k$, and similarly to the previous case, we have

$$\sigma_k^2 + \delta^2 \leq (1 + \eta^2)\delta^2. \tag{35}$$

By applying these two estimates to the formula (32) for $\rho_\delta(0)$ we are able to obtain a two-sided estimate

$$M(1 + \eta^2)^{-D/2}\delta^{d-D} \leq \rho_\delta(0) \leq M\delta^{d-D}, \tag{36}$$

with $M = (2\pi)^{-D/2} \prod_{k=1}^{d} \sigma_k^{-1}$ independent of δ. Finally, after taking $\log$ we can see that

$$\log \rho_\delta(0) = (d - D) \log \delta + \log M - \frac{D}{2} \log(1 + \eta^2), \tag{37}$$

and the last term is positive and bounded from above by $D\eta^2/2$, yielding the desired estimate by substituting η. □

From the above Proposition we can see that if there is a large gap between σ_d and σ_{d+1}, then for δ in the neighborhood of their geometric mean, the LID estimate obtained through linear regression should be approximately d, with approximation error decreasing, and the range of viable δ increasing with the growth of the gap size, expressed by the ratio σ_{d+1}/σ_d.

C.3. Points along a line

Consider a zero-dimensional manifold M, consisting of N points, endowed with uniform probability measure. Suppose M is embedded into $\mathbb{R}^D$ in such a way that its image $\{x_1, \ldots, x_N\}$ is actually contained in $\mathbb{R} \subset \mathbb{R}^D$, and has the form $x_k = (\xi_k, 0, \ldots, 0)$, where $\xi_{k+1} \geq \xi_k + \eta$ for some $\eta > 0$, i.e. the indexing is chosen in such a way that the points x_k are ordered along $\mathbb{R}$, and the distances between them are at least η.

In this setting, we will study the quantity $\rho_\delta(x_n)$ more closely, and attempt to understand its relationship with the perturbation magnitude for any δ, not just sufficiently small ones. We have

$$\rho_\delta(x_0) = \frac{1}{N} \sum_{k=1}^{N} \phi_\delta^D(x_n - x_k) = \frac{\phi_\delta^D(0)}{N} \left(1 + \sum_{\substack{k=1 \\ k \neq n}}^{N} \frac{\phi_\delta^D(x_n - x_k)}{\phi_\delta^D(0)} \right) = M\delta^{-D} \left(1 + \epsilon_\delta\right), \tag{38}$$

where $M = (N(2\pi)^{D/2})^{-1}$, and

$$\epsilon_\delta = \sum_{\substack{k=1 \\ k \neq n}}^{N} \frac{\phi_\delta^D(x_n - x_k)}{\phi_\delta^D(0)} = \sum_{\substack{k=1 \\ k \neq n}}^{N} \exp\left[-\frac{1}{2} \left(\frac{\xi_n - \xi_k}{\delta} \right)^2 \right]. \tag{39}$$

After taking $\log$, we get

$$\log \rho_\delta(x_0) = -D \log \delta + \log M + \log(1 + \epsilon_\delta), \tag{40}$$

where the term $\log M$ is independent of δ, and $0 \le \log(1 + \epsilon_\delta) \le \epsilon_\delta$.

Proposition C.2. *Let $\lambda \ge 1$. If $\delta < \eta/(\sqrt{2}\lambda)$ then $\epsilon_\delta \le 4e^{-\lambda^2}$. In particular, for $\epsilon > 0$, we have $\epsilon_\delta < \epsilon$ provided that*

$$\delta < \frac{\eta}{(-2\log(\epsilon/4))^{1/2}}, \tag{41}$$

i.e. the threshold value for δ depends logarithmically on ϵ.

Proof. We have $|\xi_i - \xi_j| \ge \eta\, |i - j|$, and therefore

$$\epsilon_\delta \le \sum_{\substack{k=1 \\ k \ne n}}^{N} \exp\left[-\frac{1}{2}\left(\frac{\eta(n-k)}{\delta}\right)^2\right] \le \sum_{\substack{k=1 \\ k \ne n}}^{N} e^{-\lambda^2 (n-k)^2}. \tag{42}$$

For an upper estimate, we may also extend the summation over all integers except n, obtaining

$$\epsilon_\delta \le \sum_{k \ne n} e^{-\lambda^2 (n-k)^2} = 2\sum_{j=1}^{\infty} e^{-\lambda^2 j^2} \le 2\sum_{j=1}^{\infty} e^{-\lambda^2 j} = \frac{2}{1 - e^{-\lambda^2}} e^{-\lambda^2}. \tag{43}$$

For $\lambda \ge 1$ we have $(1 - e^{-\lambda^2})^{-1} \le 2$, so in the end $\epsilon_\delta \le 4e^{-\lambda^2}$. By solving $\epsilon = 4e^{-\lambda^2}$ for λ we obtain $\lambda = (-\log(\epsilon/4))^{1/2}$, yielding the last assertion. □

D. Ideal LIDL for normal distribution on a line

Suppose our submanifold S is the image of the standard embedding $\mathbb{R} \subset \mathbb{R}^D$, and let $p_S = \mathcal{N}(0, 1)$. In this case, the perturbed distribution is $\mathcal{N}(0, \Sigma)$, where Σ is a diagonal matrix with entries $(1 + \delta^2, \delta^2, \ldots, \delta^2)$. The density ρ_δ at a point $x = (t, 0, \ldots, 0) \in S$ is therefore

$$\rho_\delta(x) = \frac{\delta^{1-D}}{(2\pi)^{D/2}(1+\delta^2)^{1/2}} \exp\left(-\frac{t^2}{2(1+\delta^2)}\right), \tag{44}$$

and its logarithm can be decomposed into the following sum

$$\log \rho_\delta(x) = (1 - D)\log \delta - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log(1+\delta^2) - \frac{t^2}{2(1+\delta^2)}. \tag{45}$$

Let us now apply the trivial case of linear regression involving only two points, amounting to computing the slope of the line passing through two points. We have

$$\frac{\log \rho_{\delta_1}(x) - \log \rho_{\delta_2}(x)}{\log \delta_1 - \log \delta_2} = 1 - D - \epsilon(t), \tag{46}$$

where the error term expands to

$$\epsilon(t) = \frac{1}{2(\log \delta_1 - \log \delta_2)}\left[\log \frac{1+\delta_1^2}{1+\delta_2^2} - t^2\left(\frac{1}{1+\delta_1^2} - \frac{1}{1+\delta_2^2}\right)\right], \tag{47}$$

yielding a LID estimate $\hat{d}_x = 1 - \epsilon(t)$ at x. We can see that the error ϵ decomposes into two terms of opposite signs. The first term depends only on δ, and the second one, grows quadratically with t.

