ISSN 1346-5597

National Institute of Informatics

NII Technical Report

Inlierness, Outlierness, Hubness and Discriminability: an Extreme-Value-Theoretic Foundation

Michael E. Houle

NII-2015-002E
Mar. 2015

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

Noname manuscript No.
(will be inserted by the editor)

Inlierness, Outlierness, Hubness and Discriminability: an Extreme-Value-Theoretic Foundation

Michael E. Houle

March 2015

Abstract For many large-scale applications in data mining, machine learning, and multimedia, fundamental operations such as similarity search, retrieval, classification, clustering, and anomaly detection generally suffer from an effect known as the 'curse of dimensionality'. As the dimensionality of the data increases, distance values tend to become less discriminative due to their increasing relative concentration about the mean of their distribution. For this reason, researchers have considered the analysis of similarity applications in terms of measures of the intrinsic dimensionality (ID) of the data sets. This theory paper is concerned with a generalization of a discrete measure of ID, the expansion dimension, to the case of continuous distance distributions. This notion of the ID of a distance distribution is shown to precisely coincide with a natural notion of the indiscriminability of distances, thereby establishing a theoretically-founded relationship among probability density, the cumulative density (cumulative probability divided by distance), intrinsic dimensionality, and discriminability. The indiscriminability function proposed in this paper is shown to completely determine an extreme-value-theoretic representation of the distance distribution. From this representation, a characterization in terms of continuous ID is derived for the notions of outlierness and inlierness of data, as well as the hubness phenomenon in data sets.

The author acknowledges the financial support of JSPS Kakenhi Kiban (A) Research Grant 25240036, the JST ERATO Kawarabayashi Large Graph Project, and the JST ERATO Minato Discrete Structure Manipulation System Project.

M. E. Houle
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
Tel.: +81-3-4212-2538
Fax: +81-3-4212-2120
E-mail: meh@nii.ac.jp

Keywords Distance distribution · discriminability · intrinsic dimensionality · extreme value theory · outlierness · hubness

1 Introduction

In such areas as search and retrieval, data mining, machine learning, multimedia, recommendation systems, and bioinformatics, the efficiency and efficacy of many fundamental operations commonly depend on the interplay between measures of data similarity and the choice of features by which objects are represented. Similarity search, perhaps the most fundamental operation involving similarity measures, is ubiquitous in data analysis tasks such as clustering, k-nearest-neighbor classification, and anomaly detection, as well as content-based multimedia applications.

One of the most common strategies employed in similarity search is that of neighborhood expansion, in which the radius of the search (or, equivalently, the number of points visited) is increased until a neighborhood of the desired size has been identified. Even when this radius is known in advance, the actual number of points visited can be considerably larger than the target neighborhood size, particularly if the similarity measure is not discriminative. A highly indiscriminative similarity measure is more susceptible to measurement error, and (in the case of distance metrics) is less suited to classical methods for search path pruning based on the triangle inequality.

1.1 Discriminability and Dimensionality

Much anecdotal and empirical evidence exists linking the discriminability of similarity measures and the dimensionality of data sets [5,6,36,79,80]. When the number of features (the 'representational dimension') is high, similarity values tend to concentrate strongly about their respective means, a phenomenon widely referred to as the 'curse of dimensionality'. In the face of this concentration effect, a very slight relative increase in the search radius can result in the discovery of an unmanageably-large number of new objects. Consequently, as the dimensionality increases, the discriminative ability of similarity measures can diminish to the point where methods that depend on them lose their effectiveness altogether [6,59,80]. Most theoretical studies of the relationship between discriminability and dimensionality have hitherto been restricted to demonstrations of the concentration effect for certain classes of distributions and distances.

In the design and analysis of similarity applications, measures or criteria often directly or indirectly express some notion of the discriminability of similarity measures within neighborhoods. Of the many examples in the literature, three of the most prominent are as follows. Some spectral feature selection criteria [55], such as the Laplacian score [31], measure the discriminative power of candidate feature in terms of the variance of feature values.

The distance ratio (aspect ratio), defined as the ratio between the largest and smallest pairwise distances within a data set, has been applied to the analysis of nearest-neighbor search [12]. Disorder inequalities, relaxations of the usual metric triangle inequality, have been proposed for the analysis of combinatorial search algorithms making use of rankings of data points with respect to a query [23]. The degree of relaxation of the disorder inequality can be regarded as a measure of the discriminability of the data.

The curse of dimensionality has also been linked to the phenomenon of hubness in data sets [62,63]. Given a finite dataset $S \subseteq \mathbb{R}^m$ and an integer parameter k, the hubness of a point $x \in S$ can be defined as the number of objects of S whose k-nearest neighbor (k-NN) set contains x — or equivalently, as the size of the reverse k-NN set of x. The hubs of the data set are defined to be those objects having unusually high hubness values; the data set as a whole is said to have high hubness if it contains many hubs. The hubness phenomenon has been shown to have an influence on a wide range of tasks in data mining, machine learning, and indexing [63], including clustering [73], image data analysis [72], collaborative filtering [48,58], text retrieval [64], time series classification [65], and content-based music retrieval [19,47]. It has been empirically observed that hubness tends to increase with both the dimensionality of the data set, and variations in density within the data set [63]; as such, high hubness can thus be regarded as an indicator of the degree of difficulty of indexing and analyzing data.

In an attempt to alleviate the effects of high dimensionality, and thereby improve the discriminability of data, simpler representations of the 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. Feature selection, the elimination of redundant or irrelevant features, produces a projective subspace that is axis-aligned [54]. In feature extraction, a new (smaller) feature set is generated via a linear transformation of the original features, resulting in an arbitrarily-oriented projective subspace; the most well-known feature extraction methods are PCA and its variants [10,20,60,68,76]. Multidimensional scaling determines a projection that approximately preserves the local distances within the data [67,71]. Manifold learning [67,70,75] and other non-linear extensions [46] resemble projective methods in that they attempt to fit a more complex (yet still lower-dimensional) manifold to the data. With Kernel PCA [70], the data is implicitly transformed into a higher-dimensional setting within which PCA can be applied. In Locally-Linear Embedding (LLE) [67], the data is modeled in terms of a collection of linear tangent spaces to an underlying manifold. Related to manifold learning, regression-based similarity learning [81] uses regression techniques on pairs of data objects in order to learn a simpler, more discriminative similarity function with the greatest possible level of agreement with the original similarity measure. For a recent survey of distance metric learning for data mining applications, see [78].

In general, dimensional reduction requires that an appropriate dimension for the reduced space (or approximating manifold) must 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 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.

1.2 Characterizations of Intrinsic Dimensionality

Over the past decades, many characterizations of the intrinsic dimensionality of sets have been proposed. The earliest theoretical measures of intrinsic dimensionality, such as the classical Hausdorff dimension, Minkowski-Bouligand or 'box counting' dimension, and packing dimension, associate a non-negative real number to metric spaces in terms of their covering or packing properties (for a general reference, see [17]). Although they are of significant theoretical importance, they are impractical for direct use in data mining 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 [24], and 'fractal' methods which estimate intrinsic dimensionality from the space-filling capacity or self-similarity properties of the data [11,18,26]. Other practical techniques for the estimation of intrinsic dimensionality include the topological approaches, which estimate the basis dimension of the tangent space of a data manifold from local samples [10,20,60,68,76]. 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 intrinsic dimension of the data. Parametric modeling and estimation of distribution often allow for estimators of intrinsic dimension to be derived [50,52].

An important family of dimensional models, including the expansion dimension (ED) [45], generalized expansion dimension (GED) [35], and minimum neighbor distance (MiND) models [68], quantify the intrinsic dimensionality 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 oper-

ations are required for the computation of estimates. Expansion models have seen applications in the design and analysis of index structures for similarity search [7,37–40,45], and heuristics for anomaly detection [77], as well as in manifold learning.

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 an underlying distance distribution. 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. EVT has seen widespread applications in such areas as civil engineering [29], operations research [14,57,74], risk assessment [51], material sciences [16,25], bioinformatics [66], geophysics [51], and multimedia [21]. One of the pillars of EVT, a theorem independently proven by Balkema and de Haans [2] and by Pickands [61], 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) [13]. In an equivalent (and much earlier) formulation of EVT due to Karamata [43], 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 [13]; 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 [13]. Many practical methods have been developed for the estimation of the index, including the well-known Hill estimator and its variants [4,32,42] (for a survey, see [22]).

