Intrinsic Dimensionality Estimation within Tight Localities

Laurent Amsaleg, Oussama Chelly, Michael E Houle, Ken-Ichi Kawarabayashi, Miloš Radovanović, Weeris Treeratanajaru

Intrinsic Dimensionality Estimation within Tight Localities*

Laurent Amsaleg† laurent.amsaleg@irisa.fr
Oussama Chelly‡ v-ouchel@microsoft.com
Michael E. Houle† meh@nii.ac.jp
Ken-ichi Kawarabayashi† k_keniti@nii.ac.jp
Miloš Radovanović§ radacha@dmi.uns.ac.rs
Weeris Treeratanajaru¶ weeris.t@gmail.com

Abstract

Accurate estimation of Intrinsic Dimensionality (ID) is of crucial importance in many data mining and machine learning tasks, including dimensionality reduction, outlier detection, similarity search and subspace clustering. However, since their convergence generally requires sample sizes (that is, neighborhood sizes) on the order of hundreds of points, existing ID estimation methods may have only limited usefulness for applications in which the data consists of many natural groups of small size. In this paper, we propose a local ID estimation strategy stable even for 'tight' localities consisting of as few as 20 sample points. The estimator applies MLE techniques over all available pairwise distances among the members of the sample, based on a recent extreme-value-theoretic model of intrinsic dimensionality, the Local Intrinsic Dimension (LID). Our experimental results show that our proposed estimation technique can achieve notably smaller variance, while maintaining comparable levels of bias, at much smaller sample sizes than state-of-the-art estimators.

1 Introduction

In high-dimensional contexts where data is represented by many features, the performance of data analysis techniques often greatly depends on the inherent complexity of the data model. Although this complexity is often taken to be the number of features themselves (that is, the 'representational dimension' or the 'ambient dimension'), simply counting the number of features does not take into account the relationships among them: some features may be redundant, while others may be irrelevant, while yet others may exhibit various degrees of dependency. A better measure of the complexity of the data model is to determine the intrinsic dimensionality (ID) of the data according to some natural criterion, such as the number of latent variables required to describe the data, or the number of basis vectors needed to describe a manifold that closely approximates the data set.

Over the decades, many characterizations of intrinsic dimensionality were proposed, each with its own estimators [9]. Topological models estimate the basis dimension of the tangent space of the data manifold [53, 39, 8, 62]. This class of estimators includes Principal Component Analysis (PCA) and its variants [53, 39, 41, 61], and multidimensional scaling (MDS) [18, 58, 44, 3, 14, 20, 43]. Graph-based methods attempt to preserve the k-NN graph [17]. Fractal models, popular in physics applications, are used to estimate the dimension of nonlinear systems [52] — these include popular estimators due to Camastra & Vinciarelli [10], Fan, Qiao & Zhang [22], Grassberger & Procaccia [25], Hein & Audibert [28], Kégl [42], Raginsky & Lazebnik [54] and Takens [59]. Statistical estimators such as IDEA [57] and DANCo [13] estimate the dimension from the concentration of norms and angles.

The aforementioned estimators can be described as 'global', in that they provide a single ID measurement for the full data set, as opposed to 'local' ID estimators that assign a different dimensionality to each point or region in the data set. Commonly-used local estimators of ID include: topological methods that model dimensionality as that of a locally tangent subspace to a manifold, such as PCA-based approaches [24, 8, 22, 47, 48], and ID estimation from Expected Simplex Skewness [38]; Local Multidimensional Scaling methods [18] such as Isometric Mapping [60] Locally Linear Embedding [55], Laplacian and Hessian eigenmaps [21, 2], and Brand's Method [7]; distance-based measures such as the Expansion Dimension (ED), that assess the rate of expansion of the neighborhood size with increasing radius [40, 32, 27], as well as other probabilistic methods that view the data as a sample from a hidden distance distribution, such as the Hill estimator [29], the Manifold-Adaptive Dimension [23], Levina and Bickel's algorithm [46], the minimum neighbor dis-

*M. E. H. supported by JSPS Kakenhi Kiban (B) Research Grant 18H03296. K. K. supported by JST ERATO Kawarabayashi Large Graph Project JPMJER1201 and by JSPS Kakenhi JP18H05291. M. R. thanks Serbian nat'l project OI174023.

†CNRS-IRISA, France.
‡Microsoft IoT & AI Insider Lab, Germany.
§University of Novi Sad, Faculty of Sciences, Serbia.
¶Chulalongkorn University, Thailand.

tance (MiND) framework [56], and the local intrinsic dimensionality (LID) framework [1, 30].

Distance-based estimators of local ID infer the data dimensionality solely from the distribution of distances from a reference point to its nearest neighbors; they generally do not require any assumptions as to whether the data can be modeled as a manifold. This convenience allows distance-based measures of local ID to be used in the context of similarity search, where they are used to assess the complexity of a search query [40], to control the early termination of search [12, 33, 34], or the interleaving of feature sparsification with the construction of a neighborhood graph [31, 35]. Distance-based measures have also found applications in deep neural network classification [50]; the characterization and detection of adversarial perturbation [49]; and outlier detection, in the analysis of a projection-based heuristic [19], and in the estimation of local density [36]. The efficiency and effectiveness of the algorithmic applications of local intrinsic dimensional estimation (such as [12, 33, 34, 35]) depends heavily on the quality of the estimators employed, and on their ability to provide a trustworthy estimate despite the limited number of available samples. Distance-based local estimators are well-suited for many of the applications in question, since nearest neighbor distances are often precomputed and are thus readily available for use in estimation.