If we put $\delta_1 = \eta\delta$, and $\delta_2 = \delta$, the coefficient of t^2 can be further rewritten as

$$\frac{1}{2(\log \delta_1 - \log \delta_2)}\left(\frac{1}{1+\delta_1^2} - \frac{1}{1+\delta_2^2}\right) = \frac{\delta^2(1-\eta^2)}{2\log \eta (1+\delta^2)(1+(\delta\eta)^2)} \asymp \frac{\delta^2(1-\eta^2)}{2\log \eta}, \tag{48}$$

where the estimate holds uniformly in δ if δ is bounded δ from above. Although for fixed δ and η the error is unbounded as a function of t, if we were allowed to adjust δ based on t (with fixed η), for the error $\epsilon(t)$ to be bounded in t it is necessary and sufficient that $\delta \leq C/t$ for some constant C.

Finally, the expected error for the LID estimate (computed in the above manner) at a random x drawn from our distribution can be computed

$$\begin{aligned}\int_{\mathbb{R}} \epsilon(t)\phi^1(t)\,dt &= \frac{1}{2(\log\delta_1 - \log\delta_2)}\left[\log\frac{1+\delta_1^2}{1+\delta_2^2} - \int_{\mathbb{R}} t^2\phi^1(t)\,dt\left(\frac{1}{1+\delta_1^2} - \frac{1}{1+\delta_2^2}\right)\right] = \\ &= \frac{1}{2(\log\delta_1 - \log\delta_2)}\left[\log\frac{1+\delta_1^2}{1+\delta_2^2} - \left(\frac{1}{1+\delta_1^2} - \frac{1}{1+\delta_2^2}\right)\right] = \epsilon(1),\end{aligned} \tag{49}$$

where the last integral is just the variance of $\mathcal{N}(0,1)$, i.e. 1.

E. Normalizing Flows

NF are very flexible tools for approximating probability distributions. They use parametrized nonlinear invertible transformation f_θ and change of variable formula to transform a simple density $\pi(z)$ into a more complicated one. NF are trained using gradient-based methods (e.g. SGD) to maximize log-likelihood of the data

$$\max_\theta \sum_i \log q(x_i)$$

where

$$q(x) = \pi(f_\theta(x))\left|\det\frac{\partial f_\theta(x)}{\partial x}\right|.$$

We used MAF (Papamakarios et al., 2017), RQ-NSF (Durkan et al., 2019) and Glow (Kingma & Dhariwal, 2018) models in our experiments. More detailed introduction to normalizing flows can be found in (Dinh et al., 2014).

Table 1: Relative bias of LID estimates. All algorithm names explained in Table 3

Distribution	LID	LIDL$_M$	LIDL$_R$	COR	ESS	KNN	LPC	MAD	MIN	MLE	MOM	TLE	TWO
Lollipop in $\mathbb{R}^2$	0	**0.00**	**0.00**	1.67	1.67	1.60	1.67	1.82	1.67	1.80	1.74	-	1.65
Lollipop in $\mathbb{R}^2$	1	**0.00**	**0.01**	**0.00**	**0.00**	0.58	**0.01**	0.10	**0.00**	**0.05**	**0.00**	-	**0.00**
Lollipop in $\mathbb{R}^2$	2	**-0.00**	**-0.00**	**-0.01**	**-0.00**	-0.07	**0.00**	0.10	**-0.00**	0.08	**0.01**	-	**-0.03**
$\mathcal{U}$ on helix in $\mathbb{R}^3$	1	**0.01**	**0.00**	**0.00**	**0.00**	0.68	**0.00**	0.12	**0.00**	0.06	**0.00**	**0.00**	**0.00**
$\mathcal{U}$ on $S^7 \subseteq \mathbb{R}^8$	7	**-0.00**	**0.00**	-0.28	**0.00**	-0.37	**0.00**	**0.03**	-0.18	**0.02**	**-0.04**	0.08	-0.13
Swiss roll in $\mathbb{R}^3$	2	**0.04**	**0.01**	**-0.00**	**0.00**	**-0.05**	**0.00**	0.06	**0.00**	0.05	**0.00**	0.05	**-0.01**
$\mathcal{N}_{10} \subseteq \mathbb{R}^{10}$	10	**0.00**	**0.00**	-0.40	**-0.00**	-0.47	**0.00**	**0.02**	-0.25	**0.01**	-0.07	**0.01**	-0.16
$\mathcal{N}_{100} \subseteq \mathbb{R}^{100}$	100	**-0.00**	**0.00**	-0.78	**-0.01**	-0.66	-0.28	-0.51	-0.90	-0.50	-0.57	-0.56	-0.60
$\mathcal{N}_{1000} \subseteq \mathbb{R}^{1000}$	1000	**0.00**	**0.00**	-0.93	-0.09	-0.74	-0.90	-0.83	-0.99	-0.82	-0.85	-0.85	-0.86
$\mathcal{N}_{4000} \subseteq \mathbb{R}^{4000}$	4000	**-0.00**	-	-0.96	-0.29	-0.77	-0.98	-0.91	-1.00	-0.91	-0.92	-0.93	-0.93
$\mathcal{N}_{10} \subseteq \mathbb{R}^{20}$	10	**0.00**	**0.01**	-0.40	**-0.00**	-0.25	**0.00**	**0.02**	-0.25	**0.01**	-0.07	**0.01**	-0.16
$\mathcal{N}_{100} \subseteq \mathbb{R}^{200}$	100	**0.04**	**0.03**	-0.78	**-0.01**	-0.46	-0.28	-0.51	-0.90	-0.50	-0.57	-0.56	-0.60
$\mathcal{N}_{1000} \subseteq \mathbb{R}^{2000}$	1000	0.11	0.30	-0.93	-0.09	-0.52	-0.90	-0.83	-0.99	-0.82	-0.85	-0.85	-0.86
$\mathcal{N}_{2000} \subseteq \mathbb{R}^{4000}$	2000	0.11	-	-0.95	-0.17	-0.55	-0.95	-0.88	-0.99	-0.87	-0.89	-0.90	-0.90
$\mathcal{U}_{10} \subseteq \mathbb{R}^{10}$	10	**-0.04**	**-0.04**	-0.39	-0.07	-0.47	**0.00**	-0.10	-0.29	-0.11	-0.17	-0.07	-0.22
$\mathcal{U}_{100} \subseteq \mathbb{R}^{100}$	100	**0.00**	**0.01**	-0.75	**-0.02**	-0.67	-0.28	-0.50	-0.90	-0.49	-0.56	-0.53	-0.59
$\mathcal{U}_{1000} \subseteq \mathbb{R}^{1000}$	1000	**-0.00**	**0.00**	-0.92	-0.09	-0.76	-0.90	-0.81	-0.99	-0.81	-0.84	-0.84	-0.85
$\mathcal{U}_{4000} \subseteq \mathbb{R}^{4000}$	4000	**-0.00**	-	-0.96	-0.29	-0.78	-0.98	-0.90	-1.00	-0.90	-0.92	-0.92	-0.92

