A Wiener Process Perspective on Local Intrinsic Dimension Estimation Methods

Piotr Tempczyk[*1,2,3], **Łukasz Garncarek**[3,4], **Dominik Filipiak**[5,6,7], **Adam Kurpisz**[*3,8,9]
[1]Institute of Informatics, University of Warsaw, Poland
[2]NASK National Research Institute, Warsaw, Poland
[3]PL4AI, Poland
[4]Snowflake, USA
[5]Adam Mickiewicz University, Poznań, Poland
[6]Perelyn, Warsaw, Poland
[7]University of Innsbruck, Austria
[8]BFH Bern Business School, Switzerland
[9]ETH Zurich, Switzerland
tempczyk.piotr@gmail.com, lukgar@gmail.com, df@amu.edu.pl, adam.kurpisz@ifor.math.ethz.ch

Abstract

Local intrinsic dimension (LID) estimation methods have received a lot of attention in recent years thanks to the progress in deep neural networks and generative modeling. In opposition to old non-parametric methods, new methods use generative models to approximate diffused dataset density to scale the methods to high-dimensional datasets (e.g. images). In this paper, we investigate the recent state-of-the-art parametric LID estimation methods from the perspective of the Wiener process. We explore how these methods behave when their assumptions are not met. We give an extended mathematical description of those methods and their error as a function of the probability density of the data.

Extended version of paper — arxiv.org/abs/2406.17125

1 Introduction

LID estimation has gained increasing attention in recent years as part of the fast-growing field of topological data analysis. It was able to progress from non-parametric to parametric methods thanks to the latest progress in the field of generative modeling. LID estimation methods are algorithms for estimating the manifold's dimension from which the data point x was sampled. In these methods, we assume that the data lie on the union of one or more manifolds that may be of different dimensions.

The estimation of intrinsic dimension is substantial for data analysis and machine learning (Ansuini et al. 2019; Li et al. 2018; Rubenstein, Schoelkopf, and Tolstikhin 2018) and was investigated in relation to dimensionality reduction and clustering (Vapnik 2013; Kleindessner and Luxburg 2015; Camastra and Staiano 2016), analyzing the training and representation learning processes within deep neural networks (Li et al. 2018; Ansuini et al. 2019; Pope et al. 2020; Loaiza-Ganem et al. 2024), verifying the union of manifolds hypothesis for images (Brown et al. 2022), and used to improve out-of-distribution detection algorithm (Kamkari et al. 2024a).

Prior to introducing LIDL (Tempczyk et al. 2022), two approaches to solve the intrinsic dimension estimation problem were presented. The first employs global methods; see, e.g., Fukunaga and Olsen (1971); Minka (2000); Fan et al. (2010). These methods are known to suffer from issues related to a manifold curvature and non-uniformity of the data distribution. They also assume that data lies on a single manifold of constant dimension, so dimensionality is the same for all x. The second approach is based on local non-parametric methods that explore the geometric properties of neighborhoods (Johnsson, Soneson, and Fontes 2014; Levina and Bickel 2004) calculating some statistics using points from the neighborhood of x. Although all aforementioned methods perform reasonably well for a small number of dimensions, the higher dimensionality negatively affects their performance (Tempczyk et al. 2022; Campadelli et al. 2015; Camastra and Staiano 2016).

Very recently, a new approach emerged for the local methods, which is based on a two-step procedure. In the first step, the dataset is perturbed with some noise, and in the second step, its dimensionality is estimated using various techniques. These include generative models to analyze changes in density (Tempczyk et al. 2022; Kamkari et al. 2024b) or in singular values of the Jacobian (Horvat and Pfister 2022, 2024) for different noise magnitudes, and analyzing rank of scores from diffusion model as recently presented by Stanczuk et al. (2024).

While new algorithms provide state-of-the-art results for large, real-life datasets, it is vital for their performance that adding noise to the dataset should be as close as possible to an injective process. This means that the resulting densities after adding noise should uniquely identify the original distribution of the dataset. Otherwise, it is impossible to neither uniquely reverse the process nor analyze the original structure of the dataset. One example in the theoretical model that does not admit such a problem is a flat manifold with a uniform distribution of data points. In such cases these algorithms perform best, as shown in experimental results.

[*]These authors contributed equally.

In other cases, a hypothetical line of research would be to decrease the magnitude of noise added to the data set so that the manifold is approximately flat and has locally uniform data density in the scale of the noise magnitude. This, however, is not possible in practice for several reasons. For instance, data is often a set of discrete points (especially in audio and visual modality), and considering the noise of magnitude much smaller than the minimum distance between points does not lead to any meaningful process. Moreover, neural net training is done with finite precision and the stochastic gradient descent procedure introduces noise to the density estimates, so very small changes in density cannot be observed in practice, which may lead to poor quality estimates of LID.

In this paper, we point out and exploit the fact that adding Gaussian noise of varying magnitudes can be seen as studying the evolution of the Wiener process describing the diffusion of particles (points of the dataset) in the ambient space. This point of view enables us to employ Fick's Second Law of Diffusion to eliminate time derivatives from mathematical descriptions of state-of-the-art LID algorithms (Tempczyk et al. 2022; Kamkari et al. 2024b), and replace them with spatial derivatives. Such considerations can be taken into account in the second step of the considered algorithms, leading to more accurate results. We encourage reader to get familiar with Appendix A from the extended version of the paper (`arxiv.org/abs/2406.17125`), which contains important definitions and clarifications regarding this work.

