# Local Intrinsic Dimensionality I: An Extreme-Value-Theoretic Foundation for Similarity Applications

Michael E. Houle(✉)

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

**Abstract.** Researchers have long considered the analysis of similarity applications in terms of the intrinsic dimensionality (ID) of the data. This theory paper is concerned with a generalization of a discrete measure of ID, the expansion dimension, to the case of smooth functions in general, and distance distributions in particular. A local model of the ID of smooth functions is first proposed and then explained within the well-established statistical framework of extreme value theory (EVT). Moreover, it is shown that under appropriate smoothness conditions, the cumulative distribution function of a distance distribution can be completely characterized by an equivalent notion of data discriminability. As the local ID model makes no assumptions on the nature of the function (or distribution) other than continuous differentiability, its extreme generality makes it ideally suited for the non-parametric or unsupervised learning tasks that often arise in similarity applications. An extension of the local ID model is also provided that allows the local assessment of the rate of change of function growth, which is then shown to have potential implications for the detection of inliers and outliers.

## 1 Introduction

In an attempt to alleviate the effects of high dimensionality, and thereby improve the discriminability of data, simpler representations of data are often sought by means of a number of supervised or unsupervised learning techniques. One of the earliest and most well-established simplification strategies is dimensional reduction, which seeks a projection to a lower-dimensional subspace that minimizes the distortion of the data according to a given criterion. In general, dimensional reduction requires that an appropriate dimension for the reduced space (or approximating manifold) be either supplied or learned, ideally so as to minimize the error or loss of information incurred. The dimension of the surface that best approximates the data can be regarded as an indication of the intrinsic dimensionality (ID) of the data set, or of the minimum number of latent variables needed to represent the data. Intrinsic dimensionality thus serves as an important natural measure of the complexity of data.


C. Beecks et al. (Eds.): SISAP 2017, LNCS 10609, pp. 64–79, 2017.
DOI: 10.1007/978-3-319-68474-1_5

### 1.1 Characterizations of Intrinsic Dimensionality

Over the past decades, many characterizations of ID have been proposed. The earliest theoretical measures of ID such as the classical Hausdorff dimension, Minkowski-Bouligand or 'box counting' dimension, and packing dimension, all associate a non-negative real number to metric spaces in terms of their covering or packing properties (for a general reference, see [1]). Although they are of significant theoretical importance, they are impractical for direct use in similarity applications, as the value of such measures is zero for any finite set. However, these theoretical measures have served as the foundation of practical methods for finite data samples, including the correlation dimension [2], and 'fractal' methods which estimate ID from the space-filling capacity or self-similarity properties of the data [3,4]. Other practical techniques for the estimation of ID include the topological approaches, which estimate the basis dimension of the tangent space of a data manifold from local samples (see for example [5]). In their attempt to determine lower-dimensional projective spaces or surfaces that approximate the data with minimum error, projection-based learning methods such as PCA can produce as a byproduct an estimate of the ID of the data. Parametric modeling and estimation of distribution often allow for estimators of ID to be derived [6].

An important family of dimensional models, including the minimum neighbor distance (MiND) models [5], the expansion dimension (ED) [7], generalized expansion dimension (GED) [8], and the local intrinsic dimension (LID) [9], quantify the ID in the vicinity of a point of interest in the data domain. More precisely, expansion models of dimensionality assess the rate of growth in the number of data objects encountered as the distance from the point increases. For example, in Euclidean spaces the volume of an $m$-dimensional set grows proportionally to $r^m$ when its size is scaled by a factor of $r$ — from this rate of volume growth with distance, the dimension $m$ can be deduced. Expansion models of dimensionality provide a local view of the dimensional structure of the data, as their estimation is restricted to a neighborhood of the point of interest. They hold an advantage over parametric models in that they require no explicit knowledge of the underlying global data distribution. Expansion models also have the advantage of computational efficiency: as they require only an ordered list of the neighborhood distance values, no expensive vector or matrix operations are required for the computation of estimates. Expansion models have seen applications in the design and analysis of index structures for similarity search [7,10–14], and heuristics for anomaly detection [15], as well as in manifold learning.

### 1.2 Local Intrinsic Dimensionality and Extreme Value Theory

With one exception, the aforementioned expansion models assign a measure of intrinsic dimensionality to specific sets of data points. The exception is the local intrinsic dimension ('local ID', or 'LID'), which extends the GED model to a statistical setting that assumes an underlying (but unknown) distribution of distances from a given reference point [9]. Here, each object of the data set induces a distance to the reference point; together, these distances can be regarded as

samples from the distribution. The only assumptions made on the nature of the distribution are those of smoothness.

In [9], the local intrinsic dimension is shown to be equivalent to a notion of discriminability of the distance measure, as reflected by the growth rate of the cumulative distribution function. For a random distance variable $\mathbf{X}$, with a continuous cumulative distribution function $F_{\mathbf{X}}$, the $k$-nearest neighbor distance within a sample of $n$ points is an estimate of the distance value $r$ for which $F_{\mathbf{X}}(r) = k/n$. If $k$ is fixed, and $n$ is allowed to tend to infinity, the indiscriminability of $F_{\mathbf{X}}$ at the $k$-nearest neighbor distance tends to the local intrinsic dimension. The local intrinsic dimension can thus serve to characterize the degree of difficulty in performing similarity-based operations within query neighborhoods using the underlying distance measure, asymptotically as the sample size (that is, the data set size) scales to infinity.

