




Extreme-value-theoretic estimation of local intrinsic dimensionality

Laurent Amsaleg¹ · Oussama Chelly²  · Teddy Furon³ · Stéphane Girard⁴ · Michael E. Houle² · Ken-ichi Kawarabayashi² · Michael Nett⁵

Received: 16 January 2016 / Accepted: 14 March 2018 / Published online: 27 July 2018
© The Author(s) 2018

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, the method of moments, probability weighted moments, and regularly varying functions. An experimental evaluation is also provided, using both real and artificial data.

Keywords Intrinsic dimension · Indiscriminability · Manifold learning · Curse of dimensionality · Maximum likelihood estimation · Extreme value theory

1 Introduction

Both the efficiency and efficacy of fundamental operations in areas such as search and retrieval, data mining, and machine learning commonly depend on the interplay between measures of data similarity and the choice of features by which objects are represented. In settings where the number of features (the so-called representational dimension) is high, similarity values tend to concentrate strongly about their

Responsible editor: Jieping Ye.

L.A. and T.F. supported by French Project Secular ANR-12-CORD-0014. O.C., M.E.H. and K.K. supported by JST ERATO Kawarabayashi Project. M.E.H. supported by JSPS Kakenhi Kiban (A) Research Grant 25240036. O.C. and M.E.H. supported by JSPS Kakenhi Kiban (B) Research Grant 15H02753.

Extended author information available on the last page of the article

respective means, a phenomenon widely referred to as ‘the curse of dimensionality’. Consequently, as the dimensionality increases, the discriminative ability of similarity measures diminishes to a point where methods that depend on them lose their effectiveness (Weber et al. 1998; Beyer et al. 1999; Pestov 2000).

The representational dimension alone cannot explain the curse of dimensionality. This can be seen from the fact that the number of degrees of freedom within a subspace or manifold is independent of the dimension of the space in which it is embedded. This number is often described as the ‘intrinsic dimensionality’ of the manifold or subspace.

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 (Jolliffe 1986; Bouveyron et al. 2011). Non-linear dimensionality reduction methods—also known as manifold learning techniques—include Isometric Mapping (Tenenbaum et al. 2000), Multi-Dimensional Scaling (Tenenbaum et al. 2000; Venna and Kaski 2006), Locally Linear Embedding and its variants (Roweis and Saul 2000), Hessian Eigenmapping Spectral Embedding (Donoho and Grimes 2003), Local Tangent Space Alignment (Zhang and Zha 2004), and Non-Linear Component Analysis (Schölkopf et al. 1998). Most dimensional reduction techniques require that a target dimension be provided by the user, although some attempt to determine an appropriate dimension automatically. Ideally, the supplied dimension should depend on the intrinsic dimensionality (ID) of the data. This has served as a prime motivation for the development of models of ID, as well as accurate estimators.

Over the past few decades, many practical models of the intrinsic dimensionality of datasets have been proposed. Examples include the previously mentioned Principal Component Analysis and its variants (Jolliffe 1986; Bouveyron et al. 2011), as well as several manifold learning techniques (Schölkopf et al. 1998; Roweis and Saul 2000; Venna and Kaski 2006; Karhunen and Joutsensalo 1994). Topological approaches to ID estimate the basis dimension of the tangent space of the data manifold from local samples (Fukunaga and Olsen 1971; Bruske and Sommer 1998; Pettis et al. 1979; Verveer and Duin 1995). Fractal measures such as the Correlation Dimension (CD) estimate ID from the space-filling capacity of the data (Faloutsos and Kamel 1994; Camastra and Vinciarelli 2002; Gupta et al. 2003). Graph-based methods use the k -nearest neighbors graph along with density in order to measure ID (Costa and Hero 2004). Parametric modeling and estimation of distribution often allow for estimators of intrinsic dimension to be derived (Larrañaga and Lozano 2002; Levina and Bickel 2004).

The aforementioned intrinsic dimensionality measures can be described as ‘global’, in that they consider the dimensionality 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) (Karger and Ruhl 2002), the generalized expansion dimension (GED) (Houle et al. 2012a), the minimum neighbor

distance (MiND) (Rozza et al. 2012), and local continuous intrinsic dimension (which we will refer to here as LID) (Houle 2013). 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.

In general, machine learning techniques that rely too strongly on local information can be accused of overfitting the data. This has motivated the development of global techniques for manifold learning such as Local Tangent Space Alignment (Zhang and Zha 2004), which first identifies manifolds restricted to neighborhoods of selected points, and then optimizes the alignment of these local structures in order to produce a more complex description of the data. The alignment process often involves an explicit penalty for overfitting. In general, local learning can compensate for overfitting by accounting for it in the final optimization process for the alignment of the local manifolds.

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 (Karger and Ruhl 2002), or to control the early termination of search (Houle et al. 2012b, 2014). They have also found applications in outlier detection, in the analysis of a projection-based heuristic (Vries et al. 2012), and in the estimation of local density (von Brünken et al. 2015). The efficiency and effectiveness of the algorithmic applications of intrinsic dimensional estimation (such as Houle et al. 2012b, 2014) depends greatly on the quality 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, where the exponent can be viewed as a form of dimension (Coles et al. 2001). Specifically, 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 in terms of a regularly-varying (RV) function whose dominant factor is a polynomial in the distance (Coles et al. 2001; Houle 2015); 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 (Coles et al. 2001). Maximum likelihood estimation of the index leads to the well-known Hill estimator for power-law distributions (Hill 1975).

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 (Houle 2015). The LID model introduces a function that assesses the discriminative power of the distribution at any given distance value (Houle 2013, 2015). 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 (Houle 2015). 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. Moreover, the interpretation of continuous ID in terms of the indiscriminability of the distance measure naturally lends itself to the design of outlier detection techniques (von Brünken et al. 2015), and in the understanding of density-related phenomena such as the hubness of data (Radovanović et al. 2010a, b; Houle 2015).

The original contributions of this paper can be summarized as:

- 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 (Hill 1975).
- 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.
- An experimental study showing the effectiveness of LID estimation when using approximate nearest neighbors.
- Profiles of several real-world datasets 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.

A preliminary version of this work was published in Amsaleg et al. (2015).

The remainder of the paper is structured as follows. The next section provides a brief introduction of the framework of continuous ID. Subsequently, in Sect. 3, we explain the relationship between continuous ID and central results from the statistical discipline of extreme value theory. Using the theory, in Sect. 4 we propose and analyze several estimators of continuous ID, using maximum likelihood estimation (MLE, which yields the Hill estimator), the method of moments (MoM), probability weighted moments (PWM), and regularly varying functions (RV). In Sect. 5 we

present our experimental study, and discuss the practical performance of our proposed estimators. We conclude the paper in Sect. 7 with a discussion of potential future applications.

2 Continuous intrinsic dimension

In this section, we survey local intrinsic dimensionality (LID), an extension of a well-studied model of intrinsic dimensionality to continuous distributions of distances proposed in Houle (2013). LID aims to quantify the local ID of a feature space exclusively in terms of the distribution of inter-point distances. Formally, let $(\mathbb{R}^m, \text{dist})$ be a domain equipped with a non-negative distance function dist . 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

$$\Pr[a \leq \mathbf{X} \leq b] = \int_a^b f_{\mathbf{X}}(x) dx,$$

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

$$F_{\mathbf{X}}(x) = \Pr[\mathbf{X} \leq x] = \int_0^x f_{\mathbf{X}}(u) du.$$

Accordingly, whenever \mathbf{X} is absolutely continuous at x , $F_{\mathbf{X}}$ is differentiable at x and its first-order derivative is $f_{\mathbf{X}}(x)$. For such settings, the *local intrinsic dimension* is defined as follows.

Definition 1 (Houle 2013) 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

$$\text{ID}_{\mathbf{X}}(x) \triangleq \lim_{\epsilon \rightarrow 0^+} \frac{\ln F_{\mathbf{X}}((1 + \epsilon)x) - \ln F_{\mathbf{X}}(x)}{\ln(1 + \epsilon)},$$

wherever the limit exists.

With respect to the *generalized expansion dimension* (Houle et al. 2012a), a precursor of LID, the above definition of $\text{ID}_{\mathbf{X}}(x)$ 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 2013) *Let \mathbf{X} be an absolutely continuous random distance variable. If $F_{\mathbf{X}}$ is both positive and differentiable at x , then*

$$ID_{\mathbf{X}}(x) = \frac{x f_{\mathbf{X}}(x)}{F_{\mathbf{X}}(x)}.$$

Local ID has potential for wide application thanks to its very general treatment of distances as continuous random variable. Direct estimation of $ID_{\mathbf{X}}$, however, requires the knowledge of the distribution of \mathbf{X} . Extreme value theory, which we survey in the following section, allows the estimation of the limit of $ID_{\mathbf{X}}(x)$ as x tends to 0 without any explicit assumptions of the data distribution other than continuity.

3 Extreme value theory

Extreme value theory is concerned with the modeling of what can be regarded as the extreme behavior of stochastic processes. It has seen applications in areas such as civil engineering (Harris 2001), operations research (Tryon and Cruse 2000; McNulty et al. 2000; Dahan and Mendelson 2001), risk assessment (Lavenda and Cipollone 2000), material sciences (Grimshaw 1993), bioinformatics (Roberts 2000), geophysics (Lavenda and Cipollone 2000), and multimedia (Furon and Jégou 2013).

We begin the introduction of extreme value theory with the following definition.