Table 2: Relative MAE of LID estimates. All algorithm names explained in Table 3

Distribution	LID	LIDL$_M$	LIDL$_R$	COR	ESS	KNN	LPC	MAD	MIN	MLE	MOM	TLE	TWO
Lollipop in $\mathbb{R}^2$	0	**0.00**	**0.00**	1.67	1.67	1.60	1.67	1.82	1.67	1.80	1.74	-	1.65
Lollipop in $\mathbb{R}^2$	1	**0.00**	**0.01**	**0.00**	**0.00**	0.58	**0.01**	0.12	**0.00**	**0.05**	**0.00**	-	**0.00**
Lollipop in $\mathbb{R}^2$	2	**0.01**	**0.01**	**0.01**	**0.00**	0.62	**0.00**	0.43	**0.00**	0.25	**0.03**	-	**0.05**
$\mathcal{U}$ on helix in $\mathbb{R}^3$	1	**0.01**	**0.00**	**0.00**	**0.00**	0.68	**0.00**	0.14	**0.00**	0.06	**0.00**	**0.00**	**0.00**
$\mathcal{U}$ on $S^7 \subseteq \mathbb{R}^8$	7	**0.00**	**0.00**	0.28	**0.00**	0.44	**0.00**	0.27	0.18	0.18	0.08	0.17	0.15
Swiss roll in $\mathbb{R}^3$	2	0.06	**0.01**	**0.00**	**0.00**	0.37	**0.00**	0.25	**0.00**	0.14	**0.02**	0.06	**0.03**
$\mathcal{N}_{10} \subseteq \mathbb{R}^{10}$	10	**0.00**	**0.01**	0.40	**0.00**	0.47	**0.00**	0.27	0.25	0.19	0.11	0.16	0.17
$\mathcal{N}_{100} \subseteq \mathbb{R}^{100}$	100	**0.00**	**0.02**	0.78	**0.02**	0.66	0.28	0.52	0.90	0.50	0.57	0.56	0.60
$\mathcal{N}_{1000} \subseteq \mathbb{R}^{1000}$	1000	**0.00**	**0.01**	0.93	0.09	0.74	0.90	0.83	0.99	0.82	0.85	0.85	0.86
$\mathcal{N}_{4000} \subseteq \mathbb{R}^{4000}$	4000	**0.01**	-	0.96	0.29	0.77	0.98	0.91	1.00	0.91	0.92	0.93	0.93
$\mathcal{N}_{10} \subseteq \mathbb{R}^{20}$	10	**0.00**	**0.01**	0.40	**0.00**	0.68	**0.00**	0.27	0.25	0.19	0.11	0.16	0.17
$\mathcal{N}_{100} \subseteq \mathbb{R}^{200}$	100	**0.04**	**0.03**	0.78	**0.02**	0.87	0.28	0.52	0.90	0.50	0.57	0.56	0.60
$\mathcal{N}_{1000} \subseteq \mathbb{R}^{2000}$	1000	0.12	0.30	0.93	0.09	0.95	0.90	0.83	0.99	0.82	0.85	0.85	0.86
$\mathcal{N}_{2000} \subseteq \mathbb{R}^{4000}$	2000	0.12	-	0.95	0.17	0.96	0.95	0.88	0.99	0.87	0.89	0.90	0.90
$\mathcal{U}_{10} \subseteq \mathbb{R}^{10}$	10	**0.04**	**0.04**	0.39	0.07	0.47	**0.00**	0.27	0.29	0.20	0.18	0.17	0.22
$\mathcal{U}_{100} \subseteq \mathbb{R}^{100}$	100	**0.00**	**0.02**	0.75	**0.02**	0.67	0.28	0.50	0.90	0.49	0.56	0.53	0.59
$\mathcal{U}_{1000} \subseteq \mathbb{R}^{1000}$	1000	**0.00**	**0.02**	0.92	0.09	0.76	0.90	0.81	0.99	0.81	0.84	0.84	0.85
$\mathcal{U}_{4000} \subseteq \mathbb{R}^{4000}$	4000	**0.01**	-	0.96	0.29	0.78	0.98	0.90	1.00	0.90	0.92	0.92	0.92

Table 3: Algorithms used for comparison. All implementations from scikit-dimension library (Bac et al., 2021).