Local estimators of ID can potentially have significant impact when used in subspace outlier detection, subspace clustering, or other applications in which the intrinsic dimensionality is assumed to vary from location to location. However, in practical settings, the localities assumed in data analysis are often too 'tight' to provide the number of neighborhood samples needed by current estimators of ID. State-of-the-art outlier detection algorithms, for example, typically use neighborhoods of 20 or fewer data points as localities within which to assess the outlierness of test examples [11], whereas estimators of local intrinsic dimension generally require sample sizes in excess of 100 in order to converge [1]. Simply choosing a number of samples sufficient for the convergence of the estimator would very often lead to a violation of the locality constraint, and to ID estimates that are consequently less reliable, as the samples would consist of points from several different natural data groups with different local intrinsic dimensional characteristics.

Global estimators are sometimes adapted for local estimation of ID simply by applying them to the subset of the data lying within some region surrounding a point of interest. Global methods such as PCA generally make use of most (if not all) of the (quadratic) pairwise relationships within the data, giving them an apparent advantage over expansion-based local estimators, which use only the (linear) number of distances from the reference point to each sample point. However, embedding-based and projective-based ID estimators can be very sensitive to noise; moreover, they can be greatly misled when the chosen region is larger than the targeted locality, or when the data distribution does not conform to a linear manifold [1]. Also, as we shall argue in Section 3, 'clipping' of the data set to a region can also give rise to boundary effects that have the potential for extreme bias when estimating ID, whenever the region shape is not properly accounted for in the ID model or estimation strategy. With these issues in mind, locally-restricted application of global ID estimation should not automatically be regarded as valid for local ID estimation. On the other hand, expansion-based estimators of local ID can be seen to avoid clipping bias, since they model (explicitly or implicitly) the restriction of the distribution of distances to a fixed-radius neighborhood centered at the reference point [1].

As the sizes of the natural groupings (clusters) of the data set are generally not known in advance, in order to ensure that the majority of the points are drawn from the same local distribution, it is highly desirable to use local estimators that can cope with the smallest possible sample sizes [1, 56]. One possible strategy for improving the convergence properties of estimation without violating locality is to draw more measurements from smaller data samples. For the case of distance-based local estimation from a neighborhood sample of size k, this would require the use of distances between pairs of neighbors (potentially quadratic in k), and not merely the distances from the reference point to its neighbors (linear in k). This presents a challenge in that any additional measurements must be used in a way that locally conforms to the underlying data distribution without introducing bias due to clipping.

In this paper, we develop an effective estimator of local intrinsic dimension suitable for use in tight localities — that is, within neighborhoods of small size that are often employed in such applications as outlier detection and nearest-neighbor classification. Given a sample of k points drawn from some target locality (generated by restricting the data set to a spherical region of radius r centered at $\mathbf{q}$), our estimator can be regarded as an aggregation of $2k$ expansion-based LID estimation processes, each taking either a distinct sample point $\mathbf{v}$ or its symmetric reflection relative to $\mathbf{q}$ (that is, the point $2\mathbf{q}-\mathbf{v}$) as the origin of its expansion. To ensure that these processes use only such information that is available within the locality, the expansion processes are skewed, in that their centers are allowed to shift gradually towards $\mathbf{q}$ as the radius of expansion approaches r. Under the modeling assumption that

the underlying local intrinsic dimensionality is uniform throughout the region, an estimator resulting from the aggregation of $2k$ skewed expansion processes will be shown to use all $O(k^2)$ pairwise distances within the sample, without introducing clipping bias.

The main contributions of this paper include:

- an explanation of the clipping effect in skewed expansion-based local ID estimation, and an illustration of its impact on bias;
- a description of the proposed LID-based estimator for tight localities, as well as a justification under the assumption of continuity of local ID;
- an experimental investigation showing that our proposed tightly-local estimation technique can achieve notably smaller variance, while maintaining comparable levels of bias, at much smaller sample sizes than state-of-the-art estimators.

The remainder of the paper is structured as follows. In the next section, we review the LID model of local intrinsic dimensionality [30, 31] and its estimators [1]. In Section 3, we discuss the issue of the reuse of neighborhood samples for the estimation of expansion-based local intrinsic dimensionality. We also show how clipping bias can result when the expansion originates at a point which is not the center of the neighborhood. In Section 4 we introduce our proposed tight LID-based estimator together with its justification under the assumption of continuity of local ID. In Section 5, we provide the details of our experimental framework, and in Section 6 we present an experimental comparison of our proposed estimator with existing local and global ID estimators. We conclude the discussion in Section 7.

2 LID and Extreme Value Theory

In the theory of intrinsic dimensionality, classical expansion models (such as the expansion dimension and generalized expansion dimension [40, 32]) measure the rate of growth in the number of data objects encountered as the distance from the reference sample increases. As an intuitive example, in Euclidean space, the volume of an m-dimensional ball grows proportionally to r^m, when its size is scaled by a factor of r. From this rate of volume growth with distance, the expansion dimension m can be deduced as:

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^m \Rightarrow m = \frac{\ln(V_2/V_1)}{\ln(r_2/r_1)}. \tag{2.1}$$

By treating probability mass as a proxy for volume, classical expansion models provide a *local view* of the dimensional structure of the data, as their estimation is restricted to a neighborhood around the sample of interest. Transferring the concept of expansion dimension to the statistical setting of continuous distance distributions leads to the formal definition of LID [30].

Definition 1. (Local Intrinsic Dimensionality)

Given a data sample $\mathbf{x} \in X$, let $R > 0$ be a random variable denoting the distance from $\mathbf{x}$ to other data samples. If the cdf $F(r)$ of R is positive and continuously differentiable at distance $r > 0$, the LID of $\mathbf{x}$ at distance r is given by:

$$\mathrm{ID}_F(r) \triangleq \lim_{\epsilon \to 0} \frac{\ln\big(F((1+\epsilon)\cdot r)/F(r)\big)}{\ln(1+\epsilon)} = \frac{r \cdot F'(r)}{F(r)}, \tag{2.2}$$

whenever the limit exists.