Definition 2 Let $\mu, \xi \in \mathbb{R}$ and $\sigma > 0$. The family of *generalized extreme value distributions* \mathcal{F}_{GEV} covers distributions whose cumulative distribution functions have the form

$$\mathcal{F}_{\text{GEV}} = \left\{ \begin{array}{ll} \exp\left(-\left[1 + \xi \left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right) & \text{if } \xi \neq 0 \\ \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) & \text{if } \xi = 0 \end{array} \right\}.$$

A distribution $G \in \mathcal{F}_{\text{GEV}}$ has support $\text{supp}(G) = [\mu - \frac{\sigma}{\xi}, \infty)$ whenever $\xi > 0$ and $\text{supp}(G) = (-\infty, \mu - \frac{\sigma}{\xi}]$ when $\xi < 0$. If $\xi = 0$, the support covers the complete real line.

Its best known theorem, attributed in parts to Fisher and Tippett (1928), and Gnedenko (1943), 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.

Theorem 2 (Fisher–Tippett–Gnedenko) *Let $(\mathbf{X}_i)_{i \in \mathbb{N}}$ be a sequence of independent identically-distributed random variables and let $\mathbf{M}_n = \max_{1 \leq i \leq n} \mathbf{X}_i$. If there exist a sequence of positive constants $(a_i)_{i \in \mathbb{N}}$, and a sequence of constants $(b_i)_{i \in \mathbb{N}}$, such that*

$$\lim_{n \rightarrow \infty} Pr \left[\frac{\mathbf{M}_n - b_n}{a_n} \leq x \right] = F(x),$$

for any $x \in [0, 1]$, where F is a non-degenerate distribution function, then $F \in \mathcal{F}_{\text{GEV}}$.

Extreme value theory mainly draws its power from two major results due to Fisher, Tippett and Gnedenko, as well as Balkema, de Haan and Pickands (Fisher and Tippett 1928; Gnedenko 1943; Balkema and De Haan 1974; Pickands 1975). Consider the following parametric family of distributions. This theorem is useful when observing several samples, each containing N occurrences, to estimate the distribution of the sample maxima. However, in our setting, we have only one sample; we are not interested in its extremum, but only in the $n > 1$ extreme values. For this reason, we switch to two alternative approaches to modeling extremal behavior which are more suitable to our application: threshold excesses, and regularly-varying functions.

3.1 Threshold excesses

Consider the following two definitions.

Definition 3 Let $\xi \in \mathbb{R}$ and $\sigma > 0$. The family of *generalized Pareto distributions* \mathcal{F}_{GPD} is defined by its cumulative distribution function

$$\mathcal{F}_{\text{GPD}} = \left\{ \begin{array}{ll} 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - e^{-\frac{x}{\sigma}} & \text{if } \xi = 0 \end{array} \right\}.$$

Every distribution $G \in \mathcal{F}_{\text{GPD}}$ has support $\text{supp}(G) = (\max\{0, -\frac{\sigma}{\xi}\}, \infty)$.

Definition 4 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 (1974), Pickands (1975), which can be regarded as the counterpart to the central limit theorem for extremal statistics.

Theorem 3 (Pickands–Balkema–de Haan) *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 $w \rightarrow x^+$, $F_{\mathbf{X},w}(x)$ converges to a distribution in \mathcal{F}_{GPD} .*

In the following we demonstrate a direct relation between local ID and extreme value theory, which arises as an implication of Theorem 3. 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 Coles et al. (2001), Theorem 3 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}_{GPD} , as w tends to the lower endpoint of $F_{\mathbf{X}}$ located at zero (Nett 2014). Accordingly, as w tends to zero, the distribution in the tail $[0, w)$ can be restated as follows Coles et al. (2001).

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 $\xi < 0$,*

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

Note that the distribution of excess distance $w - \mathbf{X}$ is bounded from above by w which, according to Coles et al. (2001), enforces that $\xi < 0$.

Proof Consider the distribution of threshold excess $w - \mathbf{X}$ with \mathbf{X} restricted to $[cw, w)$. According to Theorem 3, $w - \mathbf{X}$ asymptotically follows a generalized Pareto distribution:

$$\Pr[w - \mathbf{X} \leq y] \rightarrow 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}},$$

with $\sigma > 0$ and $\xi < 0$, so that

$$\Pr[\mathbf{X} \leq w - y] \rightarrow \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}}.$$

Since a distance x corresponds to a threshold excess of $w - y$,

$$F_{\mathbf{X},w}(x) = \Pr[\mathbf{X} \leq x] \rightarrow \left(1 + \frac{\xi(w - x)}{\sigma}\right)^{-\frac{1}{\xi}}.$$

We see that $F_{\mathbf{X},w}(0) = 0$ holds if and only if

$$\left(1 + \frac{\xi(w - x)}{\sigma}\right)^{-\frac{1}{\xi}} = 0,$$

implying that $\sigma = -\xi w$. With this additional constraint, the distribution of distances in the tail $[cw, w)$ simplifies to

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

□

To summarize, whenever Theorem 3 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 4 *Let \mathbf{X} be an absolutely continuous random distance variable with support $[0, \infty)$, satisfying the conditions of Theorem 3, and $w > 0$ be a distance threshold. Then, as w tends to zero,*

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

Proof (Sketch only. For a more detailed and rigorous treatment, see Houle 2015.) Lemma 1 states that under the conditions of Theorem 4, the cumulative excess distribution $F_{\mathbf{X},w}$ follows

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

as the threshold w approaches zero. The probability density $f_{\mathbf{X},w}$ in the tail of the distribution is obtained by taking the derivative with respect to x :

$$f_{\mathbf{X},w}(x) \approx \frac{\partial}{\partial x} \left(\frac{x}{w}\right)^{-\frac{1}{\xi}} = -\frac{1}{\xi w} \left(\frac{x}{w}\right)^{-\frac{1}{\xi}-1}.$$

This formulation relies on the smoothness properties of slowly varying functions (*c.f.* Sect. 3.2 and Bingham et al. 1989).

Applying Theorem 1 gives

$$\text{ID}_{\mathbf{X}}(x) \approx \frac{x \cdot f_{\mathbf{X},w}(x)}{F_{\mathbf{X},w}(x)} \rightarrow -\frac{1}{\xi}.$$

□

Note that together Lemma 1 and Theorem 4 allow us to restate the asymptotic cumulative distribution of distances in the tail $[cw, w)$ as

$$\frac{(x/w)^{\text{ID}_{\mathbf{X}}}}{F_{\mathbf{X},w}(x)} \rightarrow 1. \quad (1)$$

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

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

$F_{\mathbf{X}}$ restricted to $[0, w)$ is itself said to be *regularly varying* with index κ . In particular, a cumulative distribution $F \in \mathcal{F}_{\text{GEV}}$ 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 Bingham et al. (1989).

The following corollary is a straightforward extension of the examples given in Sect. 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 $\text{ID}_{\mathbf{X}}$.*

Proof The probability density function associated with $F_{\mathbf{X}}$ is

$$f_{\mathbf{X}}(x) = \kappa x^{\kappa-1} \ell_{\mathbf{X}}(1/x) - x^{\kappa-2} \ell'_{\mathbf{X}}(1/x).$$

From Theorem 1, we have

$$\text{ID}_{\mathbf{X}}(x) = \frac{x f_{\mathbf{X}}(x)}{F_{\mathbf{X}}(x)} = \kappa - \frac{1}{x} \frac{\ell'_{\mathbf{X}}(1/x)}{\ell_{\mathbf{X}}(1/x)}.$$

The slowly varying property of $\ell_{\mathbf{X}}$ ensures that

$$\lim_{x \rightarrow 0} \frac{1}{x} \frac{\ell'_{\mathbf{X}}(1/x)}{\ell_{\mathbf{X}}(1/x)} = 0.$$

Therefore, the index κ converges to $\text{ID}_{\mathbf{X}}$ as w tends to zero. □

The following section is concerned with deriving estimators of the intrinsic dimension based on the asymptotic distribution of distances within neighborhoods stated in Eq. 1.

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 Sect. 4.1) and moment based estimators (in Sects. 4.2, 4.3), and propose a new family of estimators based on regularly varying functions (in Sect. 4.4).

For the remainder of this discussion, we assume that we are given a distance threshold $w > 0$ and a sequence x_1, \dots, x_k of independent observations of a random distance variable \mathbf{X} with support $[0, w)$. Without loss of generality, we assume that the observations are given in ascending order—that is, $x_1 \leq x_2 \leq \dots \leq x_k$.

4.1 Maximum likelihood estimation

Maximization of the likelihood function is one of the most widely used parameter estimation techniques in statistics. The Maximum Likelihood Estimator (MLE) has no optimality guarantees for finite samples, but has the advantage of being asymptotically consistent, optimal, and efficient (in that it achieves the Cramer-Rao bound).

Definition 5 Given a random variable X with parameter θ , the likelihood of θ as a function of observations x_1, x_2, \dots, x_k is defined as

$$L(\theta | x_1, \dots, x_k) = \prod_{i=1}^k f(x_i | \theta).$$

Note that θ can be multivariate. In the case of our study, we are interested in a single parameter of the distribution, namely the shape parameter ξ of the distribution of X .

Maximizing the likelihood function is mathematically equivalent to maximizing its logarithm. It is often more convenient to work with the 'log-likelihood' function defined as follows:

Definition 6 Given a random variable X with parameter θ , the log-likelihood of θ as a function of observations x_1, x_2, \dots, x_k is defined as

$$\mathcal{L}(\theta | x_1, \dots, x_k) = \ln L(\theta | x_1, \dots, x_k).$$

Definition 7 Given a random variable X with parameter θ , and a set of observations x_1, x_2, \dots, x_k , the Maximum Likelihood Estimator (MLE) of θ is the value for which $L(\theta | x_1, \dots, x_k)$ is maximized:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta | x_1, \dots, x_k) = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta | x_1, \dots, x_k).$$

For convenience, when the sample x_1, x_2, \dots, x_k is understood, we will denote the likelihood and log-likelihood of θ by $L(\theta)$ and $\mathcal{L}(\theta)$, respectively.

