ISSN 1346-5597

National Institute of Informatics

NII Technical Report

Enhanced Estimation of Local Intrinsic Dimensionality Using Auxiliary Distances

Oussama Chelly, Michael E. Houle, Ken-ichi Kawarabayashi

NII-2016-007E
Aug. 2016

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

Enhanced Estimation of Local Intrinsic Dimensionality Using Auxiliary Distances

Oussama Chelly, Michael E. Houle, Ken-ichi Kawarabayashi
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda
Tokyo 103-8430, Japan
chelly@nii.ac.jp, meh@nii.ac.jp, k_keniti@nii.ac.jp

Abstract

Estimating Intrinsic Dimensionality (ID) is of high interest in many machine learning tasks, including dimensionality reduction, outlier detection, similarity search and subspace clustering. Our proposed estimation strategy, ALID, 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. We provide a theoretical analysis of the properties of the ALID estimator, and a thorough experimental framework that shows its faster convergence, smaller bias, and smaller variance compared with state-of-the-art estimators of ID.

1 Introduction

Over the past decades, many characterizations of the intrinsic dimensionality (ID) of data sets have been proposed, each with its own estimators. Topological models estimate the basis dimension of the tangent space of the data manifold [8, 14, 30]. This class of estimators includes Principal Component Analysis (PCA) and its variants [8, 14, 22], and multidimensional scaling (MDS) [3, 9]. Graph-based methods attempt to preserve the k-nearest neighbor graph [9, 12]. Fractal models, very popular in physics applications, are used to estimate the dimension of nonlinear systems — these include the popular estimators due to Hein [17], Takens [29], and Grassberg & Procaccia [15]. In addition to more traditional areas such as manifold learning and feature extraction, ID has found application in the design and analysis of similarity search methods [20, 23] and outlier detection methods [13].

The aforementioned estimators can be described as 'global', in that they provide a single ID measurement for the full dataset, as opposed to 'local' ID estimators that assign a different dimensionality to each point in the dataset. Commonly-used local estimators of ID include: topological methods that measure the dimension of the space locally tangent to the manifold, such as locally linear embedding [27], Laplacian and Hessian eigenmaps, and Brand's Method [7]; measures of the rate of expansion of the neighborhood size with increasing radius [19, 23]; and probabilistic methods that view the data as a sample from a hidden distance distribution, such as the Hill estimator [18], Levina and Bickel's algorithm [25], the minimum neighbor distance (MiND) framework [28], and the local intrinsic dimensionality (LID) framework [2].

A global estimator can be adapted for local estimation of ID simply by applying it to the subset of the data lying within some region surrounding a point of interest. Global methods typically make use of many (if not most or all) of the pairwise relationships within the data; however, 'clipping' of the data set to a region, by discounting some of these relationships while preserving others, may lead to estimation bias whenever the boundary shape is not properly accounted for in the ID model or estimation strategy. On the other hand, implicit in their design, local estimators of ID avoid the negative affect of clipping, by considering only the direct relationships between a reference point and

its nearest neighbors. The sample boundary is usually set to the distance from the reference point to the farthest object in the neighborhood. With this distinction in mind, application of global estimators within the neighborhood of a given reference point should not be regarded as truly 'local'.

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 natural groups within the data are often too small to provide the number of samples necessary for accurate estimation of ID — in the LID framework, for example, approximately one hundred distance values are usually required for convergence [2]. Simply choosing a number of samples sufficient for the convergence of the estimator can lead to a violation of the locality constraint, as the sample could consist of points from several different natural groups, each with their own intrinsic dimensionalities. When the cluster memberships and size are not known in advance, in order to ensure that the majority of the points are drawn from the same group, it is necessary to use estimators that can cope with the smallest possible sample sizes [2, 28]. Thus, the development of local ID estimators with faster convergence properties is essential for the effectiveness and the efficiency of subspace-based applications.

One possible strategy for improving the convergence properties of estimation without violating locality is to draw more measurements from smaller data samples — however, for the case of distance-based local estimation from neighborhood samples, this would require the use of distances between pairs of neighbors, and not merely the distances from the reference point to its neighbors. Indeed, the global distance-based correlation dimension (CD) [29], if restricted to a neighborhood, would use all pairwise distances within the neighborhood to achieve its estimate. Although for a given neighborhood size this local use of CD would be expected to converge much faster than true local ID estimators, the result would be biased due to the clipping.

In this paper, we show that the convergence properties of LID estimation can be improved by augmenting it with distance measurements from members of the neighbor set to their own nearest neighbors. The sizes of these 'auxiliary' neighborhoods is restricted so that they are completely contained within the original, 'primary' neighborhood, thus preserving the locality of the estimation. Within a given primary neighborhood of k elements, the number of distance measurements thus could range between a minimum of k and a maximum of $k(k+1)/2$. We show that under certain assumptions, the number of measurements available depends on the local ID itself, with the greatest number of auxiliary distance measurements being available when the ID is small. The main contributions of this paper include:

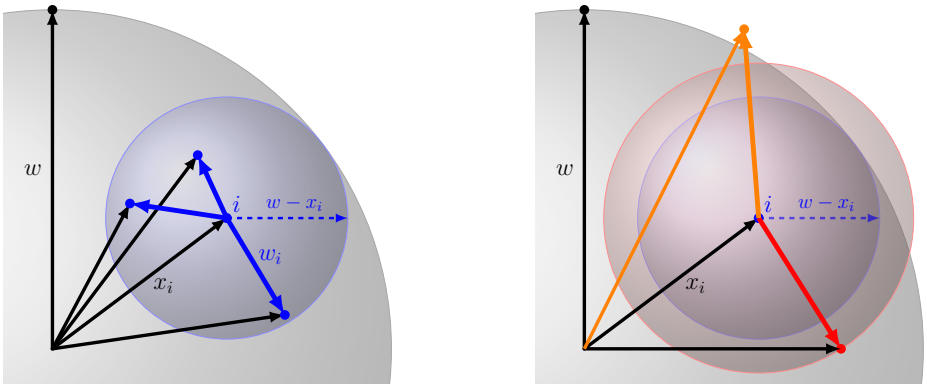
- the augmented local ID estimator, ALID;
- for the case of uniform data distributions in Euclidean space, a theoretical analysis of the expected number of auxiliary distances available in terms of ID;
- an experimental comparison of the bias, variation, and convergence properties of ALID with LID and other local and global estimators of ID, on both synthetic and real data sets.

The remainder of the paper is structured as follows. In Section 2 we introduce our proposed estimator, and present a theoretical analysis relating the expected number of auxiliary distances with the intrinsic dimensionality. In Section 3 our experimental framework is described in detail, and in Section 4 we present our experimental comparison of ALID with existing local and global ID estimators. In this latter section we also validate our theoretical analysis empirically, by showing the number of auxiliary measurements available and comparing them to the numbers predicted by the theory. We conclude the discussion in Section 5.

2 Augmented Local ID Estimation

2.1 Local ID estimation and extreme value theory

The LID model has its foundations in extreme value theory (EVT), which is concerned with the asymptotic behavior of continuous distributions in their extreme tails [2]. The choice of a neighborhood of radius w based at the test point is equivalent to the lower tail of the distribution of distances on $[0, w)$. Given a distance variable X and using the transformation $\mathbf{Y} = -\mathbf{X}$, the distribution of the distance excess $\mathbf{Y} - (-w)$ (conditioned on $\mathbf{Y} > -w$) over the threshold w tends to a distribution in the family of Generalized Pareto Distributions $\mathcal{F}_{\text{GPD}}$, as w tends to the lower endpoint of the cumulative distribution function F_X. Accordingly, as w tends to zero, the distribution in the tail $[0, w)$ can be restated as follows [2, 11].



(a) Pairwise distances that remain within internally tangent balls can be used in the ID estimation without introducing distortions. In this figure we consider only one nearest neighbor (i) and the corresponding usable auxiliary distances (in blue).

(b) Neither auxiliary nor direct distances (in orange) to neighbors that are outside the locality can be used. Moreover, auxiliary distances where the corresponding ball (in red) crosses over the original locality can not be used for the estimation.

Figure 1: State-of-the-art local ID estimators use only direct distances (in black). The proposed estimator $\mathrm{ID}_{\mathrm{ALID}}$ uses additional distances between pairs of neighbors. Some of these distances (in blue) can be used, while others (in orange and red) cannot.

Lemma 1. *Let* $\mathbf{X}$ *be an absolutely continuous random distance variable with support* $[0, \infty)$ *and cumulative distribution function* F_X *such that* $F_X(x) > 0$ *if* $x > 0$*. Let* $c \in (0, 1)$ *be an arbitrary constant. Let* $w > 0$ *be a distance threshold, and consider* x *restricted to the range* $[cw, w)$*. As* w *tends to zero, the distribution of* $\mathbf{X}$ *restricted to the tail* $[cw, w)$ *satisfies, for some fixed* $\xi < 0$*:*

$$\frac{(x/w)^{-\frac{1}{\xi}}}{F_{X,w}(x)} \to 1$$

The parameter $\xi < 0$ is related to the LID through the following theorem [2]:

Theorem 1. *Let* $\mathbf{X}$ *be an absolutely continuous random distance variable with support* $[0, \infty)$*, and* $w > 0$ *be a distance threshold. Then, as* w *tends to zero,*

$$\mathrm{ID}_{\mathbf{X}}(w) \to -\frac{1}{\xi} =: \mathrm{ID}_{\mathbf{X}}.$$

Lemma 1 and Theorem 1 allow us to approximate the asymptotic cumulative distribution of distances in the tail $[cw, w)$ as

$$\frac{(x/w)^{\mathrm{ID}_{\mathbf{X}}}}{F_{X,w}(x)} \to 1. \tag{1}$$

Maximum Likelihood Estimation (MLE) is a popular method in statistics for estimating parameters of probability distributions. The MLE has no optimality guarantees for finite samples, but is shown to be asymptotically consistent, optimal, and efficient. For a given sample of neighborhood distances $x_1, x_2, \ldots, x_k$ following the asymptotic distance distribution given in Equation 1, the MLE (or Hill) estimate $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ is [2, 18]

$$\widehat{\mathrm{ID}}_{\mathrm{MLE}} = -\left(\frac{1}{k}\sum\nolimits_{i=1}^{k} \ln\frac{x_i}{w}\right)^{-1}.$$

The variance is asymptotically given by the inverse of the Fisher information, defined as

$$I = E[-\frac{\partial^2 \mathcal{L}(\mathrm{ID}_{\mathbf{X}})}{\partial\, \mathrm{ID}_{\mathbf{X}}^2}] = \frac{n}{\mathrm{ID}_{\mathbf{X}}^2},$$

where $E[\cdot]$ denotes the expectation.