1.3 Contributions

In this theoretical paper, a general EVT framework for continuous distance distributions is proposed that formally unites the notions of intrinsic dimensionality and data discriminability, in a manner suitable for the non-parametric or unsupervised learning applications that often arise in data mining contexts and other applications of content-based similarity. The specific original contributions of the paper include the following:

- In Section 2, the (generalized) expansion dimension is extended to a statistical setting, in which the distribution of distances to a query point is modeled in terms of a continuous random variable $\mathbf{X}$. This continuous model of intrinsic dimensionality (here referred to simply as 'ID') is shown to be equivalent to that of a model of the indiscriminability of the underlying continuous distance distribution, expressed as a function $\mathrm{ID}_{F_{\mathbf{X}}}(x)$ of the distance $x \in [0, \infty)$. A preliminary version of this work (Sections 2.1 to 2.6) previously appeared as [34]. The relationship between ID and the Hausdorff dimension is also established here, in Section 2.7.
- In Section 3, a representation theorem is derived under the ID model, which shows that under reasonable assumptions, every continuous distance distri-

bution is completely determined by its associated ID function. A second-order formulation of ID is then introduced, and shown to be a natural measure of the inlierness or outlierness of the reference point of the distance distribution.
- In Section 4, the theory developed in Section 3 is revealed to be a reworking of extreme value theory for the case of continuous distance distributions. The representation theorem is shown to be a special case of the Karamata representation for short-tailed distributions, with all elements of the Karamata representation being given an interpretation in terms of ID. A well-studied second-order EVT parameter governing the convergence rate of extreme values is also given an interpretation in terms of higher-order ID.
- In Section 5, an explanation of the hubness phenomenon is offered in terms of ID. First, a definition is stated of the hubness of a collection of distance distributions, for which the original formulation of data hubness can be seen as an estimator. Second, the ID representation of the distribution is applied to establish a relationship between ID and local density on the one hand, and the distributional hubness on the other.
- In Section 6, the paper concludes with a discussion of potential applications of ID, as well as future research directions.

2 Data Discriminability and Intrinsic Dimensionality

In this section, we shall see how the (generalized) expansion dimension proposed for finite data sets can be extended to the setting of continuous distributions of distance values. The measure of intrinsic dimensionality proposed for this distributional model will be shown to be equivalent to a natural measure of the indiscriminability of the distance function. Other fundamental properties of this measure will also be derived. We begin the discussion with an overview of the expansion dimension and its applications.

2.1 Expansion Dimension

In [45], Karger and Ruhl introduced a measure of intrinsic dimensionality as a way of analyzing the performance of a local search strategy for handling nearest neighbor queries. The complexity of their method depended heavily on the rate at which the number of visited elements grew as the search expanded. For their analysis, they limited their attention to data sets satisfying the following smooth-growth property. Formally, let $(\mathbb{R}^m, d)$ be a domain for which some non-negative distance function d is defined. Given a set of objects S drawn from the domain, let

$$B_S(v, r) = \{w \in S \mid d(v, w) \leq r\}$$

be the set of elements of S contained in the closed ball of radius r centered at $v \in S$. Given a query set $\mathcal{U}$, S is said to have (b, Δ)-expansion if for all $q \in \mathcal{U}$ and $r > 0$,

$$|B_S(q,r)| \geq b \Longrightarrow |B_S(q,2r)| \leq \Delta \cdot |B_S(q,r)| \,.$$

The *expansion rate* δ of S is the minimum value of Δ such that the above condition holds over all $q \in \mathcal{U}$, subject to the choice of some minimum ball set size b. For cases where d is not a distance metric, the expansion rate has been shown to be closely related to the degree of relaxation of the disorder inequality [53].

Imagine now that the dimension m has been hidden from us. For the Euclidean distance metric in $\mathbb{R}^m$, doubling the radius of a sphere would increase its volume by a factor of 2^m. Were we inclined to measure the volumes of these two spheres, applying the base-2 logarithm to their ratio would recover the dimension m.

This way of discovering the true dimension of the vector space is of course neither necessary or practical. However, by estimating the volumes of spheres by the numbers of points of S that they contain, the value $\mathrm{ED} = \log_2 \delta$ would then serve as a measure of the intrinsic dimensionality of S. In other words, ED can be regarded as the smallest dimension that could accommodate a uniformly-distributed data set having the same maximum growth rate as S. The quantity ED, known as the *expansion dimension* of S, can differ greatly from the representational dimension m: as pointed out in [45], subsets in very high-dimensional spaces can have very low expansion dimensions, whereas even for one-dimensional data the expansion rate can be logarithmic in the size of S. The expansion dimension has been shown to be related to another measure of intrinsic dimensionality based on ball coverage, the *doubling dimension* [26].

The earliest uses of the expansion rate and expansion dimension have been in the complexity analyses of several similarity search structures [7, 39, 40, 45, 49]; in each case, the number of distance comparisons performed was shown to depend on ED and not on m. Of these structures, the Cover Tree [7] and Rank Cover Tree (RCT) [39, 40] indices are noteworthy in that their query execution costs depend only on a relatively small constant number of factors of the expansion rate: for the Cover Tree, δ^{12}, and for the RCT, as low as δ^4 or δ^5.

Generalizations of the expansion dimension have also been proposed in which the radii of the enclosing balls could be chosen with a ratio other than 2. In [77], a heuristic for outlier detection was presented in which approximations of the well-known local outlier factor (LOF) score [9, 83] 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 was taken to be a function of the targeted approximation error bound.

The generalized expansion dimension (GED) has very recently found application not only in complexity analysis, but also as a basis for guiding algorithmic decisions at runtime for a form of adaptive search, the so-called *multi-step*

similarity search problem [37,38]. Multi-step search accepts a user-supplied 'target' distance function together with the query object, making use of a search index constructed according to a second distance function; the only constraint on the target distance function is that it must be bounded from below by the distance used to precompute the index. In [37], a heuristic for the multi-step search problem was proposed, utilizing generalized expansion dimension values as the basis of an early termination condition. More information on the estimation of generalized expansion dimension can be found in [35].

2.2 Intrinsic Dimensionality of Distance Distributions

The generalized expansion dimension can be adapted for the continuous distributions discussed in Section 2.3, by replacing the notion of ball set size by that of probability measure. With this substitution, we can quantify the local intrinsic dimensionality of a feature space exclusively in terms of a distribution of a non-negative random variable $\mathbf{X}$ with support $[0,\infty)$, without knowledge of the features themselves. Henceforth, we will restrict our attention to the case where $\mathbf{X}$ is continuous 'almost everywhere' — that is, throughout the support range $[0,\infty)$ with the possible exception of a set of measure zero.

The variable $\mathbf{X}$ is said to have probability density $f_{\mathbf{X}}$, where $f_{\mathbf{X}}$ is a non-negative Lebesgue-integrable function, if and only if

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

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

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

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

As a motivating example from m-dimensional Euclidean space, consider the situation which the volumes V_1 and V_2 are given for two balls of differing radii r_1 and r_2, respectively. The dimension m can be deduced from the ratios of the volumes and the distances, as follows:

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

For finite data sets, GED formulations are obtained by estimating the volume of balls by the numbers of points they enclose. In contrast, for continuous random distance variables, the notion of volume is naturally analogous to that of probability measure. 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 $\epsilon \to 0^+$.

Definition 1 Let $\mathbf{X}$ be an absolutely continuous random distance variable. For any x such that $F_{\mathbf{X}}(x) > 0$, the *(continuous) intrinsic dimensionality* of $\mathbf{X}$ at x is given by

$$\begin{aligned}\mathrm{IntrDim}_{\mathbf{X}}(x) &\triangleq \lim_{\epsilon \to 0^+} \frac{\ln F_{\mathbf{X}}((1+\epsilon)x) - \ln F_{\mathbf{X}}(x)}{\ln((1+\epsilon)x) - \ln x} \\ &= \lim_{\epsilon \to 0^+} \frac{\ln F_{\mathbf{X}}((1+\epsilon)x) - \ln F_{\mathbf{X}}(x)}{\ln(1+\epsilon)},\end{aligned}$$

wherever the limit exists.

2.3 Indiscriminability of Distance Distributions

A natural way of assessing the discriminability of a random distance variable $\mathbf{X}$ is in terms of the relative rate at which probability measure increases as the distance increases. If $\mathbf{X}$ is discriminative at a given distance r, then expanding the distance by some factor should incur a small increase in the probability measure (or, equivalently for data sets, the expected number of data points in the neighborhood of the reference point). Conversely, if $\mathbf{X}$ is indiscriminative at distance r, then the increase in probability measure would be large.

For applications in which we are concerned with relative increases in the cost of exploring a neighborhood set (such as similarity search), or when measurement error is a concern, or when the scale of distances is not in itself meaningful, it is appropriate to consider relative increases in distance values and probability measure, rather than absolute increases. Accordingly, we propose the following definition of the indiscriminability of $\mathbf{X}$.