From the perspective of a given query point, the smallest distances encountered in a query result could be regarded as 'extreme events' associated with the lower tail of the underlying distance distribution [16]. The modeling of neighborhood distance values can thus be investigated from the viewpoint of extreme value theory (EVT), a statistical discipline concerned with the extreme behavior of stochastic processes. One of the pillars of EVT, a theorem independently proven by Balkema and de Haans [17] and by Pickands [18], states that under very reasonable assumptions, the tails of continuous probability distributions converge to a form of power-law distribution, the Generalized Pareto Distribution (GPD) [19]. In an equivalent (and much earlier) formulation of EVT due to Karamata [20], the cumulative distribution function of a tail distribution can be represented in terms of a 'regularly varying' function whose dominant factor is a polynomial in the distance [19]; the degree (or 'index') of this polynomial factor determines the shape parameter of the associated GPD. The index has been interpreted as a form of dimension within statistical contexts [19]. Many practical methods have been developed for the estimation of the index, including the well-known Hill estimator and its variants (for a survey, see [21]).

In a recent paper, Amsaleg et al. [22] developed estimators of local ID through a heuristic approximation of the true underlying distance distribution by a transformed GPD. The scale parameter of the GPD was shown to determine the local ID value. Estimators of the scale parameter of the GPD were then considered as candidates for the heuristic estimation of the local ID of the true distance distribution. Of these, the Hill estimator [23] has recently been used for ID estimation in the context of reverse $k$-NN search [14] and the analysis of non-functional dependencies among data features [24].

### 1.3 Contributions

In this paper, we revisit the intrinsic dimensionality model proposed in [9] so as to establish a firm theoretical connection between LID and EVT. The specific contributions of the paper include the following:

1. In Sect. 2.2, an overview of the LID model, extended so as to cover not only the cumulative distribution functions of distance distributions, but also a more general class of functions satisfying certain smoothness conditions.
2. In Sect. 3, a theoretical result demonstrating that any smooth functions can be fully represented in terms of an associated LID discriminability function. When applied to distance distributions, the result implies that the cumulative distribution function can be characterized entirely in terms of its discriminability, with no explicit knowledge of probability densities.
3. In Sect. 4, the development of a second-order theory of local intrinsic dimensionality that captures the growth rates within the discriminability measure itself. In the context of distance distributions, the second-order LID is shown to be a natural measure of the inlierness or outlierness of the underlying data distribution.
4. In Sect. 5, the theory developed in Sect. 3 is revealed to be a reworking of extreme value theory from first principles, for the growth rate of smooth functions from the origin. Rather than relying on the heuristic asymptotic connection to the generalized Pareto distribution that was identified in [22], we show that the LID characterization theorem is a more precise statement of the Karamata representation for the case of short-tailed distributions, with all elements of the Karamata representation being given an interpretation in terms of LID. A well-studied second-order EVT parameter governing the convergence rate of extreme values is also given an interpretation in terms of higher-order LID.

## 2 Background and Preliminaries

In this section, we give an overview of the LID model of [9], extended to account for a more general class of smooth functions (and not just cumulative distribution functions over the non-negative real domain). We begin the discussion with an overview of the expansion dimension and its applications.

### 2.1 Expansion Dimension

For the Euclidean distance metric in $\mathbb{R}^m$, increasing the radius of a ball by a factor of $\Delta$ would increase its volume by a factor of $\Delta^m$. Were we inclined to measure the volumes $V_1$ and $V_2$ of two balls of radii $r_1$ and $r_2$, with $r_2 > r_1 > 0$, taking the logarithm of their ratios would reveal the dimension $m$:

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^m \implies m = \frac{\ln(V_2/V_1)}{\ln(r_2/r_1)}\,. \tag{1}$$

The *generalized expansion dimension* (GED) can be regarded as the smallest upper bound on the values of $m$ that would be produced over a set of allowable ball placements and ball radii [8]; Karger and Ruhl's original expansion dimension (ED) further constrained $r_2$ to be double the value of $r_1$ [7].

The ED and GED have also appeared in the complexity analyses of several other similarity search structures [10,12,25]. The GED has also been successfully applied to guide algorithmic decisions at runtime for a form of adaptive search, the so-called *multi-step* similarity search problem [11,13,14,26]. In [15], a heuristic for outlier detection was presented in which approximations of the well-known local outlier factor (LOF) score [27] were calculated after projection to a lower-dimensional space. The quality of the approximation was shown to depend on a measure of expansion dimension, in which the ratio of the ball radii relates to a targeted error bound.

### 2.2 Intrinsic Dimensionality of Distance Distributions

If one accepts the observed data set as indicative of an underlying generation process, the generalized expansion dimension can be regarded as an attempt to model the worst-case growth characteristics of the distribution of distances to generated objects, as measured from a reference object drawn from $\mathcal{U}$. When the reference object $q \in \mathcal{U}$ is fixed, a supplied data set $S$ thus gives rise to a sample of values drawn from the distance distribution associated with $q$.

For finite data sets, GED formulations are obtained by estimating the volume of balls by the numbers of points they enclose [8]. In contrast, for continuous real-valued random distance variables, the notion of volume is naturally analogous to that of probability measure. As shown in [9], the generalized expansion dimension can thus be adapted for distance distributions by replacing the notion of ball set size by that of the probability measure of lower tails of the distribution. As in Eq. 1, intrinsic dimensionality can then be modeled as a function of distance $\mathbf{X} = x$, by letting the radii of the two balls be $r_1 = x$ and $r_2 = (1+\epsilon)x$, and letting $\epsilon \to 0$. The following definition (adapted from [9]) generalizes this notion even further, to any real-valued function that is non-zero in the vicinity of $x \neq 0$.