Contribution

1. We recognize and define new categories (isolated and holistic algorithms) for the Wiener process-based parametric LID estimation algorithm family and categorize the existing algorithms accordingly.
2. We explore the first step of existing algorithms in the language of Wiener processes and calculate important cases of diffusion from lower-dimensional manifolds with non-uniform probability density into ambient space.
3. We derive closed-form expressions for important parameters used in two state-of-the-art isolated LID estimation algorithms as a function of on-manifold density and manifold dimensionality, which can be viewed as closed-form expressions for deviation from a flat manifold with uniform distribution case.

2 Related Work

The review of the non-parametric methods for local and global intrinsic dimension estimation can be found in the work of Campadelli et al. (2015), or Camastra and Staiano (2016). Tempczyk et al. (2022) compared these methods on bigger datasets in terms of their dimensionality.

Although we do not analyze non-parametric methods in this paper, it is worth mentioning a recent work on non-parametric LID estimation, in which Bailey, Houle, and Ma (2022) explore the connection of LID to other well-known measures for complexity: entropy and statistical divergences, and develop new analytical expressions for these quantities in terms of LID. Consequently, Bailey, Houle, and Ma (2023) establish the relationships for cumulative Shannon entropy, entropy power, Bregman formulation of cumulative Kullback-Leibler divergence, and generalized Tsallis entropy variants, and propose four new estimators of LID, based on nearest neighbor distances.

During the last few years many parametric methods for estimating LID emerged. Zheng et al. (2022) prove that VAE are capable of recovering the correct manifold dimension and demonstrate methods to learn manifolds of varying dimensions across the data sample. Yeats et al. (2023) connect the adversarial vulnerability of score models with the geometry of the underlying manifold they capture. They show that minimizing the Dirichlet energy of learned score maps boosts their robustness while revealing the LID of the underlying data manifold.

Wiener process-based algorithms. Regarding parametric methods, there is a group of algorithms, that have one thing in common: they simulate a Wiener process on a dataset and directly use some properties of time-evolving density to estimate LID. We can divide those algorithms into the following three groups:

1. **LIDL** (Tempczyk et al. 2022) and its efficient and most accurate implementation using diffusion models called **FLIPD** (Kamkari et al. 2024b). These algorithms use the rate of change of the probability density at point x during the Wiener process to estimate LID at x. For small diffusion times t the logarithm of a density is a linear function of a logarithm of t, and the proportionality coefficient is equal $d - D$ for small t, where d is manifold density and D is ambient space dimensionality. Our experiments with ID-NF (described below) and FLIPD show, that the latter is more scalable than ID-NF (and ID-DM) due to the high memory and computational complexity of SVD (which has to be calculated for each data point). Experiments from (Kamkari et al. 2024b) show that FLIPD is more accurate than NB (described below), which led us to the conclusion, that FLIPD is state-of-the-art in LID estimation among the most scalable algorithms.
2. **ID-NF** (Horvat and Pfister 2022) (using normalizing flows), and the diffusion-using follow-up paper **ID-DM** (Horvat and Pfister 2024) analyze how singular values of a Jacobian of a function transforming a standard normal distribution into a diffused dataset density at x evolves during the Wiener process. Horvat and Pfister observed that when transforming a d-dimensional manifold into a D-dimensional Gaussian we have to expand space more in the normal direction to manifold, especially for small diffusion times t.
3. **NB** (Stanczuk et al. 2024), which is an abbreviation from Normal Bundle (name used in (Kamkari et al. 2024b)). Stanczuk et al. observed that for small diffusion times t the gradients of the logarithm of a diffused data density (score function) close to x lies in the normal space to the manifold and use this fact to estimate LID at x.

3 Isolated and Holistic Algorithms

In this work, we take a closer look at the Wiener process-based algorithms described in the last paragraph of Sec-

tion 2. Although all these algorithms apply a Wiener process to the dataset during their first phase, when looking at their second step, we can divide them into two groups: isolated (LIDL, FLIPD, NB) and holistic (ID-NF, ID-DM). The intermediate results of the first group that are used for LID calculation use only the information about the local shape of the data probability density function (without normalization constant). We assume that the generative model approximates diffused data distribution ρ_t perfectly. Their estimates depend on the proximity of $\nabla \log \rho_t$ to x (NB) or $d \log \rho_t / d \log t$ at x (LIDL, FLIPD).

This is a consequence of ρ_t at x being a function that – in practice – depends on the values of original data distribution p_S in a ball of radius $r \approx 4\sqrt{t}$ around x in the data space. The values of p_S outside this ball do not matter in practice for isolated algorithms, because the diffused particles in the Wiener process can travel longer distances than r with very low probability (less than $3.7 \cdot 10^{-5}$). When we add some new data points to the dataset far away from x, it does not change the shape of ρ_t close to x. This operation only changes a normalization constant, which becomes 0 after taking a logarithm of ρ_t and taking a derivative either w.r.t time or spatial variables.

The holistic algorithms work quite differently. In the case of ID-NF and ID-DM, they calculate singular values of the Jacobian of the function $\zeta : \mathrm{R}^N \to \mathrm{R}^N$ transforming ρ_t into a standard normal distribution. This ζ function strongly depends on the entire shape of ρ_t. As mentioned before, when we change ρ_t far away from x we do not change the shape of ρ_t close to x. The same is not true for $\zeta(x)$. When we add many data points to the dataset far away from x while we keep the latent distribution fixed, we have to change the way we compress and stretch the space to match the new distribution. This property makes the analysis of the behavior of ID-NF and ID-DM much harder (maybe even impossible) if we want to take into account only the data density in the neighbourhood of x.

