Estimating Local Intrinsic Dimensionality

Laurent Amsaleg
Equipe LINKMEDIA,
CNRS/IRISA Rennes, France
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
laurent.amsaleg@irisa.fr

Oussama Chelly
National Institute of Informatics, Japan
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
chelly@nii.ac.jp

Teddy Furon
Equipe LINKMEDIA,
INRIA/IRISA Rennes, France
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
teddy.furon@inria.fr

Stéphane Girard
Equipe MISTIS, INRIA
Grenoble, France
Inovallée, 655, Montbonnot
38334 Saint-Ismier Cedex, France
stephane.girard@inria.fr

Michael E. Houle
National Institute of Informatics, Japan
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
meh@nii.ac.jp

Ken-ichi Kawarabayashi
National Institute of Informatics, Japan
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
k_keniti@nii.ac.jp

Michael Nett
Google, Japan
6-10-1 Roppongi, Minato-ku
Tokyo 106-6126, Japan
mnett@google.com

ABSTRACT

This paper is concerned with the estimation of a local measure of intrinsic dimensionality (ID) recently proposed by Houle. The local model can be regarded as an extension of Karger and Ruhl's expansion dimension to a statistical setting in which the distribution of distances to a query point is modeled in terms of a continuous random variable. This form of intrinsic dimensionality can be particularly useful in search, classification, outlier detection, and other contexts in machine learning, databases, and data mining, as it has been shown to be equivalent to a measure of the discriminative power of similarity functions. Several estimators of local ID are proposed and analyzed based on extreme value theory, using maximum likelihood estimation (MLE), the method of moments (MoM), probability weighted moments (PWM), and regularly varying functions (RV). An experimental evaluation is also provided, using both real and artificial data.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Distribution Functions*; I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Learning, Parameter Learning*

KDD '15, August 10-13, 2015, Sydney, Australia

DOI: http://dx.doi.org/10.1145/2783258.2783405.

Keywords

intrinsic dimension, indiscriminability, manifold learning

1. INTRODUCTION

In an attempt to improve the discriminability of similarity measures, and the scalability of methods that depend on them, much attention has been given in the areas of machine learning, databases, and data mining to the development of dimensional reduction techniques. Linear techniques for dimensionality reduction include Principal Component Analysis (PCA) and its variants [4, 24]. Non-linear dimensionality reduction methods — also known as manifold learning techniques — include Isometric Mapping [36], Multi-Dimensional Scaling [35,37], Locally Linear Embedding and its variants [30], and Non-Linear Component Analysis [32]. Most reduction techniques require that a target dimension be provided by the user, although some attempt to determine the dimension automatically. Ideally, the supplied dimension should depend on the intrinsic dimensionality (ID) of the data. This has served to motivate the development of models of ID, as well as accurate estimators.

Over the past few decades, many practical models of the intrinsic dimensionality of data sets have been proposed. Examples include the previously mentioned Principal Component Analysis and its variants [4, 24], as well as several manifold learning techniques [26, 30, 32, 37]. Topological approaches to ID estimate the basis dimension of the tangent space of the data manifold from local samples [5,38]. Fractal methods such as the Correlation Dimension (CD) estimate an intrinsic dimension from the space-filling capacity of the data [6, 14]. Graph-based methods use the k-nearest neighbors graph along with density in order to estimate ID [8].

The aforementioned intrinsic dimensionality measures can be described as 'global', in that they consider the dimension-

ality of a given set as a whole, without any individual object being given a special role. In contrast, 'local' ID measures are defined in this paper as those that involve only the k-nearest neighbor distances of a specific location in the space. Several local intrinsic dimensionality models have been proposed recently, such as the expansion dimension (ED) [25], the generalized expansion dimension (GED) [19], the minimum neighbor distance (MiND) [31], and local continuous intrinsic dimension (which we will refer to here as LID) [17]. These models quantify ID in terms of the rate at which the number of encountered objects grows as the considered range of distances expands from a reference location.

Local approaches can be very useful when data is composed of heterogeneous manifolds. In addition to applications in manifold learning, measures of local ID have been used in the context of similarity search, where they are used to assess the complexity of a search query [22, 25], or to control the early termination of search [20, 21]. They have also found applications in outlier detection, in the analysis of a projection-based heuristic [9], and in the estimation of local density [39]. The efficiency and effectiveness of the algorithmic applications of intrinsic dimensional estimation (such as [20, 21]) depends greatly on the quality of of the estimators employed.

Distances from a query point can be seen as realizations of a continuous positive random variable. In this case, the smallest distances encountered would be 'extreme events' associated with the lower tail of the underlying distance distribution. In Extreme Value Theory (EVT), a discipline of statistics concerned with the study of tails of continuous probability distributions, the random variable associated with nearest neighbor distances can be assumed to follow a power-law distribution [7]. Continuous lower-bounded random variables are known to asymptotically converge to the Weibull distribution as the sample size grows, regardless of the original distance measure and its distribution. In an equivalent formulation of EVT due to Karamata, the cumulative distribution function of a tail distribution can be represented as a regularly-varying (RV) function whose dominant factor is a polynomial in the distance [7,18]; the degree (or 'index') of this polynomial factor determines the shape parameter of the associated Weibull distribution, or equivalently the exponent of the associated power law. The index has been interpreted as a form of intrinsic dimension [7]. Maximum likelihood estimation of the index leads to the well-known Hill estimator for power-law distributions [16].

While EVT provides an asymptotic description of tail distributions, in the case of continuous distance distributions, the distribution can be exactly characterized in terms of LID [18]. The LID model introduces a function that assesses the discriminative power of the distribution at any given distance value [17,18]. A distance measure is described as 'discriminative' when an expansion in the distance results in a relatively small increase in the number of observations. This function is shown to fully characterize the cumulative distribution function without the explicit involvement of the probability density [18]. The limit of this function yields the skewness of the Weibull distribution (or equivalently, the Karamata representation index, or power law exponent) associated with the lower tail. It is the estimation of this limit that is the main focus of this paper.

In addition to the more traditional applications stated earlier, LID has the potential for wide application in many machine learning and data mining contexts, as it makes no assumptions on the nature of the data distribution other than continuity.

The main original contributions of this paper are:

- a framework for the estimation of local continuous intrinsic dimension (LID) using well-established techniques: the maximum likelihood estimation (MLE), the method of moments (MoM), and the method of probability-weighted moments (PWM). In particular, we verify that applying MLE to LID leads to the well-known Hill estimator [16].
- a new family of estimators based on the extreme-value-theoretic notion of regularly varying functions. Several existing dimensionality models (ED, GED, and MiND) are shown to be special cases of this family.
- confidence intervals for the variance and convergence of the estimators we propose.
- an experimental study using artificial data and synthetic distance distributions, in which we compare our estimators with state-of-the-art global and local estimators. We also show that the empirical variance and convergence rates of the MLE (Hill) and MoM estimators are superior to those of the other local estimators studied.
- experiments showing that local estimators are more robust than global ones in the presence of noise in nonlinear manifolds. Our experiments show that our approaches are very competitive in this regard with other methods, both local and global.
- profiles of several real-world data sets in terms of LID, illustrating the degree of variability of complexity from region to region within a dataset. The profiles demonstrate that a single 'global' ID value is in general not sufficient to fully characterize the complexity of real-world data.

2. CONTINUOUS INTRINSIC DIMENSION

LID [17] aims to quantify the local ID of a feature space exclusively in terms of the distribution of inter-point distances. Formally, let $(\mathbb{R}^m, \mathrm{d})$ be a domain equipped with a non-negative distance function d. Let us consider the distribution of distances within the domain with respect to some fixed point of reference. We model this distribution in terms of a random variable $\mathbf{X}$ with support $[0, \infty)$. $\mathbf{X}$ is said to have probability density $f_{\mathbf{X}}$, where $f_{\mathbf{X}}$ is a non-negative Lebesgue-integrable function, if and only if

$$\mathsf{Pr}[a \leq \mathbf{X} \leq b] = \int_{x=a}^{b} f_{\mathbf{X}}(x)\,\mathrm{d}x,$$

for any $a, b \in [0, \infty)$ such that $a \leq b$. The corresponding cumulative density function $F_{\mathbf{X}}$ is canonically defined as

$$F_{\mathbf{X}}(x) = \mathsf{Pr}[\mathbf{X} \leq x] = \int_{u=0}^{x} f_{\mathbf{X}}(u)\,\mathrm{d}u.$$

Accordingly, whenever $\mathbf{X}$ is absolutely continuous at x, $F_{\mathbf{X}}$ is differentiable at x and its first-order derivative is $f_{\mathbf{X}}(x)$.