Name	Shortcut	citation
CorrInt	COR	(Gassberger & Procaccia, 1983)
MADA	MAD	(Farahmand et al., 2007)
MLE	MLE	(Levina & Bickel, 2004)
lPCA	LPC	(Cangelosi & Goriely, 2007)
KNN	KNN	(Carter et al., 2009)
DANCo	DAN	(Ceruti et al., 2014)
MiND_ML	MIN	(Rozza et al., 2012)
ESS	ESS	(Johnsson et al., 2014)
MOM	MOM	(Amsaleg et al., 2018)
FisherS	FIS	(Albergante et al., 2019)
TwoNN	TWO	(Facco et al., 2017)
TLE	TLE	(Amsaleg et al., 2019)

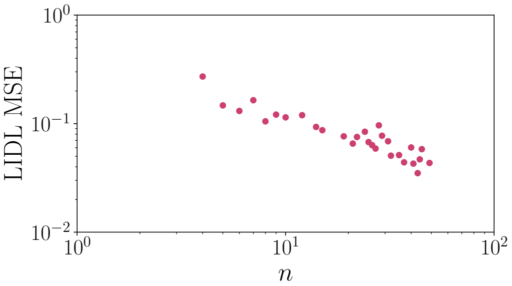

Figure 12: The dependence of mean-square error (MSE) on the number of models used in LIDL (n from Algorithm1). We can observe monothonic decrease of the estimate error with the increase of n.

F. Experimental details

In this section we present some results of additional experiments, some details and other observations.

When using LIDL with parametric density estimators on non-synthetic datasets, choosing hyperparameters is a challenge. We cannot directly estimate the error of the algorithm because we does not have access to ground truth LID. However, we observed in our experiments that choosing the hyperparameters leading to models minimizing negative log-likelihood on the validation set is a good strategy for minimizing the error of the LID estimate. We apply this approach in all our experiments; as density estimators we employ MAF (Papamakarios et al., 2017), RQ-NSF (Durkan et al., 2019) and Glow (Kingma & Dhariwal, 2018).

In scalability experiments we used 3 types of datasets. Uniform distribution on interval $(0, 1)$ on a hypercube (denoted by $\mathcal{U}_N$, where N is dimensionality of a cube), multivariate Gaussian ($\mathcal{N}_N \subseteq \mathbb{R}^N$) where N is dimensionality of a distribution and data space, and ($\mathcal{N}_N \subseteq \mathbb{R}^{2N}$), where we embedded N-dimensional Gaussian in $2N$-dimensional space by duplicating each coordinate. In each experiment we used 11 δs between 0.025 and 0.1.

F.1. Reducing the error of the density estimate

Because model ensemble methods (Opitz & Maclin, 1999) often reduces prediction error in many machine learning models, and most of LIDL error comes from the imperfect density estimators, we applied it to our problem by increasing the number of models n used in LIDL. We were able to reduce an error of each estimate by simply adding more models between the same range of δs. An example of this behavior for 10-dimensional Gaussian embedded in 20-dimensional space is plotted in Fig.12 in Appendix F.

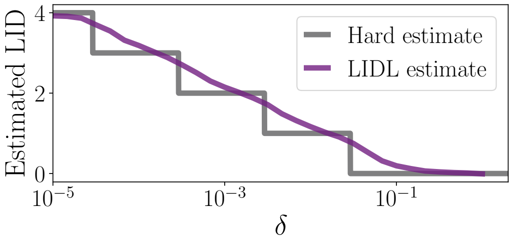

Figure 13: LIDL and hard estimate for different values of δ for 4-dimensional multiscale uniform distribution. We can see that LIDL ignores dimensions that are much smaller than δ even with imperfect density estimators.

F.2. Image datasets

We present cumulative distribution function (CDF) for MNIST and FMNIST in Fig. 14 and 15. More samples from MNIST, FMNIST, Celeb-A sorted by their LID can be found in Fig. 20, 21, and 22.

We can observe, that dimensionality estimates obtained from LIDL on MNIST are higher than those reported in (Pope et al., 2020) or (Kleindessner & Luxburg, 2015) and obviously depend on choosing the range of δs. We want to present some

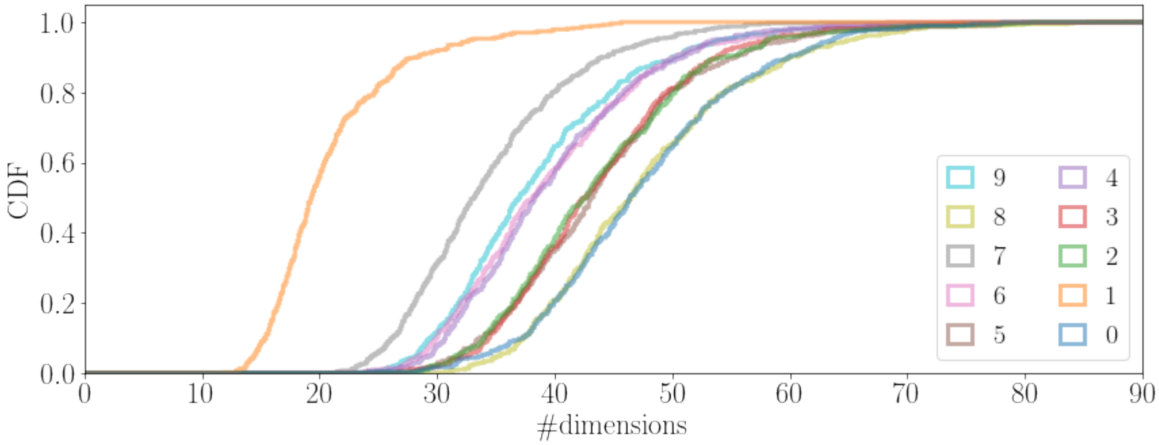

Figure 14: Empirical CDF of 5000 examples from MNIST dataset. Each line represents CDF for separate class in the dataset. Class number (which also is a represented digit in this case) can be found in the legend.