To give an illustrative example of this behavior, one needs to imagine that our dataset is one-dimensional and consists of 10K points sampled from $\mathcal{N}(0, 1)$. Typically, we are transforming it into $\mathcal{N}(0, 1)$, so ζ is just an identity function. Now let's add to the dataset another 10K points sampled from $\mathcal{N}(100, 1)$. As a consequence, we have to stretch our space in some areas to transform one density into another, whereas ζ changes along with its Jacobian.

4 Wiener Process Perspective on LID Estimation

Wiener process is a stochastic process modeling particle diffusion. Its increments over disjoint time intervals are independent and normally distributed, with variance proportional to time increments. Since in the machine learning community the term *diffusion* is already overloaded, we will stick to Wiener process when speaking of particle diffusion process.

In this section we present a new perspective on perturbing datasets, unifying the approaches seen in the algorithms presented by Tempczyk et al. (2022); Stanczuk et al. (2024); Horvat and Pfister (2022, 2024); Kamkari et al. (2024b). As already mentioned, all these algorithms consist of two stages, the first of which amounts to perturbing the dataset with normally distributed random noise of fixed variance t. In the second stage, each of the algorithms utilizes the behavior of the perturbed density in the neighborhood of a fixed point under changes in the noise variance.

The first phase of each algorithm can be interpreted as applying the Wiener process to the points in the dataset. Afterward, the resulting set of points is used to train some type of generative model (or models) to estimate the distribution of the dataset undergoing the Wiener process at time t. From the point of view of differential equations, the distribution density function of the diffused dataset is described by Fick's Second Law of Diffusion.

Fick's Second Law of Diffusion. *Let $\rho_t : \mathbb{R}^D \mapsto \mathbb{R}$ denote the probability density function modeling particles undergoing diffusion at time t. Then ρ_t satisfies the differential equation*

$$\frac{d}{dt}\rho_t = C\Delta\rho_t, \tag{1}$$

where $C \in \mathbb{R}$, and Δ stands for the standard Laplacian in $\mathbb{R}^D$.

Now, given a dataset embedded in $\mathbb{R}^D$, we assume that it has been drawn from some latent union of submanifolds S endowed with a probability measure p_S (which can be naturally treated as a probability measure on $\mathbb{R}^D$). The goal of Local Intrinsic Dimension estimation is to find out the dimension of S at any point of the dataset.

To model the Wiener process with initial distribution p_S (which is not a function on $\mathbb{R}^D$), let us first define

$$\phi_t^D(x) = (2\pi t)^{-D/2} e^{-\|x\|^2/2t}. \tag{2}$$

This is the density of normal distribution on $\mathbb{R}^D$ with covariance matrix tI. It is the fundamental solution of the differential equation given by Fick's Second Law of Diffusion (1) with $C = 1/2$. Here, this means that the convolution

$$\rho_t = p_S * \phi_t^D \tag{3}$$

is the solution of (1) for $t > 0$ and hence it describes the Wiener process starting from the initial probability distribution p_S.

To limit the complexity introduced by curvature, from now on we will consider only flat manifolds. This means that, without loss of generality, we may assume that S is the first factor in product decomposition $\mathbb{R}^D = \mathbb{R}^d \times \mathbb{R}^{D-d}$. We will denote the coordinates of $\mathbb{R}^d$ and $\mathbb{R}^{D-d}$ by x and y, respectively. We will moreover assume that p_S, now a probability distribution on $\mathbb{R}^d$, has a density $\psi \colon \mathbb{R}^d \to \mathbb{R}$.

In Appendix B we discuss the Laplacian of ρ_t and derive the following result:

Lemma 4.1 (Lemma B.1). *For $t > 0$ and $(x, y) \in \mathbb{R}^D$ we have*

$$\begin{aligned}\Delta\rho_t(x, y) = &\left(\frac{\|y\|^2}{t^2} + \frac{d - D}{t}\right)\rho_t(x, y) \\ &+ \phi_t^{D-d}(y)\Delta_x(\psi * \phi_t^d)(x).\end{aligned} \tag{4}$$

As a consequence, by putting $y = 0$ and using

$$\phi_t^{D-d}(0) = (2\pi t)^{(d-D)/2} \tag{5}$$

we obtain the following.

Corollary 4.2. *For $t > 0$ and $x \in \mathbb{R}^d$ we have*

$$\Delta\rho_t(x,0) = \frac{d-D}{t}\rho_t(x,0) + (2\pi t)^{(d-D)/2}\Delta_x(\psi * \phi_t^d)(x). \tag{6}$$

5 From Wiener Process to LID Estimation

The findings from the last section can be used to analyze how the first family of algorithms (Tempczyk et al. 2022; Kamkari et al. 2024b) behaves for some particular cases. The main contribution of Kamkari et al. (2024b) is a substantial improvement on the side of density estimation. Therefore, when dealing with perfect density estimators and very small noise differences, both algorithms estimate the same quantity and give the same results from the theoretical perspective. Due to this fact from now on we will be analyzing LIDL, as we want to analyze the aspects of those implementations that do not depend on the problems with density estimation itself.

Reformulating LIDL

Given a point $x \in S$ and a set of times $t_1, \ldots, t_n$, LIDL estimates the linear regression coefficient α of the set of points $(\log \delta_i, \log \rho_{t_i}(x))$, where $\delta_i = \sqrt{t_i}$. Tempczyk et al. (2022) proved that

$$\log \rho_t(x) = (d-D)\log\sqrt{t} + O(1), \tag{7}$$

and therefore $\alpha \approx d - D$. The authors show that if t is small enough, this estimate is accurate.