2.2 MLE estimation for ALID

Global estimators based on correlation dimension use the smallest pairwise distances within the data in order to measure the global ID. In particular, the Takens estimator [29] uses all pairwise distances within balls of a fixed radius and evaluates ID using the same Hill estimator. With this approach,

intracluster distances are likely to dominate intercluster distances that may occur whenever the radius is too high.

Restricting the computation of correlation dimension to a neighborhood is not a satisfactory estimation strategy for local ID. Consider a neighbor i at distance x_i from the center, where the radius of the neighborhood is w. To avoid the negative effects of clipping, ALID makes use of an auxiliary distance measurement $x_{i,j}$ from i only if the ball of radius $x_{i,j}$ centered at i is entirely contained within the original neighborhood (see Figure 1). This condition can be stated as $x_{i,j} \leq w - x_i$.

The proposed auxiliary-distance estimator ($\mathrm{ID}_{\mathrm{ALID}}$), like the Hill estimator ($\mathrm{ID}_{\mathrm{MLE}}$), is based on the method of maximum likelihood estimation. We assume that the ID is constant within a small radius w around the supplied test point.

Let X_i be the random distance variable from the neighbor i in the range $[0, w - x_i)$, and let $f_{X_i,w-x_i}$ and $F_{X_i,w-x_i}$ be respectively the pdf and cdf associated with X_i. To simplify the notation, we assign the rank $i = 0$ to the test point. The log-likelihood function is:

$$\begin{aligned}
\mathcal{L}(\mathrm{ID}_{\mathbf{X}}) &= \ln\left[\prod_{\substack{x_{i,j}+x_i<w \\ i,j\in[0,k]}} f_{X_i,w-x_i}(x_{i,j})\right] \\
&= \ln\left[\prod_{\substack{x_{i,j}+x_i<w \\ i,j\in[0,k]}} \mathrm{ID}_{\mathbf{X}} \frac{F_{X_i,w-x_i}(w-x_i)}{w-x_i}\left(\frac{x_{i,j}}{w-x_i}\right)^{\mathrm{ID}_{\mathbf{X}}-1}\right] \\
&= (k+\rho(w))\cdot \mathrm{ID}_{\mathbf{X}} \\
&\quad + (\mathrm{ID}_{\mathbf{X}}-1)\sum_{\substack{x_{i,j}+x_i<w \\ i,j\in[0,k]}} \ln\left[\frac{x_{i,j}}{w-x_i}\right] + \sum_{\substack{x_{i,j}+x_i<w \\ i,j\in[0,k]}} \ln\left[\frac{F_{X_i,w-x_i}(w-x_i)}{w-x_i}\right],
\end{aligned}$$

where $\rho(w) = \sum_{i,j\in[1,k]} \mathbb{1}[x_{i,j} + x_i < w]$ denotes the number of auxiliary distances used in the estimation. Accordingly, our auxiliary-distance MLE estimator is

$$\widehat{\mathrm{ID}}_{\mathrm{ALID}} = -\left(\frac{1}{k+\rho(w)}\sum_{\substack{x_{i,j}<w-x_i \\ i,j\in[0,k]}} \ln\left[\frac{x_{i,j}}{w-x_i}\right]\right)^{-1}. \tag{2}$$

The number of available auxiliary distance measurements $\rho(w)$ varies from data set to data set, and even from one locality within the set to another. However, under certain simplifying assumptions, it is possible to show that this quantity depends on the local intrinsic dimensionality. If the data distribution is locally uniform in the vicinity of the test point, the expected number of points within a volume would be proportional to the volume itself. Accordingly, the following theorem determines the cumulative volume of all maximal ball placements centered at locations within a neighborhood ball — or in other words, the cumulative volume of all internally tangent balls.

Theorem 2. *In a Euclidean manifold of dimensionality α, let us consider a ball of radius w, and volume $V_\alpha(w)$. The total volume of all internally tangent balls is:*

$$\rho_\alpha(w) = \frac{V_\alpha(w)^2}{2}\cdot\frac{\Gamma(\alpha)\Gamma(\alpha+1)}{\Gamma(2\alpha)}.$$

Proof. In order to measure the total volume of all internally tangent balls, it is possible to integrate the volumes of all balls of volume $V_\alpha(w-r)$ with centers located on the surface of a sphere of radius r, over values of $r \in [0, w]$. The total volume is given by

$$\rho_\alpha(w) = \int_0^w A_\alpha(r)\cdot V_\alpha(w-r)\,dr,$$

where $A_\alpha(r)$ is the surface area of a sphere of radius r in a manifold of intrinsic dimensionality α.