Definition 1 (Houle [17]). *Given an absolutely continuous random distance variable $\mathbf{X}$, for any distance threshold x such that $F_{\mathbf{X}}(x) > 0$, the local continuous intrinsic*

dimension of $\mathbf{X}$ *at distance* x *is given by*

$$\mathrm{ID}_{\mathbf{X}}(x) \triangleq \lim_{\epsilon \to 0^+} \frac{\ln F_{\mathbf{X}}\left((1+\epsilon)x\right) - \ln F_{\mathbf{X}}(x)}{\ln(1+\epsilon)}$$

wherever the limit exists.

With respect to the *generalized expansion dimension* [19], a precursor of LID, the above definition of $\mathrm{ID}_{\mathbf{X}}(x)$ is the outcome of a dimensional test of neighborhoods of radii x and $(1+\epsilon)x$ in which the neighborhood cardinalities are replaced by the expected number of neighbors. LID also turns out to be equivalent to a formulation of the (lack of) discriminative power of a distance measure, as both formulations have the same closed form:

THEOREM 1 (HOULE [17]). *Let* $\mathbf{X}$ *be an absolutely continuous random distance variable. If* $F_{\mathbf{X}}$ *is both positive and differentiable at* x*, then*

$$\mathrm{ID}_{\mathbf{X}}(x) = \frac{x f_{\mathbf{X}}(x)}{F_{\mathbf{X}}(x)}.$$

3. EXTREME VALUE THEORY

Extreme value theory is concerned with the modeling of what can be regarded as the extreme behavior of stochastic processes. Its best known theorem, attributed in parts to Fisher and Tippett [10], and Gnedenko [13], states that the maximum of N independent identically-distributed random variables (after proper renormalization) converges in distribution to a generalized extreme value distribution as N goes to infinity.

3.1 Threshold excesses

Consider the following two definitions.

DEFINITION 2. *Let* $\xi \in \mathbb{R}$ *and* $\sigma > 0$*. The family of* generalized Pareto distributions *is defined by its cumulative distribution function:*

$$F_X(x) = 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}}.$$

DEFINITION 3. *Let* $\mathbf{X}$ *be a random variable whose distribution* $F_{\mathbf{X}}$ *has the upper endpoint* $x^+ \in \mathbb{R} \cup \{\infty\}$*. Given* $w < x^+$*, the conditional excess distribution* $F_{\mathbf{X},w}$ *of* $\mathbf{X}$ *is the distribution of* $\mathbf{X} - w$ *conditioned on the event* $\mathbf{X} > w$*:*

$$F_{\mathbf{X},w}(x) = \frac{F_{\mathbf{X}}(w+x) - F_{\mathbf{X}}(w)}{1 - F_{\mathbf{X}}(w)}.$$

We are now in a position to introduce a powerful theorem due to Balkema and de Haan [1], and Pickands [28], which can be regarded as the counterpart to the central limit theorem for extremal statistics.

THEOREM 2 (BALKEMA-DE HAAN [1], PICKANDS [28]). *Let* $(\mathbf{X}_i)_{i \in \mathbb{N}}$ *be a sequence of independent random variables with identical distribution function* $F_{\mathbf{X}}$ *satisfying the conditions of the Fisher-Tippett-Gnedenko Theorem. As* $w \to x^+$*,* $F_{\mathbf{X},w}(x)$ *converges to a distribution in* $\mathcal{F}_{\mathrm{GPD}}$*.*

In the following we demonstrate a direct relation between local ID and extreme value theory, which arises as an implication of Theorem 2. Note that any choice of distance threshold w corresponds to a neighborhood of radius w based at the reference point, or equivalently, to the tail of the distribution of distances on $[0, w)$. As discussed in [7], Theorem 2 also applies to lower tails: one can reason about minima using the transformation $\mathbf{Y} = -\mathbf{X}$. The distribution of the excess $\mathbf{Y} - (-w)$ (conditioned on $\mathbf{Y} > -w$) then tends to a distribution in $\mathcal{F}_{\mathrm{GPD}}$, as w tends to the lower endpoint of $F_{\mathbf{X}}$ located at zero. Accordingly, as w tends to zero, the distribution in the tail $[0, w)$ can be restated as follows [7].

LEMMA 1. *Let* $\mathbf{X}$ *be an absolutely continuous random distance variable with support* $[0, \infty)$ *and cumulative distribution function* $F_{\mathbf{X}}$ *such that* $F_{\mathbf{X}}(x) > 0$ *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* $\xi < 0$*:*

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

Note that the distribution of excess distance $w - \mathbf{X}$ is bounded from above by w which, according to [7], enforces that $\xi < 0$.

To summarize, whenever Theorem 2 applies to a distance variable $\mathbf{X}$, the cumulative distribution of distances within a radius-w neighborhood is asymptotically determined by a single parameter $\xi < 0$. We can prove the following statement concerning LID.

THEOREM 3. *Let* $\mathbf{X}$ *be an absolutely continuous random distance variable with support* $[0, \infty)$*, satisfying the conditions of Theorem 2, and* $w > 0$ *be a distance threshold. Then, as* w *tends to zero,*

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

PROOF. Omitted due to space limitations. □

Note that together Lemma 1 and Theorem 3 allow us to restate the asymptotic cumulative distribution of distances in the tail $[cw, w)$ as

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

3.2 Regularly-varying functions

The Fisher-Tippett-Gnedenko Theorem and the Pickands-Balkema-de Haan Theorem have been shown to be equivalent to a third characterization of the tail behavior, in terms of regularly-varying (RV) functions. The asymptotic cumulative distribution of $\mathbf{X}$ in the tail $[0, w)$ can be expressed as $F_{\mathbf{X}}(x) = x^{\kappa} \ell_{\mathbf{X}}(1/x)$, where $\ell_{\mathbf{X}}$ is differentiable and *slowly varying*; that is, for all $c > 0$, $\ell_{\mathbf{X}}$ satisfies

$$\lim_{t \to \infty} \frac{\ell_{\mathbf{X}}(ct)}{\ell_{\mathbf{X}}(t)} = 1.$$

$F_{\mathbf{X}}$ restricted to $[0, w)$ is itself said to be *regularly varying* with index κ. In particular, a cumulative distribution $F \in \mathcal{F}_{\mathrm{GEV}}$ has $\xi < 0$ if and only if F is RV and has a finite endpoint. Note that the slowly-varying component $\ell_{\mathbf{X}}(1/x)$ of $F_{\mathbf{X}}$ is not necessarily constant as x tends to zero. For a detailed account of RV functions, we refer the reader to [2].

The following corollary is a straightforward extension of the examples given in Section 2.

COROLLARY 1. *Let* $\mathbf{X}$ *be a random distance variable restricted to* $[0, w)$ *with distribution* $F_{\mathbf{X}}(x) = x^{\kappa} \ell_{\mathbf{X}}(1/x)$*. As* w *tends to zero, the index* κ *converges to* $\mathrm{ID}_{\mathbf{X}}$*.*

4. ESTIMATION

This section is concerned with practical methods for the estimation of the local intrinsic dimension of a random distance variable $\mathbf{X}$. In particular, we adapt known GPD parameter estimators such as the maximum-likelihood estimator (in Section 4.1) and moment based estimators (in Sections 4.2 and 4.3), and propose a new estimator based on regularly varying functions (in Section 4.4).