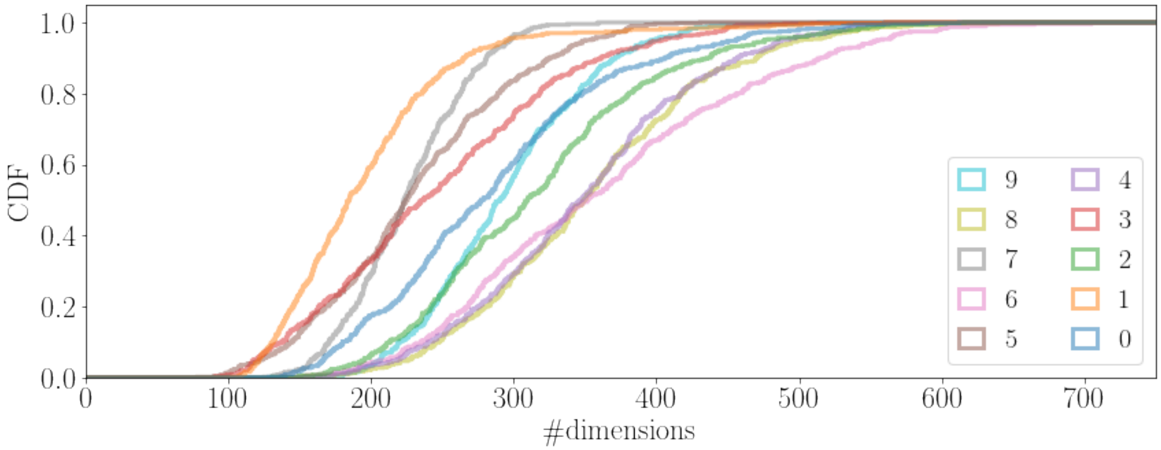

Figure 15: Empirical CDF of 5000 examples from FMNIST dataset. Each line represents CDF for separate class in the dataset. Class number can be found in the legend.

argument for our estimates: (Cavallari et al., 2018) used auto-encoder representation of MNIST as an input to SVN digit classifier and they achieved the best classification results for an auto-encoder with latent space size greater than 100. This means that we need more than 100 dimensions to encode an average MNIST digit to preserve all the information about it. Of course the compression with autoencoder is not ideal, but we can argue that true ID lies somewhere between.

Relation between examples for different range of δs We observed that for image dataset LID estimates for two disjoint sets of 4 δs have similar ranks (they on average differ between 10%-15%), and relations between points in each set (i.e. if LID estimate for x_j is lower than LID estimate for x_i) are preserved in 80-90% of cases. This of course depends strongly on δ range and the dataset.

LID estimate dependence on δ and effect of dequantization We present MNIST and FMNIST LID estimates (averaged per class) dependence on δ in Fig. 18 and 19. Images present wide range of δs (from 10^{-4} to 10^{1}) for original datasets and datasets with dequantization used after training. Black dashed line indicates a theoretical δ, above which LIDL should not calculate dequantization dimensions into LIDL estimate. This is 10 times standard deviation of dequantization noise $\mathcal{U}(0, 1/255)$. We can see that slightly above this threshold estimates for quantized and dequantized datasets aligh with each other. We can also observe, that for dequantized datasets and very small δ LID estimate is close to the dimensionality of the space, and for very big δs, LID estimates are close to 0 as expected.

F.3. LID and model performance

In this section, we show, that LID estimates are connected with model behavior on some benchmark datasets for autoencoder and classification deep neural networks. Our result suggests, that the connection between LID and model performance is significant, so LIDL estimates can potentially be used in problems like semi-supervised learning, active learning, uncertainty estimation, and curriculum learning.

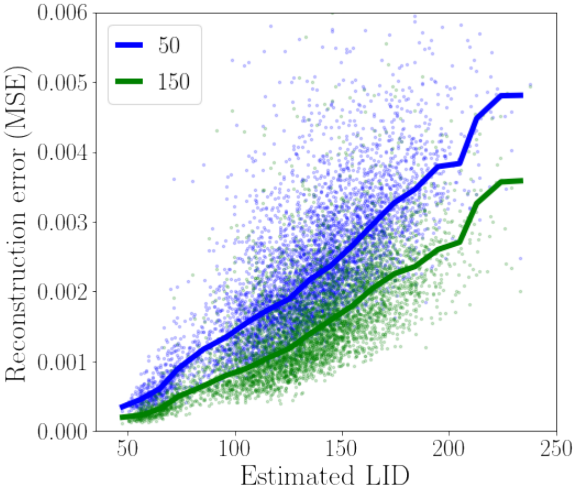

Figure 16: LIDL estimates and VAE MSE scatterplot for a sample from MNIST dataset. Lines are running medians for those point clouds. Values in legend are VAE latent space size.

Reconstruction error vs LID for autoencoders Results from the previous section led us to next experiment, where we wanted to investigate if the estimate from LIDL is correlated with reconstruction error for the image in VAE (Kingma & Welling, 2014). We trained VAE on MNIST with latent space sizes 50 and 150, and observed that there is a high correlation (Pearsons $R > 0.7$ in both cases) between MSE and LID. We plotted LIDL estimates against MSE for 5K images in Fig. 16. We can see an almost linear relationship between those quantities.

LID and classification accuracy We observed that classifiers can achieve better accuracy on data points with lower LID estimates. We trained neural networks on a subset of MNIST (300 images) and FMNIST (50K images) datasets and noticed a negative correlation of LID and an accuracy on a test set. Results are presented in Fig 17. What is more, we observed similar behavior inside a majority of classes (9 classes for MNIST, and 8 for FMNIST).

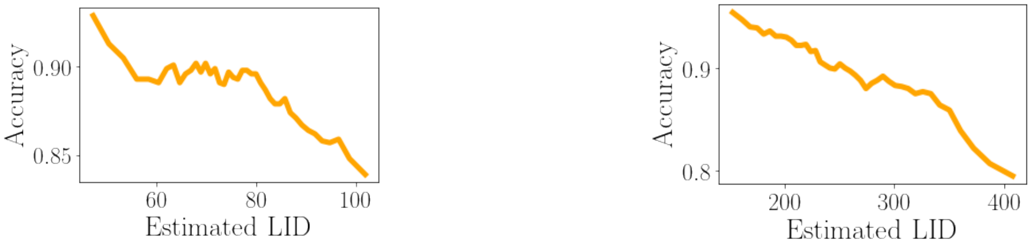

Figure 17: Average classifier accuracy per LID value on MNIST(left) and FMNIST(right) datasets. We can see a very strong negative correlation between those values.