Given that the manifold is Euclidean, the surface area and volume formulas are:

$$A_\alpha(r) = \frac{2\pi^{\alpha/2}}{\Gamma(\frac{\alpha}{2})}r^{\alpha-1}, \quad \text{and} \quad V_\alpha(r) = \frac{\pi^{\alpha/2}}{\Gamma(\frac{\alpha}{2}+1)}r^\alpha.$$

By replacing $A_\alpha(r)$ and $V_\alpha(w-r)$ in the expression of $\rho_\alpha(w)$, we obtain:

$$\begin{aligned}\rho_\alpha(w) &= \frac{2\pi^\alpha}{\Gamma(\frac{\alpha}{2})\cdot\Gamma(\frac{\alpha}{2}+1)}\cdot\int_0^w r^{\alpha-1}\cdot(w-r)^\alpha\,dr \\ &= \frac{\alpha\pi^\alpha}{\Gamma(\frac{\alpha}{2}+1)^2}\cdot\int_0^w r^{\alpha-1}\cdot(w-r)^\alpha\,dr \\ &= \frac{\alpha V_\alpha(w)^2}{w^{2\alpha}}\cdot\int_0^w r^{\alpha-1}\cdot(w-r)^\alpha\,dr.\end{aligned}$$

It therefore suffices to show that

$$\int_0^w r^{\alpha-1}\cdot(w-r)^\alpha\,dr = \frac{w^{2\alpha}}{2}\cdot\frac{\Gamma(\alpha)^2}{\Gamma(2\alpha)}.$$

The variable change $r = u + \frac{w}{2}$ yields

$$\begin{aligned}\int_0^w r^{\alpha-1}\cdot(w-r)^\alpha\,dr &= \int_{-\frac{w}{2}}^{\frac{w}{2}}\left(\frac{w}{2}+u\right)^{\alpha-1}\cdot\left(\frac{w}{2}-u\right)^\alpha du \\ &= \int_{-\frac{w}{2}}^{\frac{w}{2}}\left(\frac{w^2}{4}-u^2\right)^{\alpha-1}\cdot\left(\frac{w}{2}-u\right)du \\ &= \int_{-\frac{w}{2}}^{\frac{w}{2}}\frac{w}{2}\left(\frac{w^2}{4}-u^2\right)^{\alpha-1}du - \int_{-\frac{w}{2}}^{\frac{w}{2}}u\left(\frac{w^2}{4}-u^2\right)^{\alpha-1}du.\end{aligned}$$

Noting that the function in the first integral is even while the function in the second integral is odd, we use the variable change $u = \frac{w}{2}\sin\theta$ to obtain

$$\begin{aligned}\int_0^w r^{\alpha-1}\cdot(w-r)^\alpha\,dr &= w\int_0^{\frac{w}{2}}\left(\frac{w^2}{4}-u^2\right)^{\alpha-1}du \\ &= \frac{w^2}{2}\int_0^{\frac{\pi}{2}}\left(\frac{w^2}{4}-\frac{w^2}{4}\sin^2\theta\right)^{\alpha-1}\cos\theta\,d\theta \\ &= 2\left(\frac{w}{2}\right)^{2\alpha}\int_0^{\frac{\pi}{2}}\cos^{2\alpha-1}\theta\,d\theta.\end{aligned}$$

Using the following relationship between Euler's β and Γ functions [1],

$$\beta(x,y) \triangleq \int_0^{\frac{\pi}{2}}\cos^{2x-1}\theta\cdot\sin^{2y-1}\theta\,d\theta = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

we conclude that:

$$\begin{aligned}\int_0^w r^{\alpha-1}\cdot(w-r)^\alpha\,dr &= \left(\frac{w}{2}\right)^{2\alpha}\frac{\Gamma(\alpha)\Gamma(\frac{1}{2})}{\Gamma(\alpha+\frac{1}{2})} \\ &= \sqrt{\pi}\left(\frac{w}{2}\right)^{2\alpha}\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})} \\ &= \frac{w^{2\alpha}}{2}\cdot\frac{\Gamma(\alpha)^2}{\Gamma(2\alpha)}.\end{aligned}$$

□

Under the assumption that the expected number of points in a volume is proportional to the volume itself, Theorem 2 implies that the expected number of distances $\rho(w)$ in a neighborhood of radius $w = x_k$ is $\frac{k^2}{2}\frac{\Gamma(\mathrm{ID})\Gamma(\mathrm{ID}+1)}{\Gamma(2\mathrm{ID})}$.

2.3 Complexity of the auxiliary-distance ID estimator

The auxiliary-distance estimator has a higher complexity than most local estimators of ID. From Theorem 2 and under the same assumptions, we can infer that $C_{\mathrm{ID}_{\mathrm{ALID}}} = O(k \cdot (1 + k\frac{\Gamma(\mathrm{ID})\Gamma(\mathrm{ID}+1)}{\Gamma(2\mathrm{ID})}))$. Thus, the complexity is linear when the estimated ID is high, matching the complexity of $\mathrm{ID}_{\mathrm{MLE}}$ and $\mathrm{ID}_{\mathrm{MoM}}$. When the estimated ID is low, $C_{\mathrm{ID}_{\mathrm{ALID}}}$ becomes quadratic in the number of neighbors like $\mathrm{ID}_{\mathrm{GED}}$ or Levina & Bickel estimator.

Complexity considerations may restrict the use of our estimator in situations that do not require the computation of ID for all data but for specific query points only. Nonetheless even within algorithms where ID has to be estimated for all data, our estimator can be used without increasing the overall asymptotic computational costs as long as these costs are higher than linear in the data volume. In particular, such algorithms include all applications where the full distance matrix is evaluated (for example in order to compute the nearest neighbors).