This procedure can be seen as approximating the asymptotic slope of the parametric curve $(\log\sqrt{t}, \log\rho_t(x))$. In other words, the graph of $s \mapsto \log \rho_{e^{2s}}(x)$ for $s \to -\infty$. Another approach would consider the its derivative. Let us define its reparameterized derivative (with $t = e^{2s}$)

$$\beta_t(x) = \frac{2t}{\rho_t(x)}\frac{d}{dt}\rho_t(x) = \frac{t\Delta\rho_t(x)}{\rho_t(x)}, \tag{8}$$

where the last equality comes from the diffusion equation (1) with $C = 1/2$. Moreover, denote the asymptotic slope of the aforementioned curve by

$$\beta(x) = \lim_{s\to-\infty}\frac{d}{ds}\log\rho_{e^{2s}}(x) = \lim_{t\to 0^+}\beta_t(x). \tag{9}$$

The results presented below are proved in Appendix C. The next Proposition shows that the two approaches discussed above are equivalent.

Proposition 5.1 (Proposition C.1)**.** *Given a strictly positive differentiable function $f\colon (0,a) \to (0,\infty)$ and a positive real number $\alpha > 0$, the following conditions are equivalent.*

1. *The function f explodes at 0 like $t^{-\alpha}$, i.e. for some positive constants $c, C > 0$ one has $c < t^\alpha f(t) < C$ for some $\epsilon > 0$ and $t \in (0,\epsilon)$.*
2. $\log f(t) = -\alpha \log t + O(1)$.
3. $\lim_{t\to 0^+} \log f(t)/\log t = -\alpha$.
4. $\lim_{t\to 0^+} tf'(t)/f(t) = -\alpha$.

As a consequence, the estimation of Local Intrinsic Dimension using LIDL can be achieved by computing $\beta(x)$, yielding $d = D + \beta(x)$.

Proposition 5.2 (Proposition C.2)**.** *For t near 0 the following estimate holds*

$$\log\rho_t(x) = \beta(x)\log\sqrt{t} + O(1). \tag{10}$$

The next proposition provides an elegant expression for $\beta_t(x)$, and consequently for $\beta(x)$, expressed in terms of the density ψ on $\mathbb{R}^d$.

Proposition 5.3 (Proposition C.3)**.** *For $t > 0$ and $x \in S = \mathbb{R}^d \subseteq \mathbb{R}^D$ we have*

$$\beta_t(x) = d - D + \frac{\Delta_x(\psi * \phi_t^d)(x)}{\psi * \phi_t^d(x)} \cdot t. \tag{11}$$

LIDL Examples

From the theoretical considerations of Tempczyk et al. (2022) it follows that $\beta(x) = d - D$ if ψ is sufficiently regular and positive near x. In other words,

$$\lim_{t\to 0^+}\frac{\Delta_x(\psi * \phi_t^d)(x)}{\psi * \phi_t^d(x)} \cdot t = 0. \tag{12}$$

Now, we will try to obtain this conclusion directly and calculate bias of LIDL for $t > 0$ in a few special cases by analyzing the behavior of $\beta_t(x)$.

The "uniform distribution" on Euclidean space. There is no such thing as the uniform distribution on $\mathbb{R}^d$. However, from a purely theoretical viewpoint, in our differential equation approach we don't need the assumption of ϕ being a probability density; it could be any function. And since constant functions are usually the simplest examples, we will now investigate what happens if we put $\psi(x) \equiv 1$ on the whole $\mathbb{R}^d$ space.

Using Proposition 5.3 and the fact that ψ has bounded derivatives, this case leaves us with

$$\beta_t(x) = d - D + \frac{\Delta_x\psi * \phi_t^d(x)}{\psi * \phi_t^d(x)} \cdot t = d - D, \tag{13}$$

since $\Delta_x\psi \equiv 0$. This expression is constant in t, and in particular its limit at 0 is $\beta(x) = d - D$. In this case, LIDL estimator is not biased for all $t > 0$.

Normal distribution. Now consider the normal distribution on $\mathbb{R}^d$ with covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, and denote its density function by ψ. The convolution $\psi * \phi_t^d$ is the density of the normal distribution with covariance matrix $\Sigma + tI$. If we simplify notation be putting $\phi_i = \phi^1_{\sigma_i^2+t}$, we get

$$\psi * \phi_t^d(x) = \prod_{i=1}^d \phi_i(x_i). \tag{14}$$

To compute the Laplacian of this convolution, note that

$$\psi * \phi_t^d(x) = \frac{\psi * \phi_t^d(x)}{\phi_i(x_i)} \cdot \phi_i(x_i), \tag{15}$$

where the first factor does not depend of x_i, and therefore

$$\frac{\partial^2(\psi * \phi_t^d)}{\partial x_i^2}(x) = \psi(x) * \phi_t^d(x) \cdot \frac{1}{\phi_i(x_i)} \frac{\partial^2 \phi_i}{\partial x_i^2}(x_i), \tag{16}$$

leading to

$$\begin{aligned} \beta_t(x) &= d - D + t \frac{\Delta(\psi * \phi_t^d)(x)}{\psi * \phi_t^d(x)} = \\ &= d - D + t \sum_{i=1}^{d} \frac{1}{\phi_i(x_i)} \frac{\partial^2 \phi_i}{\partial x_i^2}(x_i) \\ &= d - D + t \sum_{i=1}^{d} \frac{x_i^2 - (\sigma_i^2 + t)}{(\sigma_i^2 + t)^2}. \end{aligned} \tag{17}$$

It is easy to see that the second derivatives of ϕ_i are continuous in $t > -\sigma_i^2$, so the sum in the above expression has finite limit for $t \to 0$, and therefore $\beta(x) = d - D$.