$F(r)$ is analogous to the volume V in Equation (2.1); however, we note that the underlying distance measure need not be Euclidean. The last equality of Equation (2.2) follows by applying L'Hôpital's rule to the limits [30]. The local intrinsic dimension at $\mathbf{x}$ is in turn defined as the limit, when the radius r tends to zero:

$$\mathrm{ID}_F^* \triangleq \lim_{r \to 0} \mathrm{ID}_F(r). \tag{2.3}$$

ID_F^* describes the relative rate at which its cdf $F(r)$ increases as the distance r increases from 0, and can be estimated using the distances of $\mathbf{x}$ to its k nearest neighbors within the sample [1].

Estimation of LID: According to the branch of statistics known as extreme value theory, the smallest k nearest neighbor distances could be regarded as extreme events associated with the lower tail of the underlying distance distribution. Under very reasonable assumptions, the tails of continuous probability distributions converge to the Generalized Pareto Distribution (GPD), a form of power-law distribution [16]. From this, [1] developed several estimators of LID to heuristically approximate the true underlying distance distribution by a transformed GPD; among these, the Maximum Likelihood Estimator (MLE) — which coincides with the Hill estimator [29] for the scale parameter of untransformed GPDs — exhibited a useful trade-off between statistical efficiency and complexity. Given a reference sample $\mathbf{x} \sim \mathcal{P}$, where $\mathcal{P}$ represents the global data distribution, the MLE estimator of the LID at $\mathbf{x}$ is:

$$\widehat{\mathrm{ID}}_{\mathrm{MLE}}(\mathbf{x}) = -\left(\frac{1}{k}\sum_{i=1}^{k} \ln \frac{r_i(\mathbf{x})}{r_k(\mathbf{x})}\right)^{-1}. \tag{2.4}$$

Here, $r_i(\mathbf{x})$ denotes the distance between $\mathbf{x}$ and its i-th nearest neighbor within a sample of points drawn from $\mathcal{P}$, where $r_k(\mathbf{x})$ is the maximum of the neighbor distances. In practice, the sample set is drawn uniformly at random from the available training data (omitting $\mathbf{x}$ itself), which itself is presumed to have been randomly drawn from $\mathcal{P}$.

3 LID Estimation and the Clipping Effect

In practice, the LID model is typically applied to the cdf F induced by some global distribution of data with respect to a reference location $\mathbf{q}$. In the ideal case where the data in the vicinity of $\mathbf{q}$ is distributed uniformly within a subspace (or manifold), ID_F^* equals the dimension of the subspace. In general, however, these distributions are not ideal, the subspace model of data does not perfectly apply, and ID_F^* is not necessarily an integer. Instead, by characterizing the growth rate of the distribution of distances from $\mathbf{q}$, it naturally takes into account the effect of variation within the subspace, and error relative to the subspace, all in one value. Nevertheless, the local intrinsic dimensionality does give some indication of the dimension of the subspace containing $\mathbf{q}$ that would best fit the data distribution in the vicinity of $\mathbf{q}$, provided that the distribution of distances to $\mathbf{q}$ is smooth. We refer readers to [30, 31] for more details concerning the LID model.

From the global perspective, it is not necessarily the case that ID^* exists for every possible reference location in the domain. However, if the distribution is in some sense smooth in the vicinity of $\mathbf{q}$, it is reasonable to assume that for some point $\mathbf{v}$ sufficiently close to $\mathbf{q}$, the underlying value of $\mathrm{ID}_{F_\mathbf{v}}^*$ could be a close approximation of ID_F^*, where $F_\mathbf{v}$ is the cdf for the distance distribution induced relative to $\mathbf{v}$. We therefore adopt the following definition of the continuity of LID.

Definition 2. *The local intrinsic dimensionality will be said to be* (uniformly) continuous *at* $\mathbf{q} \in \mathcal{S}$ *if the following conditions hold:*

1. *There exists a distance* $\rho > 0$ *for which all points* $\mathbf{v} \in \mathcal{S}$ *with* $\|\mathbf{v} - \mathbf{q}\| \leq \rho$ *admit a distance distribution whose cdf* $F_\mathbf{v}$ *is continuously differentiable and positive within some open interval with lower bound* 0.
2. *For each* $\mathbf{v}$ *satisfying Condition 1,* $\mathrm{ID}_{F_\mathbf{v}}^*$ *exists.*
3. $\lim_{s\to 0} \mathrm{ID}_{F_{\psi(s)}}^*$ *converges uniformly to* $\mathrm{ID}_{F_\mathbf{q}}^*$*, where* $\psi(s) = s\mathbf{v} + (1-s)\mathbf{q}$ *interpolates* $\mathbf{q}$ *and* $\mathbf{v}$.

For the data model underlying our proposed estimator, we will assume that the local intrinsic dimensionality is continuous in the vicinity of the test point $\mathbf{q}$. Under the assumption of continuity, the estimator will use estimates of $\mathrm{ID}_{F_\mathbf{v}}^*$ for points $\mathbf{v}$ close to $\mathbf{q}$ to help stabilize the estimate of ID_F^*.