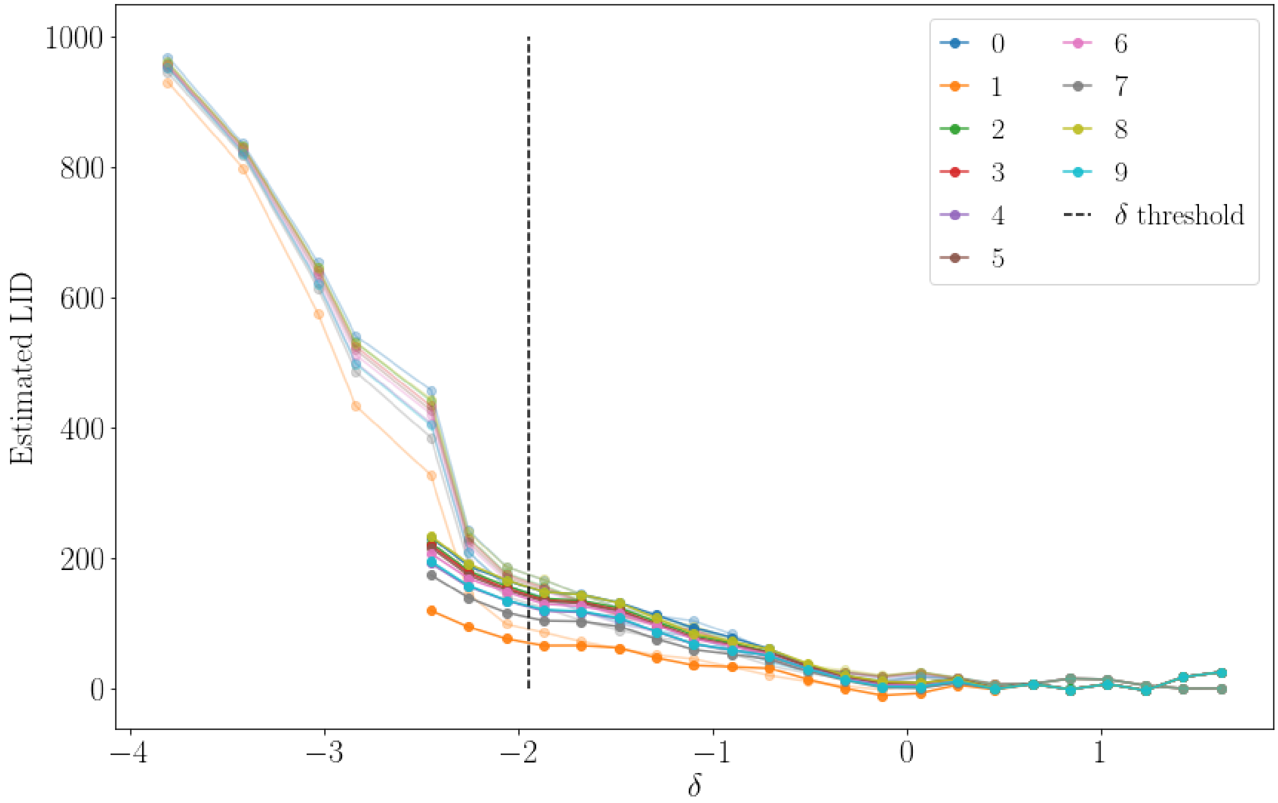

Figure 18: MNIST average LID estimates for each class for quantized (strong color) and dequantized (faded colors) as a function of δ.

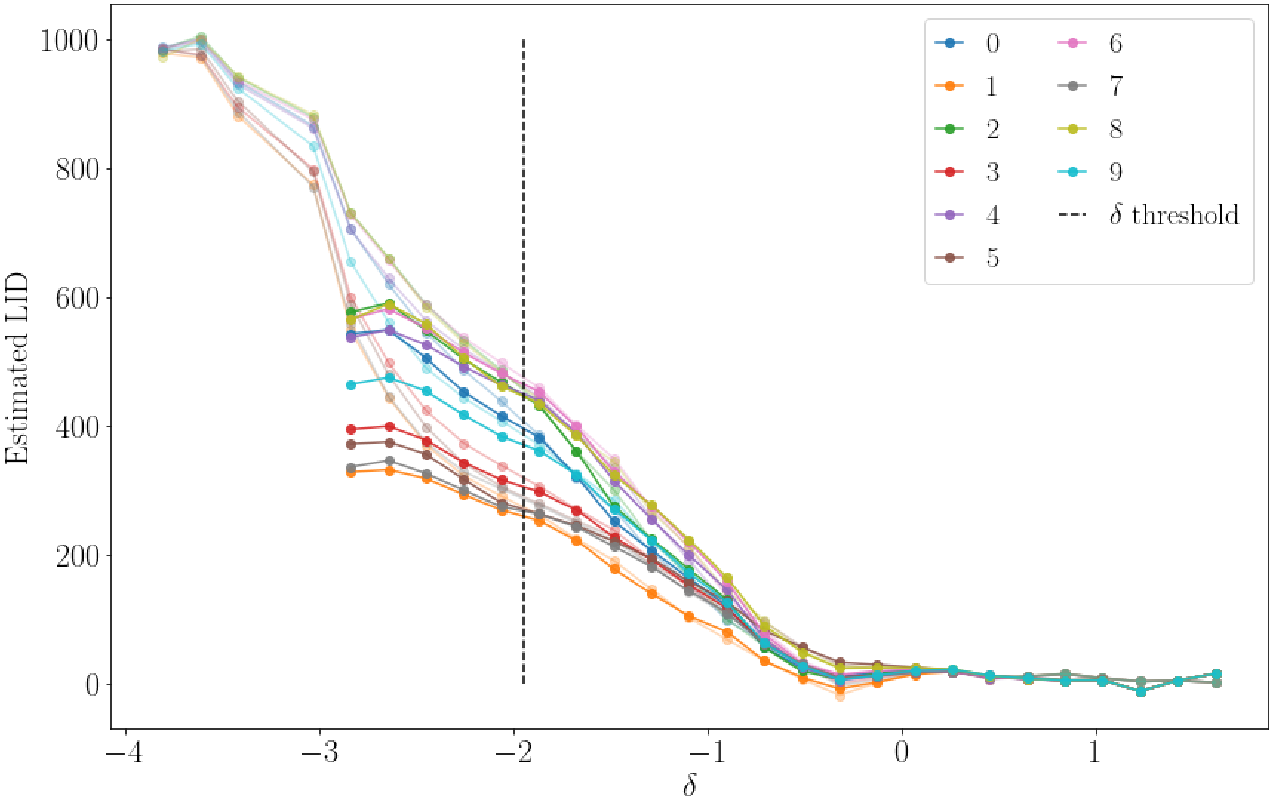

Figure 19: FMNIST average LID estimates for each class for quantized (strong color) and dequantized (faded colors) as a function of δ.