Definition 2 Let $\mathbf{X}$ be an absolutely continuous random distance variable. For any x such that $F_{\mathbf{X}}(x) > 0$, the *indiscriminability* of $\mathbf{X}$ at x is given by

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

wherever the limit exists.

Note that this definition of indiscriminability is unitless, and does not depend on statistical parameters of the distribution, such as a mean or variance.

2.4 Equivalence of Indiscriminability and Intrinsic Dimensionality

The following fundamental theorem shows that for continuous distance distributions with differentiable cumulative distribution functions, the notions of indiscriminability and intrinsic dimensionality presented earlier are one and the same.

Theorem 1 *Let* $\mathbf{X}$ *be an absolutely continuous random distance variable. If* $F_{\mathbf{X}}$ *is both positive and differentiable at* x*, then*

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

Proof Since $\mathbf{X}$ is absolutely continuous, and since $F_{\mathbf{X}}$ is differentiable at x, we may apply l'Hôpital's rule to the limits in the definitions of both the intrinsic dimensionality and the indiscriminability. In the former case we obtain

$$\begin{aligned}\mathrm{IntrDim}_{\mathbf{X}}(x) &= \lim_{\epsilon \to 0^+} \frac{\frac{\partial}{\partial \epsilon}\left(\ln F_{\mathbf{X}}((1+\epsilon)x) - \ln F_{\mathbf{X}}(x)\right)}{\frac{\partial}{\partial \epsilon}\ln(1+\epsilon)} \\ &= \lim_{\epsilon \to 0^+} \frac{(1+\epsilon) \cdot x \cdot f_{\mathbf{X}}((1+\epsilon)x)}{F_{\mathbf{X}}((1+\epsilon)x)} \\ &= \frac{x \cdot f_{\mathbf{X}}(x)}{F_{\mathbf{X}}(x)},\end{aligned}$$

and in the latter case as well,

$$\begin{aligned}\mathrm{InDiscr}_{\mathbf{X}}(x) &= \lim_{\epsilon \to 0^+} \frac{\frac{\partial}{\partial \epsilon}\left(F_{\mathbf{X}}((1+\epsilon)x) - F_{\mathbf{X}}(x)\right)}{\frac{\partial}{\partial \epsilon}\left(\epsilon \cdot F_{\mathbf{X}}(x)\right)} \\ &= \lim_{\epsilon \to 0^+} \frac{x \cdot f_{\mathbf{X}}((1+\epsilon)x)}{F_{\mathbf{X}}(x)} \\ &= \frac{x \cdot f_{\mathbf{X}}(x)}{F_{\mathbf{X}}(x)}.\end{aligned}$$

As $F_{\mathbf{X}}$ is positive and differentiable at x, the existence of both limits is guaranteed. □

$\mathrm{ID}_{F_{\mathbf{X}}}$ need not be defined for all $x \in [0, \infty)$; however, since $\mathbf{X}$ is assumed to be absolutely continuous, $\mathrm{ID}_{F_{\mathbf{X}}}$ exists almost everywhere over the range where $F_{\mathbf{X}}$ is positive. Moreover, we can extend the definition to the case where $x = 0$, by taking the limit of $\mathrm{ID}_{F_{\mathbf{X}}}(x)$ as $x \to 0^+$, whenever this limit exists:

$$\mathrm{ID}_{F_{\mathbf{X}}}(0) \triangleq \lim_{x \to 0^+} \mathrm{ID}_{F_{\mathbf{X}}}(x)\,.$$

For continuous distance distributions, the functional $\mathrm{ID}_{F_{\mathbf{X}}}$ can be viewed interchangeably as the intrinsic dimensionality or indiscriminability of $F_{\mathbf{X}}$. The acronym 'ID' can thus refer to 'Intrinsic Dimensionality' or 'InDiscriminability', without the need to make the distinction explicit. Henceforth, it will also be convenient to extend the functional notation to apply to any real-valued function $g(x)$ that is differentiable at x, as follows:

$$\mathrm{ID}_{g}(x) \triangleq \lim_{y \to x} \frac{y \cdot g'(y)}{g(y)},$$

whenever the limit exists.

2.5 Transformations of Variable

In some similarity applications, transformations of the underlying distance measures are sometimes sought so as to improve the overall performance of tasks that depend upon them. Here, we show that for reasonably well-behaved transformations, the ID of a continuous distance distribution can be decomposed into two factors: the ID of the transformed distribution, and the ID of the transform itself.

Consider now a strictly increasing transform $g : [0, \infty) \to [0, \infty)$; if g is applied to the continuous random distance variable $\mathbf{X}$, the result is a random variable $\mathbf{Y} = g(\mathbf{X})$ with cumulative distribution function $F_{\mathbf{Y}}$ such that $(F_{\mathbf{Y}} \circ g)(x) = F_{\mathbf{X}}(x)$ for all $x \geq 0$. Note that $F_{\mathbf{Y}}$ is well-defined, since $F_{\mathbf{Y}}(y) = F_{\mathbf{Y}}(g(0)) = F_{\mathbf{X}}(0) = 0$ for all $0 \leq y \leq g(0)$, and since $\lim_{x\to\infty} F_{\mathbf{Y}}(g(x)) = \lim_{x\to\infty} F_{\mathbf{X}}(x) = 1$.

Theorem 2 *Let $\mathbf{X}$ and $\mathbf{Y}$ be random distance variables with $\mathbf{Y} = g(\mathbf{X})$, for a strictly increasing function g as defined above. If $F_{\mathbf{X}}$ and g are both positive and differentiable at x, then*

$$\mathrm{ID}_{F_{\mathbf{X}}}(x) = \mathrm{ID}_{F_{\mathbf{Y}}\circ g}(x) = \mathrm{ID}_g(x) \cdot \mathrm{ID}_{F_{\mathbf{Y}}}(g(x)).$$

Proof The assumption that $F_{\mathbf{X}}$ and g are both differentiable at x implies that $F_{\mathbf{Y}} \circ g$ is differentiable at x. Using the chain rule, we obtain

$$(F_{\mathbf{Y}} \circ g)'(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_{\mathbf{Y}}(g(x)) = g'(x) \cdot f_{\mathbf{Y}}(g(x)).$$

Since by assumption $g(x) > 0$ and $(F_{\mathbf{Y}} \circ g)(x) = F_{\mathbf{X}}(x) > 0$, applying Theorem 1 yields

$$\begin{aligned}
\mathrm{ID}_{F_{\mathbf{Y}}\circ g}(x) &= \frac{x \cdot (F_{\mathbf{Y}} \circ g)'(x)}{(F_{\mathbf{Y}} \circ g)(x)} = \frac{x \cdot g'(x) \cdot f_{\mathbf{Y}}(g(x))}{F_{\mathbf{Y}}(g(x))} \\
&= \frac{x \cdot g'(x)}{g(x)} \cdot \frac{g(x) \cdot f_{\mathbf{Y}}(g(x))}{F_{\mathbf{Y}}(g(x))} \\
&= \mathrm{ID}_g(x) \cdot \mathrm{ID}_{F_{\mathbf{Y}}}(g(x)).
\end{aligned}$$

as required. □

Note that the proof of Theorem 2 does not strictly require that $F_{\mathbf{Y}}$ be a cumulative distribution function — the proof also holds when $F_{\mathbf{Y}}$ and g are replaced by any two positive-valued differentiable functions. Several interesting choices of transformation are shown in Table 1.

2.6 Joint Probability Distributions

Given a collection of continuous random distance variables $\mathfrak{X} = \{\mathbf{X}_i | \, 1 \leq i \leq m\}$ (for $m \geq 1$), their joint probability distribution is determined by a multivariate cumulative distribution function $F_{\wedge\mathfrak{X}}(\mathbf{x}) = \Pr[\bigwedge_{i=1}^{m}(\mathbf{X}_i \leq x_i)]$, with corresponding probability density function $f_{\wedge\mathfrak{X}}$. If the distance associated with vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is the vector norm $\|\mathbf{x}\|$ (not necessarily Euclidean), the

Table 1 Effects of various distance transformations on ID, where c and m are positive constants.

$g(x) = cx^m$	$\mathrm{ID}_{F_{\mathbf{X}}}(x) = m \cdot \mathrm{ID}_{F_{\mathbf{Y}}}(cx^m)$
$g(x) = e^{cx^m}$	$\mathrm{ID}_{F_{\mathbf{X}}}(x) = cmx^m \cdot \mathrm{ID}_{F_{\mathbf{Y}}}(e^{cx^m})$
$g(x) = x + c$	$\mathrm{ID}_{F_{\mathbf{X}}}(x) = \frac{x}{x+c} \cdot \mathrm{ID}_{F_{\mathbf{Y}}}(x + c)$
$g(x) = c\ln(x + 1)$	$\mathrm{ID}_{F_{\mathbf{X}}}(x) = \frac{x}{(x+1)\ln(x+1)} \cdot \mathrm{ID}_{F_{\mathbf{Y}}}(c\ln(x + 1))$

definitions of intrinsic dimensionality and indiscriminability can be extended to the multivariate case in a natural way.