3 Experimental framework

Method	Parameters
$\mathrm{ID}_{\mathrm{ALID}}$	$k = 100$
$\mathrm{ID}_{\mathrm{MLE}}$ [2]	$k = 100$
$\mathrm{ID}_{\mathrm{MoM}}$ [2]	$k = 100$
kNNG [12]	$k = 100$, $\gamma = 1$, $M = 1$, $N = 10$
l-PCA [22]	$k = 100$, $\theta = 0.025$
MiND_{ml1} [28]	None
MiND_{mli} [28]	$k = 100$
PCA [22]	$\theta = 0.025$

Table 1: Parameter choices for the methods used in the experiments.

Manifold	d	D	Description
h-d	d	d	Uniformly sampled hypercube.
m1	10	11	Uniformly sampled sphere.
m2	3	5	Affine space.
m3	4	6	Concentrated figure confusable with a 3d one.
m4	4	8	Non-linear manifold.
m5	2	3	2-d Helix
m6	6	36	Non-linear manifold.
m7	2	3	Swiss-Roll.
m8	12	72	Non-linear manifold.
m9	20	20	Affine space.
m10a	10	11	Uniformly sampled hypercube.
m10b	17	18	Uniformly sampled hypercube.
m10c	24	25	Uniformly sampled hypercube.
m11	2	3	Möbius band 10-times twisted.
m12	20	20	Isotropic multivariate Gaussian.
m13	1	13	Curve.

Table 2: Artificial datasets used in the experiments.

3.1 Competing estimation methods

In this framework, to show the advantages and limitations of ALID, we compared our proposed estimator $\mathrm{ID}_{\mathrm{ALID}}$ with other popular estimators, both local and global. The fractal methods used in our experiments (Grassberger-Procaccia's Correlation Dimension (CD), Hein, and Takens) do not require any parameters to be set, while the parameter choices for the remaining methods are summarized in Table 1. We denote by l-PCA the estimator obtained by applying PCA on the respective neighborhoods of size $k = 100$.

It must be noted that PCA variants and methods from the MiND family must be provided with knowledge of the representational dimension, which may give them an advantage in head-to-head comparison with other methods. Moreover, when applied to synthetic data sets, PCA variants and $\mathrm{MiND}_{\mathrm{mli}}$ can often return the exact dimension, since they can return only integer-valued estimates. While it may be claimed that the intrinsic dimension should ideally be an integer, for real data this is not always the case. For example, LID has been shown to be equivalent to a measure of the indiscriminability of the distance measure, which is in general not an integer [2]. Furthermore, non-integer values of ID can indicate non-linear properties of an underlying manifold, such as convexity.

3.2 Synthetic data

Our study includes two families of synthetic datasets. For each manifold we generated 20 sets of 10^3 and 10^4 points, and in each experiment we report the average ID measures over the 20 sets. The first family (h) is a set of hypercubes meant to evaluate the convergence of local ID estimators. The second (m) is a benchmark of various types of manifolds [2, 28].

3.3 Real data

The use of real-world datasets lacks the ground truth available for synthetic data. Therefore, to evaluate our proposed estimator on such sets, we must compare the convergence, bias, and variance characteristics directly against competing methods. In particular, we test the consistency of $\mathrm{ID}_{\mathrm{ALID}}$ for the same suite of experiments provided for $\mathrm{ID}_{\mathrm{MLE}}$ in [2], using the 8 real datasets listed in Table 3.

Dataset	Instances	Dimension	Classes
ALOI [6]	110250	641	1000
ANN_SIFT1B [21]	10^9	128	$3 \cdot 10^7$
BCI5 [26]	31216	96	3
CoverType [5]	581012	54	7
Gisette [16]	7000	5000	2
Isolet [10]	7797	617	26
MNIST [24]	70000	784	10
MSD [4]	515345	90	90

Table 3: Artificial datasets used in the experiments.

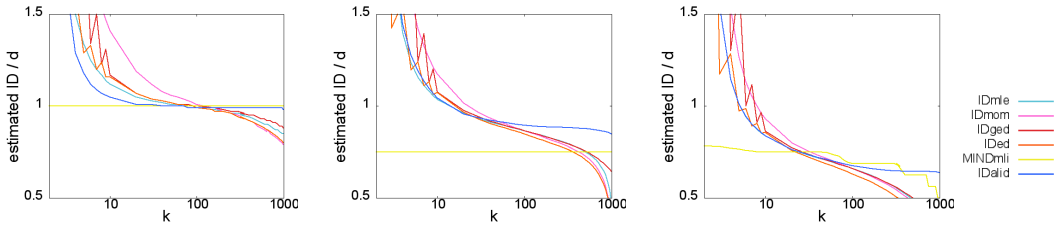
- The *ALOI* (*Amsterdam Library of Object Images*) data consists of 110250 color photos of 1000 different objects. Photos are taken from varying angles under various illumination conditions. Each image is described by a 641-dimensional vector of color and texture features [6].
- The *ANN_SIFT1B* dataset consists of 10^9 128-dimensional SIFT descriptors randomly selected from the dataset *ANN_SIFT* which contains $2.8 \cdot 10^{10}$ SIFT descriptors extracted from $3 \cdot 10^7$ images. These sets have been created for the evaluation of nearest-neighbor search strategies at very large scales [21].
- *BCI5* [26] is a brain-computer interface dataset in which the classes correspond to brain signal recordings taken while the subject contemplated one of three different actions (movement of the right hand, movement of the left hand, and the pronunciation of words beginning with the same letter).
- *CoverType* [5] consists of 581012 geographical locations (a surface of 30 by 30 meters) described by 54 attributes. each location is majorly covered by one of seven tree species.
- *Gisette* [16] is a subset of the MNIST handwritten digit image dataset [24], consisting of 50-by-50-pixel images of the highly confusable digits '4' and '9'. 2500 random features were artificially generated and added to the original 2500 features, so as to embed the data into a higher-dimensional feature space.
- *Isolet* [10] is a set of 7797 human voice recordings in which 150 subjects read each of the 26 letters of the alphabet twice. Each entry consists of 617 features representing utterances of the recording.
- The *MNIST* database [24] contains of 70000 recordings of handwritten digits. The images have been normalized and discretized to a 28×28-pixel grid. The gray-scale values of the resulting 784 pixels are used to form the feature vectors.
- *MSD* [4] is a subset of the 'Million Song Database' which is a set of radio recordings (from the years 1922 to 2011) described by 12 timbre averages and 78 timbre covariances.