By differentiating the asymptotic expression of the distance distribution given in Eq. 1, we obtain the associated probability density function $f_{\mathbf{X},w}$:

$$f_{\mathbf{X},w}(x) = \frac{F_{\mathbf{X},w}(w) \operatorname{ID}_{\mathbf{X}}}{w} \left(\frac{x}{w} \right)^{\operatorname{ID}_{\mathbf{X}}-1}.$$

Then, for a given sample of neighborhood distances x_1, x_2, \dots, x_k , we see that the log-likelihood of $\operatorname{ID}_{\mathbf{X}}$ is given by

$$\begin{aligned} \mathcal{L}(\text{ID}_{\mathbf{X}}) &= \ln \left[\prod_{i=1}^k f_{\mathbf{X},w}(x_i) \right] \\ &= \ln \left[\prod_{i=1}^k \frac{F_{\mathbf{X},w}(w) \text{ID}_{\mathbf{X}}}{w} \left(\frac{x_i}{w} \right)^{\text{ID}_{\mathbf{X}}-1} \right] \\ &= k \ln \frac{F_{\mathbf{X},w}(w)}{w} + k \ln \text{ID}_{\mathbf{X}} + (\text{ID}_{\mathbf{X}} - 1) \sum_{i=1}^k \ln \frac{x_i}{w}. \end{aligned}$$

The first- and second-order derivatives of the log-likelihood function are respectively

$$\mathcal{L}'(\text{ID}_{\mathbf{X}}) = \frac{k}{\text{ID}_{\mathbf{X}}} + \sum_{i=1}^k \ln \frac{x_i}{w} \quad \text{and} \quad \mathcal{L}''(\text{ID}_{\mathbf{X}}) = -\frac{k}{\text{ID}_{\mathbf{X}}^2}.$$

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

$$\widehat{\text{ID}}_{\mathbf{X}} = -\left(\frac{1}{k} \sum_{i=1}^k \ln \frac{x_i}{w} \right)^{-1},$$

which follows the form of the well-known Hill estimator for the scaling exponent of a power-law tail distribution (Hill 1975).

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

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

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

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

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

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

4.2 Method of moments

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

$$\mu_m = \mathbb{E}[\mathbf{X}^m] = \int_{x=0}^w x^m f_{\mathbf{X}}(x) dx = w^m \frac{\text{ID}_{\mathbf{X}}}{\text{ID}_{\mathbf{X}} + m}.$$

Solving for the intrinsic dimension gives

$$\text{ID}_{\mathbf{X}} = -m \frac{\mu_m}{\mu_m - w^m} = g\left(\frac{\mu_m}{w^m}\right),$$

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

$$\text{Var}[\hat{\mu}_m^2] = \frac{\mu_{2m} - \mu_m^2}{k} = \frac{w^{2m} \text{ID}_{\mathbf{X}} m^2}{k(\text{ID}_{\mathbf{X}} + 2m)(\text{ID}_{\mathbf{X}} + m)^2}.$$

Assuming the convergence of the empirical moments, the distribution of $\frac{\hat{\mu}_m}{w^m}$ is therefore asymptotically normal, with

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

According to Rao (2009, Th. 6a2.9), if $x \sim \mathcal{N}(\mu; \sigma^2 k^{-1})$ asymptotically, then $g(x) \sim \mathcal{N}(g(\mu); \sigma^2 k^{-1} g'(\mu)^2)$, where g' is the first-order derivative of g . Therefore, asymptotically

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

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

4.3 Probability-weighted moments

General probability-weighted moments are defined as

$$v_{m,p,q} = \mathbb{E}[F_{\mathbf{X}}(\mathbf{X})^m (1 - F_{\mathbf{X}}(\mathbf{X}))^p \mathbf{X}^q].$$

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

$$v_m \triangleq \mathbb{E}[F_{\mathbf{X}}(\mathbf{X})^m \mathbf{X}] = \int_{x=0}^w F_{\mathbf{X}}(x)^m x f_{\mathbf{X}}(x) dx = \frac{\text{ID}_{\mathbf{X}} w}{\text{ID}_{\mathbf{X}} m + \text{ID}_{\mathbf{X}} + 1};$$

solving for the intrinsic dimension yields

$$ID_{\mathbf{X}} = \frac{v_m}{w - v_m(m + 1)} = h\left(\frac{v_m}{w}\right), \quad \text{where } h(x) = \frac{x}{1 - (m + 1)x}.$$

According to Hosking and Wallis (1987) and Landwehr et al. (1979), a commonly-used estimator of the m -th probability-weighted moment of this form is

$$\hat{v}_m = \frac{1}{k} \sum_{i=1}^k \left(\frac{i - 0.35}{k}\right)^m x_i.$$

Analogously to the previous section, we can show that this estimator has variance

$$\text{Var}[\hat{v}_m] = \frac{ID_{\mathbf{X}} w^2}{(ID_{\mathbf{X}} m + ID_{\mathbf{X}} + 1)(2 ID_{\mathbf{X}} m + ID_{\mathbf{X}} + 2)}.$$

Similarly, we find that asymptotically

$$\hat{ID}_{\mathbf{X}} \sim \mathcal{N}\left(ID_{\mathbf{X}}; \frac{ID_{\mathbf{X}}^2}{k} \left(1 + \frac{(ID_{\mathbf{X}} m + 1)^2}{ID_{\mathbf{X}}(2 ID_{\mathbf{X}} m + ID_{\mathbf{X}} + 2)}\right)\right).$$

For $m = 0$, the variance is equivalent to that of the moment-based estimator with $m = 1$ (see Sect. 4.2). Since the variance increases monotonically with m for any fixed $ID_{\mathbf{X}}$, the use of lower-order probability-weighted moments is advisable.

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 Sect. 3). Consider the empirical distribution function $\hat{F}_{\mathbf{X}}$, defined as

$$\hat{F}_{\mathbf{X}}(x) = \frac{1}{k} \sum_{j=1}^k \mathbb{I}[x_j < x],$$

where $\mathbb{I}[\varphi]$ 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 8 Let \mathbf{X} be an absolutely continuous random distance variable restricted to $[0, w)$. The local intrinsic dimension $ID_{\mathbf{X}}$ can be estimated as

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

under the assumption that $x_k, \delta_k \rightarrow 0$ as $n \rightarrow \infty$, 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 $\tau_j \delta_k$ for $1 \leq j \leq J$, we may assume without loss of generality that $\tau_1 + \dots + \tau_J = 1$. The estimators are based on the observation that, for all $1 \leq j \leq J$,

$$\begin{aligned} & \ln [F_{\mathbf{X}}((1 + \tau_j \delta_k)x_k) / F_{\mathbf{X}}(x_k)] \\ &= \kappa \ln(1 + \tau_j \delta_k) + \ln [\ell_{\mathbf{X}}((1 + \tau_j \delta_k)x_k) / \ell_{\mathbf{X}}(x_k)] \\ &\simeq \kappa \ln(1 + \tau_j \delta_k). \end{aligned}$$

The RV family covers several of the known local estimators of intrinsic dimensionality. For the parameter choices $J = 1$ and $\epsilon = \tau \delta_k$, the RV estimator reduces to the GED formulation proposed in Houle et al. (2012a):

$$\widehat{\text{ID}}_{\mathbf{X}} = \frac{\ln [\hat{F}_{\mathbf{X}}((1 + \epsilon)x_k) / \hat{F}_{\mathbf{X}}(x_k)]}{\ln(1 + \epsilon)}.$$

By setting $\epsilon = 1$, Karger & Ruhl’s expansion dimension is obtained, while by setting x_k as the distance to the k -nearest neighbor and ϵ such as $(1 + \epsilon)x_k$ as the distance to the nearest neighbor, we find a special case of the MiND family (MiND_{*m*1}) (Rozza et al. 2012).

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

$$\widehat{\text{ID}}_{\mathbf{X}} = \frac{\sum_{j=1}^k \ln [j/k]}{\sum_{j=1}^k \ln [x_j/x_k]} \approx \frac{\ln \sqrt{2\pi k} - k}{\sum_{j=1}^k \ln [x_j/x_k]}.$$

As $k \rightarrow \infty$, this converges to the MLE (Hill) estimator presented in Sect. 4.1, with $w = x_k$.

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 8. 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 Alves et al. (2003b, a), 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 5 *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 $\tau_{\max} \triangleq \max_{1 \leq j \leq J} \tau_j$. Furthermore, let $\delta_k, x_k \rightarrow 0$ so that $k F_{\mathbf{X}}(x_k) \delta_k \rightarrow \infty$ and $\sqrt{k F_{\mathbf{X}}(x_k) \delta_k} \varepsilon_{\mathbf{X}}(1 / [(1 + \tau_{\max} \delta_k)x_k]) \rightarrow 0$ as k approaches*

infinity. If the auxiliary function $\varepsilon_{\mathbf{X}}$ is ultimately non-increasing, then $\sqrt{kF_{\mathbf{X}}(x_k)\delta_k} \cdot [\widehat{\text{ID}}_{\mathbf{X}} - \text{ID}_{\mathbf{X}}]$ converges to a centered Gaussian with variance

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

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

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

$$\sqrt{kF_{\mathbf{X}}(x_k)\delta_k} \varepsilon_{\mathbf{X}}(1/[r_k(1 + \tau_{\max} \delta_k)]) \rightarrow 0$$

enforces that the approximation bias $\varepsilon_{\mathbf{X}}(1/[(1 + \delta_k)x_k])$ is negligible compared to the standard deviation of the estimate, $1/\sqrt{kF_{\mathbf{X}}(x_k)\delta_k}$. We continue the analysis by proposing choices of α that minimize the variance in Theorem 5.