For the remainder of this discussion we assume that we are given a sequence $x_1, \ldots, x_n$ of observations of a random distance variable $\mathbf{X}$ with support $[0, w)$, in ascending order — that is, $x_1 \leq x_2 \leq \cdots \leq x_n$.

4.1 Maximum Likelihood Estimation

Using the asymptotic expression of the distance distribution given in Equation 1, we see that the log-likelihood of $\mathrm{ID}_{\mathbf{X}}$ for the sample is

$$\mathcal{L}(\mathrm{ID}_{\mathbf{X}}) = n \ln \frac{F_{\mathbf{X},w}(w)}{w} + n \ln \mathrm{ID}_{\mathbf{X}} + (\mathrm{ID}_{\mathbf{X}} - 1) \sum_{i=1}^{n} \ln \frac{x_i}{w}.$$

Accordingly, the maximum-likelihood estimate $\widehat{\mathrm{ID}}_{\mathbf{X}}$ is

$$\widehat{\mathrm{ID}}_{\mathbf{X}} = -\left(\frac{1}{n} \sum\nolimits_{i=1}^{n} \ln \frac{x_i}{w} \right)^{-1},$$

which follows the form of the well-known Hill estimator for the scaling exponent of a power-law tail distribution [16].

The MLE model ensures the usual regularity conditions that guarantee the consistency, the asymptotic normality and the efficiency of this estimator. The variance is asymptotically given by the inverse of the Fisher information defined as:

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

where $\mathsf{E}[\cdot]$ denotes the expectation. Therefore, if the number of samples n is sufficiently large, we have $\widehat{\mathrm{ID}}_{\mathbf{X}} \sim \mathcal{N}(\mathrm{ID}_{\mathbf{X}}, \mathrm{ID}_{\mathbf{X}}^2 / n)$. Accordingly, with probability $1 - \beta$, a sample of n distances in $[0, w)$ provides an estimate $\widehat{\mathrm{ID}}_{\mathbf{X}}$ lying within

$$\mathrm{ID}_{\mathbf{X}} \pm \frac{\mathrm{ID}_{\mathbf{X}}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\beta}{2}\right).$$

In other words, the $1 - \beta$ confidence interval is

$$\left[\frac{\widehat{\mathrm{ID}}_{\mathbf{X}}}{1 + n^{-1/2}\Phi^{-1}(1-\beta/2)}, \frac{\widehat{\mathrm{ID}}_{\mathbf{X}}}{1 - n^{-1/2}\Phi^{-1}(1-\beta/2)} \right].$$

4.2 Method of Moments

For any choice of $k \in \mathbb{N}$, the *k-th order non-central moment* μ_k of the random distance $\mathbf{X}$ is

$$\mu_k = \mathsf{E}\left[\mathbf{X}^k\right] = \int_{x=0}^{w} x^k f_{\mathbf{X}}(x)\, \mathrm{d}x = w^k \frac{\mathrm{ID}_{\mathbf{X}}}{\mathrm{ID}_{\mathbf{X}} + k}.$$

Solving for the intrinsic dimension gives

$$\mathrm{ID}_{\mathbf{X}} = -k \frac{\mu_k}{\mu_k - w^k} = g\left(\frac{\mu_k}{w^k}\right),$$

with $g(x) = k\frac{x}{1-x}$. When estimating the order-k moment by its empirical counterpart $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$, we see that $\mathsf{E}[\hat{\mu}_k] = \mu_k$ and $\mathsf{E}[\hat{\mu}_r^2] = (n\mu_{2k} + n(n-1)\mu_k^2)n^{-2}$, so that

$$\mathsf{Var}[\hat{\mu}_k^2] = \frac{\mu_{2k} - \mu_k^2}{n} = \frac{w^{2k}\mathrm{ID}_{\mathbf{X}} k^2}{n(\mathrm{ID}_{\mathbf{X}} + 2k)(\mathrm{ID}_{\mathbf{X}} + k)^2}.$$

Therefore, the distribution of $\frac{\hat{\mu}_k}{w^k}$ is asymptotically normal with

$$\frac{\hat{\mu}_k}{w^k} \sim \mathcal{N}\left(\frac{\mathrm{ID}_{\mathbf{X}}}{\mathrm{ID}_{\mathbf{X}} + k}; \frac{\mathrm{ID}_{\mathbf{X}} k^2}{n(\mathrm{ID}_{\mathbf{X}} + 2k)(\mathrm{ID}_{\mathbf{X}} + k)^2} \right).$$

According to [29, Th. 6a2.9], if $x \sim \mathcal{N}(\mu; \sigma^2 n^{-1})$ asymptotically, then $g(x) \sim \mathcal{N}(g(\mu); \sigma^2 n^{-1} g'(\mu)^2)$, where g' is the first-order derivative of g. Therefore, asymptotically

$$\widehat{\mathrm{ID}}_{\mathbf{X}} \sim \mathcal{N}\left(\mathrm{ID}_{\mathbf{X}}; \frac{\mathrm{ID}_{\mathbf{X}}^2}{n}\left(1 + \frac{(k/\mathrm{ID}_{\mathbf{X}})^2}{\mathrm{ID}_{\mathbf{X}}^2(1 + 2k/\mathrm{ID}_{\mathbf{X}})}\right)\right).$$

This variance is monotonically increasing in $k/\mathrm{ID}_{\mathbf{X}}$, which indicates that we should use moments of small order k. When $k/\mathrm{ID}_{\mathbf{X}}$ tends to zero, the variance converges to $\mathrm{ID}_{\mathbf{X}}^2/n$, the variance of the maximum-likelihood estimator (see Section 4.1). Note that an upper bound on $\mathrm{ID}_{\mathbf{X}}$ implies that the variance is bounded. In this case we can derive confidence intervals similar to Section 4.1.

4.3 Probability-Weighted Moments

General probability-weighted moments are defined as

$$m_{k,l,m} = \mathsf{E}\left[F_{\mathbf{X}}(x)^k (1 - F_{\mathbf{X}}(x))^l \mathbf{X}^m \right].$$

We restrict here our attention to a subfamily: for any choice of $k \in \mathbb{N}$, ν_k is defined as

$$\begin{aligned} \nu_k &\triangleq \mathsf{E}\left[F_{\mathbf{X}}(x)^k \mathbf{X} \right] = \int_{x=0}^{w} F_{\mathbf{X}}(x)^k x f_{\mathbf{X}}(x)\, \mathrm{d}x \\ &= \frac{\mathrm{ID}_{\mathbf{X}}\, w}{\mathrm{ID}_{\mathbf{X}}\, k + \mathrm{ID}_{\mathbf{X}} + 1}; \end{aligned}$$

solving for the intrinsic dimension yields

$$\mathrm{ID}_{\mathbf{X}} = \frac{\nu_k}{w - \nu_k(k+1)} = h\left(\frac{\nu_k}{w}\right),$$

where $h(x) = \frac{x}{1-(k+1)x}$.

4.4 Estimation Using Regularly Varying Functions

In this section we introduce an ad hoc estimator for the intrinsic dimensionality based on the characterization of distribution tails as regularly varying functions (as discussed in Section 3). Consider the empirical distribution function $\hat{F}_{\mathbf{X}}$, defined as

$$\hat{F}_{\mathbf{X}}(x) = \frac{1}{n} \sum\nolimits_{j=1}^{n} [\![x_j < x]\!],$$

where $[\![\varphi]\!]$ refers to the Iverson bracket which evaluates to 1 if φ is true, and 0 otherwise. We propose the following estimator for the index κ of $F_{\mathbf{X}}$.