In the special case where $\Sigma = \sigma^2 I$, these calculations simplify further, as $\psi * \phi_t^d = \phi_{\sigma^2+t}^d$, and since

$$\Delta_x \phi_{\sigma^2+t}^d(x) = \left(\frac{\|x\|^2}{(\sigma^2 + t)^2} - \frac{d}{\sigma^2 + t} \right) \phi_{\sigma^2+t}^d(x), \tag{18}$$

we have

$$\beta_t(x) = d - D + \left(\frac{\|x\|^2}{(\sigma^2 + t)^2} - \frac{d}{\sigma^2 + t} \right) t. \tag{19}$$

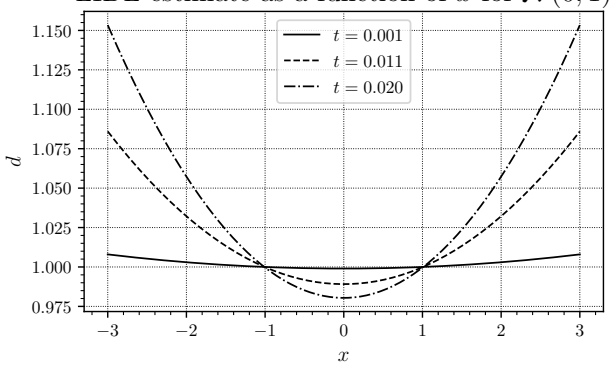
These results express analytically the experimental observations from LIDL and FLIPD papers, as can be verified by looking at Fig. 1a. We can observe, that if we move to the regions of very low probability for a Gaussian, it generates very high positive bias, which may highly overestimate the true LID (also observed as a *bump* at $t = 10^{-12}$ in Fig. 1b). Luckily, most of the points in our dataset come from the region of high probability, but we should be less certain of the estimates for points from low probability regions.

It is worth noting, that the values of t used to generate Fig. 1a are the same as values of δ, which is equal to $\sqrt{t}$ in our convention. After double-checking our results we argue that the most probable cause of this is that the authors of LIDL used squared values of δ by mistake. Additionally, one can observe that curves plotted by Tempczyk et al. (2022) are somewhat flatter than one in this study. The fact that in their paper the derivative was approximated by linear regression on numerically calculated densities – which may lead to slightly different results – might be a possible reason.

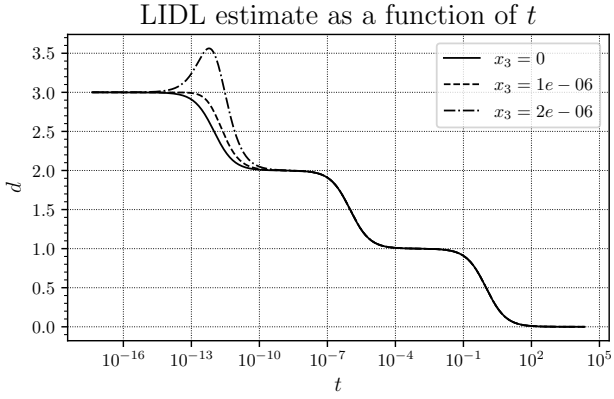
Arbitrary distribution with sufficiently *nice* density. By this point, the notion of *nice* density is shall be more clear. We want to be able to use the equality

$$\Delta_x(\psi * \phi_t^d) = \Delta_x \psi * \phi_t^d. \tag{20}$$

To do so, we need ψ to be bounded, twice differentiable, and have bounded first and second-order partial derivatives.



(a) Example of the bias of a LIDL estimate for different points from $\mathcal{N}(0, 1)$ and for different values of t. This plot recreates a numerical calculations presented in Figure 4 from (Tempczyk et al. 2022), with two minor differences described in Sec. 5



(b) Plot of a LIDL estimate as a function of t for the distribution $\mathcal{N}(\mathbf{0}, \mathrm{diag}(1, 10^{-6}, 10^{-12}))$ and for three different points $\mathbf{x} = (0, 0, x_3)$, which represents a distance of 0, 1 and 2 σ_3 from 0 on 3rd dimension (compare with Fig. 8 from FLIPD and Fig. 2 from LIDL).

Figure 1: LIDL estimates for Gaussian distributions.

We will also require ψ to have *continuous* second-order partial derivatives. This is not a severe restriction, as numerous distributions satisfy these properties – including the normal distribution or more generally, mixtures of Gaussians.

In this case, we have

$$\beta_t(x) = d - D + \frac{\Delta_x \psi * \phi_t^d(x)}{\psi * \phi_t^d(x)} \cdot t, \tag{21}$$

however this time $\Delta_x \psi$ is some arbitrary continuous function. Being differentiable, ψ is also continuous, and we can use the general fact that for a bounded continuous function, f on $\mathbb{R}^d$ one has

$$\lim_{t \to 0^+} f * \phi_t^d(x) = f(x). \tag{22}$$

This gives us, for x such that $\psi(x) > 0$,

$$\lim_{t \to 0^+} \frac{\Delta_x \psi * \phi_t^d(x)}{\psi * \phi_t^d(x)} \cdot t = \frac{\Delta_x \psi(x)}{\psi(x)} \lim_{t \to 0^+} t = 0, \tag{23}$$

and again $\beta(x) = d - D$. It has been already proven that in this case β yields a correct estimate of dimension, circumventing complexities of Tempczyk et al. (2022) proofs.

It is worth noting, that when $\Delta_x \psi = 0$, the estimate is accurate. It is the case for the aforementioned "uniform distribution" on $\mathbb{R}^d$, but it is also true if locally the density is a linear function of x. In Fig. 1a we can observe that for $x \approx \pm 1$ (Laplacian of a Gaussian density equals 0 at these points) and small values of t, the estimate is accurate.