For all of these datasets, we set the locality parameter k of both $\mathrm{ID}_{\mathrm{MLE}}$ and $\mathrm{ID}_{\mathrm{ALID}}$ to 100. We extend our experiment using the ALOI dataset by studying the effect of varying k in $\{50, 100, 200, 400\}$ on the estimated ID values.

4 Results

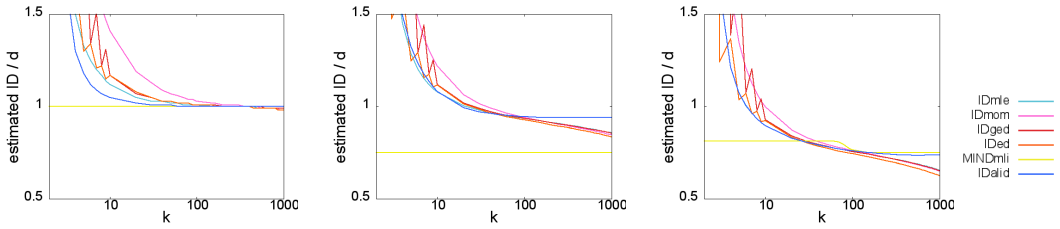
4.1 Experiments with synthetic data

In Figures 2 and 3, we examine the convergence properties of the local ID estimators on two artificial data sets. As the neighborhood size k increases, $\mathrm{ID}_{\mathrm{ALID}}$ is the first estimator to stabilize. For the



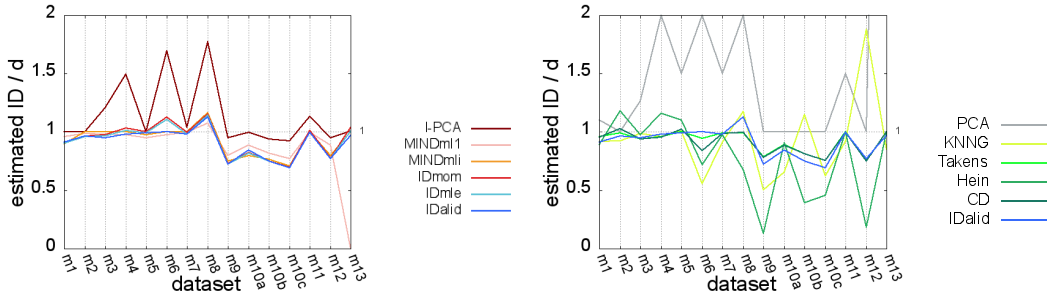
(a) h1 (b) h4 (c) h16

Figure 2: Convergence of local ID estimators in 1000-point-sets uniformly sampled from d-dimensional hypercubes.



(a) h1 (b) h4 (c) h16

Figure 3: Convergence of local ID estimators in 10000-point-sets uniformly sampled from d-dimensional hypercubes.



(a) Local ID estimators (b) Global ID estimators

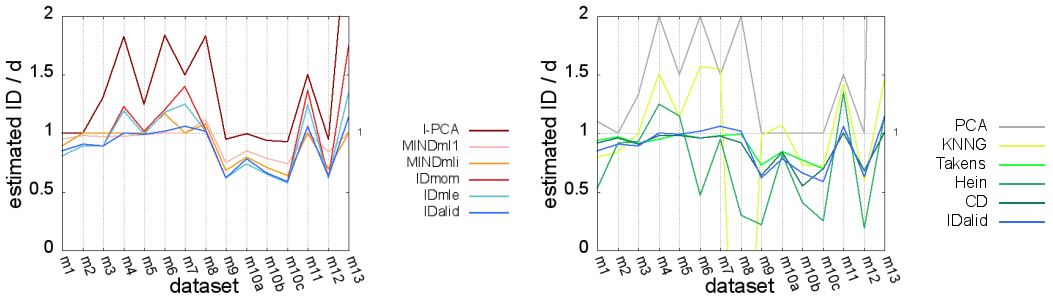
Figure 4: Comparison of $\mathrm{ID}_{\mathrm{ALID}}$ with state-of-the-art ID estimators on 10000-point manifolds of various dimensionalities.

lower-dimensional manifolds, $\mathrm{ID}_{\mathrm{MLE}}$ requires in the order of 100 neighbors to converge [2], whereas the many auxiliary distance measurements allow $\mathrm{ID}_{\mathrm{ALID}}$ to converge much faster — it requires fewer than 10 neighbors to draw within 10% of the true dimensionality. Meanwhile, as predicted by Theorem 2, as the dimensionality increases , the performance of $\mathrm{ID}_{\mathrm{ALID}}$ tends to that of $\mathrm{ID}_{\mathrm{MLE}}$.