Definition 4. *Let $\mathbf{X}$ be an absolutely continuous random distance variable restricted to $[0, w)$. The local intrinsic dimension $\mathrm{ID}_{\mathbf{X}}$ can be estimated as*

$$\widehat{\mathrm{ID}}_{\mathbf{X}} = \hat{\kappa} = \frac{\sum_{j=1}^{J} \alpha_j \ln\left[\hat{F}_{\mathbf{X}}((1+\tau_j\delta_n)x_n)/\hat{F}_{\mathbf{X}}(x_n)\right]}{\sum_{j=1}^{J} \alpha_j \ln(1+\tau_j\delta_n)},$$

under the assumption that $x_n, \delta_n \to 0$ as $n \to \infty$, where $(\alpha_j)_{1\leq j\leq J}$ and $(\tau_j)_{1\leq j\leq J}$ are sequences.

We will refer to this family of estimators as RV, for 'regularly varying'. Note that since RV estimators involve only the products $\tau_j\delta_n$ for $1 \leq j \leq J$, we may assume without loss of generality that $\tau_1 + \cdots + \tau_J = 1$. The estimators are based on the observation that, for all $1 \leq j \leq J$,

$$\ln\left[F_{\mathbf{X}}((1+\tau_j\delta_n)x_n)/F_{\mathbf{X}}(x_n)\right]$$
$$= \kappa\ln(1+\tau_j\delta_n) + \ln\left[\ell_{\mathbf{X}}((1+\tau_j\delta_n)x_n)/\ell_{\mathbf{X}}(x_n)\right]$$
$$\simeq \kappa\ln(1+\tau_j\delta_n).$$

The RV family covers several of the known local estimators of intrinsic dimensionality. For the parameter choices $J = 1$ and $\epsilon = \tau\delta_n$, the RV estimator reduces to the GED formulation proposed in [19]:

$$\widehat{\mathrm{ID}}_{\mathbf{X}} = \frac{\ln\left[\hat{F}_{\mathbf{X}}((1+\epsilon)x_n)/\hat{F}_{\mathbf{X}}(x_n)\right]}{\ln(1+\epsilon)},$$

By setting $\epsilon = 1$, Karger & Ruhl's expansion dimension is obtained, while by setting x_n as the distance to the k-nearest neighbor and ϵ such as $(1+\epsilon)x_n$ as the distance to the nearest neighbor, we find a special case of the MiND family (MiND_{ml1}) [31].

Alternatively, by setting $J = n$, $\alpha_i = 1$ for all $i \in [1..n]$, and choosing the vector τ such that $1 + \tau_i\delta_n = \frac{x_i}{x_n}$, the RV estimator becomes

$$\widehat{\mathrm{ID}}_{\mathbf{X}} = \frac{\sum_{j=1}^{n}\ln[j/n]}{\sum_{j=1}^{n}\ln[x_j/x_n]} \approx \frac{\ln\sqrt{2\pi n} - n}{\sum_{j=1}^{n}\ln[x_j/x_n]}$$

As $n \to \infty$, this converges to the MLE (Hill) estimator presented in Section 4.1, with $w = x_n$.

We now turn our attention to an analysis of the variation of RV estimators. First, we introduce an auxiliary function which drives the speed of convergence of the estimator proposed in Definition 4. For $x \in \mathbb{R}$ let $\varepsilon_{\mathbf{X}}(x)$ be defined as

$$\varepsilon_{\mathbf{X}}(x) \triangleq \frac{x\ell'_{\mathbf{X}}(x)}{\ell_{\mathbf{X}}(x)}.$$

In [11, 12], the auxiliary function is assumed to be regularly varying, and the estimation of the corresponding regular variation index is addressed. Within this article, so as to prove the following results, we limit ourselves to the assumption that $\varepsilon_{\mathbf{X}}$ is ultimately non-increasing.

THEOREM 4. *Let $\mathbf{X}$ be a random distance variable over $[0, w)$ with distribution function $F_{\mathbf{X}}(x) = x^{\kappa}\ell_{\mathbf{X}}(1/x)$, and let $\tau_{\max} \triangleq \max_{1\le j\le J}\tau_j$. Furthermore, let $\delta_n, x_n \to 0$ so that $n\,F_{\mathbf{X}}(x_n)\delta_n \to \infty$ and $\sqrt{nF_{\mathbf{X}}(x_n)\delta_n}\varepsilon_{\mathbf{X}}(1/[(1+\tau_{\max}\delta_n)x_n]) \to 0$ as n approaches infinity. If the auxiliary function $\varepsilon_{\mathbf{X}}$ is ultimately non-increasing, then $\sqrt{nF_{\mathbf{X}}(x_n)\delta_n}\cdot[\mathrm{ID}_{\mathbf{X}} - \widehat{\mathrm{ID}}_{\mathbf{X}}]$ converges to a centered Gaussian with variance*

$$\mathrm{ID}_{\mathbf{X}}V_{\alpha,\tau} = \mathrm{ID}_{\mathbf{X}}\frac{\alpha^{\top}S\alpha}{(\alpha^{\top}\tau)^2},$$

where $S_{a,b} = (|\tau_a| \wedge |\tau_b|)[\![\tau_a\tau_b > 0]\!]$ for $(a,b) \in \{1,\dots,J\}^2$. ($A \wedge B$ denotes the minimum of A and B.)

Note that the requirement $nF_{\mathbf{X}}(x_n)\delta_n \to \infty$ can be interpreted as a necessary and sufficient condition for the almost sure presence of at least one distance sample in the interval $[x_n, (1+\tau_j\delta_n)x_n)]$. In addition, the condition

$$\sqrt{nF_{\mathbf{X}}(x_n)\delta_n}\varepsilon_{\mathbf{X}}(1/[r_n(1+\tau_{\max}\delta_n)]) \to 0$$

enforces that the approximation bias $\varepsilon_{\mathbf{X}}(1/[(1+\delta_n)x_n])$ is negligible compared to the standard deviation of the estimate, $1/\sqrt{nF_{\mathbf{X}}(x_n)\delta_n}$. We continue the analysis by proposing choices of α that minimize the variance in Theorem 4.

LEMMA 2. *The weight vector $\alpha = (\alpha_1,\dots,\alpha_J)^{\top}$ minimizing $V_{\alpha,\tau}$ is proportional to $\alpha_0 = S^{-1}\tau = (1,0,\dots,0)^{\top}$, and the associated optimal variance is given by $V_0(\tau) = \left(\tau^{\top}S^{-1}\tau\right)^{-1}$.*

PROOF. Omitted due to space limitations. □

For the case $J = 1$, we see that $\tau = (1)^{\top}$ and $V_0(1) = 1$. This indicates that the GED minimizes the variance of estimation. However, different choices can be made regarding the weight vector τ and regarding the criterion to use in order to optimize the choice of α. Minimizing variance is one choice explored in this paper, but other criteria can be used. In general, however, the following confidence interval holds for RV estimators:

LEMMA 3. *Let $\beta \in (0,1)$, and assume that the assumptions of Theorem 4 hold with $\alpha = S^{-1}\tau$. Let $u_\beta = \Phi^{-1}((1+\beta)/2)$, where Φ is the cumulative distribution function of the standard Gaussian distribution. Then*

$$\mathrm{ID}_{\mathbf{X}} \pm u_\beta\left(n\delta_n V_0(\tau)\widehat{\mathrm{ID}}_{\mathbf{X}}\hat{F}_{\mathbf{X}}(x_n)\right)^{-1/2}$$

are the boundaries of the asymptotic confidence interval of level β for $\widehat{\mathrm{ID}}_{\mathbf{X}}$.