Definition 3 Let $\mathfrak{X} = \{\mathbf{X}_i | 1 \leq i \leq m\}$ be a collection of $m \geq 1$ absolutely continuous random distance variables. For any $\mathbf{x}$ such that $F_{\wedge\mathfrak{X}}(\mathbf{x}) > 0$, the (*joint*) *intrinsic dimensionality* of $\mathfrak{X}$ at $\mathbf{x}$ is given by

$$\begin{aligned}\mathrm{IntrDim}_{\wedge\mathfrak{X}}(\mathbf{x}) &= \lim_{\epsilon\to 0^+} \frac{\ln F_{\wedge\mathfrak{X}}((1+\epsilon)\mathbf{x}) - \ln F_{\wedge\mathfrak{X}}(\mathbf{x})}{\ln((1+\epsilon)\|\mathbf{x}\|) - \ln\|\mathbf{x}\|} \\ &= \lim_{\epsilon\to 0^+} \frac{\ln F_{\wedge\mathfrak{X}}((1+\epsilon)\mathbf{x}) - \ln F_{\wedge\mathfrak{X}}(\mathbf{x})}{\ln(1+\epsilon)},\end{aligned}$$

wherever the limit exists.

Definition 4 Let $\mathfrak{X} = \{\mathbf{X}_i | 1 \leq i \leq m\}$ be a collection of $m \geq 1$ absolutely continuous random distance variables. For any $\mathbf{x}$ such that $F_{\wedge\mathfrak{X}}(\mathbf{x}) > 0$, the (*joint*) *indiscriminability* of $\mathfrak{X}$ at x is given by

$$\begin{aligned}\mathrm{InDiscr}_{\wedge\mathfrak{X}}(\mathbf{x}) &= \lim_{\epsilon\to 0^+} \left(\frac{F_{\wedge\mathfrak{X}}((1+\epsilon)\mathbf{x}) - F_{\wedge\mathfrak{X}}(\mathbf{x})}{F_{\wedge\mathfrak{X}}(\mathbf{x})} \Big/ \frac{(1+\epsilon)\|\mathbf{x}\| - \|\mathbf{x}\|}{\|\mathbf{x}\|} \right) \\ &= \lim_{\epsilon\to 0^+} \frac{F_{\wedge\mathfrak{X}}((1+\epsilon)\mathbf{x}) - F_{\wedge\mathfrak{X}}(\mathbf{x})}{\epsilon \cdot F_{\wedge\mathfrak{X}}(\mathbf{x})},\end{aligned}$$

wherever the limit exists.

Theorem 3 *Let $\mathfrak{X} = \{\mathbf{X}_i | 1 \leq i \leq m\}$ be a collection of $m \geq 1$ absolutely continuous random distance variables. If $F_{\wedge\mathfrak{X}}$ is both positive and differentiable at $\mathbf{x}$, then*

$$\mathrm{IntrDim}_{\wedge\mathfrak{X}}(\mathbf{x}) = \mathrm{InDiscr}_{\wedge\mathfrak{X}}(\mathbf{x}) = \sum_{i=1}^{m} \mathrm{ID}_{F_{\mathbf{X}_i}}(x_i) \triangleq \mathrm{ID}_{\wedge\mathfrak{X}}(\mathbf{x}).$$

Proof Theorem 1 implies that the result holds for the case where $m = 1$. For the remainder of the proof, we will assume that $m > 1$.

Since $F_{\wedge\mathfrak{X}}$ is both positive and differentiable at $\mathbf{x}$, we may apply l'Hôpital's rule to the limits of Definitions 3 and 4 (using the chain rule for multivariate differentiation). Letting $\mathbf{v} = (1+\epsilon)\mathbf{x}$, the indiscriminability of $\mathfrak{X}$ thus becomes

$$\begin{aligned}\mathrm{InDiscr}_{\wedge\mathfrak{X}}(\mathbf{x}) &= \lim_{\epsilon\to 0^+} \frac{\sum_{i=1}^{m} \frac{\partial F_{\wedge\mathfrak{X}}(\mathbf{v})}{\partial v_i} \cdot \frac{\partial v_i}{\partial \epsilon}}{\frac{\partial(\epsilon \cdot F_{\wedge\mathfrak{X}}(\mathbf{x}))}{\partial \epsilon}} \\ &= \frac{1}{F_{\wedge\mathfrak{X}}(\mathbf{x})} \cdot \lim_{\epsilon\to 0^+} \sum_{i=1}^{m} \frac{\partial F_{\wedge\mathfrak{X}}(\mathbf{v})}{\partial v_i} x_i \qquad (1)\end{aligned}$$

To determine formulae for the partial derivatives, we first express $F_{\wedge\mathfrak{X}}$ as a multiple integral

$$F_{\wedge\mathfrak{X}}(\mathbf{b}) = \int_0^{b_1} \int_0^{b_2} \cdots \int_0^{b_m} f_{\wedge\mathfrak{X}}(x_1, x_2, \ldots, x_m)\, \mathrm{d}x_m \ldots \mathrm{d}x_2\, \mathrm{d}x_1.$$

The partial derivative with respect to x_m is thus

$$\begin{aligned}
&\left.\frac{\partial F_{\wedge\mathfrak{X}}(\mathbf{x})}{\partial x_m}\right|_{\mathbf{b}} \\
&= \int_0^{b_1} \int_0^{b_2} \cdots \int_0^{b_{m-1}} f_{\wedge\mathfrak{X}}(x_1, x_2, \ldots, x_{m-1}, b_m)\, \mathrm{d}x_{m-1} \ldots \mathrm{d}x_2\, \mathrm{d}x_1 \\
&= f_{\mathbf{X_m}}(b_m) \cdot \\
&\quad \int_0^{b_1} \int_0^{b_2} \cdots \int_0^{b_{m-1}} f_{\wedge\mathfrak{X}\setminus\mathbf{X_m}\,|\,\mathbf{X_m}}(x_1, x_2, \ldots, x_{m-1} |\, b_m)\, \mathrm{d}x_{m-1} \ldots \mathrm{d}x_2\, \mathrm{d}x_1 \\
&= f_{\mathbf{X_m}}(b_m) \cdot F_{\wedge\mathfrak{X}\setminus\mathbf{X_m}\,|\,\mathbf{X_m}}(b_1, b_2, \ldots, b_{m-1} |\, b_m) \\
&= f_{\mathbf{X_m}}(b_m) \cdot \frac{F_{\wedge\mathfrak{X}}(\mathbf{b})}{F_{\mathbf{X_m}}(b_m)};
\end{aligned}$$

the other partial derivatives are of the same form. Substituting into Equation (1) yields

$$\begin{aligned}
\mathrm{InDiscr}_{\wedge\mathfrak{X}}(\mathbf{x}) &= \frac{1}{F_{\wedge\mathfrak{X}}(\mathbf{x})} \cdot \lim_{\epsilon\to 0^+} \sum_{i=1}^{m} x_i \cdot f_{\mathbf{X}_i}(v_i) \cdot \frac{F_{\wedge\mathfrak{X}}(\mathbf{v})}{F_{\mathbf{X}_i}(v_i)} \\
&= \sum_{i=1}^{m} \frac{x_i \cdot f_{\mathbf{X}_i}(x_i)}{F_{\mathbf{X}_i}(x_i)} \;=\; \sum_{i=1}^{m} \mathrm{ID}_{F_{\mathbf{X}_i}}(x_i).
\end{aligned}$$

To complete the proof, we note that the joint intrinsic dimensionality can be derived in a similar way:

$$\begin{aligned}
&\mathrm{IntrDim}_{\wedge\mathfrak{X}}(\mathbf{x}) \\
&= \lim_{\epsilon\to 0^+} \frac{\sum_{i=1}^{m} \frac{\partial \ln F_{\wedge\mathfrak{X}}(\mathbf{v})}{\partial v_i} \cdot \frac{\partial v_i}{\partial \epsilon}}{\frac{\partial \ln(1+\epsilon)}{\partial \epsilon}} \\
&= \lim_{\epsilon\to 0^+} (1+\epsilon) \cdot \sum_{i=1}^{m} \frac{1}{F_{\wedge\mathfrak{X}}(\mathbf{v})} \cdot \left(f_{\mathbf{X}_i}(v_i) \cdot \frac{F_{\wedge\mathfrak{X}}(\mathbf{v})}{F_{\mathbf{X}_i}(v_i)} \right) \cdot x_i \\
&= \sum_{i=1}^{m} \mathrm{ID}_{F_{\mathbf{X}_i}}(x_i).
\end{aligned}$$

□

Perhaps surprisingly, Theorem 3 indicates that for absolutely continuous random distance variables, the ID of a joint distribution is equal to the sum of the IDs of the individual distributions, even when the distributions are not independent. However, the correctness of the proof relies heavily on the fact

that the joint distribution is differentiable, and that the region of integration is the Cartesian product of the intervals $[0, x_i]$ for $1 \leq i \leq m$. Theorem 3 should therefore not be taken to mean that ID is necessarily additive for other local regions of interest. For example, for a distribution of Euclidean distances to a reference point in the the two-dimensional XY-plane, the ID of the distribution is not necessarily the sum of the IDs of the two distance distributions obtained by projection to the X-axis and Y-axis.