However, straightforwardly estimating and aggregating values of $\mathrm{ID}_{F_\mathbf{v}}^*$ over all neighbors $\mathbf{v}$ of $\mathbf{q}$ can give rise either to clipping bias, or a violation of locality, or both. To see this, consider the situation shown in Figure 1, in which we have a sample $V = \{v_i | 1 \leq i \leq k\}$ of points

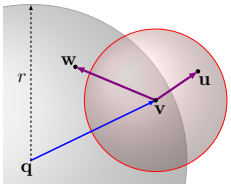

Figure 1: When estimating the local ID at $\mathbf{v}$ to support the estimation of the local ID at $\mathbf{q}$, using the sample $\mathbf{u}$ violates the locality restriction for $\mathbf{q}$, but ignoring it in favor of $\mathbf{w}$ can lead to clipping bias.

drawn from the restriction of the global distribution to a neighborhood $B(\mathbf{q}, r)$ of radius r. For a given neighbor $\mathbf{v} \in V$, many if not most of its own neighbors may lie well outside the vicinity of $\mathbf{q}$. If the estimation makes use of a neighbor $\mathbf{u}$ external to $B(\mathbf{q}, r)$, then locality is violated, which can have drastic consequences for any task that makes use of the estimates.

On the other hand, if the locality of $\mathbf{q}$ is not to be violated, then straightforward use of the sample set V to estimate the characteristics of the distribution of distances from $\mathbf{v}$ would suffer from clipping bias — close neighbors of $\mathbf{v}$ (such as $\mathbf{u}$ in Figure 1) will have effectively been replaced by other members of V (e.g. $\mathbf{w}$) that are farther from $\mathbf{v}$. As will be shown in Section 6, for LID this typically has the effect of introducing a strong negative bias on the estimated values.

4 Tight LID Estimation

For our estimator, which we will refer to as $\widehat{\mathrm{ID}}_{\mathrm{TLE}}$, we limit the sample points to those points of the data set that lie within a tight neighborhood of the test point $\mathbf{q}$. In order to avoid clipping bias, we adjust the distributions of distances computed from a nearby point $\mathbf{x}$ by taking advantage of the assumption of uniform continuity of local ID, as laid out in Definition 2.

4.1 LID Estimation from Moving Centers.

Let r be the radius of the neighborhood V, and let $\mathbf{x}$ be a point within distance r of $\mathbf{q}$. The distribution of distances based at $\mathbf{x}$ is generated through a smooth interpolative process involving an expanding circle whose center is smoothly transformed from $\mathbf{x}$ to $\mathbf{q}$ as its radius is increased from 0 to r. The radii of these circles, together with the probability measure associated with their interiors, determine a distribution of distance values. More formally, if r is the radius of the neighborhood V, the point $\mathbf{x}$ can be associated with a distribution whose c.d.f. $F_{\mathbf{q},\mathbf{x},r}$ is defined as

$$\begin{aligned} \phi_{\mathbf{q},\mathbf{x},r}(t) &\triangleq (t/r)\cdot\mathbf{q} + (1-t/r)\cdot\mathbf{x} \\ F_{\mathbf{q},\mathbf{x},r}(t) &\triangleq F_{\phi_{\mathbf{q},\mathbf{x},r}(t)}(t), \end{aligned}$$

where the interpolation point $\phi_{\mathbf{q},\mathbf{x},r}(t)$ is defined over the range $t \in [0, r]$, and $F_{\phi_{\mathbf{q},\mathbf{x},r}(t)}$ is the c.d.f. of the distribution of distances from $\phi_{\mathbf{q},\mathbf{x},r}(t)$. For any $t \in [0, r]$, the value $F_{\phi_{\mathbf{q},\mathbf{x},r}(t)}$ is the probability of a sample point lying inside the unique circle with center $\phi_{\mathbf{q},\mathbf{x},r}(t)$ and radius t.

THEOREM 4.1. *If the local intrinsic dimensionality* $\mathrm{ID}^*_{F_\mathbf{q}}$ *is uniformly continuous, then* $\mathrm{ID}^*_{F_{\mathbf{q},\mathbf{x},r}} = \mathrm{ID}^*_{F_\mathbf{q}}$.

Proof. Under the assumption of continuity, there exists $\rho > 0$ such that for any $0 \leq s \leq \rho$, the Moore-Osgood theorem implies that:

$$\begin{aligned}
\mathrm{ID}^*_{F_\mathbf{q}} &= \lim_{t\to 0^+} \mathrm{ID}^*_{F_{\phi_{\mathbf{q},\mathbf{x},r}(t)}} \\
&= \lim_{t\to 0^+} \lim_{s\to 0^+} \mathrm{ID}_{F_{\phi_{\mathbf{q},\mathbf{x},r}(t)}}(s) \\
&= \lim_{s\to 0^+} \lim_{t\to 0^+} \mathrm{ID}_{F_{\phi_{\mathbf{q},\mathbf{x},r}(t)}}(s) \\
&= \lim_{s\to 0^+} \mathrm{ID}_{F_{\phi_{\mathbf{q},\mathbf{x},r}(s)}}(s) \\
&= \lim_{s\to 0^+} \mathrm{ID}_{F_{\mathbf{q},\mathbf{x},r}}(s) \;=\; \mathrm{ID}^*_{F_{\mathbf{q},\mathbf{x},r}} .
\end{aligned}$$

Under the assumption of continuity, the local ID at q can therefore be estimated from the distribution $F_{\mathbf{q},\mathbf{x},r}$ for any location $\mathbf{x}$ falling within a sufficiently small neighborhood of $\mathbf{q}$. For the purpose of estimation, the distance value associated with a sample point $\mathbf{v} \in V$ is determined by the radius of the expanding circle at the time its boundary encounters $\mathbf{v}$, and not the actual distance from $\mathbf{x}$ to $\mathbf{v}$. This distance is given by t, s. t.

$$\begin{aligned}
\|\phi_{\mathbf{q},\mathbf{x},r}(t) - \mathbf{x}\| &= \|\phi_{\mathbf{q},\mathbf{x},r}(t) - \mathbf{v}\|, \text{ or equivalently,} \\
\|t(\mathbf{q}-\mathbf{x})\| &= \|t(\mathbf{q}-\mathbf{x}) + r(\mathbf{x}-\mathbf{v})\| ,
\end{aligned}$$

which for any inner-product norm has the solution

$$d_{\mathbf{q},r}(\mathbf{x},\mathbf{v}) \;\triangleq\; \frac{r(\mathbf{v}-\mathbf{x})\cdot(\mathbf{v}-\mathbf{x})}{2(\mathbf{q}-\mathbf{x})\cdot(\mathbf{v}-\mathbf{x})} \;=\; t\,. \tag{4.5}$$

The MLE estimator for a single moving center is obtained by using Equation 2.4 with adjusted distances of the form of Equation 4.5, for all samples of $V \setminus \{\mathbf{x}\}$.