PROOF. Lemma 3 is a direct consequence of the asymptotic distribution established in Theorem 4 and the convergence of $\hat{F}_{\mathbf{X}}(x_n)$ to $F_{\mathbf{X}}(x_n)$ as $n \to \infty$. □

5. EXPERIMENTAL FRAMEWORK

5.1 Methods

The methods used in this study include MLE, MoM, PWM, and RV. The RV estimators are evaluated for the choices $J = 1$ and $J = 2$, as follows:

$$\widehat{\mathrm{ID}}_{\mathrm{RV}} = \begin{cases} \frac{\ln n - \ln\lfloor n/2\rfloor}{\ln x_n - \ln x_{\lfloor n/2\rfloor}}, & \text{if} \quad J = 1 \\ \frac{\ln\lfloor n/j\rfloor - (p-1)\ln\lfloor i/j\rfloor}{\ln x_n/x_j + (p-1)\ln x_i/x_j}, & \text{if} \quad J = 2, \end{cases}$$

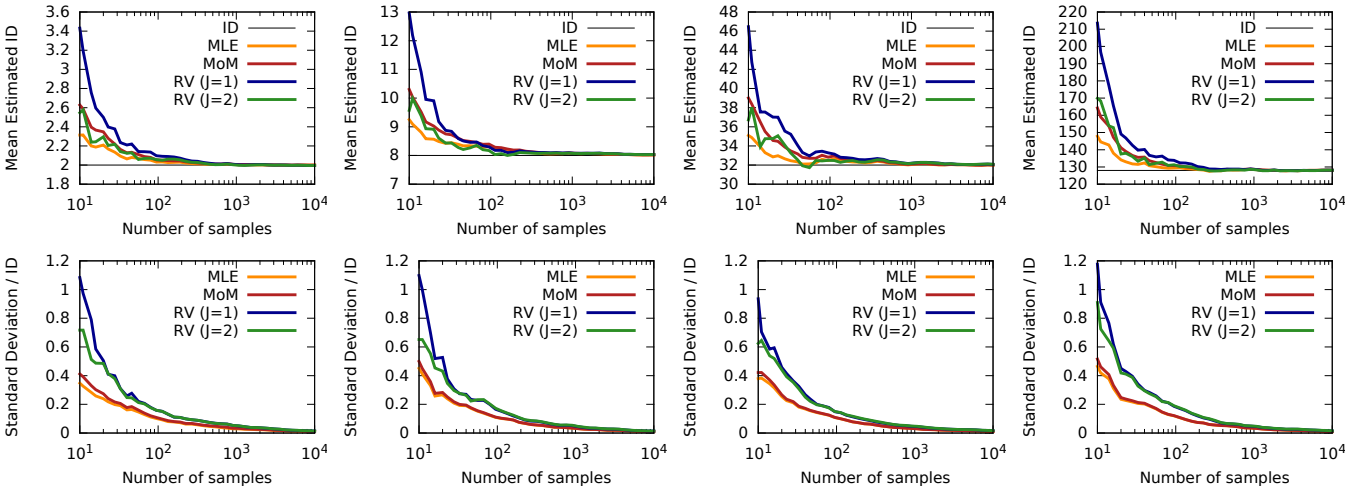
where $p = (x_i - 2x_j + x_n)/(x_n - x_j)$, $i = \lfloor n/2\rfloor$, and $j = \lfloor 3n/4\rfloor$. Note that the estimator RV for $J = 1$ is a form of generalized expansion dimension (GED) [19]. For every dataset, we report the average of ID estimates across all the points in the dataset. All estimators in our study can be computed in time linear in the number of sample points.

Method	Parameters
PCA	threshold = 0.025
kNNG_1	$k = 100$, $\gamma = 1$, $M = 1$, $N = 10$
kNNG_2	$k = 100$, $\gamma = 1$, $M = 10$, $N = 1$
MiND_{ml1}	None
MiND_{mli}	$k = 100$

Table 1: Parameter choices used in the experiments.

Our experimental framework includes several state-of-the-art intrinsic dimensionality measures. The global estimators consist of a projection method (PCA), fractal methods (CD [6], Hein [15], Takens [34]), and graph-based methods (kNNG_1, kNNG_2 [8]). The local distance-based estimators are MiND_{ml1} and MiND_{mli} [31]. Table 1 summarizes the parameter choices for every method, except for the fractal methods, which do not involve any parameter.

The MiND variants makes more restrictive assumptions than our methods: they assume the data to be uniformly distributed on a hypersphere, with a locally isometric smooth



(a) ID = 2 (b) ID = 8 (c) ID = 32 (d) ID = 128

Figure 1: Comparison of the mean and standard deviation of LID estimates provided by MLE, MoM and RV (for $J = 1$ and $J = 2$) on increasingly large samples drawn from artificially-generated distance distributions. The results cover target dimensionality values of 2, 8, 32, and 128. The values are marked in the corresponding plots.

map between the hypersphere and the representational space. MiND uses only the two extreme samples (smallest and largest), and requires knowledge of the dimension of the space (D). In contrast, our approach assumes only that the nearest neighbor distances are in the lower tail of the distance distribution, where EVT estimation can be performed.

5.2 Artificial Distance Distributions

In the following we propose a set of experiments concerning artificial data, and describe the method employed for the generation of test data.

First, consider a point $\mathbf{P}$ drawn uniformly at random from within the m-dimensional unit sphere, for some choice of $m \in \mathbb{N}$. According to the method of normal variates, we define $\mathbf{P} = \mathbf{Z}^{1/m}\mathbf{Y}\|\mathbf{Y}\|^{-1}$, where $\mathbf{Z}$ is uniformly distributed on $[0, 1]$, and $\mathbf{Y}$ is a random vector in $\mathbb{R}^m$ whose coefficients follow the standard normal distribution. The distance of $\mathbf{P}$, with respect to our choice of reference point at location $0 \in \mathbb{R}^m$, is distributed as follows.

$$\mathbf{X} = \frac{\|\mathbf{Z}^{1/m}\mathbf{Y}\|}{\|\mathbf{Y}\|} = \mathbf{Z}^{1/m}.$$

Note that, by measuring LID purely based on distance values with respect to a reference point, the model does not require that the data have an underlying spatial representation. As such, non-integer values of $m \in \mathbb{R}$ can be selected for the generation of distances, if desired.

For choices of $m \in \{2, 8, 32, 128\}$, we draw 100 independent sequences of sample distance values from the distribution described above, and record the estimates produced by each of our methods for sample sizes n between 10 and 10^4.

5.3 Artificial Data

The data sets used in our experiments have been proposed in [31]. They consist of 15 manifolds of various stuctures and intrinsic dimensionalities (d) represented in spaces of different dimensions (D). They are summarized in Table 2.

These datasets were generated in different sizes (10^3, 10^4, and 10^5 points) in order to evaluate the effect of the number of points on the quality of the different estimators. For each dataset and for each of the three sizes, we average the estimates over 20 instances.

Manifold	d	D	Description
1	10	11	Uniformly sampled sphere.
2	3	5	Affine space.
3	4	6	Concentrated figure confusable with a 3d one.
4	4	8	Non-linear manifold.
5	2	3	2-d Helix
6	6	36	Non-linear manifold.
7	2	3	Swiss-Roll.
8	12	72	Non-linear manifold.
9	20	20	Affine space.
10a	10	11	Uniformly sampled hypercube.
10b	17	18	Uniformly sampled hypercube.
10c	24	25	Uniformly sampled hypercube.
11	2	3	Möbius band 10-times twisted.
12	20	20	Isotropic multivariate Gaussian.
13	1	13	Curve.

Table 2: Artificial datasets used in the experiments.

In order to evaluate the robustness of the estimators, we also prepared versions of these datasets with noise added. For each attribute f, we added normally-distributed noise with mean equal to zero and standard deviation $\sigma_n = p \cdot \sigma_f$ where σ_f is the standard deviation of the attribute itself, and $p \in \{0.01, 0.04, 0.16, 0.64\}$. For attributes with $\sigma_f = 0$, the noise was generated with standard deviation $\sigma_n = p \cdot \sigma_f^*$ where σ_f^* is the minimum of the nonzero standard deviations over all attributes.