2.7 ID and the Hausdorff Dimension

Hausdorff dimension (which we will refer to here as 'HD') was introduced in 1919 as a measure of the local size of a set S, in terms of the distance associated with an underlying metric space [30]. When S is a subspace of dimension m, the Hausdorff dimension of S is m, as one would expect. However, HD is of particular use in accounting for the complexity of more complex shapes, such as fractals. In such situations, HD can take non-integer values. Despite the wide theoretical importance of HD, it is very difficult to estimate in practice (although for finite sets, the Hausdorff dimension is always 0).

The concept of Hausdorff dimension relates to the properties of cover sets of S, defined as (possibly infinite) collections of balls whose union contains S. The radius of a cover set is taken to be the largest radius of any ball in the collection. If $\mathbb{A}$ is a cover of S, then consider the quantity

$$\alpha(m) = \lim_{\epsilon \to 0} \inf_{\mathbb{A}:\, \rho(\mathbb{A}) \leq \epsilon} \sum_{A \in \mathbb{A}} \left(\rho(A)\right)^m \ ,$$

where $\rho(A)$ is the radius of set A. The Hausdorff dimension is the unique real value $m_0 \geq 0$ such that

$$\begin{aligned} m < m_0 &\Longrightarrow \alpha(m) = \infty \\ m > m_0 &\Longrightarrow \alpha(m) = 0 \, . \end{aligned}$$

The Hausdorff dimension is a 'global' measure of intrinsic dimensionality, in that it describes the overall complexity of S. The following theorem (a proof of which can be found in [82]) provides bounds on HD in terms of local contributions across (almost) all points of a set S having positive measure in a probability space.

Theorem 4 ([82]) *Let the random variable* $\mathbf{X}$ *represent the distribution of distances from* $\mathbf{x} \in \mathcal{M}$*, where* $\mathcal{M}$ *is a manifold with positive measure. Here,* $F_{\mathbf{X}}(x)$ *represents the probability measure within distance* x *of* $\mathbf{x}$*. If*

$$\delta_0 \leq \liminf_{x \to 0^+} \frac{\log F_{\mathbf{X}}(x)}{\log x} \leq \limsup_{x \to 0^+} \frac{\log F_{\mathbf{X}}(x)}{\log x} \leq \delta_1$$

for almost every point $\mathbf{x} \in \mathcal{M}$*, then the Hausdorff dimension of* $\mathcal{M}$ *is in the range* $[\delta_0, \delta_1]$.

The following theorem makes use of Theorem 4 to show that under certain conditions of continuity and differentiability of the distance distributions based at the points of manifold $\mathcal{M}$, the Hausdorff dimension falls within the range of values of $IDX(0)$ attained over these points of $\mathcal{M}$.

Theorem 5 *Let the random variable* $\mathbf{X}$ *represent the distribution of distances from* $\mathbf{x} \in \mathcal{M}$*, where* $\mathcal{M}$ *is a manifold with positive measure, and let* $F_{\mathbf{X}}$ *be the cumulative distribution function associated with* $\mathbf{x}$*. Suppose that almost everywhere in* $\mathcal{M}$*, we have that* $\mathbf{X}$ *is absolutely continuous, and that the limit* $\mathrm{ID}_{F_{\mathbf{X}}}(0)$ *exists. Then the Hausdorff dimension* $\mathrm{HD}(\mathcal{M})$ *of* $\mathcal{M}$ *satisfies*

$$\inf_{\mathbf{x}\in\mathcal{M}_*} \mathrm{ID}_{F_{\mathbf{X}}}(0) \leq \mathrm{HD}(\mathcal{M}) \leq \sup_{\mathbf{x}\in\mathcal{M}_*} \mathrm{ID}_{F_{\mathbf{X}}}(0),$$

where $\mathcal{M}_* \in \mathcal{M}$ *is the subset of* $\mathcal{M}$ *for which the assumptions are satisfied.*

Proof Let $\mathbf{x} \in \mathcal{M}_*$ be a point satisfying the assumptions of the theorem. Since $\mathbf{X}$ is absolutely continuous, $F_{\mathbf{X}}$ is differentiable almost everywhere. Let the probability density function of $\mathbf{X}$ be $f_{\mathbf{X}}$. Using l'Hôpital's rule together with Theorem 1, we see that

$$\lim_{x\to 0^+} \frac{\log F_{\mathbf{X}}(x)}{\log x} = \lim_{x\to 0^+} \frac{x \cdot f_{\mathbf{X}}(x)}{F_{\mathbf{X}}(x)} = \lim_{x\to 0^+} \mathrm{ID}_{F_{\mathbf{X}}}(x) = \mathrm{ID}_{F_{\mathbf{X}}}(0).$$

The result then follows by applying Theorem 4 with $\delta_0 = \inf_{\mathbf{x}\in\mathcal{M}_*} \mathrm{ID}_{F_{\mathbf{X}}}(0)$ and $\delta_1 = \sup_{\mathbf{x}\in\mathcal{M}_*} \mathrm{ID}_{F_{\mathbf{X}}}(0)$. □

In contrast with HD, ID is a 'local' measure of dimensionality — when the distribution of distances from a fixed point is continuous, $\mathrm{ID}_{F_{\mathbf{X}}}(0)$ measures the intrinsic dimensionality in the vicinity of that point. Theorem 5 essentially states that the global Hausdorff dimension falls within the range of variation of the local ID values associated with the distance distributions determined within the set.

3 ID-Based Characterization of Distance Distributions

The ID formula $\mathrm{ID}_{F_{\mathbf{X}}}(x) = x \cdot f_{\mathbf{X}}(x)/F_{\mathbf{X}}(x)$ established in Theorem 1 simultaneously expresses the notions of intrinsic dimensionality and indiscriminability of distance. The formula also suggests an interpretation of ID as a normalization of the probability density $f_{\mathbf{X}}(x)$ with respect to the cumulative density $F_{\mathbf{X}}(x)/x$. In this section we will see that the 'normalized' probability density function $\mathrm{ID}_{F_{\mathbf{X}}}(x)$ fully characterizes a continuous distance distribution. Moreover, we show that the second-order ID function $\mathrm{ID}_{\mathrm{ID}_{F_{\mathbf{X}}}(x)} = x \cdot \mathrm{ID}'_{F_{\mathbf{X}}}(x)/\mathrm{ID}_{F_{\mathbf{X}}}(x)$ (the 'indiscriminability of the indiscriminability') in turn fully characterizes $\mathrm{ID}_{F_{\mathbf{X}}}(x)$. We conclude by showing that second-order ID naturally expresses the inlierness or outlierness of a data point with respect to its locality.

Later on, in Section 4, we will see that the characterizations presented in this section are in fact a reworking of extreme value theory (EVT) for short-tailed distributions.

3.1 ID Characterization Theorem

Our goal here is to show that under reasonable assumptions, every continuous distance distribution is completely determined by its associated ID function. We begin by presenting a characterization of a slightly more general class of functions.

Theorem 6 *Let $F : (0, z) \to \mathbb{R}$ be a function over the range $(0, z)$, for some choice of $z > 0$ (possibly infinite), such that F is absolutely continuous, and positive everywhere. Let $v \in [0, z)$ be a value for which $\mathrm{ID}_F(v)$ exists. Then for any $x, w \in (0, z)$,*

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

Proof For any $x \in (0, z)$

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

since the absolute continuity of F implies that F is differentiable almost everywhere.

Since $F(x)$ and $F(w)$ are assumed to be positive, and since $\mathrm{ID}_F(t)$ exists almost everywhere,

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

as required. □

The representation formula in Theorem 6 reveals the behavior of the function at values close to some reference value. Let us consider the values of $F(x)$ for those values of x contained in a shrinking range $[v, w]$, where $0 \leq v < w$ and w tends to v from above. In this situation, the following lemma shows that the exponential factor $G_{F,v,w}(x)$ tends to 1, with some further restrictions on x when $v = 0$.