**Definition 1.** *Let $F$ be a real-valued function that is non-zero over some open interval containing $x \in \mathbb{R}$, $x \neq 0$. The* intrinsic dimensionality of $F$ at $x$ *is defined as*

$$\mathrm{IntrDim}_F(x) \triangleq \lim_{\epsilon \to 0} \frac{\ln\left(F((1+\epsilon)x)/F(x)\right)}{\ln\left((1+\epsilon)x/x\right)} = \lim_{\epsilon \to 0} \frac{\ln\left(F((1+\epsilon)x)/F(x)\right)}{\ln(1+\epsilon)},$$

*whenever the limit exists.*

Using the same assumptions on the distance distribution, [9] also proposed a natural measure of the discriminability of a random distance variable $\mathbf{X}$, in terms of the relative rate at which its cumulative distance function $F_{\mathbf{X}}$ increases as the distance increases. If $\mathbf{X}$ is discriminative at a given distance $r$, then expanding the distance by some small factor should incur a small increase in probability measure as a proportion of the value of $F_{\mathbf{X}}(r)$ (or, expressed in terms of a data sample, the proportional expansion in the expected number of data points in the neighborhood of the reference point $q$). Conversely, if the distance variable $\mathbf{X}$ is indiscriminative at distance $r$, then the proportional increase in probability

measure would be large. Accordingly, [9] defined the indiscriminability of the distance variable as the limit of the ratio of two quantities: the proportional rate of increase of probability measure, and the proportional rate of increase in distance. As with the intrinsic dimensionality formulation of Definition 1, we generalize the notion of a cumulative distribution function to any real-valued function $F(x)$ that is non-zero in the vicinity of $x$.

**Definition 2.** *Let $F$ be a real-valued function that is non-zero over some open interval containing $x \in \mathbb{R}$, $x \neq 0$. The* indiscriminability of $F$ at $x$ *is defined as*

$$\begin{aligned}\mathrm{InDiscr}_F(x) &\triangleq \lim_{\epsilon \to 0} \left[ \frac{(F((1+\epsilon)x) - F(x))}{F(x)} \Big/ \frac{(1+\epsilon)x - x}{x} \right] \\ &= \lim_{\epsilon \to 0} \frac{F((1+\epsilon)x) - F(x)}{\epsilon \cdot F(x)},\end{aligned}$$

*whenever the limit exists.*

When $F$ satisfies certain smoothness conditions in the vicinity of $x > 0$, the intrinsic dimensionality and the indiscriminability of $F$ both exist at $x$, and are equivalent. Once again, we generalize the original statement appearing in [9] so as to apply not only to distance distributions, but also to any general function $F : \mathbb{R} \to \mathbb{R}$ at values for which $F$ is both non-zero and continuously differentiable. The proof follows from applying l'Hôpital's rule to the numerator and denominator in the limits of $\mathrm{IntrDim}_F$ and $\mathrm{InDiscr}_F$; since it is essentially the same as the version in [9], we omit it here.

**Theorem 1.** *Let $F$ be a real-valued function that is non-zero over some open interval containing $x \in \mathbb{R}$, $x \neq 0$. If $F$ is continuously differentiable at $x$, then*

$$\mathrm{IntrDim}_F(x) = \mathrm{InDiscr}_F(x) = \frac{x \cdot F'(x)}{F(x)}.$$

This equivalence can be extended to those cases where $x = 0$ or $F(x) = 0$ by taking the limit of $\mathrm{IntrDim}_F(t) = \mathrm{InDiscr}_F(t)$ as $t \to x$, wherever the limit exists.

**Corollary 1.** *Let $F$ be a real-valued function that is non-zero and continuously differentiable over some open interval containing $x \in \mathbb{R}$, except perhaps at $x$ itself. Then*

$$\mathrm{ID}_F(x) \triangleq \lim_{t \to x} \frac{t \cdot F'(t)}{F(t)} = \lim_{t \to x} \mathrm{IntrDim}_F(t) = \lim_{t \to x} \mathrm{InDiscr}_F(t),$$

*whenever the limits exist.*

For values of $x$ at which $\mathrm{ID}_F(x)$ exists, we observe that $\mathrm{ID}_F(x) = \mathrm{ID}_{-F}(x)$; the LID model therefore expresses the local growth rate relative to the magnitude of $F$, regardless of its sign. Although in general $\mathrm{ID}_F$ is negative whenever $|F|$ is

decreasing, if $F$ is a cumulative distribution function, $\mathrm{ID}_F$ must be non-negative whenever it exists.

$\mathrm{ID}_F$ can be viewed interchangeably as both the intrinsic dimensionality and the indiscriminability of $F$ at $x$. However, we will henceforth refer to $\mathrm{ID}_F(x)$ as the *indiscriminability of* $F$ at $x$ whenever $x \neq 0$, and to $\mathrm{ID}_F^* \triangleq \mathrm{ID}_F(0)$ as the *local intrinsic dimension of* $F$.

## 3 ID-Based Representation of Smooth Functions

The LID formula $\mathrm{ID}_F(x) = x \cdot F'(x)/F(x)$ established in Corollary 1 simultaneously expresses the notions of local intrinsic dimensionality and indiscriminability. In general, the formula measures the instantaneous rate of change $F'(x)$ normalized by the cumulative rate of change $F(x)/x$. When $F$ is the cumulative distribution function of a distance distribution, the formula can be interpreted as a normalization of the probability density $F'(x)$ with respect to the cumulative density $F(x)/x$. The following theorem states conditions for which the indiscriminability $\mathrm{ID}_F$ fully characterizes $F$.

**Theorem 2 (Local ID Representation).** *Let $F : \mathbb{R} \to \mathbb{R}$ be a real-valued function, and let $v \in \mathbb{R}$ be a value for which $\mathrm{ID}_F(v)$ exists. Let $x$ and $w$ be values for which $x/w$ and $F(x)/F(w)$ are both positive. If $F$ is non-zero and continuously differentiable everywhere in the interval $[\min\{x,w\}, \max\{x,w\}]$, then*

$$\frac{F(x)}{F(w)} = \left(\frac{x}{w}\right)^{\mathrm{ID}_F(v)} \cdot G_{F,v,w}(x), \text{ where}$$
$$G_{F,v,w}(x) \triangleq \exp\left(\int_x^w \frac{\mathrm{ID}_F(v) - \mathrm{ID}_F(t)}{t}\,\mathrm{d}t\right),$$