5.4 Real Data

Not only can a reliable estimation of ID greatly benefit the practical performance of many applications, it also serves as a characterization of high-dimensional data sets and the potential problems associated with their use in practice. To this end, we investigate the distribution of LID estimates on

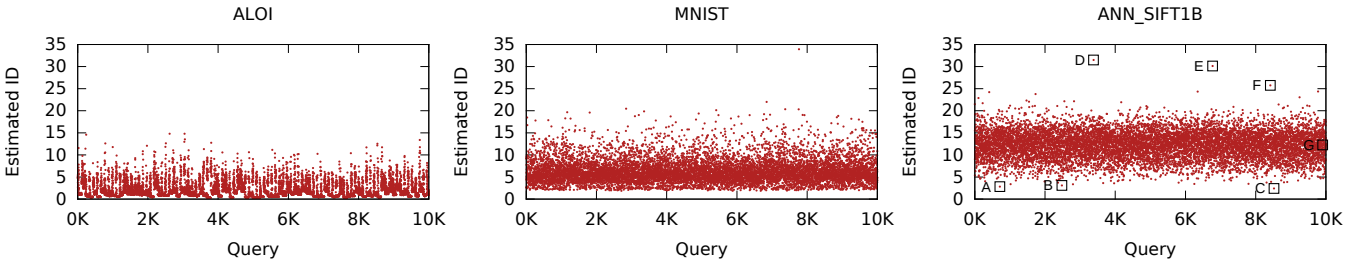


Figure 2: Plots of the distribution of LID values across 10^4 distinct query locations for each data set. The LID values were obtained using the MLE estimator on the size-1000 neighborhoods of the individual reference points.

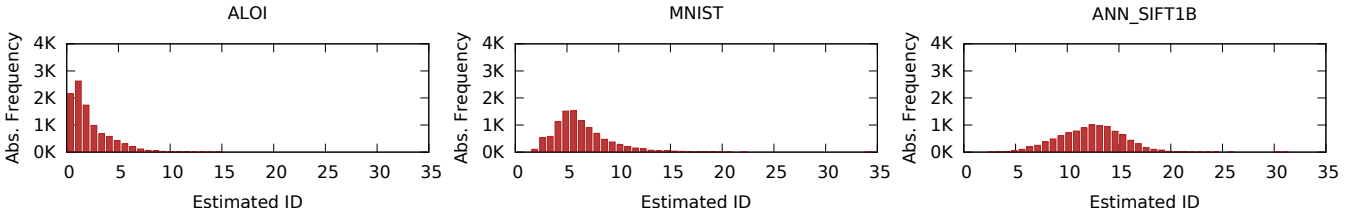


Figure 3: Histograms of LID values across 10^4 distinct query locations for each data set, obtained using the MLE estimator on the size-1000 neighborhoods of the individual reference points.

the following data sets, each taken from a real-world application scenario.

The *ALOI* (*Amsterdam Library of Object Images*) data set contains a total of 110250 color photos of 1000 different objects taken from varying viewpoints under various illumination conditions. Each image is described by a 641-dimensional vector of color and texture features [3].

The *MNIST* database [27] 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.

The *ANN_SIFT1B* data set consists of 128-dimensional SIFT descriptors extracted from a collection of $\sim 10^9$ images. This set has been created for the evaluation of nearest-neighbor search strategies at very large scales [23].

For each data set, we estimate LID with respect to 10^4 distinct reference points, based on the distribution of distances to their respective 10^3-nearest neighbors. For *ANN_SIFT1B* we use a selection of 10^4 query points that is provided with the data. In the case of *ALOI* and *MNIST*, we computed distance samples with respect to 10^4 points selected uniformly at random.

6. EXPERIMENTAL RESULTS

6.1 Artificial Distance Distributions

We begin our experimental study with an assessment — in terms of bias, variance, and convergence — of the ability of each estimator to identify the ID of a sample of distance values generated according to different choices of target ID. Note that for these trials, the distributional model asserted in Lemma 1 holds everywhere on the range $[0, w)$ by construction (with $w = 1$).

Fig. 1 shows the behavior of MLE, MoM, and RV (for choices of $J = 1$ and $J = 2$). The convergence to the target ID value observed in every case empirically confirms the consistency of these estimators. Likewise, PWM is consistent however, one should beware of PWM's susceptibility to the effects of numerical instability.

We also note that the RV estimator with $J = 1$ (GED) — which asymptotically minimizes variance according to Lemma 2 — is not the choice that minimizes variance when the number of samples is limited. Faster initial convergence favors the choice of MLE and MoM for applications where the number of available query-to-neighbor distances is limited, or where time complexity is an issue.

6.2 Artificial Data

In Tables 3 and 4, due to space limitations, we present only a representative selection of the experimental results, averaged over 20 runs each. It should be noted that as PCA and MiND_{mli} estimates are restricted to integer values, their bias is lower for examples having integer ground-truth intrinsic dimension, especially when this dimensionality is small. Also, unlike the other estimators tested, MiND estimators also require that an upper bound on the ID be supplied (set to D in these experiments). PCA requires a threshold parameter to be supplied, the value of which can greatly influence the estimation.

The experimental results indicate that local estimators tend to over-estimate dimensionality in the case of non-linear manifolds (sets m3, m4, m5, m6, m7, m8, m11 and m13) and to under-estimate it in the case of linear manifolds (sets m1, m2, m9, m10a, m10b, m10c and m12). For highly non-linear manifolds, such as the Swiss Roll (m7), global estimators have difficulty in identifying the intrinsic dimension. The experimental results with higher sampling rates confirm the reduction in bias that would be expected with smaller k-nearest-neighbor distances, as the local manifold structure more closely approximates the tangent space.

To show the effects of noise on the estimators, we display in Tables 5, 6 and 7 for each method the deviation of every estimate in the presence of noise as a proportion of the estimate obtained in the absence of noise. On the one hand, we note that global methods, k-NNG in particular, are

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{ml1}$	$MiND_{mli}$	CD	Hein	Takens	kNNG$_1$	kNNG$_2$	PCA
m1	10	11	8.07	8.08	8.14	7.91	7.79	9.50	8.95	9.24	5.35	9.44	7.96	7.02	11.00
m2	3	5	2.67	2.67	2.68	2.65	2.60	2.94	3.00	2.87	2.75	2.91	2.53	2.52	3.00
m3	4	6	3.56	3.56	3.59	3.55	3.49	3.88	4.00	3.63	3.70	3.66	4.00	2.88	5.30
m7	2	3	2.49	2.80	3.04	3.22	3.12	2.00	2.00	1.95	1.90	1.95	3.10	2.86	3.00
m8	12	72	12.29	12.33	12.51	11.97	11.79	13.49	13.00	11.00	3.60	11.85	14.28	12.56	24.00
m9	20	20	12.39	12.40	12.50	11.96	11.79	15.03	13.50	12.84	4.30	14.68	19.68	10.84	20.00
m10a	10	11	7.39	7.40	7.47	7.28	7.16	8.50	8.00	8.42	8.15	8.45	10.69	6.65	10.00
m10c	24	25	14.05	14.07	14.22	13.52	13.32	17.69	15.35	16.82	6.05	16.90	17.31	29.77	24.00
m11	2	3	2.49	2.74	2.94	3.05	2.97	2.01	2.00	1.99	2.70	2.00	2.83	2.59	3.00
m12	20	20	12.48	12.46	12.43	11.85	11.67	16.79	14.00	13.69	3.70	13.64	11.71	5.13	20.00

Table 3: ID estimates for 1000 points.

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{ml1}$	$MiND_{mli}$	CD	Hein	Takens	kNNG$_1$	kNNG$_2$	PCA
m1	10	11	9.04	9.10	9.32	9.06	8.92	9.61	9.00	9.56	8.95	9.59	9.20	9.87	11.00
m2	3	5	2.88	2.90	2.94	2.90	2.85	2.96	3.00	3.08	3.55	2.98	2.77	2.44	3.00
m3	4	6	3.86	3.90	3.97	3.92	3.85	3.92	4.00	3.75	3.90	3.76	3.94	3.94	5.05
m7	2	3	1.96	1.99	2.02	1.99	1.95	1.99	2.00	1.97	1.95	1.98	1.83	1.83	3.00
m8	12	72	13.72	13.86	14.50	13.91	13.69	12.91	14.00	11.95	8.10	11.92	14.08	14.08	24.00
m9	20	20	14.47	14.56	15.08	14.41	14.18	15.95	15.00	15.69	2.65	15.74	10.11	10.11	20.00
m10a	10	11	8.20	8.25	8.43	8.21	8.08	8.86	8.00	8.87	9.10	8.92	6.55	6.55	10.00
m10c	24	25	16.66	16.77	17.45	16.54	16.28	18.50	17.00	18.08	10.90	18.13	15.00	15.00	24.00
m11	2	3	1.99	2.03	2.06	2.04	2.00	1.99	2.00	1.99	2.00	2.00	1.84	1.84	3.00
m12	20	20	15.46	15.54	16.03	15.23	15.00	17.74	16.00	15.04	3.70	15.00	37.63	37.63	20.00

Table 4: Dimensionality estimates for 10000 points.