For the experiment shown in Figure 3, we evaluated the cumulative absolute error $e = \int_{k=2}^{k=1000} (\widehat{\mathrm{ID}}/d)\ d \log k$ (the normalized difference between the estimate and the true ID value). For data set h1, $\mathrm{ID}_{\mathrm{ALID}}$ has the smallest error (8.78), with $\mathrm{ID}_{\mathrm{MLE}}$ coming in second (9.28). As the dimensionality increases, $\mathrm{ID}_{\mathrm{ALID}}$ converges to $\mathrm{ID}_{\mathrm{MLE}}$, since fewer legal auxiliary distances are encountered in the neighborhood. This is reflected in the respective errors achieved for h4 (7.50 and 7.54) and h16 (7.46 and 7.54).

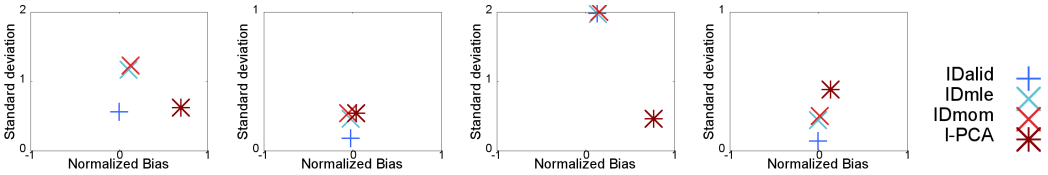
Overall, the results lead us to two conclusions: (i) our estimator converges faster than its competitors, and (ii) our estimator is amongst the least affected when the neighborhood size k is large.

In the second experiment, we estimated the ID on various types of manifolds, with different dimensionalities as summarized in Table 3. As shown in Figure 5, local estimators consistently underestimate the dimensionality on linear manifolds (m1, m2, m9, m10, and m12), due to clipping bias. However, local estimators tend to overestimate the dimensionality of nonconvex manifolds (m7, m11, and m13). In both cases, this bias is reduced as the sample sizes increase (Figure 4a). As shown in Figure 6, on nonlinear and nonconvex manifolds, $\mathrm{ID}_{\mathrm{ALID}}$ has the smallest bias and variance, with the exception



(a) Local ID estimators (b) Global ID estimators

Figure 5: Comparison of $\mathrm{ID}_{\mathrm{ALID}}$ with state-of-the-art ID estimators on 1000-point manifolds of various dimensionalities.



(a) m6 (b) m7 (c) m8 (d) m11

Figure 6: Bias and standard deviation of local ID estimators on nonconvex and nonlinear manifolds.

of $\mathrm{MiND}_{\mathrm{mli}}$ on linear manifolds, due to its advantage in having been provided the representational dimension.

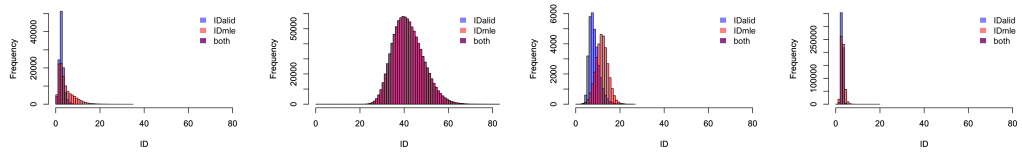
In convex and linear manifolds, l-PCA appears to provide consistently accurate estimates with the least bias and variance. However, in realistic data where the manifolds are not convex, probabilistic local ID methods provide the best trade-off (*c.f.* Figure 6). When PCA is used locally, the variance along a given component coincides with the global variance only if the manifold is linear and homogeneous. Whenever the manifold is nonlinear or nonconvex, the local components are very likely to be different from the global components, due to clipping.

Global estimators can be split into two groups based on the experimental results shown in figure 4b. Topological estimators (PCA) return the exact dimensionality only when the manifold is linear. However they always overestimate the ID on nonlinear manifolds, and are easily mistaken when the manifold is nonconvex. The remaining global estimators tend to behave similarly to local estimators in both their dependency on the linearity and convexity and their better estimation with larger samples.

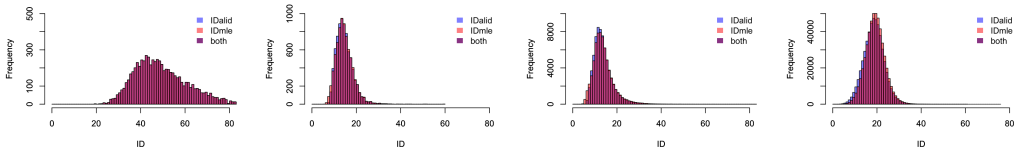
4.2 Experiments with real-world data

As a first step, we evaluated ID on 8 publicly available datasets using $\mathrm{ID}_{\mathrm{MLE}}$ and $\mathrm{ID}_{\mathrm{ALID}}$ (see Figure 7). In all of the real-world datasets, the results are consistent with the theory, in that the estimates of $\mathrm{ID}_{\mathrm{ALID}}$ are much sharper than those of $\mathrm{ID}_{\mathrm{MLE}}$ when the ID is small, but tend to those of $\mathrm{ID}_{\mathrm{MLE}}$ as ID increases.