*whenever the integral exists.*

*Proof.* For any $x$ and $w$ for which $x/w$ and $F(x)/F(w)$ are both positive,

$$\begin{aligned}
F(x) &= F(w) \cdot \exp\left(\ln(F(x)/F(w))\right) \\
&= F(w) \cdot \exp\left(\mathrm{ID}_F(v)\ln(x/w) + \mathrm{ID}_F(v)\ln(w/x) + \ln(F(x)/F(w))\right) \\
&= F(w) \cdot \left(\frac{x}{w}\right)^{\mathrm{ID}_F(v)} \cdot \exp\left(\mathrm{ID}_F(v)\ln(w/x) - \ln(F(w)/F(x))\right) \\
&= F(w) \cdot \left(\frac{x}{w}\right)^{\mathrm{ID}_F(v)} \cdot \exp\left(\mathrm{ID}_F(v)\int_x^w \frac{1}{t}\,\mathrm{d}t - \int_x^w \frac{F'(t)}{F(t)}\,\mathrm{d}t\right),
\end{aligned}$$

since $F$ is differentiable within the range of integration. Furthermore, since $F$ is also non-zero over the range, and since $F'$ is continuous, Corollary 1 implies that $F'(t)/F(t)$ can be substituted by $\mathrm{ID}_F(t)/t$. Combining the two integrals, the result follows. □

The representation formula in Theorem 2 can be used to characterize the behavior of the function $F$ in the vicinity of a given reference value $v$.

To see why, let us consider the value of the function at a point $w$ that is tending towards $v$. The following theorem shows that when $x$ is restricted to lie not too far from $w$, the exponential factor $G_{F,v,w}(x)$ eventually vanishes: in other words, the relationship stated in Theorem 2 tends asymptotically towards $F(x)/F(w) = (x/w)^{\mathrm{ID}_F(v)}$. This asymptotic relationship fits the intuition presented in Eq. 1 of Sect. 2.1, where the dimension is revealed by the ratios of the volumes and the radii of two balls. Here, as per the definitions of local intrinsic dimensionality and indiscriminability in Sect. 2, the role of volume is played by probability measure, and the dimension is the local ID. The asymptotic relationship is formalized in the following theorem.

**Theorem 3.** *Let $F : \mathbb{R} \to \mathbb{R}$ be a real-valued function, and let $v \in \mathbb{R}$ be a value for which $\mathrm{ID}_F(v)$ exists. Assume that there exists an open interval containing $v$ for which $F$ is non-zero and continuously differentiable, except perhaps at $v$ itself. For any fixed $c > 1$, if $v \neq 0$, then*

$$\lim_{\substack{w \to v \\ |x-v| \leq c|w-v|}} G_{F,v,w}(x) \;=\; 1\,;$$

*otherwise, if $v = 0$, then*

$$\lim_{\substack{w \to 0^+ \\ 0 < 1/c \leq x/w \leq c}} G_{F,0,w}(x) \;=\; \lim_{\substack{w \to 0^- \\ 0 < 1/c \leq x/w \leq c}} G_{F,0,w}(x) \;=\; 1\,.$$

*Proof.* For each case, it suffices to show that $\int_x^w (\mathrm{ID}_F(v) - \mathrm{ID}_F(t))/t \,\mathrm{d}t \to 0$. First we consider the case where $v = 0$. Since $\mathrm{ID}_F(v)$ is assumed to exist, for any real value $\epsilon > 0$ there must exist a value $0 < \delta < 1$ such that $|t - v| < \delta$ implies that $|\mathrm{ID}_F(t) - \mathrm{ID}_F(v)| < \epsilon$. Therefore, when $|w - v| < \delta$,

$$\left| \int_x^w \frac{\mathrm{ID}_F(v) - \mathrm{ID}_F(t)}{t} \,\mathrm{d}t \right| \;\leq\; \epsilon \cdot \left| \int_x^w \frac{1}{t} \,\mathrm{d}t \right| \;=\; \epsilon \ln \frac{w}{x}\,. \tag{2}$$

Since we have that $0 < 1/c \leq w/x \leq c$, $\ln(w/x)$ is bounded from above and below by constants. Therefore, since $\epsilon$ can be made arbitrarily small, the limit is indeed 0, and the result follows for the case $v = 0$.

Next, we consider the case where $v \neq 0$. The argument is the same as when $v = 0$, except that $\delta$ is chosen such that $0 < \delta < |v|/c$. Again, when $|w - v| < \delta$, Inequality 2 holds. Moreover, since by assumption $|x - v| \leq c|w - v| < c\delta < |v|$, we have $||v| - |x|| < c\delta$ and $||w| - |v|| < \delta$. Together, these inequalities imply that

$$0 < \frac{\delta(c-1)}{2|v|} < \frac{|v| - \delta}{|v| + c\delta} < \frac{w}{x} = \frac{|w|}{|x|} < \frac{|v| + \delta}{|v| - c\delta}\,.$$

Since $\ln(w/x)$ is once again bounded from above and below by positive constants, the limits in this case exist and are 0, and the result follows. □

For the case when $v = 0$, $x$ can be allowed to range over an arbitrarily large range relative to the magnitude of $w$, by choosing $c$ sufficiently large. However,

$x$ and $w$ must be of the same sign (either both strictly positive or both strictly negative). When $v \neq 0$, the separation between $x$ and $v$ can be much greater than that between $w$ and $v$, provided that the ratio of the two separations remains bounded by a constant — the constant can be chosen to be arbitrarily large, but once fixed, it cannot be changed.

Given a random distance variable $\mathbf{X}$, its cumulative distribution function satisfies the conditions of Theorem 2 with $v = 0$, provided that it is strictly positive and continuously differentiable within some open interval of distances with lower endpoint 0. The ID representation expresses the behavior of the entire distribution in terms of the local intrinsic dimensionality and the indiscriminability function, without the explicit involvement of a probability density function. In this sense, the indiscriminability function holds all the information necessary to reconstruct the distribution.