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 $V_0(\tau) = (\tau^\top S^{-1} \tau)^{-1}$.*

Proof The maximum of the Rayleigh functional $\alpha^\top \tau \tau^\top \alpha (\alpha^\top S \alpha)^{-1}$ is known to be attained when α is proportional to the eigenvector associated with the largest eigenvalue of $S^{-1} \tau \tau^\top$. Since $S^{-1} \tau \tau^\top$ is a rank-one matrix, the eigenvector corresponding to the unique non-zero eigenvalue is $S^{-1} \tau$. Without any loss of generality, we permute the entries of the vector τ such that $\tau_a < \tau_b$ for all $a < b$. Asymptotically, we have $0 < \tau_1 < \dots < \tau_J$. Noting that the first column of the matrix S is $(\tau_1, \tau_2, \dots, \tau_J)^\top$, we can infer that the vector $(1, 0, \dots, 0)^\top$ is a solution of the equation $S \alpha_0 = \tau$. Since S is invertible, the solution α_0 must be unique. We therefore conclude that $\alpha_0 = (1, 0, \dots, 0)^\top$. □

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 5 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*

$$\text{ID}_{\mathbf{X}} \pm u_\beta \left(k \delta_k V_0(\tau) \widehat{\text{ID}}_{\mathbf{X}} \widehat{F}_{\mathbf{X}}(x_k) \right)^{-1/2}$$

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

Proof Lemma 3 is a direct consequence of the asymptotic distribution established in Theorem 5 and the convergence of $\hat{F}_X(x_k)$ to $F_X(x_k)$ as $k \rightarrow \infty$. \square

5 Experimental framework

As part of our evaluation of our estimators of local intrinsic dimension, we investigate their performance (as well as those of competing estimators) on a series of data distributions, both real and artificially generated. While trials involving real application data are primarily of practical interest, the study of artificial data allows to systematically assess the ability of the individual methods to identify data dimensionality.

5.1 Methods

The methods used in this study include MLE, MOM, PWM, and RV. For all estimators, the neighborhood size is set to $k = 100$. The RV estimators are evaluated for the choices $J = 1$ and $J = 2$, as follows:

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

where $p = (x_i - 2x_j + x_k)/(x_k - x_j)$, $i = \lfloor k/2 \rfloor$, and $j = \lfloor 3k/4 \rfloor$. Note that the estimator RV for $J = 1$ is a form of generalized expansion dimension (GED) (Houle et al. 2012a). 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.

Our experimental framework includes several state-of-the-art estimators of intrinsic dimensionality, both local and global. The global estimators consist of a projection method (PCA), fractal methods (CD Camastra and Vinciarelli 2002; Hein and Audibert 2005; Takens 1985), and graph-based methods (kNNG₁, kNNG₂ Costa and Hero 2004). The local distance-based estimators are MiND_{m11} and MiND_{mli} (Rozza et al. 2012). 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 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.

Table 1 Parameter choices used in the experiments

Method	Parameters
CD	None
Hein	None
kNNG ₁	$k = 100, \gamma = 1, M = 1, N = 10$
kNNG ₂	$k = 100, \gamma = 1, M = 10, N = 1$
MiND _{<i>ml</i>1}	None
MiND _{<i>ml</i><i>i</i>}	$k = 100$
PCA	Threshold = 0.025
Takens	None
GED	$k = 100, J = 1$
MLE	$k = 100$
MoM	$k = 100$
PWM	$k = 100$
RV	$k = 100, J = 2$

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

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

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 $d \in \mathbb{R}$ can be selected for the generation of distances, if desired.

For choices of $d \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, 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 datasets used in our experiments have been proposed in Rozza et al. (2012). They consist of 15 manifolds of various structures 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.

Table 2 Artificial datasets used in the experiments

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

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 (Karger and Ruhl 2002; Beygelzimer et al. 2006; Houle et al. 2012b), it also serves as a characterization of high-dimensional datasets and the potential problems associated with their use in practice. To this end, we investigate the distribution of LID estimates on the following datasets, each taken from a real-world application scenario.

- The *ALOI* (*Amsterdam Library of Object Images*) data set contains a total of 110,250 color photos of 1000 different objects taken from varying view-points under various illumination conditions. Each image is described by a 641-dimensional vector of color and texture features (Boujemaa et al. 2001).
- The *ANN_SIFT1B* dataset consists of one billion 128-dimensional SIFT descriptors randomly selected from the dataset *ANN_SIFT*, consisting of 2.8×10^{10} SIFT descriptors extracted from 3×10^7 images. These sets have been created for the evaluation of nearest-neighbor search strategies at very large scales (Jégou et al. 2011).

Algorithm 1: NN-Descent

```

input : dataset  $D$ , distance function  $\text{dist}$ , neighborhood size  $k$ 
output:  $k$ -NN graph  $G$ 
1 foreach data point  $q \in D$  do
2   Initialize  $G$  by randomly generating a tentative  $k$ -NN list for  $q$  with an assigned distance of  $+\infty$ ;
3   Compute the RNN (reverse nearest neighbor) lists for  $q$ .
4 end
5 repeat
6   foreach data point  $q \in D$  do
7     Check different pairs of  $q$ 's neighbors  $(u, v)$  in  $q$ 's  $k$ -NN and RNN lists, and compute
        $\text{dist}(u, v)$ ;
8     Use  $(u, \text{dist}(u, v))$  to update  $v$ 's  $k$ -NN list, and use  $(v, \text{dist}(u, v))$  to update  $u$ 's  $k$ -NN list;
9   end
10 until  $G$  converges;
11 Return  $G$ .

```

- *BCI5* (Millán 2004) is a brain-computer interface dataset in which the classes correspond to brainwave readings taken while the subject contemplated one of three different actions (movement of the right hand, movement of the left hand, and the subvocalization of words beginning with the same letter).
- *Gisette* (Guyon et al. 2004) is a subset of the MNIST handwritten digit image dataset (LeCun et al. 1998), 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. As the dataset was created for the NIPS 2003 feature selection challenge, the precise generation mechanism of the random features was not made public.
- *Isolet* (Cole and Fanty 1990) is a set of 7797 human voice recordings in which 150 subjects recite each of the 26 letters of the alphabet twice. Each entry consists of 617 features representing selected utterances of the recording.
- The *MNIST* database (LeCun et al. 1998) contains of 70,000 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.

5.5 Approximate nearest neighbors

For many datasets, various approximate nearest neighbor (ANN) methods can generate neighborhood sets much faster than would be possible using an exact indexing method. As a rule, with ANN indexing methods it is possible to influence the trade-off between accuracy and time complexity by means of parameter choices at query time, design choices at construction time, or both. However, the use of approximate neighborhood information can lead to a degradation in the quality of data statistics that rely on it. In particular, the question arises as to how the quality of LID estimators are affected when applied to distance samples generated from approximate neighborhoods of diminishing accuracy. In this part of the experimental study, we investigate

the relationship between the accuracy of neighborhood sets and the accuracy of LID estimates. Here, accuracy is measured as the proportion of distance samples in the exact neighborhood that also appear in the approximate neighborhood under consideration. Under the assumption that the exact and approximate neighborhoods all have the same size k , this notion of accuracy coincides with those of both recall and precision.

For any given dataset, we can generate approximate k -NN sets with carefully controlled levels of accuracy, through the sparsification of exact neighbor sets of size greater than k . The sparsification is done in two steps. In the first step, we randomly select a proportion of the exact k nearest neighbors at the desired level of accuracy. In the second step, we complete the new approximate list with nearest neighbors drawn from outside the exact k -NN list, in a way that the selection rate matches the accuracy. More precisely, let r be the target level of accuracy, expressed as a proportion between 0 and 1. Initially, the approximate neighborhood distance sample is constructed by randomly selecting $\lfloor rk \rfloor$ elements of the approximate neighborhood (without replacement) from among the first k elements in the exact k -NN set. Next, an additional $k - \lfloor rk \rfloor$ elements are randomly selected from among those ranked between $k + 1$ and $K = \lceil k/r \rceil$ in the exact K -NN set, and add their distances to the sample. With this choice of K , the accuracy of the approximate k -NN query result is almost identical to that of the K -NN query result:

- for neighbors ranked between 1 and $\lfloor rk \rfloor$, the accuracy is $\lfloor rk \rfloor/k$, where

$$r \cdot \left(1 - \frac{1}{k}\right) < \frac{\lfloor rk \rfloor}{k} \leq r,$$

- for neighbors ranked between 1 and $\lceil k/r \rceil$, the accuracy is $k/\lceil k/r \rceil$, where

$$r \cdot \left(1 - \frac{1}{k+1}\right) < \frac{k}{\lceil k/r \rceil} \leq r.$$

As k increases, these upper and lower bounds converge to r .

In our experiments, to observe the effect of using ANN on LID values, we use MLE estimation with $k = 100$. The accuracy r is chosen from the range 0.5 to 1.0, since for these values, the maximum size of the exact neighborhoods required for the experimentation is a manageable $2k = 200$.

5.6 Nearest neighbor descent

The computational and storage costs associated with the construction of an exact k -nearest neighbor graph (similarity graph) is a limitation in many machine learning algorithms. Particularly in high-dimensional settings, the cost of generating all exact k -nearest neighbor lists can be quadratic in the number of data objects, which for large datasets can be prohibitively high. Many approximation methods exist for the construction of nearest neighbor (ANN) with computation costs much less than those of exact methods, though at the expense of accuracy.