Uniform distribution supported on an interval. Now consider an example where the density is not differentiable – the uniform distribution on an interval $[a, b] \subset \mathbb{R}$, i.e.

$$\psi(x) = \frac{1}{b-a}\chi_{[a,b]}(x), \tag{24}$$

where $\chi_A(s)$ is the indicator function of the set A, equal to 1 on A and 0 outside A. In the next example, we will generalize this to a hypercube, but the core observations can be made in this simpler 1-dimensional case.

The difficulty introduced by the non-differentiability of ψ is we are no longer allowed to move the Laplacian inside the convolution to get

$$\Delta_x(\psi * \phi_t) = \Delta_x \psi * \phi_t \tag{25}$$

(we omit the superscript $d = 1$ from ϕ_t) – as tempting as it might be. Therefore, a different manner of proceeding is needed. We may still move the Laplacian to ϕ_t. In the 1-dimensional case, Δ_x is simply the second derivative, and since

$$\phi_t'(u) = -u\phi_t(u)/t \tag{26}$$

we have

$$\begin{aligned}\Delta_x(\psi * \phi_t)(x) &= \frac{1}{b-a}\int_{x-b}^{x-a} \phi_t''(u)\,du \\ &= \frac{\phi_t'(x-a) - \phi_t'(x-b)}{b-a} = \\ &= \frac{(x-b)\phi_t(x-b) - (x-a)\phi_t(x-a)}{t(b-a)},\end{aligned} \tag{27}$$

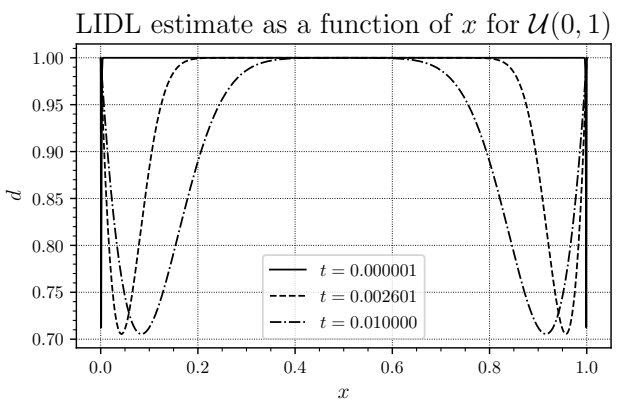
Expanding the denominator in a similar fashion yields

$$\beta_t(x) = d - D + \frac{(x-b)\phi_t(x-b) - (x-a)\phi_t(x-a)}{\Phi_t(x-a) - \Phi_t(x-b)}, \tag{28}$$

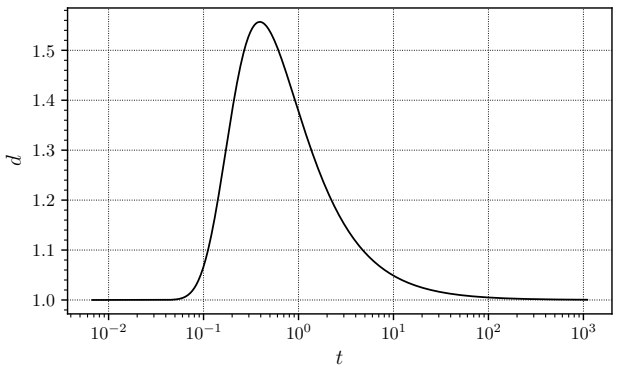
where Φ_t is the cumulative distribution function corresponding to the density ϕ_t. In particular for $x \in (a, b)$ we see that since $x - b < 0 < x - a$, when $t \to 0^+$, the denominator tends to 1, while both terms of the numerator tend to 0, leaving us with $d - D$. LIDL estimate curves for this case for different values of t are plotted in Fig. 2a.

Uniform distribution supported on a hypercube. Let us now consider a more general case – the uniform distribution on a hypercube $[a_1, b_1] \times \cdots \times [a_d, b_d] \subset \mathbb{R}^d$. We have

$$\psi(x) = \prod_{i=1}^{d} \frac{1}{b_i - a_i}\chi_{[a_i,b_i]}(x_i), \tag{29}$$



(a) Example of the bias of a LIDL estimate for different points from $\mathcal{U}(0, 1)$ and values of t. This plot recreates a numerical calculations presented in Figure 3 from (Tempczyk et al. 2022)



(b) LIDL estimate as a function of t for a point from parallel 1D manifolds separated by a distance of 1 with uniform distribution on them. Similar to result from Fig. 6 in (Tempczyk et al. 2022).

Figure 2: LIDL estimates.

Denote

$$\psi_i(s) = \frac{1}{b_i - a_i}\chi_{[a_i,b_i]}(s), \tag{30}$$

and observe that since $\phi_t^d(x)$ is the product of $\phi_t(x_i)$, we have

$$\psi * \phi_t^d(x) = \prod_{i=1}^{d} \psi_i * \phi_t(x_i). \tag{31}$$

By directly computing the derivatives, we obtain

$$\frac{\Delta_x \psi * \phi_t^d(x)}{\psi * \phi_t^d(x)} = \sum_{i=1}^{d} \frac{(\psi_i * \phi_t)''(x_i)}{\psi_i * \phi_t(x_i)} = \sum_{i=1}^{d} \frac{\psi_i * \phi_t''(x_i)}{\psi_i * \phi_t(x_i)}, \tag{32}$$

reducing our problem to the 1-dimensional variant we have dealt with in the preceding example. Summing up, we have

$$\begin{aligned}\beta_t(x) &= d - D \\ &+ \sum_{i=1}^{d} \frac{(x_i - b_i)\phi_t(x_i - b_i) - (x_i - a_i)\phi_t(x_i - a_i)}{\Phi_t(x_i - a_i) - \Phi_t(x_i - b_i)},\end{aligned} \tag{33}$$

and the asymptotic behavior with $t \to 0^+$ follows the 1-dimensional case.