Taken together, Theorems 2 and 3 show that within the extreme lower tail of a smooth distance distribution, ratios of probability measure tend to a polynomial function of the corresponding ratios in distance, with degree equal to the local ID of the cumulative distribution function. If the distances were generated from a reference point in the relative interior of a local manifold to points selected uniformly at random within the manifold, the polynomial growth rate would simply be the dimension of the manifold. However, it should be noted that in general, data distributions may not be perfectly modelled by a manifold, in which case the growth rate (and intrinsic dimensionality) may not necessarily be an integer.

## 4 Second-Order Local ID

In the previous section, we saw that a smooth function $F$ can be represented in terms of its indiscriminability function $\mathrm{ID}_F$. Here, we show that a representation formula for $\mathrm{ID}_F$ can be obtained for the second-order LID function $\mathrm{ID}_{\mathrm{ID}_F}(x)$ from the first-order representation formulae for $F$ and $F'$.

### 4.1 Second-Order ID Representation

For the proof of the representation formula for $\mathrm{ID}_F$, we require two technical lemmas. The first of the two lemmas shows that the second-order LID function $\mathrm{ID}_{\mathrm{ID}_F}(x)$ can be expressed in terms of the difference between the indiscriminabilities of $F$ and $F'$. The proof is omitted due to space limitations.

**Lemma 1.** *Let $F$ be a real-valued function over the interval $I = (0, z)$, for some choice of $z > 0$ (possibly infinite). If $F$ is twice differentiable at some distance $x \in I$ for which $F(x) \neq 0$ and $F'(x) \neq 0$, then $\mathrm{ID}_F(x)$, $\mathrm{ID}_{F'}(x)$ and $\mathrm{ID}'_F(x)$ all exist, and*

$$\mathrm{ID}_{\mathrm{ID}_F}(x) = \frac{x \cdot \mathrm{ID}'_F(x)}{\mathrm{ID}_F(x)} = \mathrm{ID}_{F'}(x) + 1 - \mathrm{ID}_F(x).$$

The next technical lemma shows that the second-order LID converges to 0 as $x \to 0$. Again, the proof is omitted due to space limitations.

**Lemma 2.** *Let $F$ be a real-valued function over the interval $I = (0, z)$, for some choice of $z > 0$ (possibly infinite). If $F$ and $F'$ are twice-differentiable and either positive everywhere or negative everywhere on $I$, if $F(x) \to 0$ as $x \to 0$, and if $\mathrm{ID}^*_F$ exists, then $\mathrm{ID}^*_{F'}$ also exists, and*

$$\mathrm{ID}^*_{\mathrm{ID}_F} = \mathrm{ID}^*_{F'} + 1 - \mathrm{ID}^*_F = 0.$$

We are now in a position to state and prove a characterization of the first-order LID function in terms of the second-order LID function.

**Theorem 4 (Second-Order ID Representation).** *Let $F$ be a real-valued function over the interval $I = (0, z)$, for some choice of $z > 0$ (possibly infinite). Also, assume that $F$ and $F'$ are twice-differentiable and either positive everywhere or negative everywhere on $I$. Given any distance values $x, w \in (0, z)$, $\mathrm{ID}_F(x)$ admits the following representation:*

$$\mathrm{ID}_F(x) \;=\; \mathrm{ID}_F(w) \cdot \exp\left( - \int_x^w \frac{\mathrm{ID}_{\mathrm{ID}_F}(t)}{t} \, \mathrm{d}t \right) .$$

*Furthermore, if $F(x) \to 0$ as $x \to 0$, and if $\mathrm{ID}^*_F$ exists and is non-zero, then the representation is also valid for $x = 0$.*

*Proof.* The assumptions on $F$ and $F'$, together with Lemma 1, imply that $\mathrm{ID}_F$, $\mathrm{ID}_{F'}$, $\mathrm{ID}'_F$ and $\mathrm{ID}_{\mathrm{ID}_F}$ exist everywhere, and that $\mathrm{ID}_F(x)$ and $\mathrm{ID}_F(w)$ are non-zero and share the same sign. We can therefore establish the result for the case where $x > 0$, as follows:

$$\begin{aligned} \mathrm{ID}_F(x)/\mathrm{ID}_F(w) &= \exp \ln \left( \mathrm{ID}_F(x)/\mathrm{ID}_F(w) \right) \\ &= \exp\left( - \int_x^w \frac{\mathrm{ID}'_F(t)}{\mathrm{ID}_F(t)} \, \mathrm{d}t \right) \;=\; \exp\left( - \int_x^w \frac{\mathrm{ID}_{\mathrm{ID}_F}(t)}{t} \, \mathrm{d}t \right) , \end{aligned}$$

where the last step follows from Theorem 1. If $F(x) \to 0$ as $x \to 0$, and if $\mathrm{ID}^*_F$ exists and is non-zero, by Lemma 2 we have that $\mathrm{ID}^*_{F'}$ exists, and that $\mathrm{ID}^*_{\mathrm{ID}_F} = 0$. Since $\mathrm{ID}_F(w)$ is also non-zero, the integral in the representation formula must converge, and therefore the representation is valid for $x = 0$ as well. □

### 4.2 Inlierness, Outlierness and LID

Local manifold learning techniques such as Locally-Linear Embedding [28] typically model data dimensionality as the dimension of a manifold that well approximates the data within a region of interest. Under these assumptions, with respect to given reference point $\mathbf{q}$ on the manifold, the model assumes that the data distribution within a neighborhood of $\mathbf{q}$ tends to uniformity as the radius of the neighborhood tends to zero. The local ID of the manifold at $\mathbf{q}$ is simply the

value of $\mathrm{ID}_F^*$, where $F$ is the cumulative distribution function for the induced distance distribution from $\mathbf{q}$. In addition, the indiscriminability function $\mathrm{ID}_F$ can indicate whether $\mathbf{q}$ should be regarded as an inlier or as an outlier relative to its locality within the manifold, as the following argument shows.