In a second experiment, we show the stability and robustness of $\mathrm{ID}_{\mathrm{ALID}}$ across various values of k, as compared to $\mathrm{ID}_{\mathrm{MLE}}$. Figure 8 shows the ID estimates on the ALOI data set, which consists of 1000 image classes of size approximately 110. The proportion of $\mathrm{ID}_{\mathrm{ALID}}$ estimates smaller than 4 consistently increases with k from 83% when $k = 50$ to 94% when $k = 400$. Meanwhile, $\mathrm{ID}_{\mathrm{MLE}}$ estimates in the range $[0, 4]$ decrease from 61% when $k = 50$ down to to 27% when $k = 400$. While 50 neighbors are probably not sufficient for the convergence of $\mathrm{ID}_{\mathrm{MLE}}$, using more than 110 neighbors results in using points from outside the cluster. For example with 400 neighbors, distances to neighbors from at least 4 different clusters are used in the estimation process. $\mathrm{ID}_{\mathrm{MLE}}$ estimates use only direct distances that reflect the intercluster dimensional properties of the data, whereas $\mathrm{ID}_{\mathrm{ALID}}$ uses auxiliary distances as well which predominate at low ID to enhance the detection of the local dimensional properties of the data.

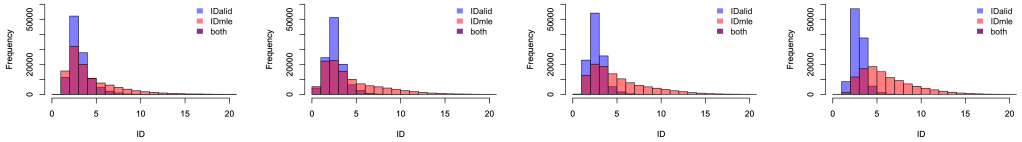


(a) ALOI (b) ANN_SIFT (c) BCI5 (d) CoverType



(e) Gisette (f) Isolet (g) MNIST (h) MSD

Figure 7: Histograms of LID values across each dataset, obtained using the $\mathrm{ID}_{\mathrm{MLE}}$ and $\mathrm{ID}_{\mathrm{ALID}}$ estimators on the size-100 neighborhoods of the individual reference points.



(a) $k = 50$ (b) $k = 100$ (c) $k = 200$ (d) $k = 400$

Figure 8: Histograms of LID values across ALOI dataset, obtained using the $\mathrm{ID}_{\mathrm{MLE}}$ and $\mathrm{ID}_{\mathrm{ALID}}$ estimators on the size-100 neighborhoods of the individual reference points.

5 Conclusion

Pairwise distances have been used in order to estimate dimensionality in some global models such as the Correlation Dimension. However, to the best of our knowledge none of the local models proposed in the literature takes advantage of the distances from neighbors to their nearest neighbors in order to increase the size of the distance sample. The proposed estimator which uses auxiliary distances converges faster, and thus can be used when the depth of the nearest neighbor graph is limited. Moreover, it has a smaller bias and variance than state-of-the-art estimators especially on nonlinear subspaces. Consequently, this 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.

As future work, it is possible to develop a similar auxiliary-distance estimator using the Method of Moments instead of the MLE. In cases where the neighborhood is very small is also possible to develop a heuristic using points from outside the locality in order to increase the distance sample size, but that would be at the cost of bias.

Acknowledgements

O. Chelly, M. E. Houle and K. Kawarabayashi supported by JST ERATO Kawarabayashi Project. O. Chelly and M. E. Houle supported by JSPS Kakenhi Kiban (B) Research Grant 15H02753. M. E. Houle supported by JSPS Kakenhi Kiban (A) Research Grant 25240036.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[2] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K. Kawarabayashi, and M. Nett. Estimating local intrinsic dimensionality. In *KDD*, pages 29–38. ACM, 2015.

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

[4] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *ISMIR 2011*, pages 591–596. University of Miami, 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 *MMCBIR 2001*, 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. *PAMI*, 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] R. Cole and M. Fanty. Spoken letter recognition. In *DARPA Speech and Natural Language Workshop*, pages 385–390, 1990.

[11] S. Coles, J. Bawa, L. Trenner, and P. Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[12] J. A. Costa and A. O. Hero III. Entropic graphs for manifold learning. In *Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 316–320. IEEE, 2004.

[13] T. de Vries, S. Chawla, and M. E. Houle. Finding local anomalies in very high dimensional space. In *ICDM*, pages 128–137, 2010.

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

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

[16] 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.

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

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

[19] M. E. Houle, H. Kashima, and M. Nett. Generalized expansion dimension. In *ICDMW*, pages 587–594. IEEE, 2012.

[20] M. E. Houle and M. Nett. Rank-based similarity search: Reducing the dimensional dependence. *PAMI*, 37(1):136–150, 2015.

[21] H. Jégou, R. Tavenard, M. Douze, and L. Amsaleg. Searching in one billion vectors: re-rank with source coding. In *ICASSP 2011*, pages 861–864. IEEE, 2011.

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

[23] D. R. Karger and M. Ruhl. Finding nearest neighbors in growth-restricted metrics. In *STOC*, pages 741–750. ACM, 2002.

[24] 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.

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

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

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

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

[29] F. Takens. *On the numerical determination of the dimension of an attractor*. Springer, 1985.

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