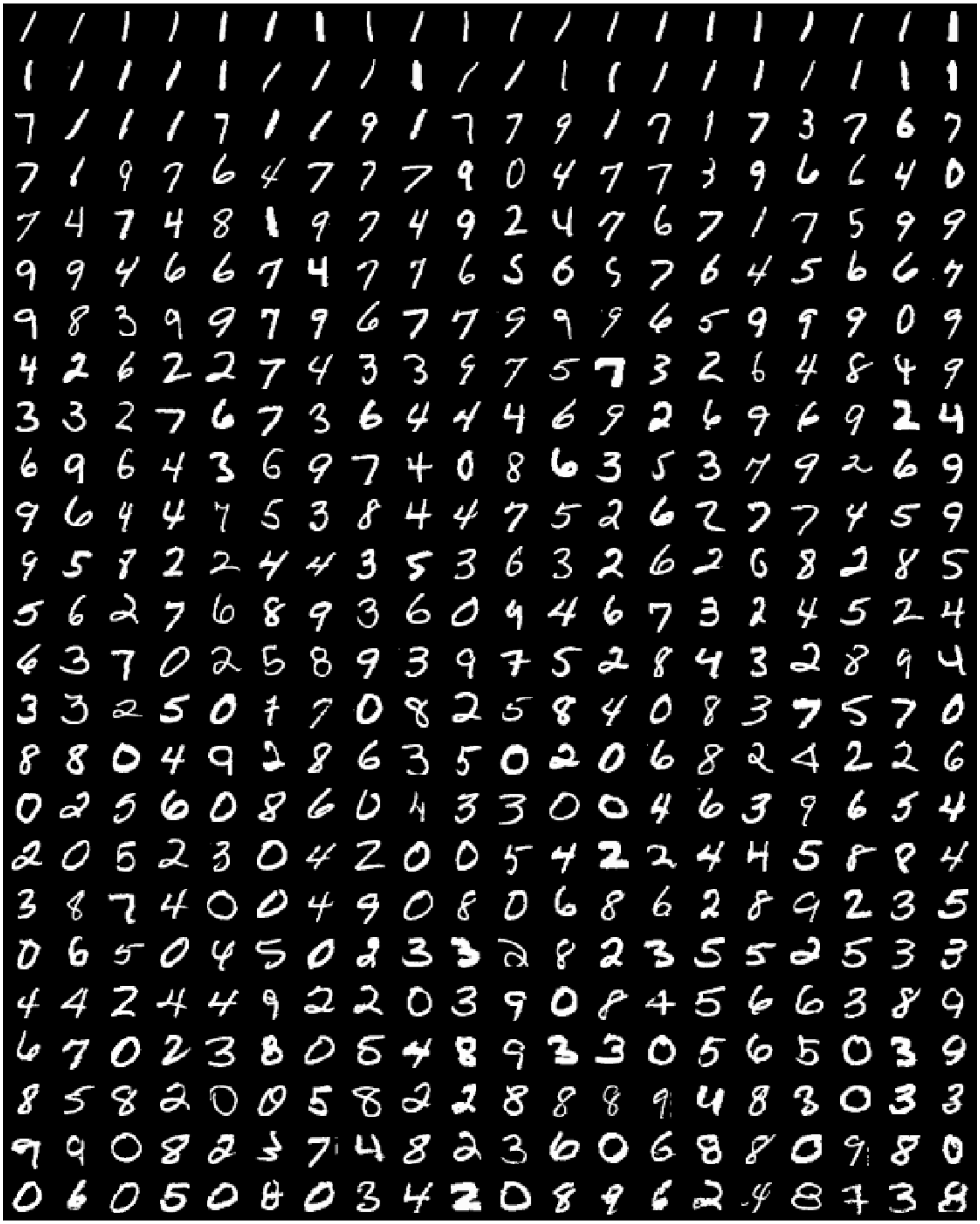

Figure 20: Samples from MNIST sorted by their LID estimates.