If $\mathrm{ID}_F(x) < \mathrm{ID}_F^*$ throughout a neighborhood of $\mathbf{q}$ of radius $0 < x < \epsilon$ (where $\epsilon > 0$ is chosen to be sufficiently small), then from the local ID representation formula of Theorem 2, we observe that $G_{F,0,\epsilon}(x)$ is greater than 1, and that $F(\epsilon)/F(x) < (\epsilon/x)^{\mathrm{ID}_F^*}$. Consequently, the growth rate in probability measure within distance $x$ from $\mathbf{q}$ is less than would be expected for a locally-uniform distribution of points within a manifold of dimension $\mathrm{ID}_F^*$. The drop in indiscriminability (or rise in discriminability) indicates a decrease in local density as the distance from $\mathbf{q}$ increases. Under this interpretation, the relationship between $\mathbf{q}$ and its neighborhood can be deemed to be that of an *inlier*.

By similar arguments, if instead $\mathrm{ID}_F(x) > \mathrm{ID}_F^*$, then a rise in indiscriminability (or drop in discriminability) would indicate an increase in local density as the distance from $\mathbf{x}$ increases, in which case $\mathbf{q}$ would be an *outlier* with respect to its neighborhood.

Within a small local neighborhood $0 < x < \epsilon$, the condition $\mathrm{ID}_F(x) < \mathrm{ID}_F^*$ is equivalent to that of $\mathrm{ID}'_F(x) < 0$, and the condition $\mathrm{ID}_F(x) > \mathrm{ID}_F^*$ is equivalent to that of $\mathrm{ID}'_F(x) > 0$. The strength of the inlierness or outlierness of $\mathbf{q}$ can be judged according to the magnitude $|\mathrm{ID}'_F(x)|$. However, for ease of comparison across manifolds of different intrinsic dimensions, and across different distances $x$, $|\mathrm{ID}'_F(x)|$ should be normalized with respect to these two quantities. The second-order LID function $\mathrm{ID}_{\mathrm{ID}_F}(x) = x \cdot \mathrm{ID}'_F(x)/\mathrm{ID}_F(x)$ can thus be viewed as a natural measure of the inlierness (when negative) or outlierness (when positive) of $\mathbf{q}$, one that normalizes the relative rate of change of the LID function with respect to the average (radial) rate of change of LID within distance $x$ of $\mathbf{q}$, namely $\mathrm{ID}_F(x)/x$.

As an illustration of the ability of second-order LID to naturally determine the inlierness or outlierness of a point with respect to a data distribution, let us consider a Gaussian distribution in $\mathbb{R}^m$ generated as a vector of normally distributed random variables with means $\mu_i$ and variances $\sigma_i^2$, for $1 \leq i \leq m$. Then the normalized distance from the origin to a point $\mathbf{X} = (X_1, X_2, \ldots, X_m)$, defined as $Z = \sqrt{\sum_{i=1}^m (X_i^2/\sigma_i^2)}$, follows a noncentral chi distribution. Although the details are omitted due to the complexity of the derivations, Theorem 1 can be applied to the probability density function for $Z$ to show that

$$\mathrm{ID}_{F_Z}^* = m\,, \quad \text{and} \quad \mathrm{ID}_{\mathrm{ID}_{F_Z}}^* = 2$$

where $\lambda = \sqrt{\sum_{i=1}^m (\mu_i^2/\sigma_i^2)}$ is a distributional parameter representing the normalized distance between the Gaussian mean and the origin. Moreover, as $z$ tends to 0, the sign of $\mathrm{ID}_{\mathrm{ID}_{F_Z}}(z)$ is positive when $\lambda > \sqrt{m}$, and negative when $0 \leq \lambda < \sqrt{m}$, indicating 'outlierness' of the tail region of the Gaussian beyond the inflection boundary $\lambda = \sqrt{m}$, and 'inlierness' of the central region. It is worth noting that the strength of outlierness or inlierness is a constant value,

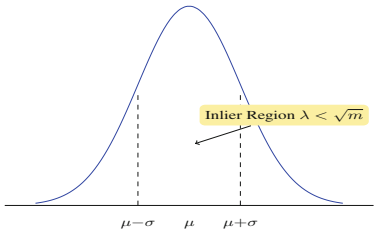


**Fig. 1.** The inlier region of a 1-dimensional Gaussian distribution. The boundary between the inlier (central) region and the outlier (tail) regions is at normalized distance $\lambda = \sqrt{m} = 1$, or equivalently at $|x - \mu| = \sigma$.

regardless of the actual dimension $m$, or of the (normalized) distance $\lambda$ to the Gaussian center. The 1-dimensional case is illustrated in Fig. 1.

## 5 Local ID and Extreme Value Theory

The characterization of continuous distance distributions established from first principles in Sect. 3 can be regarded as an elucidation of extreme value theory (EVT) in the setting of short-tailed distributions. Several mutually-equivalent formulations of EVT exist; here, the formulation with which we will concern ourselves is that of regularly varying functions, pioneered by Karamata in the 1930s. There is a vast literature on EVT and its applications, the majority of which involve the upper tails of distributions. For a detailed account of regular variation and EVT, see (for example) [29].

### 5.1 First-Order EVT

Let $F$ be a function that is continuously differentiable and strictly positive over the open interval $I = (0, z)$ for some $z > 0$. Although Karamata's original representation theorem [20] deals with the behavior of smooth functions as they diverge to (positive) infinity, the theorem can be reformulated by applying a reciprocal transformation of the function domain $(1/z, \infty)$ into the interval $I$; this yields the result that the function $F$ restricted to $I$ can be expressed in the form $F(x) = x^{\gamma}\ell(1/x)$ for some constant $\gamma$, where $\ell$ is differentiable and *slowly varying* (at infinity); that is, for all $c > 0$, $\ell$ satisfies

$$\lim_{u \to \infty} \frac{\ell(cu)}{\ell(u)} = 1.$$

The function $F$ restricted to $I$ is itself said to be *regularly varying* with index $\gamma$.