Theorem 7 *Let $F : (0, z) \to \mathbb{R}$ be a function over the range $(0, z)$, for some choice of $z > 0$ (possibly infinite), such that F is absolutely continuous, and positive everywhere. Let $c \in (0, 1)$ be a constant, and let $v \in [0, z)$ be a value for which $\mathrm{ID}_F(v)$ exists. Then for any $x \in (0, z)$,*

$$\lim_{\substack{w \to v^+ \\ \max\{v, cw\} \leq x \leq w}} G_{F,v,w}(x) \;=\; 1\,.$$

Proof It suffices to show that

$$\lim_{\substack{w \to v^+ \\ \max\{v, cw\} \leq x \leq w}} \int_x^w \frac{\mathrm{ID}_F(v) - \mathrm{ID}_F(t)}{t}\,\mathrm{d}t \;=\; 0\,.$$

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 $w - v < \delta$,

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

If $v = 0$, we have that $w/x = 1/c$. On the other hand, if $v > 0$,

$$1 \leq \frac{w}{x} \leq \frac{w}{\max\{v, cw\}} \leq \min\left\{ \frac{v+1}{v}, \frac{1}{c} \right\} \leq \frac{1}{c}.$$

In either case, $\ln(w/x)$ is bounded from above and below by positive constants. Therefore, since ϵ can be made arbitrarily small, the limit is indeed 0, and the result follows. □

For the case when $v = 0$, x can be allowed to range over an arbitrarily large proportion of the interval $[v, w]$, by choosing c sufficiently close to zero. When $v > 0$, for any valid choice of c, x becomes free to range over the entire interval $[v, w]$ as $w \to v^+$, once $w \leq v/c$ holds.

The proof of Theorem 7 can easily be adapted to show that the limit also holds when w tends to v from below.

Corollary 1 *Let $F : (0, z) \to \mathbb{R}$ be a function over the range $(0, z)$, for some choice of $z > 0$ (possibly infinite), such that F is absolutely continuous, and positive everywhere. Let $v \in (0, z)$ be a value for which $\mathrm{ID}_F(v)$ exists. Then for any $x \in (0, z)$,*

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

Given an absolutely continuous random distance variable $\mathbf{X}$, its cumulative distribution function $F_{\mathbf{X}}$ satisfies the conditions of Theorem 6 provided that it is strictly positive over $(0, \infty)$. The ID characterization expresses the behavior of the entire distribution in terms of the ID function. For the special case when $v = 0$, Theorem 6 addresses the behavior of $F_{\mathbf{X}}$ as distances tend toward zero.

The theorem, together with Theorem 7, shows that within the extreme lower tail of essentially any smooth distribution of distances, the relative increase in probability measure tends to a polynomial function of the relative increase in distance, of degree equal to the limit of the continuous ID at distance 0:

$$\frac{F_{\mathbf{X}}(x)}{F_{\mathbf{X}}(w)} \to \left(\frac{x}{w}\right)^{\mathrm{ID}_{F_{\mathbf{X}}}(0)}.$$

This is precisely the growth rate that would be expected if the distances were generated from a reference point to a uniform distribution of points, restricted to the relative interior of a manifold of dimension $\mathrm{ID}_{F_{\mathbf{X}}}(0)$.

3.2 Second-Order ID

A characterization formula for $\mathrm{ID}_{F_{\mathbf{X}}}$ can be obtained for the second-order ID function $\mathrm{ID}_{\mathrm{ID}_{F_{\mathbf{X}}}}(x)$ from the characterization formulas for $F_{\mathbf{X}}$ and $f_{\mathbf{X}}$. Note that Theorems 6 and 7, and Corollary 1, can be applied to the probability density $f_{\mathbf{X}}$ to yield a characterization in terms of $\mathrm{ID}_{f_{\mathbf{X}}}$, provided that $f_{\mathbf{X}}$ is both non-zero and differentiable almost everywhere.

For the proof of the characterization of $\mathrm{ID}_{F_{\mathbf{X}}}$, we require two technical lemmas. The first of the two lemmas shows that the second-order ID function $\mathrm{ID}_{\mathrm{ID}_F}(x)$ can be expressed in terms of the difference between the indiscriminabilities of F and F'.

Lemma 1 *Let $F : (0, z) \to \mathbb{R}$ be a function over the range $(0, z)$, for some choice of $z > 0$ (possibly infinite). If F is twice differentiable at some distance $x > 0$ 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*

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

Proof Since F is doubly differentiable at x, $F''(x)$ must exist. Together with the assumption that $F(x) \neq 0$ and $F'(x) \neq 0$, we have that $\mathrm{ID}_F(x) = x \cdot F'(x)/F(x) \neq 0$, and that $\mathrm{ID}_{F'}(x) = x \cdot F''(x)/F'(x)$ must exist. $\mathrm{ID}_F(x)$ can therefore be differentiated to obtain

$$\mathrm{ID}'_F(x) = \frac{F(x) \cdot (xF''(x) + F'(x)) - x\,(F'(x))^2}{(F(x))^2}.$$

Since $x/\mathrm{ID}_F(x) = F(x)/F'(x)$, multiplying yields

$$\begin{aligned}
\frac{x \cdot \mathrm{ID}'_F(x)}{\mathrm{ID}_F(x)} &= \frac{F(x) \cdot (xF''(x) + F'(x)) - x\,(F'(x))^2}{(F(x))^2} \cdot \frac{F(x)}{F'(x)} \\
\mathrm{ID}_{\mathrm{ID}_F(x)} &= \frac{x \cdot F''(x)}{F'(x)} + 1 - \frac{x \cdot F'(x)}{F(x)} \\
&= \mathrm{ID}_{F'}(x) + 1 - \mathrm{ID}_F(x).
\end{aligned}$$

□

The next technical lemma shows that the the second-order ID converges to 0 as $x \to 0$.

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

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

Proof Lemma 1 implies that $\mathrm{ID}_F(x)$, $\mathrm{ID}_{F'}(x)$ and $\mathrm{ID}'_F(x)$ all exist, for all $x \in (0, z)$. We may then apply l'Hôpital's rule to obtain

$$\begin{aligned}\mathrm{ID}_F(0) &= \lim_{x\to 0^+} \mathrm{ID}_F(x) = \lim_{x\to 0^+} \frac{x \cdot F'(x)}{F(x)} = \lim_{x\to 0^+} \frac{x \cdot F''(x) + F'(x)}{F'(x)} \\ &= 1 + \lim_{x\to 0^+} \mathrm{ID}_{F'}(x) = 1 + \mathrm{ID}_{F'}(0).\end{aligned}$$

By applying Lemma 1 and letting $x \to 0$, the result follows. □

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

Theorem 8 *Let $F : (0, z) \to \mathbb{R}$ be a twice-differentiable function, for some choice of $z > 0$ (possibly infinite). Also, assume that F and F' are positive everywhere or negative everywhere. 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 $\mathrm{ID}_F(0)$ 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 ID_F, $\mathrm{ID}_{F'}$, ID'_F and $\mathrm{ID}_{\mathrm{ID}_F}$ exist everywhere, and that $|\mathrm{ID}_F(x)|$ is positive everywhere. We can therefore establish the result for the case where $x > 0$, as follows:

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

If $F(x) \to 0$ as $x \to 0$, and if $|\mathrm{ID}_F(0)|$ exists and is positive, by Lemma 2 we have that $\mathrm{ID}_{F'}(0)$ exists, and that $\mathrm{ID}_{\mathrm{ID}_F}(0) = 0$. Since $|\mathrm{ID}_F(w)|$ is also positive, the integral in the representation formula must converge, and therefore the representation is valid for $x = 0$ as well. □

Theorem 8, as well as Lemmas 1 and 2 can be applied so as to obtain representation formulae for $\mathrm{ID}_{F_{\mathbf{x}}}$ and $|\mathrm{ID}_{f_{\mathbf{x}}}|$ over intervals of the form $(0, z)$ where $F_{\mathbf{X}}$ is thrice differentiable, and $F_{\mathbf{X}}$ and its derivatives are either positive everywhere or negative everywhere. The representation for $|\mathrm{ID}_{f_{\mathbf{x}}}|$ will turn out to be useful in establishing a connection between ID and second-order extreme value theory, in Section 4.

In the remainder of this section, we will see how the second-order ID function $\mathrm{ID}_{\mathrm{ID}_{F_{\mathbf{x}}}}$ naturally expresses the inlierness or outlierness of a data point with respect to its locality.