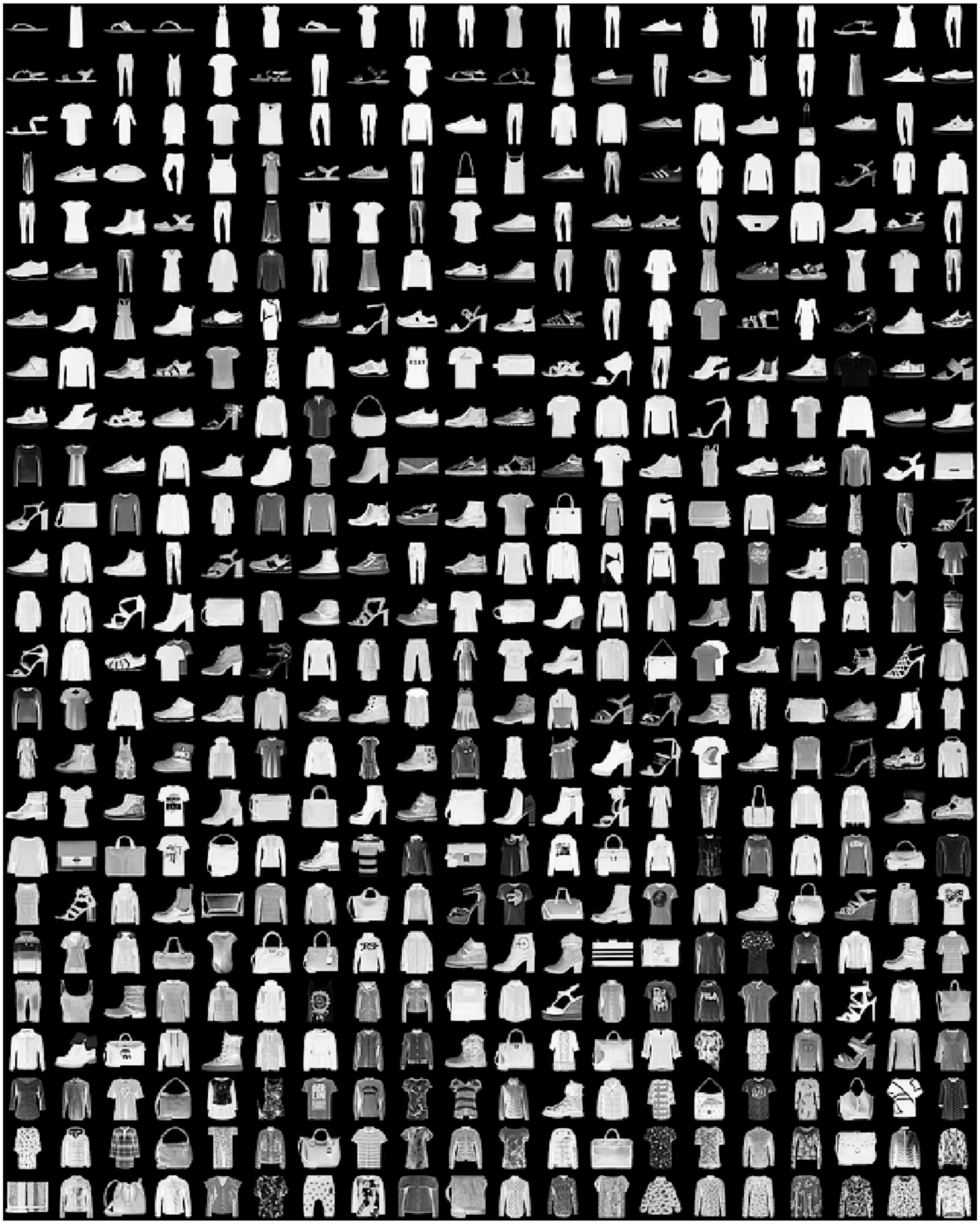

Figure 21: Samples from FMNIST sorted by their LID estimates.

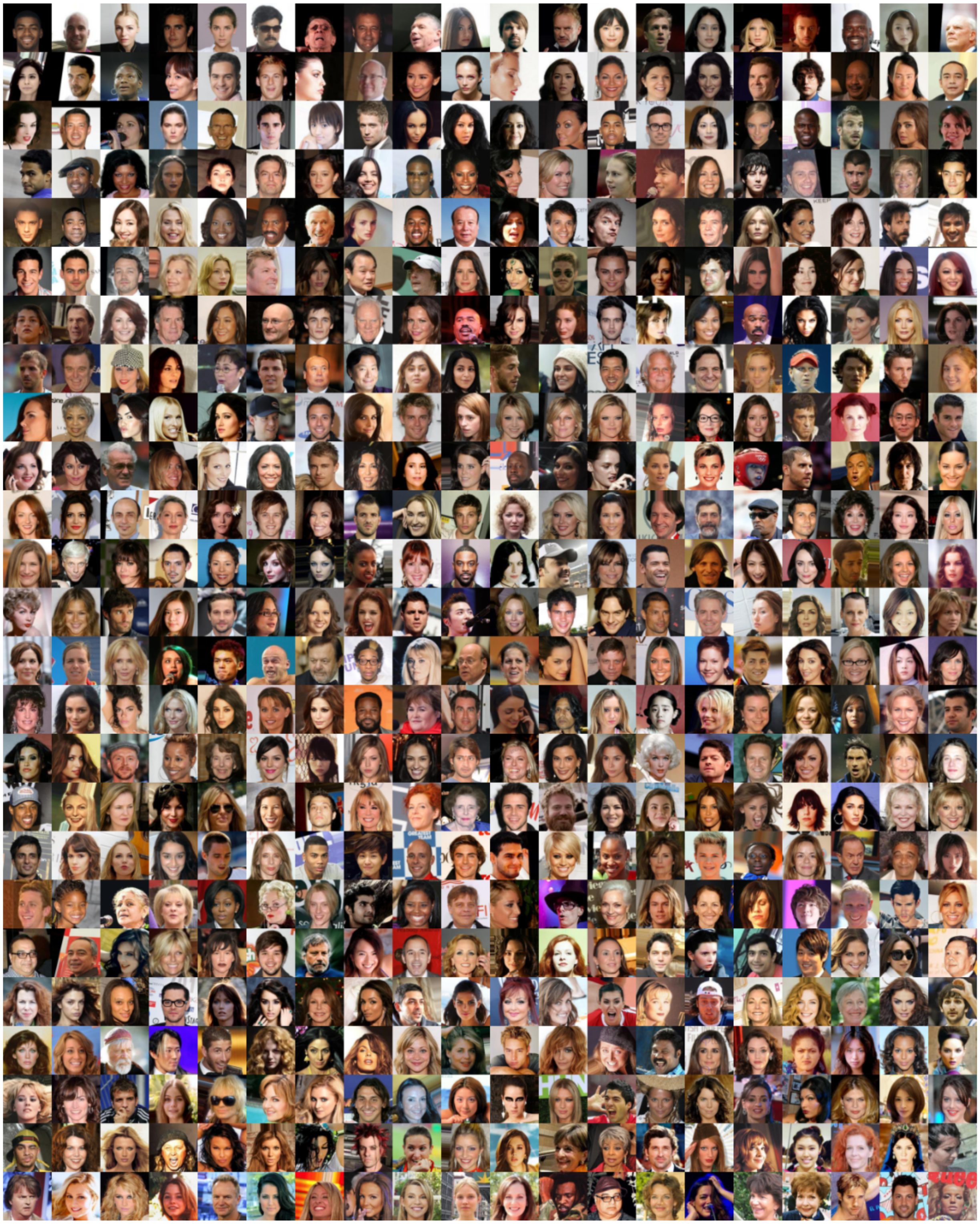

Figure 22: Samples from Celeb-A sorted by their LID estimates.