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{ml1}$	$MiND_{mli}$	CD	Hein	Takens	kNNG$_1$	kNNG$_2$	PCA
m1	10	11	-10.07	-10.55	-11.80	-11.81	-11.88	-1.56	-2.78	-11.82	-38.55	-12.10	-62.17	-64.74	-22.73
m2	3	5	2.43	-1.03	-3.06	-3.10	-3.51	36.49	0.00	12.01	-18.31	23.15	16.97	32.79	-33.33
m3	4	6	-30.83	-32.05	-33.25	-33.16	-33.25	-23.47	-25.00	-22.13	-41.03	-22.34	-35.79	-35.79	-60.40
m7	2	3	-8.67	-14.57	-16.34	-15.58	-15.90	34.17	0.00	14.21	10.26	9.60	-44.81	-44.81	-66.67
m8	12	72	44.17	43.00	39.79	35.44	35.65	115.49	60.71	86.53	25.93	85.99	93.68	93.68	95.21
m9	20	20	-21.77	-22.25	-24.34	-24.01	-23.98	-9.97	-17.00	-22.12	167.92	-22.62	157.17	157.17	-31.75
m10a	10	11	21.46	21.45	21.59	20.83	20.92	22.12	25.00	9.02	-64.29	8.07	338.17	338.17	10.00
m10c	24	25	7.98	7.87	7.45	6.83	6.88	14.76	11.76	-2.99	-74.31	-3.75	-177.73	-177.73	4.17
m11	2	3	32.16	29.56	28.64	28.43	28.50	47.74	0.00	41.21	10.00	40.50	195.65	195.65	-35.00
m12	20	20	-22.83	-23.10	-24.52	-23.90	-23.93	-16.52	-19.69	-16.22	13.51	-16.27	-84.45	-84.45	-26.00

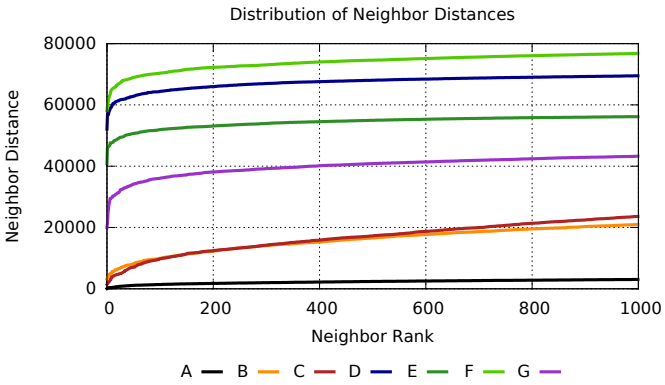
Table 5: Deviation of dimensionality estimates for 10000 manifold points with added noise (p=0.01).

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{ml1}$	$MiND_{mli}$	CD	Hein	Takens	kNNG$_1$	kNNG$_2$	PCA
m1	10	11	-10.18	-10.66	-11.91	-11.92	-12.00	-1.87	-2.78	-17.05	-63.69	-12.20	-341.09	-324.72	-23.18
m2	3	5	2.43	-1.03	-3.06	-3.45	-3.51	37.16	0.00	18.83	-9.86	22.82	-7.94	4.51	-33.33
m3	4	6	-30.57	-32.05	-33.25	-33.16	-33.25	-23.47	-25.00	-26.40	-42.31	-22.07	-31.47	-31.47	-60.40
m7	2	3	-8.67	-14.57	-16.83	-15.58	-15.90	34.17	0.00	15.74	7.69	11.62	-38.25	-38.25	-66.67
m8	12	72	44.24	43.07	39.86	35.59	35.72	116.42	60.71	86.69	46.30	86.16	9.52	9.52	95.21
m9	20	20	-21.77	-22.25	-24.27	-24.01	-23.91	-10.22	-17.33	-22.31	132.08	-22.74	15.73	15.73	-31.75
m10a	10	11	21.46	21.45	21.59	20.83	20.79	21.78	25.00	3.04	-48.35	7.96	25.65	25.65	10.00
m10c	24	25	8.04	7.87	7.51	6.83	6.94	14.49	11.76	-7.85	-59.17	-3.53	-18.80	-18.80	4.17
m11	2	3	32.16	29.56	28.64	28.43	28.50	46.73	0.00	40.20	37.50	39.00	255.43	255.43	-35.00
m12	20	20	-22.83	-23.10	-24.52	-23.90	-23.93	-16.18	-19.37	-16.16	33.78	-16.33	-174.25	-174.25	-26.00

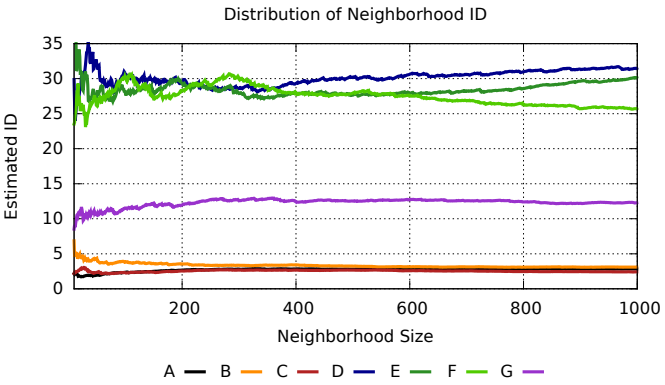
Table 6: Deviation of dimensionality estimates for 10000 manifold points with added noise (p=0.04).

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{ml1}$	$MiND_{mli}$	CD	Hein	Takens	kNNG$_1$	kNNG$_2$	PCA
m1	10	11	-10.18	-10.66	-11.80	-11.81	-11.88	-1.77	-2.78	-16.95	-35.75	-12.10	-37.61	-41.84	-22.73
m2	3	5	2.43	-1.03	-3.06	-3.10	-3.51	37.16	0.00	19.48	-18.31	23.49	-24.19	-13.93	-33.33
m3	4	6	-30.83	-32.05	-33.25	-33.42	-33.25	-22.96	-25.00	-31.20	-35.90	-22.34	-35.03	-35.03	-60.40
m7	2	3	-8.67	-14.57	-16.34	-15.58	-15.90	34.17	0.00	19.29	18.46	15.15	4.37	4.37	-66.67
m8	12	72	44.17	43.00	39.79	35.44	35.57	115.72	59.64	85.94	-11.11	85.65	-11.93	-11.93	95.21
m9	20	20	-21.77	-22.25	-24.27	-24.01	-23.98	-9.66	-17.00	-22.12	100.00	-22.68	-907.22	-907.22	-31.75
m10a	10	11	21.46	21.45	21.59	20.83	20.79	21.22	25.00	9.02	-39.56	8.18	34.35	34.35	10.00
m10c	24	25	8.04	7.93	7.51	6.89	6.88	14.43	11.76	-2.71	-30.73	-3.42	-610.20	-610.20	4.17
m11	2	3	31.66	29.06	28.64	28.43	28.50	46.73	0.00	27.64	10.00	39.00	3811.41	3811.41	-35.00
m12	20	20	-22.83	-23.17	-24.52	-23.90	-23.93	-16.52	-19.69	-16.16	6.76	-16.27	-835.80	-835.80	-26.00

Table 7: Deviation of dimensionality estimates for 10000 manifold points with added noise (p=0.16).