4.2 MLE Estimation from Multiple Centers. There are many possible ways of choosing points $\mathbf{x}$ from which to initiate a moving center for LID estimation. Here, we make use of the following candidates: (1) the neighborhood samples V, (2) the neighborhood center $\mathbf{q}$ itself, and (3) the symmetric reflections of these centers through the point $\mathbf{q}$. For a given sample $\mathbf{v} \in V$, its reflection in $\mathbf{q}$ is simply the point $2\mathbf{q} - \mathbf{v}$. The use of reflected centers is motivated by a desire to balance out whatever non-uniformity that may exist in the neighborhood samples, to obtain a more stable estimate.

$$\begin{aligned}
\widehat{\mathrm{ID}}_{\mathrm{TLE}}(\mathbf{q}) \;=\; -\Bigg(\frac{1}{|V_*|^2}\sum_{\substack{\mathbf{v},\mathbf{w}\in V_* \\ \mathbf{v}\neq\mathbf{w}}} \Bigg[\ln\frac{d_{\mathbf{q},r}(\mathbf{v},\mathbf{w})}{r} \\
+ \ln\frac{d_{\mathbf{q},r}(2\mathbf{q}-\mathbf{v},\mathbf{w})}{r}\Bigg]\Bigg)^{-1} ,
\end{aligned} \tag{4.6}$$

where $V_* = V \cup \{\mathbf{q}\}$ and r is the distance from $\mathbf{q}$ to its farthest neighbor in V.

$\widehat{\mathrm{ID}}_{\mathrm{TLE}}$ can be derived using MLE techniques in a manner analogous to that in which $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ was derived in [1]; for this reason, and for considerations of space, we omit the details here.

5 Experimental Framework

5.1 Competing Estimation Methods. To show the advantages and limitations of TLE, we compared our proposed estimator $\widehat{\mathrm{ID}}_{\mathrm{TLE}}$ with other popular estimators, focusing mostly on local methods: MLE and method of moments (MoM) estimators of LID [1], local correlation dimension (LCD, as CD [25] applied to k-neighborhoods), (generalized) expansion dimension (ED, GED) [32], and the local version of PCA (LPCA). We also (briefly) consider MiND_{ml1} and MiND_{mli} [56], two parameterless global methods. All local methods are parameterized by the neighborhood size k. See the supplement for additional notes.[1]

5.2 Synthetic Data. Our study includes two families of synthetic data sets. For each manifold we generated 20 sets of 10^4 points, and in each experiment we report the averages of observed means and standard deviations of ID measures over the 20 sets. The first family is i.i.d. Gaussian, uniform in the unit cube and multidimensional torus, meant to evaluate behavior of local ID estimators with increasing dimensionality and neighborhood size. The second family (m) is a benchmark collection of various types of manifolds [56, 1], summarized in the supplement.

5.3 Real Data. The use of real-world data sets lacks the ground truth available for synthetic data. Therefore, to evaluate TLE on such sets, we compare the bias and variance characteristics directly against competing methods using the 8 real data sets listed in Table 1. See supplement for further data set descriptions.

6 Experimental Results

6.1 Synthetic Data. As a first comparative illustration of the various ID estimators, let us consider Fi-

[1] http://perun.pmf.uns.ac.rs/radovanovic/tle/

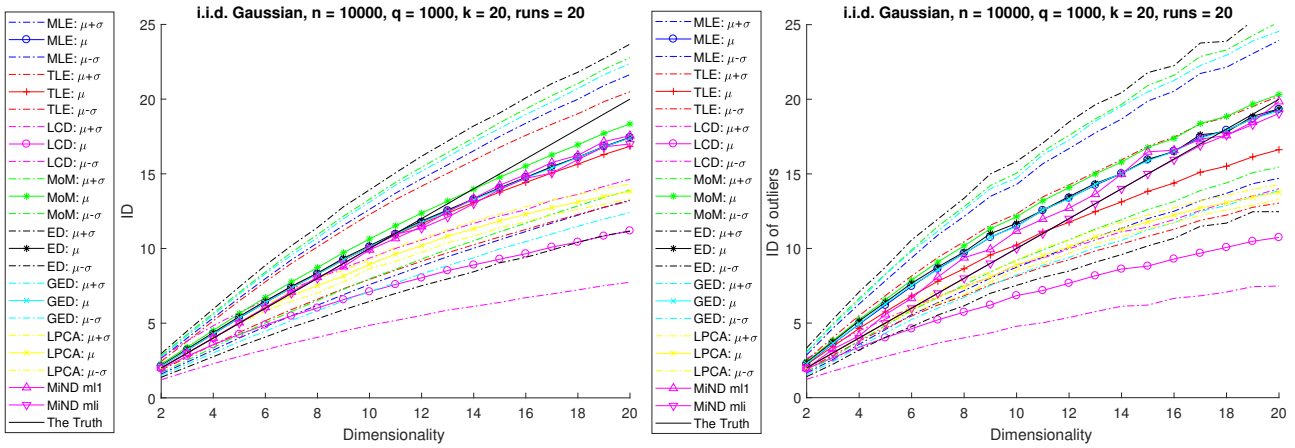

Figure 2: Means and standard deviations of estimated ID values on i.i.d. Gaussian data, for neighborhood size 20 and various dimensionalities. Left: estimates for all data points. Right: estimates for 10% of strongest outliers.