Note that the slowly-varying component $\ell(u)$ is not necessarily constant as $u \to \infty$. However, the slowly-varying condition ensures that the derivative $\ell'(u)$ is bounded, and that the following auxiliary function tends to 0:

$$\varepsilon(u) \triangleq \frac{u\ell'(u)}{\ell(u)}, \qquad \lim_{u \to \infty} \varepsilon(u) \to 0\,.$$

Slowly varying functions are also known to be representable in terms of their auxiliary function. More specifically, $\ell(1/x)$ can be shown to be slowly varying as $1/x \to \infty$ if and only if there exists some $w > 0$ such that

$$\ell(1/x) \;=\; \exp\left(\eta(1/x) + \int_{1/w}^{1/x} \frac{\varepsilon(u)}{u}\,\mathrm{d}u\right),$$

where $\eta$ and $\varepsilon$ are measurable and bounded functions such that $\eta(1/x)$ tends to a constant, and $\varepsilon(1/t)$ tends to 0, as $x$ and $t$ tend to 0. Note that under the substitution $t = 1/u$, the slowly-varying component can be expressed as

$$\ell(1/x) \;=\; \exp\left(\eta(1/x) + \int_{x}^{w} \frac{\varepsilon(1/t)}{t}\,\mathrm{d}t\right).$$

Thus the formula $F(x) = x^{\gamma}\ell(1/x)$ can easily be verified to fit the form of the representation given in Theorem 2, with the following choices:

$$\gamma \;=\; \mathrm{ID}^{*}_{F}\,; \qquad \eta(1/x) \;=\; \ln F(w) - \mathrm{ID}^{*}_{F} \ln w\,; \qquad \varepsilon(1/t) \;=\; \mathrm{ID}^{*}_{F} - \mathrm{ID}_{F}(t)\,.$$

### 5.2 Second-Order EVT

An issue of great importance and interest in the design and performance of semi-parametric EVT estimators is the speed of convergence of extreme values to their limit [30]. As is the case with first-order EVT, many approaches to the estimation of second-order parameters have been developed [21].

Here, we will follow the formulation appearing in [31] using second-order regular variation. In their paper, de Haan and Resnick proved the equivalence of two conditions regarding the derivatives of regularly varying functions, which can be stated as follows. Let $\phi : (0, \infty) \to \mathbb{R}$ be twice differentiable, with $\phi'(t)$ eventually positive as $t \to \infty$, and let $\gamma \in \mathbb{R}$. Consider a function $A(t)$ whose absolute value is regularly varying with index $\rho \leq 0$, such that $A(t) \to 0$ as $t \to \infty$ with $A(t)$ either eventually positive or eventually negative. Then the condition

$$A(t) \triangleq \frac{t \cdot \phi''(t)}{\phi'(t)} - \gamma + 1$$

is equivalent to $\phi'$ having the following representation for some non-zero constant $k$:

$$\phi'(t) = k \cdot t^{\gamma - 1} \cdot \exp\left(\int_{1}^{t} \frac{A(u)}{u}\,\mathrm{d}u\right).$$

As in the discussion of first-order EVT in Sect. 5.1, we apply a reciprocal transform of the domain to an interval of the form $I = (0, w)$, by setting $t = 1/x$ and $\phi'(t) = F'(x)$. Noting that $F''(x) = -t^2\phi''(t)$, and defining $B(x) \triangleq A(t)$, the first condition can be shown to be

$$B(x) \triangleq 1 - \gamma - \frac{x \cdot F''(x)}{F'(x)} = 1 - \gamma - \mathrm{ID}_{F'}(x),$$

and, under the substitution $u = 1/y$, the second condition can be shown to be

$$F'(x) = k \cdot x^{1-\gamma} \cdot \exp\left(\int_x^1 \frac{B(y)}{y}\,\mathrm{d}y\right) .$$

Thus these equivalent conditions can be verified to fit the form of the representation given in Theorem 2, with $w = 1$, $v = 0$, and

$$\begin{aligned} k &= F'(1); \\ \gamma &= 1 - \mathrm{ID}^*_{F'} = 2 - \mathrm{ID}^*_F\,; \\ B(x) &= 1 - \gamma - \mathrm{ID}_{F'}(x) = \mathrm{ID}^*_F - 1 - \mathrm{ID}_{F'}(x)\,. \end{aligned}$$

Second-order EVT is largely concerned with the estimation of the parameter $\rho$. The following theorem establishes that the two functions $B(x)$ and $\mathrm{ID}_{\mathrm{ID}_{F'}}(x)$ both have as their index of regular variation the non-negative value $-\rho$.

**Theorem 5.** *Let $F$ be a function that is twice differentiable over the interval $I = (0, z)$, for some choice of $z > 0$ (possibly infinite). Furthermore, assume that $F'$ and $F''$ are positive everywhere or negative everywhere over $I$, that $F'(x) \to 0$ as $x \to 0$, and that $\mathrm{ID}^*_F$ exists. Let $B(x) = \mathrm{ID}^*_F - 1 - \mathrm{ID}_{F'}(x)$. Then $B(x)$ and $B_*(x) \triangleq \mathrm{ID}_{\mathrm{ID}_{F'}}(x)$ are both regularly varying with index $-\rho$. Furthermore, if $B_*$ is continuously differentiable, then $-\rho = \mathrm{ID}^*_{B_*}$.*

The proof relies heavily on Lemma 2 and Theorem 4. However, due to space limitations, the details are omitted in this version of the paper.

## 6 Conclusion

Among the implications of the extreme-value-theoretic foundation introduced in this paper, perhaps the one with the greatest potential impact for similarity applications is that intrinsic dimensionality reveals the interchangeability between probability and distance. For distance distributions, the ID representation formula of Theorem 2 essentially states that the ratio of the expected numbers of points in neighborhoods of different radii asymptotically tends to the ratio of the neighborhood radii themselves, raised to the power of the intrinsic dimension. Knowledge of any 4 of these 5 quantities would help to determine the value of the unknown quantity. Indeed, this relationship among probability, distance and ID has already been successfully exploited to improve the accuracy/time tradeoff of certain similarity search tasks, via dimensional testing [11–14].