3.3 Inlierness, Outlierness and ID

Traditional distributional techniques for data clustering generally assume that the data can be modeled as a mixture of underlying distributions, whose nature must be decided in advance. Each distribution in the mixture represents an individual data cluster — the clustering task is to determine those parameter values that allow the best possible fit of distributions to data. Perhaps the most common example of distributional clustering is that of Gaussian mixture models, for which heuristics from the Expectation Maximization (EM) family [15] — including DENCLUE [33], and k-Means and its variants [56] — are perhaps the best known. In general, however, the underlying distributions are not known in advance, and the Gaussian assumption may not be justified.

Density-based clustering methods, such as DBSCAN [69] and OPTICS [1], avoid placing explicit assumptions on the nature of the cluster distribution. Instead, they identify clusters as regions of high local density, by means of thresholding. However, density thresholding may obscure embedded clusters in areas of relatively high density, and prune away local clusters in regions of relatively low density. Moreover, the determination of cluster borders generally depends purely on the supplied density threshold, and not on the underlying cluster distribution.

Local manifold learning techniques such as Locally-Linear Embedding [67] can be adapted to produce a data clustering, with each cluster corresponding to a tangent manifold of the embedding [41]. Although such methods have the advantage of implicitly adapting to the local intrinsic dimensionality of the data, in general the assumption of local linearity may not be warranted. Even when the data is well-described by a linear manifold, the determination of the cluster boundary or extent can be problematic.

The ID model proposed in this paper for continuous distance distributions can be used to reveal characteristics of local manifolds without the need to explicitly construct the manifold itself, without the need to learn parameters of an assumed data distribution, and without needing to set absolute thresholds on the minimum cluster density. For a given point of interest $\mathbf{x}$ on the manifold, the local intrinsic dimension of the manifold at that point is simply the value of $\mathrm{ID}_{F_{\mathbf{x}}}(0)$, where the random variable $\mathbf{X}$ follows the distribution of distances from $\mathbf{x}$. In addition, the ID function at positive distances gives an indication

as to whether $\mathbf{x}$ should be regarded as an inlier or as an outlier relative to its local neighborhood within the manifold, as the following argument shows.

If $\mathrm{ID}_{F_{\mathbf{x}}}(x) < \mathrm{ID}_{F_{\mathbf{x}}}(0)$ within a small local neighborhood $0 < x < \epsilon$ (where $\epsilon > 0$), then:

- The discriminability of the data at distance x from $\mathbf{x}$ is greater than at distances approaching 0.
- The growth rate in probability measure at distance x from $\mathbf{x}$ is less than that which would be expected within a locally-uniform distribution of points within a manifold of dimension $\mathrm{ID}_{F_{\mathbf{x}}}(0)$.
- Under the assumption that the local manifold has a fixed intrinsic dimension within a sufficiently small neighborhood of $\mathbf{x}$, the drop in indiscriminability (or rise in discriminability) indicates a decrease in local density as the distance from $\mathbf{x}$ increases.
- With this interpretation, the relationship between $\mathbf{x}$ and its neighborhood is therefore that of an *inlier*.

By similar arguments, if instead $\mathrm{ID}_{F_{\mathbf{x}}}(x) > \mathrm{ID}_{F_{\mathbf{x}}}(0)$, then the rise indiscriminability (or drop in discriminability) indicates an increase in local density as the distance from $\mathbf{x}$ increases, in which case $\mathbf{x}$ would be an *outlier* with respect to its neighborhood.

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

4 ID and Extreme Value Theory

The characterization of continuous distance distributions established in Section 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; in our treatment, the formulation that 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. For a detailed account of regular variation and EVT, see (for example) [3,8].

4.1 First-order EVT

Let $\mathbf{X}$ be an absolutely continuous random distance variable. Karamata's characterization theorem [43,44] implies that the asymptotic cumulative distribution of $\mathbf{X}$ in the lower tail $[0, w)$ can be expressed as $F_{\mathbf{X}}(x) = x^{\gamma_{\mathbf{X}}} \ell_{\mathbf{X}}(1/x)$ for some constant $\gamma_{\mathbf{x}}$, where $\ell_{\mathbf{X}}$ is differentiable and *slowly varying* (at infinity); that is, for all $c > 0$, $\ell_{\mathbf{X}}$ satisfies

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

The cumulative distribution $F_{\mathbf{X}}$ restricted to $[0, w)$ is itself said to be *regularly varying* with index $\gamma_{\mathbf{x}}$.

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

$$\begin{aligned} \varepsilon_{\mathbf{X}}(u) &\triangleq \frac{u \ell'_{\mathbf{X}}(u)}{\ell_{\mathbf{X}}(u)}, \\ \lim_{u \to \infty} \varepsilon_{\mathbf{X}}(u) &\to 0\,. \end{aligned}$$

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

$$\ell_{\mathbf{X}}(1/x) = \exp\left(\eta_{\mathbf{X}}(1/x) + \int_{1/w}^{1/x} \frac{\varepsilon_{\mathbf{X}}(u)}{u}\, \mathrm{d}u \right),$$

where $\eta_{\mathbf{X}}$ and $\varepsilon_{\mathbf{X}}$ are measurable and bounded functions such that $\eta_{\mathbf{X}}(1/x)$ tends to a constant, and $\varepsilon_{\mathbf{X}}(1/t)$ 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

$$\ell_{\mathbf{X}}(1/x) = \exp\left(\eta_{\mathbf{X}}(1/x) + \int_{x}^{w} \frac{\varepsilon_{\mathbf{X}}(1/t)}{t}\, \mathrm{d}t \right).$$

Thus the cumulative distribution formula $F_{\mathbf{X}}(x) = x^{\gamma_{\mathbf{X}}} \ell_{\mathbf{X}}(1/x)$ can easily be verified to fit the form of the representation given in Theorem 6, with

$$\begin{aligned} \gamma_{\mathbf{x}} &= \mathrm{ID}_{F_{\mathbf{X}}}(0)\,; \\ \eta_{\mathbf{X}}(1/x) &= \ln F_{\mathbf{X}}(w) - \mathrm{ID}_{F_{\mathbf{X}}}(0) \ln w\,; \\ \varepsilon_{\mathbf{X}}(1/t) &= \mathrm{ID}_{F_{\mathbf{X}}}(0) - \mathrm{ID}_{F_{\mathbf{X}}}(t)\,. \end{aligned}$$

4.2 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 [28]. As is the case with first-order EVT, many approaches to the estimation of second-order parameters have been developed [22].

Here, we will follow the formulation presented in [27] using second-order regular variation. In this paper, de Haan and Resnick presented a proof of 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 ϕ' having the representation

$$\phi'(t) = k \cdot t^{\gamma - 1} \cdot \exp\left(\int_1^t \frac{A(u)}{u} \, \mathrm{d}u \right)$$

for some non-zero constant k.

In the context of continuous distance distributions, we transform the upper-tail formulation stated above to that of the lower tail of the cumulative distribution function for $\mathbf{X}$, by setting $t = 1/x$ and $\phi'(t) = f_{\mathbf{X}}(x)$. Noting that $f'_{\mathbf{X}}(x) = -t^2 \phi''(t)$, and defining $B(x) \triangleq A(t)$, the first condition can be shown to be

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

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

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

Thus these equivalent conditions can be verified to fit the form of the representation given in Theorem 6, with $w = 1$, $v = 0$, and

$$\begin{aligned} k &= f_{\mathbf{X}}(1); \\ \gamma &= 1 - \mathrm{ID}_{f_{\mathbf{X}}}(0) = 2 - \mathrm{ID}_{F_{\mathbf{X}}}(0) = 2 - \gamma_{\mathbf{X}} \; ; \\ B(x) &= 1 - \gamma - \mathrm{ID}_{f_{\mathbf{X}}}(x) = \mathrm{ID}_{F_{\mathbf{X}}}(0) - 1 - \mathrm{ID}_{f_{\mathbf{X}}}(x) \; . \end{aligned}$$

Second-order EVT is largely concerned with the estimation of the parameter ρ. Here, we shall show that the index of regular variation of the functions $B(x)$ and $|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(x)|$ is in fact the non-negative value $-\rho$.