Data set	Instances	Dim.
ALOI [6]	110250	641
ANN_SIFT1M [37]	10^6	128
BCI5 [51]	31216	96
CoverType [5]	581012	54
Gisette [26]	7000	5000
Isolet [15]	7797	617
MNIST [45]	70000	784
MSD [4]	515345	90

Table 1: Real data sets used in the experiments.

gure 2(left), which was obtained by plotting the mean values of ID estimates, together with standard deviations (where appropriate), obtained on 10000 i.i.d. Gaussian random data points with dimensionalities 2–20, and averaged over 20 runs. All methods were executed with neighborhood size $k = 20$. Also plotted is "The Truth," the embedding dimensionality of the data space which is identical to the theoretical true ID of each data point. The general shape of the ID estimate curves suggests that estimators progressively become more biased downward as dimensionality increases compared to the ground truth. It can be seen that TLE consistently exhibits smaller variance, at the same time maintaining bias comparable to the state-of-the art methods, notably MLE and MoM. By not accounting for distances to the neighbors of neighbors that are outside the locality boundary, LCD accumulates clipping bias. Therefore, LCD has considerably more negative bias than TLE. For small to medium dimensionalities (2–10) the slightly stronger bias of TLE actually makes it somewhat more accurate, whereas in higher dimensions (>10) the bias makes it deviate more from the ground truth.

To illustrate another scenario where TLE may offer an advantage, Figure 2(right) shows the same ID estimates plotted only for the 10% of strongest outliers in the same i.i.d. Gaussian random data, where outlierness is determined by the distance to the data distribution center. Interestingly, it can be seen that the plot for TLE stays approximately the same as in the previous figure where the whole data set was used, whereas other methods become notably more biased upward (with variance still being smaller for TLE). We will leave the deeper analysis of this behavior as a point for future work. We observed similar bias/variance trends on i.i.d. uniform and torus data (see supplement).

Besides ground truth dimensionality, another important factor to consider when analyzing the performance of ID estimation methods is the neighborhood size k. Figure 3 shows analogous plots to the previous figure, but with varying k and dimensionality fixed at $d = 10$, on i.i.d. Gaussian and uniform random data. It is evident that negative bias increases with increasing k for all methods. The general trend of TLE having comparable bias and smaller variance is also exhibited here, permitting TLE to be used with smaller values of k, thus at least partially avoiding this source of bias.

6.2 Real Data. Figure 4 shows the distributions of ID estimates for real data sets, with two box plots for each data set, using neighborhood sizes $k = 20, 50$. In addition to supporting the observations made on synthetic data sets, it can be stressed that variance of TLE for $k = 20$ is usually as good as or superior to the variance of other methods for larger values of k (see supplement for box plots with $k = 100$). In the case when some other method (e.g. LCD) exhibits smaller variance, it usually has much worse bias as we

have also seen on synthetic data. An extreme example of this behavior is LPCA, which on real data often exhibits tight variance but positive bias that increases significantly with neighborhood size k.

7 Conclusion

In models such as the Correlation Dimension, pairwise distance measurements have been successfully used in order to estimate global intrinsic dimensionality. However, to the best of our knowledge, none of the existing models of local intrinsic dimensionality take advantage of distances other than those from a test point to the members of its neighborhood. Here we have shown that estimating the Correlation Dimension on small neighborhoods does not lead to a correct ID estimation if all available pairwise distances are used without accounting for the clipping of data to the respective localities.

Our proposed estimation strategy makes use of a subset of the available intra-neighborhood distances to achieve faster convergence with fewer samples, and can thus be used on applications in which the data consists of many natural groups of small size. Moreover, it has a smaller bias and variance than state-of-the-art estimators, especially on nonlinear subspaces. Consequently, the estimator can achieve more accurate ID estimates within a smaller locality than the traditional estimators. This has the potential to improve the quality of algorithms where locality is an important factor, such as subspace clustering and subspace outlier detection, which we plan to investigate in future work.

References

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

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[3] R. Bennett. The intrinsic dimensionality of signal collections. *IEEE TIT*, 15(5):517–525, 1969.

[4] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *ISMIR*, pages 591–596, 2011.

[5] J. A. Blackard and D. J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999.

[6] N. Boujemaa, J. Fauqueur, M. Ferecatu, F. Fleuret, V. Gouet, B. LeSaux, and H. Sahbi. IKONA for interactive specific and generic image retrieval. In *CBMI*, 2001.

[7] M. Brand. Charting a manifold. In *NIPS*, pages 961–968, 2002.

[8] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE TPAMI*, 20(5):572–575, 1998.

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

[10] F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE TPAMI*, 24(10):1404–1407, 2002.

[11] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *DAMI*, 30(4):891–927, 2016.

[12] G. Casanova, E. Englmeier, M. E. Houle, P. Kröger, M. Nett, , E. Schubert, and A. Zimek. Dimensional testing for reverse k-nearest neighbor search. *PVLDB*, 10(7):769–780, 2017.

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

[14] C. K. Chen and H. C. Andrews. Nonlinear intrinsic dimensionality computations. *IEEE Trans. Comput.*, 100(2):178–184, 1974.

[15] R. Cole and M. Fanty. Spoken letter recognition. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, pages 385–390, 1990.

[16] S. Coles, J. Bawa, L. Trenner, and P. Dorazio. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[17] J. A. Costa and A. O. Hero III. Entropic graphs for manifold learning. In *Asilomar Conference on Signals, Systems and Computers*, pages 316–320, 2004.