We conducted an experiment to show that the process of obtaining the neighborhoods necessary for LID estimation can be considerably accelerated using a state-of-the-art ANN method, with little or no effect on LID estimates. From among the many ANN algorithms available, we chose the state-of-the-art Nearest Neighbor Descent (NN-Descent) (Dong et al. 2011) algorithm for our experimentation. The NN-Descent algorithm is based on the assumption of transitivity of the similarity measure—in other words, that two neighbors of a given data object are also likely to be neighbors of one another. As shown in the pseudo-code description of Algorithm 1, all points are initially associated with randomly built ' k -NN lists' which are then iteratively updated. At every iteration, a pivot element q is selected, and each possible pair (u, v) of q 's neighbors is considered for mutual updates. If the distance $\text{dist}(u, v)$ is smaller than the distance to the last element in u 's k -NN list, then the list is updated by inserting v in the appropriate location. The same test is applied to the k -NN list of v . In addition, similar tests are applied to the reverse (inverted) k -NN list of q . The algorithm converges when a pivot selection round completes without updates are made to the k -NN lists. As recommended in the original paper 1, we modified the convergence condition so as to terminate after a maximum of 7 rounds of the loop in lines 5–10.

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 $w = 1$).

Figure 1 shows the behavior of MLE, MOM, and RV (for choices of $J = 1$ and $J = 2$, as stated in Sect. 5.1). 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, for each of the estimators considered in this study, we present ID estimates for the artificial datasets, averaged over 20 runs each. It should be noted that as PCA and MiND_{mi} estimates are restricted to integer values, their bias is lower

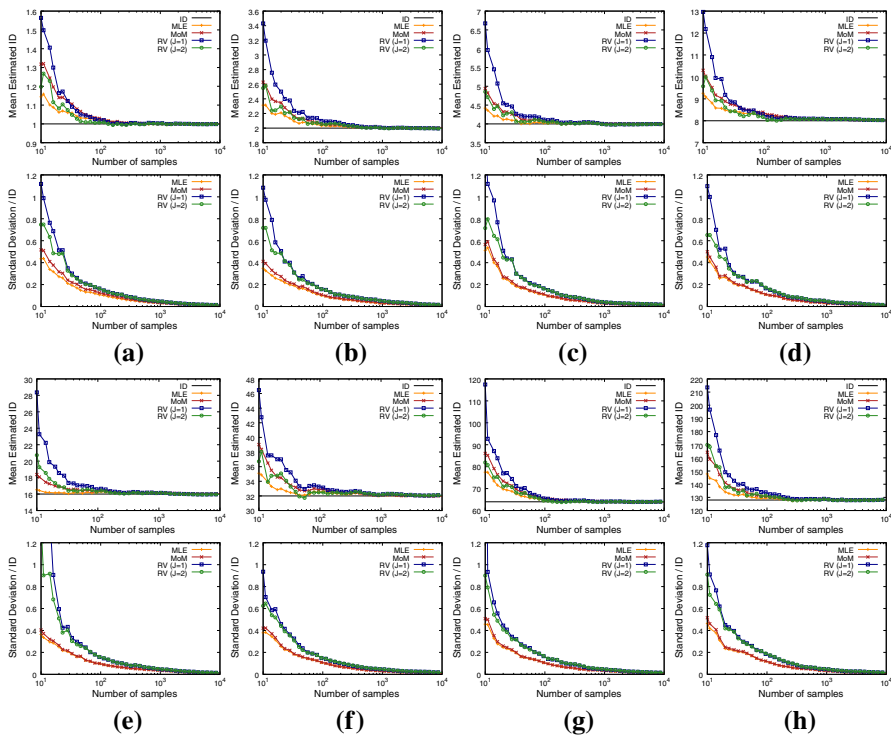


Fig. 1 Comparison of the mean and standard deviation of LID estimates provided by MLE, MOM and RV (for $J = 1$ and $J = 2$) on increasingly large samples drawn from artificially-generated distance distributions. The results cover target dimensionality values between 1 and 128. The values are marked in the corresponding plots. **a** ID = 1. **b** ID = 2. **c** ID = 4. **d** ID = 8. **e** ID = 16. **f** ID = 32. **g** ID = 64. **h** ID = 128

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). 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.

For highly non-linear manifolds, such as the Swiss Roll (m7) or the Möbius band (m11), global estimators have difficulty in identifying the intrinsic dimension. As one might expect, the local estimators ID and MiND are more accurate for such cases. Although high local curvature is reflected in the distance distribution, and consequently the local dimensional estimates as well, the effect is much smaller than for global estimators. With a higher sampling rate, k -nearest neighbor distances are diminished,

Table 3 Average ID estimates for 1000-point-manifolds using 100 nearest neighbors

Dataset	d	D	ID _{MLE}	ID _{MoM}	ID _{PWM}	ID _{GED}	ID _{RV}	MinD _{ml}	MinD _{mi}	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
m4	4	8	4.76	4.93	5.18	5.16	5.06	3.90	4.00	3.93	5.00	3.78	6.04	3.38	8.00
m5	2	3	1.98	2.03	2.07	2.03	2.00	1.97	2.00	1.95	2.30	1.98	2.27	1.99	3.00
m6	6	36	7.08	7.18	7.39	7.24	7.13	6.00	7.00	5.73	2.85	5.73	9.43	8.30	12.00
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
m10b	17	18	11.06	11.07	11.15	10.73	10.56	13.40	12.00	9.35	7.05	13.16	12.42	14.45	17.00
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
m13	1	13	1.35	1.75	2.11	2.22	2.08	1.01	1.00	1.01	1.15	1.01	1.46	1.36	7.90

Table 4 Average ID estimates for 10,000-point-manifolds using 100 nearest neighbors

Dataset	d	D	ID _{MLE}	ID _{MoM}	ID _{PWM}	ID _{GED}	ID _{RYE}	MinD _{ml}	MinD _{mi}	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
m4	4	8	4.06	4.14	4.27	4.23	4.15	3.91	4.00	3.83	4.65	3.84	3.84	3.84	8.00
m5	2	3	1.98	2.01	2.04	2.01	1.98	1.90	1.95	2.05	2.20	2.00	2.02	2.02	3.00
m6	6	36	6.64	6.78	7.11	7.01	6.89	5.85	6.00	5.05	4.30	5.66	3.34	3.34	12.00
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
m10b	17	18	12.72	12.80	13.21	12.69	12.49	13.95	13.00	13.82	6.70	13.85	19.52	19.52	17.00
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
m13	1	13	1.01	1.04	1.06	1.03	1.01	0.00	1.00	1.00	1.00	1.00	0.85	0.85	8.00

and the curvature becomes locally less significant. The local manifold structure tends to that of its tangent space, reducing the bias of local estimation. We also note that the bias is proportional to the intrinsic dimensionality of the manifold. As dimensionality increases, a higher sampling rate is required in order to reduce the bias.

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 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. However, noise in these manifolds may drive points far from their original positions, which may explain the relatively high estimates obtained by local intrinsic dimensionality estimators on these sets.

The robustness of local estimation is of great importance for many applications such as search and outlier detection. The resistance to noise seems to be generally higher in the case of manifolds of higher intrinsic dimensionality. It is important that our estimates can be trusted on these complex manifolds where the concentration effect is more important. In datasets of smaller intrinsic dimensionality, our noise model raises the dimensionality aggressively which does not happen very often in real world situations.

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.

For each of the datasets considered in this study, Fig. 2 illustrates the distribution of LID estimates based at reference points drawn from the data. Due to its large size, for the *ANN_SIFT1B* dataset, the reference set was generated by selecting 10^4 items uniformly at random. For the other datasets, the entire dataset was used as the reference set. We observe clear differences in the distribution of LID values among the datasets; for example, the center and spread of the LID estimates for *ALOI* are considerably lower than those obtained for the other datasets, whereas the LID estimates for *Gisette* are clearly higher. More precisely, we observe mean values of $\mu_{ALOI} = 4.4$, $\mu_{ANN_SIFT1B} = 12.3$, and $\mu_{GISETTE} = 49.4$. with the corresponding standard deviations of $\sigma_{ALOI} = 3.5$, $\sigma_{ANN_SIFT1B} = 3.0$, and $\sigma_{GISETTE} = 12.4$. 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.

Table 5 Deviation of ID estimates for 10,000-point-manifolds with added noise ($p = 0.01$) using 100 nearest neighbors

Dataset	d	D	ID _{MLE}	ID _{MoM}	ID _{PWM}	ID _{GED}	ID _{RVE}	MiND _{mli}	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
m4	4	8	65.02	62.32	59.25	57.21	57.59	88.75	70.00	78.85	-6.45	78.12	21.61	21.61	-18.75
m5	2	3	-48.48	-48.26	-48.04	-48.26	-48.48	-37.37	-48.72	-71.71	-54.55	-49.50	-25.74	-25.74	-66.67
m6	6	36	166.11	161.65	157.10	144.94	145.43	282.22	220.83	261.58	-30.23	221.02	219.76	219.76	131.25
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
m10b	17	18	12.89	12.73	12.49	11.66	11.69	17.35	15.38	1.37	-29.85	0.65	-36.37	-36.37	5.88
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
m13	1	13	376.24	353.85	337.74	339.81	341.58	inf	500.00	524.00	305.00	527.00	363.53	363.53	-75.00

Table 6 Deviation of ID estimates for 10,000-point-manifolds with added noise ($p = 0.04$) using 100 nearest neighbors