(a) Illustration of the distribution of k-nearest neighbor distances for $k \in [1, 1000]$ with respect to 7 points of interest.



(b) Distribution of LID estimates based on k-nearest neighbor sets for $k \in [10, 1000]$ with respect to 7 points of interest.

Figure 4: Distribution of $\mathrm{ID}_{\mathrm{MLE}}$ estimates and distance values across neighborhoods around the points of interest.

significantly affected by noise: their estimates diverge very quickly as noise is being introduced. On the other hand, the local estimators display more resistance to noise in the case of non-linear manifolds; among the local estimators, our EVT estimators tend to outperform the MiND variants.

We note that the additive noise considered in this experiment does not drastically impact the intrinsic dimensionality in the case of hypercubes. (sets m10a, m10b and m10c). That explains why PCA appears resistant to noise for the sets m10a, m10b and m10c.

6.3 Real Data

Based on our experiments on synthetic data, we expect the performance of our proposed estimators to be largely in agreement with one another. Accordingly, for clarity of presentation, for the experimentation on real data, we show results only for the MLE estimator.

Fig. 2 illustrates the distribution of LID estimates across reference points for all three data sets. The scatter plot for the *ANN_SIFT1B* data set furthermore contains several points of interest annotated with their LID values, corresponding to objects of interest which we discuss later. First, we clearly observe differences in the location of the distribution of LID values among the three data sets; for example, the mean value and standard deviation of the LID estimates for *ALOI* are considerably lower than those obtained for *ANN_SIFT1B*. More specifically, we observe mean values of $\mu_{\mathrm{ALOI}} \approx 2.2$, $\mu_{\mathrm{MNIST}} \approx 6.3$, and $\mu_{\mathrm{ANN\_SIFT1B}} \approx 12.3$, with the corresponding standard deviations of $\sigma_{\mathrm{ALOI}} \approx 1.9$, $\sigma_{\mathrm{MNIST}} \approx 2.7$, and $\sigma_{\mathrm{ANN\_SIFT1B}} \approx 3.0$. It should be noted that the measured ID within the neighborhoods that were tested is far smaller than the dimension of the full feature spaces. By plotting the same data as histograms in Fig. 3, we can furthermore see that the individual distributions of LID values differ in kurtosis and skewness as well.

The most striking difference between the individual points of interest are the distances to their respective k-nearest neighbors. Fig. 4a displays for each point of interest the specific distribution of neighbor-distances for all values of k between 1 and 1000. Interestingly, the ID measured at the points of interest appears to be associated with other properties of the respective objects. For example, distribution of neighbor-distances for objects with high corresponding dimensionality (D, E and F) indicate that these points are in some sense outliers. On the other hand, despite their distance distributions being quite dissimilar, the LID values measured at A, B, and C are nearly identical.

7. CONCLUSION

Our experimental results on synthetic data show that the estimation of LID stabilizes for sample sizes on the order of 100. However, for Theorem 2 to be applicable, one must set a sufficiently small threshold on the lower tail of the distribution, which may severely limit the number of data objects falling within the tail. Although there is a conflict between the accuracy of the estimator and the validity of the model, this conflict is resolved as the size of the dataset scales upward; it is in precisely such situations where the applications of ID have the most impact.

Estimates of local ID constitute a measure of the complexity of data. Along with other indicators such as contrast [33], LID could give researchers and practitioners more insight into the nature of their data, and therefore help them improve the efficiency and efficacy of their applications. As a tool for guiding learning processes, the proposed estimators could serve in many ways. Data collected during the retrieval processes could be automatically filtered out as noise, whenever they are associated with an unusually high ID value. In this way, the quality of query results may be enhanced as well.

The performance of content-based retrieval systems is usually assessed in terms of the precision and recall of queries on a ground truth data set. However, in high-dimensional settings it is often the case that some points are much less likely to appear in a query result than others. Unlike LID, conventional measures of complexity or performance do not account for this difficulty. LID has therefore the potential to aid in the design of fair benchmarks that truly reflect the power of retrieval systems, according to a sound, mathematically-grounded procedure.

8. ACKNOWLEDGMENTS

L. Amsaleg and T. Furon supported by French project Secular ANR-12-CORD-0014. O. Chelly, M. E. Houle and K. Kawarabayashi supported by JST ERATO Kawarabayashi Project. M. E. Houle supported by JSPS Kakenhi Kiban (A) Research Grant 25240036.

9. REFERENCES

[1] A. A. Balkema and L. de Haan. Residual Life Time at Great Age. *The Annals of Probability*, 2:792–804, 1974.

[2] N. Bingham, C. Goldie, and J. Teugels. *Regular variation*, volume 27. Cambridge University Press, 1989.

[3] N. Boujemaa, J. Fauqueur, M. Ferecatu, F. Fleuret, V. Gouet, B. LeSaux, and H. Sahbi. IKONA: Interactive Specific and Generic Image Retrieval. In *MMCBIR*, 2001.

[4] C. Bouveyron, G. Celeux, and S. Girard. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic pca. *Pattern Recogn. Lett.*, 32.

[5] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *PAMI*, 20.

[6] F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *PAMI*, 24.

[7] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. 2001.

[8] J. Costa and A. Hero. Entropic graphs for manifold learning. In *Asilomar Conf. on Signals, Sys. and Comput.*., pages 316–320 Vol.1, 2003.

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

[10] R. A. Fisher and L. H. C. Tippett. Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *Math. Proc. Cambridge Phil. Soc.*, 24:180–190, 1928.

[11] M. I. Fraga Alves, L. de Haan, and T. Lin. Estimation of the parameter controlling the speed of convergence in extreme value theory. *Math. Methods of Stat.*, 12.

[12] M. I. Fraga Alves, M. I. Gomes, and L. de Haan. A new class of semiparametric estimators of the second order parameter. *Portugalia Mathematica*, 60:193–213, 2003.

[13] B. V. Gnedenko. Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. *Ann. Math.*, 44:423–453, 1943.

[14] A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded Geometries, Fractals, and Low-Distortion Embeddings. In *FOCS*, pages 534–543, 2003.

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

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

[17] M. E. Houle. Dimensionality, Discriminability, Density & Distance Distributions. In *ICDMW*, pages 468–473, 2013.

[18] M. E. Houle. Inlierness, Outlierness, Hubness and Discriminability: an Extreme-Value-Theoretic Foundation. Technical Report 2015-002E, NII, 2015.

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

[20] M. E. Houle, X. Ma, M. Nett, and V. Oria. Dimensional Testing for Multi-Step Similarity Search. In *ICDM*, pages 299–308, 2012.

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

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

[23] H. Jégou, R. Tavenard, M. Douze, and L. Amsaleg. Searching in One Billion Vectors: Re-rank with Source Coding. In *ICASSP*, pages 861–864, 2011.

[24] I. Jolliffe. *Principal Component Analysis*. 1986.

[25] D. R. Karger and M. Ruhl. Finding Nearest Neighbors in Growth-Restricted Metrics. In *STOC*, pages 741–750, 2002.

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

[27] 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.

[28] J. Pickands, III. Statistical Inference Using Extreme Order Statistics. *Ann. Stat.*, 3:119–131, 1975.

[29] C. R. Rao. *Linear statistical inference and its applications*. 1973.

[30] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[31] 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.

[32] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998.

[33] U. Shaft and R. Ramakrishnan. Theory of nearest neighbors indexability. *ACM Trans. Database Syst.*, 31(3):814–838, 2006.

[34] F. Takens. *On the numerical determination of the dimension of an attractor*. 1985.

[35] J. Tenenbaum, V. D. Silva, and J. Langford. A global geometric framework for non linear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

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

[37] J. Venna and S. Kaski. Local Multidimensional Scaling. *Neural Networks*, 19(6–7):889–899, 2006.

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

[39] J. von Brünken, M. E. Houle, and A. Zimek. Intrinsic Dimensional Outlier Detection in High-Dimensional Data. Technical Report 2015-003E, NII, 2015.