[18] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. CRC Press, 2000.

[19] T. de Vries, S. Chawla, and M. E. Houle. Density-preserving projections for large-scale local anomaly detection. *Knowl. Inf. Syst.*, 32(1):25–52, 2012.

[20] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. Neural Netw.*, 8(1):148–154, 1997.

[21] D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[22] M. Fan, H. Qiao, and B. Zhang. Intrinsic dimension estimation of manifolds by incising balls. *Pattern Recognition*, 42(5):780–787, 2009.

[23] A. M. Farahmand, C. Szepesvári, and J.-Y. Audibert. Manifold-adaptive dimension estimation. In *ICML*, pages 265–272, 2007.

[24] K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. Comput.*, 100(2):176–183, 1971.

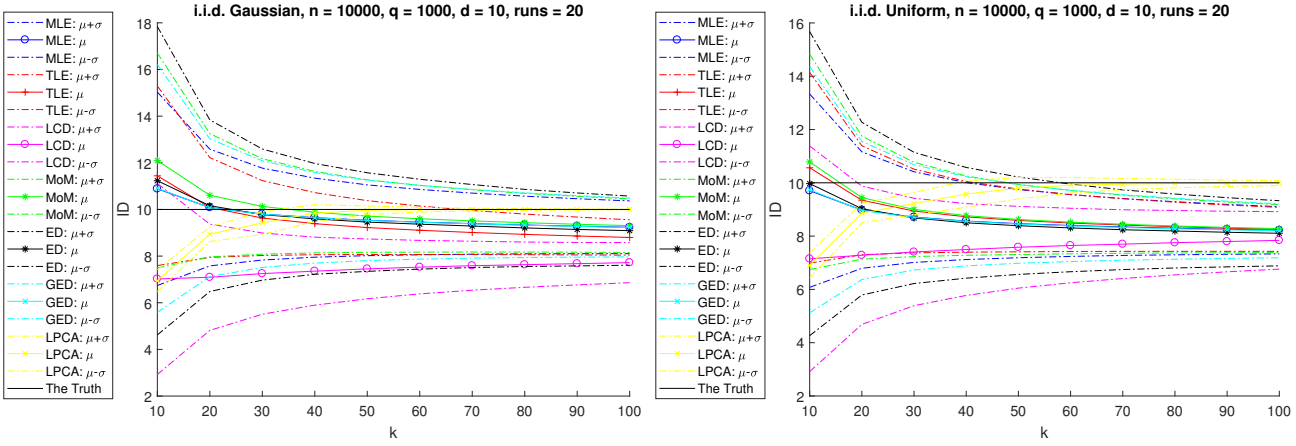

Figure 3: Means and standard deviations of estimated ID values, for dimensionality 10 and various neighborhood sizes. Left: i.i.d. Gaussian data. Right: i.i.d. uniform data.