Union of two parallel hyperplanes. Suppose that S is a union of two parallel hyperplanes, $S_1 = \mathbb{R}^d$ and $S_2 = v + \mathbb{R}^d$, where $v \perp \mathbb{R}^d$. Moreover, assume that $p_S = (1-\lambda)p_1 + \lambda p_2$ is a convex combination of probability measures p_i supported on S_i, with densities $\psi_i \colon \mathbb{R}^d \to \mathbb{R}$ (we identify S_2 with $\mathbb{R}^d$ through the map $x \mapsto x + v$). In this case, for $x \in S_1$ we have

$$\beta_t^1(x) = d - D + \frac{\Delta_x(\psi_1 * \phi_t^d)(x)}{\psi_1 * \phi_t^d(x)} \cdot t, \tag{34}$$

and by Lemma 4.1, and the observation that

$$\rho_t^2(x) = \psi_2 * \phi_t^d(x)\phi_t^{D-d}(v), \tag{35}$$

we get

$$\beta_t^2(x) = d - D + \frac{\|v\|^2}{t} + \frac{\Delta_x(\psi_2 * \phi_t^d)(x)}{\psi_2 * \phi_t^d(x)} \cdot t. \tag{36}$$

Here, we see that for an off-manifold point $x \notin S_2$, the expression for $\beta_t^2(x)$ contains a summand $\|v\|^2/t$ that explodes at 0, and if the last term is under control, $\beta^2(x)$ is infinite. However, by Lemma D.2, the coefficient of $\beta_t^2(x)$ in the expansion of $\beta_t(x)$ from Lemma D.1 decreases exponentially in $1/t$, neutralizing this divergence.

For the remainder of this example, let us assume $\psi_1 = \psi_2 = \psi$. In this case, we may apply multiple simplifications, in particular

$$\beta_t^2(x) = \beta_t^1(x) + \frac{\|v\|^2}{t}, \tag{37}$$

and moreover

$$\begin{aligned}
\frac{\rho_t^2(x)}{\rho_t(x)} =& \\
=& \frac{\psi * \phi_t^d(x)\phi_t^{D-d}(v)}{(1-\lambda)\psi * \phi_t^d(x)\phi_t^{D-d}(0) + \lambda\psi * \phi_t^d(x)\phi_t^{D-d}(v)} \\
=& \frac{\phi_t^{D-d}(v)}{(1-\lambda)\phi_t^{D-d}(0) + \lambda\phi_t^{D-d}(v)} \\
=& \frac{1}{(1-\lambda)e^{\|v\|^2/2t} + \lambda}.
\end{aligned} \tag{38}$$

Since $\frac{\lambda\rho_t^1(x)}{\rho_t(x)} = 1 - \frac{\lambda\rho_t^2(x)}{\rho_t(x)}$ we can simplify the expression for $\beta_t(x)$ from Lemma D.1, yielding

$$\begin{aligned}
\beta_t(x) &= \beta_t^1(x) + \frac{\|v\|^2}{t} \cdot \frac{\lambda\rho_t^2(x)}{\rho_t(x)} \\
&= \beta_t^1(x) + \frac{\lambda\|v\|^2}{t\left((1-\lambda)e^{\|v\|^2/2t} + \lambda\right)}.
\end{aligned} \tag{39}$$

To give a concrete example, if $\psi = \phi_{\sigma^2}^d$, then

$$\begin{aligned}
\beta_t(x) = d - D &+ \left(\frac{\|x\|^2}{(\sigma^2+t)^2} - \frac{d}{\sigma^2+t}\right) t \\
&+ \frac{\lambda\|v\|^2}{t\left((1-\lambda)e^{\|v\|^2/2t} + \lambda\right)}.
\end{aligned} \tag{40}$$

Another interesting example occurs when the data on both parallel manifolds follow uniform density functions. Although the derivation in Eq. (19) requires the density functions to be probability distributions, this scenario can be simulated by considering two Gaussian distributions with relatively large standard deviations. This yields the following formula for $\beta_t(x)$, presented in Fig. 2b for $v = 1$, $\lambda = \frac{1}{2}$:

$$\beta_t(x) = d - D + \frac{\lambda\|v\|^2}{t\left((1-\lambda)e^{\|v\|^2/2t} + \lambda\right)}. \tag{41}$$

Union of two intersecting manifolds. In this example we will consider a manifold S decomposing into a union of two components S_1 and S_2, intersecting at a point $x \in S_1 \cap S_2$. Denote by d_i the dimension of S_i. As before, let $p_S = \lambda p_1 + (1-\lambda)p_2$. Moreover, suppose that

$$\beta_t^i(x) = d_i - D + E_i(t), \tag{42}$$

where $E_i(t)$ expresses the error of β_t^i in estimating the dimension of S_i, and $\lim_{t\to 0^+} E_i(t) = 0$. By Lemma D.1 we have

$$\beta_t(x) = \left(\frac{\lambda\rho_t^1(x)}{\rho_t(x)}d_1 + \frac{(1-\lambda)\rho_t^2(x)}{\rho_t(x)}d_2\right) - D + E(t), \tag{43}$$

where the error term is a convex combination of E_1, and E_2, and thus is bounded by their maximum,

$$\begin{aligned}
E(t) &= \frac{\lambda\rho_t^1(x)}{\rho_t(x)}E_1(t) + \frac{(1-\lambda)\rho_t^2(x)}{\rho_t(x)}E_2(t) \\
&\leq \max\{E_1(t), E_2(t)\}.
\end{aligned} \tag{44}$$

In particular, it also vanishes as $t \to 0^+$.