Dataset	d	D	ID _{MLE}	ID _{MoM}	ID _{PWM}	ID _{GED}	ID _{RVE}	MiND _{mli}	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
m4	4	8	65.02	62.32	59.25	56.97	57.35	89.00	71.25	55.61	-17.20	77.34	131.25	131.25	-20.00
m5	2	3	-48.48	-48.26	-48.04	-48.26	-48.48	-37.37	-48.72	-69.76	-54.55	-49.50	-50.99	-50.99	-66.67
m6	6	36	165.96	161.50	156.82	144.79	145.28	281.20	220.83	260.79	-2.33	220.49	714.07	714.07	130.83
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
m10b	17	18	12.89	12.73	12.49	11.66	11.69	17.20	15.38	-3.84	-36.57	0.72	-38.17	-38.17	5.88
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
m13	1	13	376.24	353.85	337.74	339.81	341.58	inf	500.00	525.00	220.00	528.00	327.06	327.06	-75.00

Table 7 Deviation of ID estimates for 10,000-point-manifolds with added noise ($p = 0.16$) using 100 nearest neighbors

Dataset	d	D	ID _{MLE}	ID _{MoM}	ID _{PWM}	ID _{GED}	ID _{RVE}	MiND _{mli}	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
m4	4	8	65.02	62.08	59.02	56.97	57.35	88.75	70.00	67.10	8.60	77.86	86.98	86.98	-20.00
m5	2	3	-48.48	-48.26	-48.04	-48.26	-48.48	-37.37	-48.72	-73.66	-54.55	-49.50	-48.02	-48.02	-66.67
m6	6	36	166.11	161.65	157.10	144.94	145.57	281.71	220.83	261.19	32.56	220.67	424.55	424.55	130.83
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
m10b	17	18	12.89	12.73	12.49	11.66	11.69	17.28	15.38	1.16	-58.21	0.51	-38.78	-38.78	5.88
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
m13	1	13	376.24	352.88	336.79	339.81	340.59	inf	500.00	491.00	270.00	528.00	387.06	387.06	-75.00

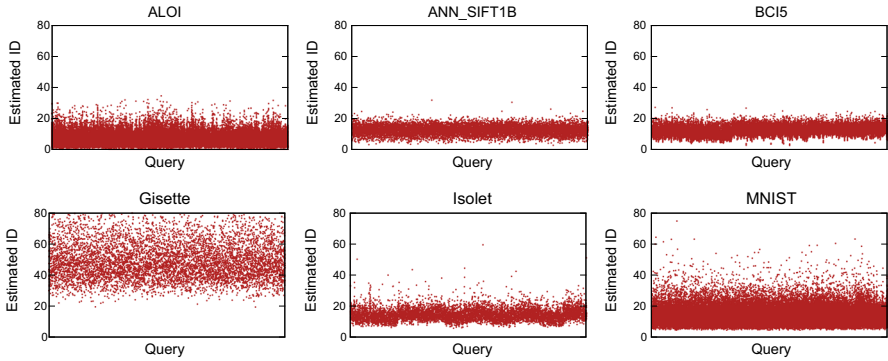


Fig. 2 Plots of the distribution of LID values across each dataset. The LID values were obtained using the MLE estimator on the size-100 neighborhoods of the individual reference points

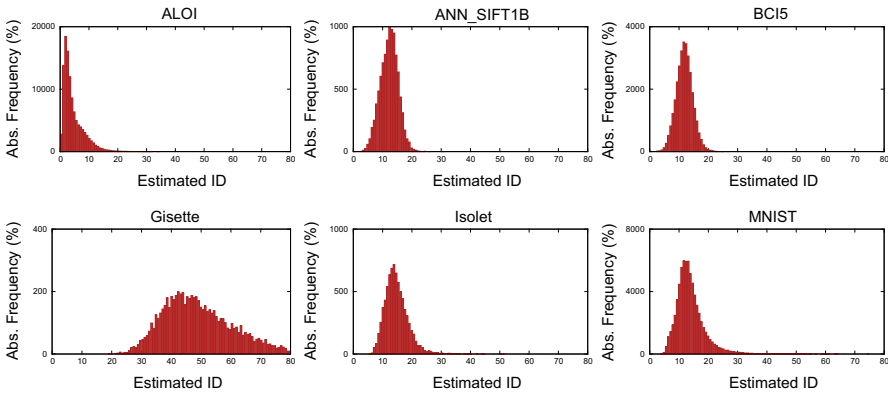


Fig. 3 Histograms of LID values across each dataset, obtained using the MLE estimator on the size-100 neighborhoods of the individual reference points

Figure 3 shows that the LID estimates for the *Gisette* dataset are very high compared to those of the other 5 sets. In particular, they are much higher than the LID values for *MNIST*, the original data set from which *Gisette* was constructed. It is clear from the LID histograms that the addition of artificial noise features in *Gisette* drastically inflates the LID values in the dataset, revealing that the generation mechanism underlying these noise features is very different from that of real-world datasets. Although this generation mechanism was not revealed by the creators of *Gisette*, local intrinsic dimension—as a measure of the subspace-filling capacity of the data—is capable of differentiating between artificial noise and natural noise.

For the *ANN_SIFT1B* dataset, from among the points of interest highlighted in the scatter plot in Fig. 2, *A*, *B* and *C* correspond to the objects for which the three lowest LID values have been estimated ($ID_A \approx 2.8$, $ID_B \approx 3.1$, and $ID_C \approx 2.4$). Likewise, the objects corresponding to *D*, *E* and *F* achieved the three greatest ID values at $ID_D \approx 31.5$, $ID_E \approx 30.1$, and $ID_F \approx 25.7$. The object *G* has been chosen as its associated dimensionality estimate ($ID_G \approx 12.3$) is closest to the mean. We

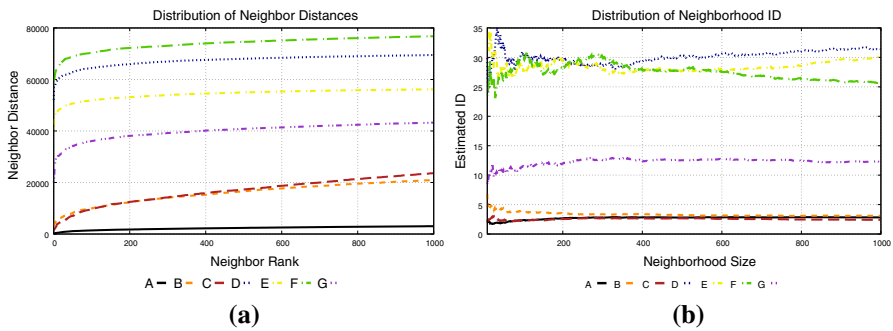


Fig. 4 Distribution of ID_{MLE} estimates and distance values across neighborhoods around the points of interest. **a** Illustration of the distribution of k -nearest neighbor distances for $k \in [1, 1000]$ with respect to 7 points of interest. **b** Distribution of LID estimates based on k -nearest neighbor sets for $k \in [10, 1000]$ with respect to 7 points of interest

subsequently investigated the distribution of distances in the neighborhoods of these points so as to gain a better understanding of why the corresponding dimensionality estimates take such low, high, or average values.

The most striking difference between the individual points of interest are the distances to their respective k -nearest neighbors. Figure 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.

6.4 Approximate nearest neighbors

This set of experiments shows that using approximate neighbors reduces the overall computation time of LID at the cost of an increase in bias. In an approximate k -NN query result, only a certain proportion of the observed distance values (equal to the accuracy of the result) correspond to distance values associated with members of an exact k -NN result. The distances associated with the approximate result can be regarded as having been generated by first sampling the dataset, and taking the distance values associated with the exact k -NN set with respect to the sample. The bias of the LID estimates for the approximate neighborhood can therefore be regarded as the result of a *sparsification* of the available distance information.

The results presented in Tables 8 and 9 show that using distances drawn from approximate neighborhoods does not lead to significant changes in estimated LID values, provided that the accuracy of the neighborhoods is reasonably high. In fact, for the datasets studied, the change in estimated LID values did not exceed 18% of the ground truth intrinsic dimension in the worst case, even with a neighborhood accuracy of 50%.

Table 8 Average ID (MLE) estimates and their standard deviation for 1000-point manifolds using 100 approximate nearest neighbors with controlled recall

Dataset	d	D	r = .5	r = .6	r = .7	r = .8	r = .9
m1	10	11	7.54 ± 0.62	7.73 ± 0.67	7.86 ± 0.72	7.93 ± 0.75	8.01 ± 0.77
m2	3	5	2.55 ± 0.31	2.60 ± 0.33	2.63 ± 0.34	2.64 ± 0.34	2.66 ± 0.35
m3	4	6	3.35 ± 0.68	3.43 ± 0.70	3.48 ± 0.71	3.50 ± 0.71	3.54 ± 0.72
m4	4	8	4.85 ± 1.10	4.87 ± 1.12	4.87 ± 1.12	4.82 ± 1.11	4.79 ± 1.10
m5	2	3	2.09 ± 0.30	2.04 ± 0.27	2.02 ± 0.25	2.00 ± 0.25	1.99 ± 0.24
m6	6	36	6.82 ± 1.18	6.94 ± 1.23	7.03 ± 1.28	7.03 ± 1.29	7.07 ± 1.30
m7	2	3	2.60 ± 0.54	2.58 ± 0.55	2.58 ± 0.55	2.54 ± 0.53	2.52 ± 0.51
m8	12	72	11.27 ± 1.44	11.62 ± 1.52	11.89 ± 1.59	12.00 ± 1.63	12.18 ± 1.67
m9	20	20	11.25 ± 0.91	11.64 ± 1.01	11.94 ± 1.09	12.08 ± 1.13	12.25 ± 1.17
m10a	10	11	6.95 ± 0.59	7.10 ± 0.64	7.22 ± 0.68	7.27 ± 0.71	7.34 ± 0.73
m10b	17	18	10.12 ± 0.83	10.44 ± 0.91	10.69 ± 0.98	10.80 ± 1.02	10.95 ± 1.05
m10c	24	25	12.70 ± 1.02	13.14 ± 1.13	13.48 ± 1.22	13.67 ± 1.27	13.88 ± 1.31
m11	2	3	2.14 ± 0.35	2.33 ± 0.42	2.45 ± 0.50	2.49 ± 0.54	2.51 ± 0.57
m12	20	20	11.08 ± 1.05	11.54 ± 1.15	11.88 ± 1.23	12.08 ± 1.27	12.32 ± 1.33
m13	1	13	1.91 ± 1.13	1.74 ± 1.02	1.63 ± 0.96	1.51 ± 0.89	1.42 ± 0.83