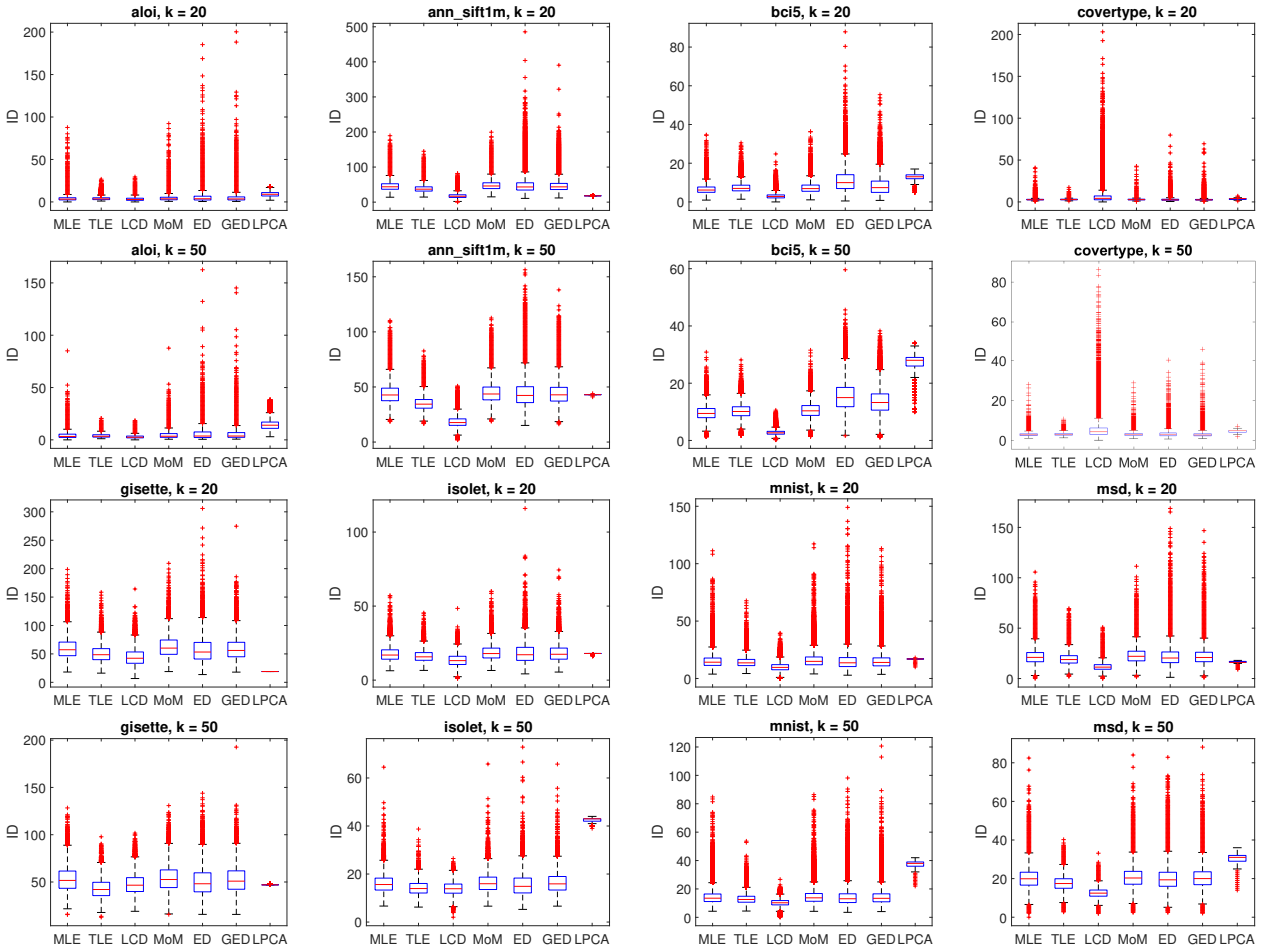

Figure 4: Box plots of estimated ID values, for neighborhood sizes 20 and 50, on real data.

[25] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. In *The Theory of Chaotic Attractors*, pages 170–189. Springer, 2004.

[26] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the NIPS 2003 feature selection challenge. In *NIPS*, pages 545–552, 2004.

[27] J. He, L. Ding, L. Jiang, Z. Li, and Q. Hu. Intrinsic dimensionality estimation based on manifold assumption. *J. Vis. Commun. Image Represent.*, 25(5):740–747, 2014.

[28] M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in R^d. In *ICML*, pages 289–296. ACM, 2005.

[29] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[30] M. E. Houle. Local intrinsic dimensionality I: An extreme-value-theoretic foundation for similarity applications. In *SISAP*, pages 64–79, 2017.

[31] M. E. Houle. Local intrinsic dimensionality II: Multivariate analysis and distributional support. In *SISAP*, pages 80–95, 2017.

[32] M. E. Houle, H. Kashima, and M. Nett. Generalized expansion dimension. In *12th International Conference on Data Mining Workshops*, pages 587–594, 2012.

[33] M. E. Houle, X. Ma, M. Nett, and V. Oria. Dimensional testing for multi-step similarity search. In *ICDM*, pages 299–308. IEEE, 2012.

[34] M. E. Houle, X. Ma, V. Oria, and J. Sun. Efficient algorithms for similarity search in axis-aligned subspaces. In *SISAP*, pages 1–12. Springer, 2014.

[35] M. E. Houle, V. Oria, and A. M. Wali. Improving k-NN graph accuracy using local intrinsic dimensionality. In *SISAP*, pages 110–124, 2017.

[36] M. E. Houle, E. Schubert, and A. Zimek. On the correlation between local intrinsic dimensionality and outlierness. In *SISAP*, pages 177–191, 2018.

[37] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE TPAMI*, 22(1):117–128, 2011.

[38] K. Johnsson, C. Soneson, and M. Fontes. Low bias local intrinsic dimension estimation from expected simplex skewness. *IEEE TPAMI*, 37(1):196–202, 2015.

[39] I. T. Jolliffe. Principal Component Analysis. *New York*, 487, 1986.

[40] D. R. Karger and M. Ruhl. Finding nearest neighbors in growth-restricted metrics. In *ACM Symposium on Theory of Computing*, pages 741–750. ACM, 2002.

[41] J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *IEEE Trans. Neural Netw.*, 7(1):113–127, 1994.

[42] B. Kégl. Intrinsic dimension estimation using packing numbers. In *NIPS*, pages 681–688, 2002.

[43] T. Kohonen. Learning vector quantization. In *Self-Organizing Maps*, pages 175–189. Springer, 1995.

[44] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[45] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[46] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *NIPS*, pages 777–784, 2004.

[47] A. V. Little, Y.-M. Jung, and M. Maggioni. Multiscale estimation of intrinsic dimensionality of data sets. In *AAAI Fall Symposium: Manifold Learning and its Applications*, page 04, 2009.

[48] A. V. Little, M. Maggioni, and L. Rosasco. Multiscale geometric methods for data sets I: Multiscale SVD, noise and curvature. *Applied and Computational Harmonic Analysis*, 2016.

[49] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. N. R. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *ICLR*, pages 1–15, 2018.

[50] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. M. Erfani, S.-T. Xia, S. N. R. Wijewickrema, and J. Bailey. Dimensionality-driven learning with noisy labels. In *ICML*, pages 3361–3370, 2018.

[51] J. del R. Millán. On the need for on-line learning in brain-computer interfaces. In *IJCNN*, volume 4, pages 2877–2882. IEEE, 2004.

[52] E. Ott. *Chaos in dynamical systems*. Cambridge University Press, 2002.

[53] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[54] M. Raginsky and S. Lazebnik. Estimation of intrinsic dimensionality using high-rate vector quantization. In *NIPS*, pages 1105–1112, 2005.

[55] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[56] A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli. Novel high intrinsic dimensionality estimators. *Machine Learning*, 89(1-2):37–65, 2012.

[57] A. Rozza, G. Lombardi, M. Rosa, E. Casiraghi, and P. Campadelli. Idea: intrinsic dimension estimation algorithm. In *ICIAP*, pages 433–442. Springer, 2011.

[58] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, 5:401–409, 1969.

[59] F. Takens. *On the Numerical Determination of the Dimension of an Attractor*. Springer, 1985.

[60] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[61] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[62] P. J. Verveer and R. P. W. Duin. An evaluation of intrinsic dimensionality estimators. *IEEE TPAMI*, 17(1):81–86, 1995.