To realize the full potential of the theory of local intrinsic dimensionality for similarity applications, it is essential that accurate and efficient estimators be available. Estimators for the first-order EVT scale parameter have been developed within the EVT community; generally, they require on the order of 100 neighborhood distance samples in order to converge [22]. However, second-order EVT estimators generally require many thousands of neighbors for convergence [32]. Reducing the sample size for both first- and second-order LID/EVT estimation would be a worthwhile target.

Another important future research direction is that of feature selection and metric learning. The LID model provides a natural measure of data discriminability that could in principle be used to guide the selection of features, or the learning of similarity measures. Towards this goal, in a companion paper [33], a theoretical investigation is made into how the local IDs of distance distributions can change as their cumulative distribution functions are combined.

**Acknowledgments.** The author gratefully acknowledges the financial support of JSPS Kakenhi Kiban (A) Research Grant 25240036 and JSPS Kakenhi Kiban (B) Research Grant 15H02753.

## References

1. Falconer, K.: Fractal Geometry: Mathematical Foundations and Applications. Wiley, Hoboken (2003)
2. Grassberger, P., Procaccia, I.: Measuring the strangeness of strange attractors. Physica D **9**(1–2), 189–208 (1983)
3. Camastra, F., Vinciarelli, A.: Estimating the intrinsic dimension of data with a fractal-based method. IEEE TPAMI **24**(10), 1404–1407 (2002)
4. Gupta, A., Krauthgamer, R., Lee, J.R.: Bounded geometries, fractals, and low-distortion embeddings. In: FOCS, pp. 534–543 (2003)
5. Rozza, A., Lombardi, G., Ceruti, C., Casiraghi, E., Campadelli, P.: Novel high intrinsic dimensionality estimators. Mach. Learn. J. **89**(1–2), 37–65 (2012)
6. Larrañaga, P., Lozano, J.A.: Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation, vol. 2. Springer, Heidelberg (2002)
7. Karger, D.R., Ruhl, M.: Finding nearest neighbors in growth-restricted metrics. In: STOC, pp. 741–750 (2002)
8. Houle, M.E., Kashima, H., Nett, M.: Generalized expansion dimension. In: ICDMW, pp. 587–594 (2012)
9. Houle, M.E.: Dimensionality, discriminability, density & distance distributions. In: ICDMW, pp. 468–473 (2013)
10. Beygelzimer, A., Kakade, S., Langford, J.: Cover trees for nearest neighbors. In: ICML, pp. 97–104 (2006)
11. Houle, M.E., Ma, X., Nett, M., Oria, V.: Dimensional testing for multi-step similarity search. In: ICDM, pp. 299–308 (2012)
12. Houle, M.E., Nett, M.: Rank-based similarity search: reducing the dimensional dependence. IEEE TPAMI **37**(1), 136–150 (2015)
13. Houle, M.E., Ma, X., Oria, V., Sun, J.: Efficient similarity search within user-specified projective subspaces. Inf. Syst. **59**, 2–14 (2016)
14. Casanova, G., Englmeier, E., Houle, M.E., Kröger, P., Nett, M., Zimek, A.: Dimensional testing for reverse $k$-nearest neighbor search. PVLDB **10**(7), 769–780 (2017)
15. de Vries, T., Chawla, S., Houle, M.E.: Density-preserving projections for large-scale local anomaly detection. Knowl. Inf. Syst. **32**(1), 25–52 (2012)
16. Furon, T., Jégou, H.: Using Extreme Value Theory for Image Detection. Research report RR-8244, INRIA, February 2013
17. Balkema, A.A., de Haan, L.: Residual life time at great age. Ann. Probab. **2**, 792–804 (1974)
18. Pickands, J.: Statistical inference using extreme order statistics. Ann. Stat. **3**, 119–131 (1975)

19. Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer, London (2001)
20. Karamata, J.: Sur un mode de croissance réguliere des fonctions. Mathematica (Cluj) **4**, 38–53 (1930)
21. Gomes, M.I., Canto e Castro, L., Fraga Alves, M.I., Pestana, D.: Statistics of extremes for IID data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions. Extremes **11**, 3–34 (2008)
22. Amsaleg, L., Chelly, O., Furon, T., Girard, S., Houle, M.E., Kawarabayashi, K., Nett, M.: Estimating local intrinsic dimensionality. In: KDD, pp. 29–38 (2015)
23. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**(5), 1163–1174 (1975)
24. Romano, S., Chelly, O., Nguyen, V., Bailey, J., Houle, M.E.: Measuring dependency via intrinsic dimensionality. In: ICPR, pp. 1207–1212 (2016)
25. Krauthgamer, R., Lee, J.R.: Navigating nets: simple algorithms for proximity search. In: SODA, pp. 798–807 (2004)
26. Houle, M.E., Ma, X., Oria, V.: Effective and efficient algorithms for flexible aggregate similarity search in high dimensional spaces. IEEE TKDE **27**(12), 3258–3273 (2015)
27. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. SIGMOD Rec. **29**(2), 93–104 (2000)
28. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
29. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: Statistics of Extremes: Theory and Applications. Wiley, Hoboken (2004)
30. de Haan, L., Stadtmüller, U.: Generalized regular variation of second order. J. Aust. Math. Soc. (Series A) **61**(3), 381–395 (1996)
31. de Haan, L., Resnick, S.: Second-order regular variation and rates of convergence in extreme-value theory. Ann. Probab. **24**(1), 97–124 (1996)
32. Fraga Alves, M.I., de Haan, L., Lin, T.: Estimation of the parameter controlling the speed of convergence in extreme value theory. Math. Methods Stat. **12**(2), 155–176 (2003)
33. Houle, M.E.: Local intrinsic dimensionality II: multivariate analysis and distributional support. In: SISAP, pp. 1–16 (2017)