Table 9 Average ID (MLE) estimates and their standard deviation for 10,000-point manifolds using 100 approximate nearest neighbors with controlled recall

Dataset	d	D	r = .5	r = .6	r = .7	r = .8	r = .9
m1	10	11	8.81 ± 0.75	8.90 ± 0.81	8.97 ± 0.85	8.97 ± 0.87	9.02 ± 0.89
m2	3	5	2.82 ± 0.31	2.85 ± 0.32	2.87 ± 0.33	2.86 ± 0.34	2.87 ± 0.34
m3	4	6	3.79 ± 0.70	3.83 ± 0.71	3.85 ± 0.71	3.84 ± 0.71	3.86 ± 0.71
m4	4	8	4.19 ± 0.68	4.16 ± 0.65	4.14 ± 0.63	4.09 ± 0.60	4.08 ± 0.58
m5	2	3	1.97 ± 0.20	1.98 ± 0.21	1.98 ± 0.21	1.98 ± 0.22	1.98 ± 0.22
m6	6	36	6.85 ± 1.21	6.82 ± 1.21	6.78 ± 1.21	6.71 ± 1.20	6.68 ± 1.19
m7	2	3	1.95 ± 0.22	1.96 ± 0.23	1.96 ± 0.23	1.96 ± 0.23	1.96 ± 0.23
m8	12	72	13.48 ± 1.80	13.60 ± 1.86	13.69 ± 1.92	13.66 ± 1.94	13.71 ± 1.96
m9	20	20	13.97 ± 1.19	14.16 ± 1.29	14.30 ± 1.36	14.32 ± 1.39	14.41 ± 1.42
m10a	10	11	8.00 ± 0.75	8.08 ± 0.80	8.15 ± 0.84	8.14 ± 0.86	8.19 ± 0.88
m10b	17	18	12.32 ± 1.06	12.47 ± 1.14	12.59 ± 1.20	12.60 ± 1.24	12.67 ± 1.26
m10c	24	25	16.02 ± 1.36	16.26 ± 1.47	16.44 ± 1.55	16.47 ± 1.59	16.59 ± 1.63
m11	2	3	2.02 ± 0.20	2.01 ± 0.21	2.01 ± 0.21	2.00 ± 0.21	2.00 ± 0.22
m12	20	20	14.72 ± 1.46	14.98 ± 1.55	15.18 ± 1.62	15.24 ± 1.66	15.37 ± 1.70
m13	1	13	1.03 ± 0.12	1.02 ± 0.11	1.02 ± 0.10	1.01 ± 0.10	1.01 ± 0.10

We observe that for each of the datasets, the observed bias is inversely proportional to the neighborhood accuracy: a higher accuracy always corresponds to a lower bias, although the relationship is not linear. We also observe that the sign of the bias depends

on the curvature of the underlying manifolds within which the datasets are distributed. This trend is clear even when only 1000 points were generated within the manifolds (see Table 8). The bias is positive for the non-convex sets (m4, m5, m7, and m13). For these sets of high curvature, distance sparsification has a proportionally greater effect on the smaller distances, as compared to when the manifolds are linear. When the loss of instances of smaller distance values is higher than for larger distance values, the estimates of LID would be expected to rise.

It is important to note that estimation over neighborhoods of size 100 within a dataset of size 1000 is not in line with the asymptotic assumptions of EVT, since the neighborhood here can hardly be viewed as being derived from an extreme lower tail of the underlying distribution. However, estimation over neighborhoods of size 100 within a dataset of size 10,000 would be expected to lead to more stable results, due to the much smaller ratio of the neighborhood set size to the full dataset size. This is borne out by the experimental results shown in Table 8, where it can be seen that the approximation of neighborhood distance values has very little effect on the quality of ID estimation.

For the artificial datasets, as a representative ANN method, NN-Descent achieves extremely high accuracies while achieving useful speedups over sequential search (especially for the larger datasets). As seen in Figs. 5 and 6, average accuracies range between 99.9982 and 100%, while average execution costs range between 3 and 8 times faster than exact k -NN computation time for sets of 10,000 points, and between 15 and 41 times faster than exact k -NN computation time for sets of 100,000 points. Under these conditions, the LID estimates for all artificial datasets included in this experiment remain unchanged. For the datasets of size 1000 or less, the execution

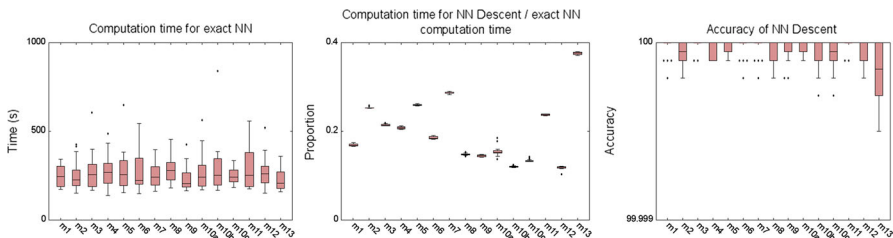


Fig. 5 Execution time and accuracy of NN-Descent compared with exact nearest neighbors' computation for 20 runs on the 10,000-point datasets

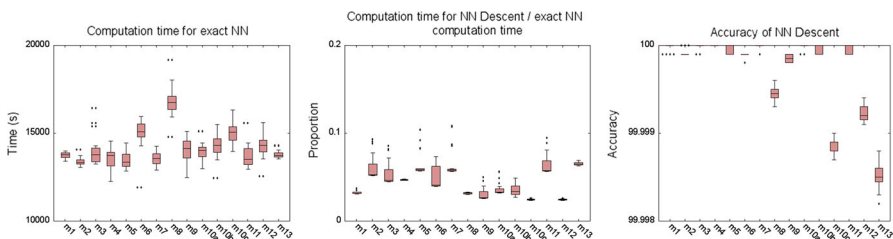


Fig. 6 Execution time and accuracy of NN-Descent compared with exact nearest neighbors' computation for 20 runs on the 100,000-point datasets

Table 10 Effect of using NN-descent

Dataset	Exact NN	NN-descent		
	Time (s)	Time (s)	Time prop.	Accuracy
ALOI	85,168	2558	0.030	0.999968
ANN_SIFT1B	2,020,520	13,305	0.007	0.945113
BCI5	1466	209	0.143	0.999995
Isolet	2523	590	0.234	1.000000
Gisette	230	111	0.486	0.999999
MNIST	50,211	2943	0.059	0.999960

cost of NN-Descent is dominated by the overheads associated with the underlying data structures. However, as shown in Fig. 6 for datasets of 100,000 points, the benefit of estimating LID with approximate neighborhoods quickly becomes apparent as the dataset size rises.

For the real-world datasets, NN-Descent achieves very high accuracies as well, while achieving important speedups over exact nearest neighbor computation. Average accuracies in all cases were at least 94.5%, as can be seen from Table 10. On the small datasets, NN-Descent accelerates the computation of nearest neighbors by no more than a factor of 2. For these small datasets, the time gain is limited by the overheads in maintaining the data structures required for NN-Descent. On the large datasets of this study, approximate nearest neighbors are obtained in up to 151 times faster than exact nearest neighbors. Due to the high accuracy of neighborhoods, LID estimates remain essentially unchanged for all datasets except for *ANN_SIFT1B*, where they deviate by only -1.82% from their original values. For most machine learning applications, such small changes in LID values would likely have little or no impact on the usefulness of the estimates.

Through these experiments, we can conclude that the use of approximate nearest neighbor computation allows LID estimation to be effectively applied at large scales. LID estimation can therefore be a viable option even for those machine learning and data mining applications where scalability is an important issue.

7 Conclusion

Our experimental results on synthetic data show that for all of the estimators of LID that we propose, the estimation stabilizes for sample sizes on the order of 100. However, for Theorem 3 to be applicable, one must set a sufficiently small threshold on the lower tail of the distribution, which may severely limit the number of samples 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.

For situations where exact neighborhood information is impractical to compute, our experimental results show that LID estimation is effective even when only approximate

neighborhood information is available. Consequently, learning machines that exploit LID values need not suffer from the high computational cost associated with the computation of exact neighborhoods.

Estimates of local ID constitute a measure of the complexity of data. Along with other indicators such as contrast (Shaft and Ramakrishnan 2006), 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 dataset. 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.