The value of LID at x estimated by $\beta_t(x)$ lies between d_1 and d_2, and is controlled by the asymptotic of $\lambda\rho_t^1(x)/\rho_t(x)$. If $d_1 = d_2 = d$, then it is also equal to d.

6 Conclusions

In this work, for the first time, we have outlined a new perspective for Wiener process-based algorithms for LID estimation and shown some results for LIDL and FLIPD, in which we found an analytical description for the phenomena observed in experiments.

The presented results open up several promising new research directions. A natural extension of this study would involve accounting for the curvature of manifolds, addressing the current assumption of manifold flatness. Moreover, an interesting direction for future research could be an extended analysis of how nearby manifolds can affect LID estimates as briefly shown in the example of parallel manifolds in Sec. 5. Another potential research avenue is to apply the proposed approach to analyze other algorithms, such as the NB algorithm, in a manner similar to LIDL.

Finally, our brief experiments show that using density models to estimate Laplacian can be beneficial in the process of improving LIDL estimate for higher values of t and non-uniform densities, leading to a promising future direction of research. Using Wiener process perspective for answering the question how dataset quantization can affect the estimate is another interesting question to answer in the area of LID estimation.

Acknowledgments

We want to thank Marek Cygan, Michał Karpowicz and Adam Goliński for their overall support during this project.

CRediT Author Statement. **Piotr Tempczyk** (32% of the work): Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project administration. **Adam Kurpisz** (32% of the work): Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision. **Łukasz Garncarek** (26% of the work): Conceptualization, Methodology, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing. **Dominik Filipiak** (10% of the work): Software, Data Curation, Writing – Review & Editing.

References

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

Bailey, J.; Houle, M. E.; and Ma, X. 2022. Local intrinsic dimensionality, entropy and statistical divergences. *Entropy*, 24(9): 1220.

Bailey, J.; Houle, M. E.; and Ma, X. 2023. Relationships between tail entropies and local intrinsic dimensionality and their use for estimation and feature representation. *Information Systems*, 118: 102245.

Brown, B. C.; Caterini, A. L.; Ross, B. L.; Cresswell, J. C.; and Loaiza-Ganem, G. 2022. Verifying the Union of Manifolds Hypothesis for Image Data. In *The Eleventh International Conference on Learning Representations*.

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

Campadelli, P.; Casiraghi, E.; Ceruti, C.; and Rozza, A. 2015. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015: 1–21.

Fan, M.; Gu, N.; Qiao, H.; and Zhang, B. 2010. Intrinsic dimension estimation of data by principal component analysis. *arXiv preprint arXiv:1002.2050*.

Fukunaga, K.; and Olsen, D. R. 1971. An Algorithm for Finding Intrinsic Dimensionality of Data. *IEEE Transactions on Computers*, C-20: 176–183.

Horvat, C.; and Pfister, J.-P. 2022. Intrinsic dimensionality estimation using normalizing flows. *Advances in Neural Information Processing Systems*, 35: 12225–12236.

Horvat, C.; and Pfister, J.-P. 2024. On gauge freedom, conservativity and intrinsic dimensionality estimation in diffusion models. *arXiv preprint arXiv:2402.03845*.

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

Kamkari, H.; Ross, B. L.; Cresswell, J. C.; Caterini, A. L.; Krishnan, R.; and Loaiza-Ganem, G. 2024a. A Geometric Explanation of the Likelihood OOD Detection Paradox. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 22908–22935. PMLR.

Kamkari, H.; Ross, B. L.; Hosseinzadeh, R.; Cresswell, J. C.; and Loaiza-Ganem, G. 2024b. A Geometric View of Data Complexity: Efficient Local Intrinsic Dimension Estimation with Diffusion Models. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*.

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

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

Li, C.; Farkhoor, H.; Liu, R.; and Yosinski, J. 2018. Measuring the Intrinsic Dimension of Objective Landscapes. In *International Conference on Learning Representations*.

Loaiza-Ganem, G.; Ross, B. L.; Hosseinzadeh, R.; Caterini, A. L.; and Cresswell, J. C. 2024. Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections. *arXiv preprint arXiv:2404.02954*.

Minka, T. 2000. Automatic choice of dimensionality for PCA. *Advances in neural information processing systems*, 13.

Pope, P.; Zhu, C.; Abdelkader, A.; Goldblum, M.; and Goldstein, T. 2020. The Intrinsic Dimension of Images and Its Impact on Learning. In *International Conference on Learning Representations*.

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

Stanczuk, J. P.; Batzolis, G.; Deveney, T.; and Schönlieb, C.-B. 2024. Diffusion Models Encode the Intrinsic Dimension of Data Manifolds. In *Forty-first International Conference on Machine Learning*.

Tempczyk, P.; Michaluk, R.; Garncarek, L.; Spurek, P.; Tabor, J.; and Golinski, A. 2022. Lidl: Local intrinsic dimension estimation using approximate likelihood. In *International Conference on Machine Learning*, 21205–21231. PMLR.

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

Yeats, E.; Darwin, C.; Liu, F.; and Li, H. 2023. Adversarial Estimation of Topological Dimension with Harmonic Score Maps. *arXiv preprint arXiv:2312.06869*.

Zheng, Y.; He, T.; Qiu, Y.; and Wipf, D. P. 2022. Learning manifold dimensions with conditional variational autoencoders. *Advances in Neural Information Processing Systems*, 35: 34709–34721.