References

- Alves MF, de Haan L, Lin T (2003a) Estimation of the parameter controlling the speed of convergence in extreme value theory. *Math Method Stat* 12(2):155–176
- Alves MIF, Gomes MI, de Haan L (2003b) A new class of semi-parametric estimators of the second order parameter. *Port Math* 60(2):193–214
- Amsaleg L, Chelly O, Furon T, Girard S, Houle ME, Kawarabayashi K, Nett M (2015) Estimating local intrinsic dimensionality. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp 29–38
- Balkema AA, De Haan L (1974) Residual life time at great age. *Ann Probab* 2(5):792–804
- Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is nearest neighbor meaningful? In: *International conference on database theory*. Springer, Berlin, pp 217–235
- Beygelzimer A, Kakade S, Langford J (2006) Cover trees for nearest neighbor. In: *International conference on machine learning*. ACM, pp 97–104
- Bingham NH, Goldie CM, Teugels JL (1989) *Regular variation*, vol 27. Cambridge University Press, Cambridge
- Boujemaa N, Fauqueur J, Ferencat M, Fleuret F, Gouet V, LeSaux B, Sahbi H (2001) IKONA for interactive specific and generic image retrieval. In: *Proceedings of international workshop on multimedia content-based indexing and retrieval*
- Bouveyron C, Celeux G, Girard S (2011) Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recogn Lett* 32(14):1706–1713
- Bruske J, Sommer G (1998) Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Trans Pattern Anal Mach Intell* 20(5):572–575
- Camastra F, Vinciarelli A (2002) Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Trans Pattern Anal Mach Intell* 24(10):1404–1407
- Cole R, Fandy M (1990) Spoken letter recognition. In: *Proceedings of the third DARPA speech and natural language workshop*, pp 385–390
- Coles S, Bawa J, Trenner L, Dorazio P (2001) *An introduction to statistical modeling of extreme values*, vol 208. Springer, Berlin
- Costa JA, Hero AO III (2004) Entropic graphs for manifold learning. In: *Asilomar conference on signals, systems and computers*, vol 1. IEEE, pp 316–320
- Dahan E, Mendelson H (2001) An extreme-value model of concept testing. *Manag Sci* 47(1):102–116

- de Vries T, Chawla S, Houle ME (2012) Density-preserving projections for large-scale local anomaly detection. *Knowl Inf Syst* 32(1):25–52
- Dong W, Moses C, Li K (2011) Efficient k-nearest neighbor graph construction for generic similarity measures. In: International World Wide Web conference. ACM, pp 577–586
- Donoho DL, Grimes C (2003) Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc Natl Acad Sci* 100(10):5591–5596
- Faloutsos C, Kamel I (1994) Beyond uniformity and independence: analysis of R-trees using the concept of fractal dimension. In: Proceedings of the 13th ACM SIGACT–SIGMOD–SIGART symposium on principles of database systems. ACM, pp 4–13
- Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Math Proc Camb Philos Soc* 24(02):180–190
- Fukunaga K, Olsen DR (1971) An algorithm for finding intrinsic dimensionality of data. *IEEE Trans Comput* 100(2):176–183
- Furon T, Jégou H (2013) Using extreme value theory for image detection. Research Report RR-8244, INRIA
- Gnedenko B (1943) Sur la distribution limite du terme maximum d'une série aléatoire. *Ann Math* 44(3):423–453
- Grimshaw SD (1993) Computing maximum likelihood estimates for the Generalized Pareto Distribution. *Technometrics* 35(2):185–191
- Gupta A, Krauthgamer R, Lee JR (2003) Bounded geometries, fractals, and low-distortion embeddings. In: Proceedings of the 44th annual IEEE symposium on foundations of computer science. IEEE, pp 534–543
- Guyon I, Gunn S, Ben-Hur A, Dror G (2004) Result analysis of the NIPS 2003 feature selection challenge. In: Neural information processing systems, pp 545–552
- Harris R (2001) The accuracy of design values predicted from extreme value analysis. *J Wind Eng Ind Aerodyn* 89(2):153–164
- Hein M, Audibert JY (2005) Intrinsic dimensionality estimation of submanifolds in R^d . In: International conference on machine learning. ACM, pp 289–296
- Hill BM et al (1975) A simple general approach to inference about the tail of a distribution. *Ann Stat* 3(5):1163–1174
- Hosking JR, Wallis JR (1987) Parameter and quantile estimation for the Generalized Pareto Distribution. *Technometrics* 29(3):339–349
- Houle ME (2013) Dimensionality, discriminability, density & distance distributions. In: 13th International conference on data mining workshops. IEEE, pp 468–473
- Houle ME (2015) Inliers, outliers, hubness and discriminability: an extreme-value-theoretic foundation. Tech. Rep. 2015-002E, National Institute of Informatics
- Houle ME, Kashima H, Nett M (2012a) Generalized expansion dimension. In: 12th international conference on data mining workshops, IEEE, pp 587–594
- Houle ME, Ma X, Nett M, Oria V (2012b) Dimensional testing for multi-step similarity search. In: 12th International conference on data mining. IEEE, pp 299–308
- Houle ME, Ma X, Oria V, Sun J (2014) Efficient algorithms for similarity search in axis-aligned subspaces. In: International conference on similarity search and applications. Springer, Berlin, pp 1–12
- Jégou H, Tavenard R, Douze M, Amsaleg L (2011) Searching in one billion vectors: re-rank with source coding. In: International conference on acoustics, speech and signal processing. IEEE, pp 861–864
- Jolliffe IT (1986) Principal component analysis and factor analysis. In: Principal component analysis. Springer, Berlin, pp 115–128
- Karger DR, Ruhl M (2002) Finding nearest neighbors in growth-restricted metrics. In: ACM symposium on theory of computing. ACM, pp 741–750
- Karhunen J, Joutsensalo J (1994) Representation and separation of signals using nonlinear PCA type learning. *IEEE Trans Neural Netw* 7(1):113–127
- Landwehr JM, Matalas N, Wallis J (1979) Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resour Res* 15(5):1055–1064
- Larrañaga P, Lozano JA (2002) Estimation of distribution algorithms: a new tool for evolutionary computation, vol 2. Springer, Berlin
- Lavenda BH, Cipollone E (2000) Extreme value statistics and thermodynamics of earthquakes: aftershock sequences. *Ann Geophys* 43(5):967–982
- LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324

- Levina E, Bickel PJ (2004) Maximum likelihood estimation of intrinsic dimension. In: Neural information processing systems, pp 777–784
- McNulty PJ, Scheick LZ, Roth DR, Davis MG, Tortora MR (2000) First failure predictions for EPROMs of the type flown on the MPTB satellite. *IEEE Trans Nucl Sci* 47(6):2237–2243
- Millán JdR (2004) On the need for on-line learning in brain-computer interfaces. In: Proceedings of the IEEE international joint conference on neural networks, vol 4. IEEE, pp 2877–2882
- Nett M (2014) Intrinsic dimensional design and analysis of similarity search. Ph.D. thesis, University of Tokyo
- Pestov V (2000) On the geometry of similarity search: dimensionality curse and concentration of measure. *Inf Process Lett* 73(1):47–51
- Pettis KW, Bailey TA, Jain AK, Dubes RC (1979) An intrinsic dimensionality estimator from near-neighbor information. *IEEE Trans Pattern Anal Mach Intell* 1:25–37
- Pickands J III (1975) Statistical inference using extreme order statistics. *Ann Stat* 3(1):119–131
- Radovanović M, Nanopoulos A, Ivanović M (2010a) Hubs in space: popular nearest neighbors in high-dimensional data. *J Mach Learn Res* 11:2487–2531
- Radovanović M, Nanopoulos A, Ivanović M (2010b) Time-series classification in many intrinsic dimensions. In: Proceedings of the 2010 SIAM international conference on data mining. Citeseer, pp 677–688
- Rao CR (2009) Linear statistical inference and its applications, vol 22. Wiley, New York
- Roberts SJ (2000) Extreme value statistics for novelty detection in biomedical data processing. *Proc Sci Meas Technol* 147:363–367
- Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326
- Rozza A, Lombardi G, Ceruti C, Casiraghi E, Campadelli P (2012) Novel high intrinsic dimensionality estimators. *Mach Learn J* 89(1–2):37–65
- Schölkopf B, Smola A, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput* 10(5):1299–1319
- Shaft U, Ramakrishnan R (2006) Theory of nearest neighbors indexability. *ACM Trans Database Syst* 31(3):814–838
- Takens F (1985) On the numerical determination of the dimension of an attractor. Springer, Berlin
- Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323
- Tryon RG, Cruse TA (2000) Probabilistic mesomechanics for high cycle fatigue life prediction. *J Eng Mater Technol* 122(2):209–214
- Venna J, Kaski S (2006) Local multidimensional scaling. *IEEE Trans Neural Netw* 19(6):889–899
- Verveer PJ, Duin RPW (1995) An evaluation of intrinsic dimensionality estimators. *IEEE Trans Pattern Anal Mach Intell* 17(1):81–86
- von Brünken J, Houle ME, Zimek A (2015) Intrinsic dimensional outlier detection in high-dimensional data. Tech. Rep. 2015-003E, National Institute of Informatics
- Weber R, Schek HJ, Blott S (1998) A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: 24th international conference on very large data bases, vol 98, pp 194–205
- Zhang Z, Zha H (2004) Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *J Shanghai Univ (English Edition)* 8(4):406–424

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Laurent Amsaleg¹ · Oussama Chelly²  · Teddy Furon³ · Stéphane Girard⁴ · Michael E. Houle² · Ken-ichi Kawarabayashi² · Michael Nett⁵

✉ Michael E. Houle
meh@nii.ac.jp

Laurent Amsaleg
laurent.amsaleg@irisa.fr

Oussama Chelly
oussama.chelly@gmail.com

Teddy Furon
teddy.furon@inria.fr

Stéphane Girard
stephane.girard@inria.fr

Ken-ichi Kawarabayashi
k_keniti@nii.ac.jp

Michael Nett
mnett@google.com

- 1 Equipe LINKMEDIA - CNRS/IRISA Rennes, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France
- 2 National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
- 3 Equipe LINKMEDIA - INRIA Rennes, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France
- 4 Equipe MISTIS - INRIA Grenoble, Inovallée, 655, Montbonnot, 38334 Saint-Ismier Cedex, France
- 5 Google Japan, 6-10-1 Roppongi, Minato-ku, Tokyo 106-6126, Japan