Theorem 9 *Let* $\mathbf{X}$ *be a random distance variable whose cumulative distribution function* $f_{\mathbf{X}}$ *is twice-differentiable over the interval* $(0, z)$*, for some choice of* $z > 0$ *(possibly infinite). Furthermore, assume that* $f_{\mathbf{X}}$ *and* $f'_{\mathbf{X}}$ *are positive everywhere or negative everywhere over* $(0, z)$*, that* $f_{\mathbf{X}}(x) \to 0$ *as* $x \to 0$*, and that* $\mathrm{ID}_{F_{\mathbf{X}}}(0)$ *exists. Let* $B(x) = \mathrm{ID}_{F_{\mathbf{X}}}(0) - 1 - \mathrm{ID}_{f_{\mathbf{X}}}(x)$*. Then* $|B(x)|$ *and* $B^*(x) \triangleq |\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(x)|$ *are both regularly varying with index* $-\rho \geq 0$*. Furthermore, if* B^* *is absolutely continuous, then* $-\rho = \mathrm{ID}_{B^*}(0)$*.*

Proof From the definition of the index of regular variation, we have that

$$c^{-\rho} = \left(\frac{1}{c}\right)^{\rho} = \lim_{t\to\infty} \frac{|A(t/c)|}{|A(t)|} = \lim_{x\to 0} \frac{|B(cx)|}{|B(x)|}$$

for any fixed $c > 0$. To prove that $|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(x)|$ is also regularly varying with index $-\rho$, it suffices to show that for all fixed $c > 0$,

$$\lim_{x\to 0} \frac{|B^*(cx)|}{|B^*(x)|} = \lim_{x\to 0} \frac{|B(cx)|}{|B(x)|} \,.$$

Since $B(x) = \mathrm{ID}_{F_{\mathbf{X}}}(0) - 1 - \mathrm{ID}_{f_{\mathbf{X}}}(x)$, Lemma 2 implies that

$$\begin{aligned}\lim_{x\to 0} \frac{|B(cx)|}{|B(x)|} &= \lim_{x\to 0} \frac{|\mathrm{ID}_{F_{\mathbf{X}}}(0) - 1 - \mathrm{ID}_{f_{\mathbf{X}}}(cx)|}{|\mathrm{ID}_{F_{\mathbf{X}}}(0) - 1 - \mathrm{ID}_{f_{\mathbf{X}}}(x)|} \\ &= \lim_{x\to 0} \frac{|\mathrm{ID}_{f_{\mathbf{X}}}(0) - \mathrm{ID}_{f_{\mathbf{X}}}(cx)|}{|\mathrm{ID}_{f_{\mathbf{X}}}(0) - \mathrm{ID}_{f_{\mathbf{X}}}(x)|} \,.\end{aligned}$$

Since $\mathrm{ID}_{f_{\mathbf{X}}}$ and $\mathrm{ID}'_{f_{\mathbf{X}}}$ do not change sign over $(0, z)$, applying the ID characterization formula of Theorem 8, we obtain

$$\lim_{x\to 0} \frac{|B(cx)|}{|B(x)|} = \lim_{x\to 0} \frac{|\mathrm{ID}_{f_{\mathbf{X}}}(0)| - |\mathrm{ID}_{f_{\mathbf{X}}}(w)| \cdot \exp\left(-\int_{cx}^{w} \frac{|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(u)|}{u}\,\mathrm{d}u\right)}{|\mathrm{ID}_{f_{\mathbf{X}}}(0)| - |\mathrm{ID}_{f_{\mathbf{X}}}(w)| \cdot \exp\left(-\int_{x}^{w} \frac{|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(u)|}{u}\,\mathrm{d}u\right)} \,.$$

The limit of the numerator and the denominator are both zero. Applying l'Hôpital's rule yields

$$\begin{aligned}\lim_{x\to 0} \frac{|B(cx)|}{|B(x)|} &= \lim_{x\to 0} \frac{-|\mathrm{ID}_{f_{\mathbf{X}}}(w)| \cdot \frac{c\cdot|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(cx)|}{cx} \cdot \exp\left(-\int_{cx}^{w} \frac{|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(u)|}{u}\,\mathrm{d}u\right)}{-|\mathrm{ID}_{f_{\mathbf{X}}}(w)| \cdot \frac{|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(x)|}{x} \cdot \exp\left(-\int_{x}^{w} \frac{|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(u)|}{u}\,\mathrm{d}u\right)} \\ &= \lim_{x\to 0} \frac{|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(cx)|}{|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(x)|} \cdot \exp\left(-\int_{cx}^{x} \frac{|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(u)|}{u}\,\mathrm{d}u\right) \,.\end{aligned}$$

Note that since $f_{\mathbf{X}}(x) \to 0$ as $x \to 0$, we have that $|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(x)| \to 0$ from Lemma 2. By means of an argument similar to that of the proof of Theorem 7, it can then be shown that

$$\lim_{x\to 0} \exp\left(-\int_{cx}^{x} \frac{|\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}(u)|}{u}\,\mathrm{d}u\right) \to 1 \,,$$

which establishes that

$$\lim_{x \to 0} \frac{|B^*(cx)|}{|B^*(x)|} = \lim_{x \to 0} \frac{|B(cx)|}{|B(x)|} .$$

To complete the proof, note that if $B^* = |\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}|$ is absolutely continuous, then $-\rho = \mathrm{ID}_{B^*}(0)$ follows from Theorem 6. □

4.3 Discussion

The connection between EVT and ID for continuous distance distributions can thus be summarized as follows:

- The first-order EVT index $\gamma_{\mathbf{X}}$ that asymptotically determines the distribution of distances within a radius-w neighborhood is precisely $\mathrm{ID}_{F_{\mathbf{X}}}(0)$, the limit of the ID function as the distance tends to zero.
- Although the Karamata representation states that $\eta_{\mathbf{X}}$ must tend to a constant as $x \to 0$, the ID representation implies that $\eta_{\mathbf{X}}$ is independent of x altogether.
- The second-order EVT index ρ that asymptotically determines the rate of convergence of extreme values is the negation of the index of the regularly varying function $B^* = |\mathrm{ID}_{\mathrm{ID}_{f_{\mathbf{X}}}}|$. If this function is itself absolutely continuous, then $-\rho$ equals $\mathrm{ID}_{B^*}(0)$.

Unlike the first-order EVT index $\gamma_{\mathbf{X}}$, for which many estimation methods are known for both short-tailed and heavy-tailed distributions, little or no research effort seems to have been devoted to the estimation of ρ for short-tailed distributions [22]. However, we have seen that the estimation of $\mathrm{ID}_{|\mathrm{ID}_{\mathrm{ID}_{F_{\mathbf{X}}}}|}(0)$ would likely be of interest in data mining applications.

5 Hubness

The characterization of continuous distance distributions in terms of ID has the potential for explaining the interactions between intrinsic dimensionality, discriminability of distances, and local density variation in similarity applications. Here, as an example, we will see how the hubness phenomenon can be better understood in terms of ID.

5.1 Hubness of Distance Distributions

As originally defined, hubness is a property of a specific point with respect to a specific data set. To demonstrate a connection between data hubness and continuous intrinsic dimensionality, we must first extend the notion of hubness to that of a data distribution. Whereas the hubness value of a reference point can be calculated directly from the neighborhood information of a given data

set, when speaking of a distribution, we will instead consider the hubness of a distance distribution $\mathcal{F}_{\mathbf{X}}$ relative to the lower tails of a collection of distance distributions $\mathcal{Y} = \{\mathcal{F}_{\mathbf{Y}}\}$.

Definition 5 (Hubness of Distributions) Let $\mathcal{Y}$ be a non-empty collection of distance distributions, and let p be a probability threshold ($0 < p < 1$) such that for all $\mathcal{F}_{\mathbf{Y}} \in \mathcal{Y}$, the cumulative probability function $F_{\mathbf{Y}}$ of $\mathcal{F}_{\mathbf{Y}}$ achieves p at a unique distance value. Given a reference distance distribution $\mathcal{F}_{\mathbf{X}}$, the (normalized) hubness of $\mathcal{F}_{\mathbf{X}}$ relative to $\mathcal{F}_{\mathbf{Y}}$ and p is

$$\mathrm{NH}(\mathcal{F}_{\mathbf{X}}, \mathcal{Y}, p) \triangleq \frac{1}{p \cdot |\mathcal{Y}|} \sum_{\mathcal{F}_{\mathbf{Y}} \in \mathcal{Y}} F_{\mathbf{X}}\left(F_{\mathbf{Y}}^{-1}(p)\right) .$$

The normalization factor in the above definition allows us to compare the hubness of distributions across different settings. If for every $\mathcal{F}_{\mathbf{Y}} \in \mathcal{Y}$ the distance distribution $\mathcal{F}_{\mathbf{Y}}$ were identical to $\mathcal{F}_{\mathbf{X}}$, then the above expression would equal 1. The distribution $\mathcal{F}_{\mathbf{X}}$ can thus be regarded as a *hub* if $\mathrm{NH}(\mathcal{F}_{\mathbf{X}}, \mathcal{Y}, p) > 1$, and as an *anti-hub* if $\mathrm{NH}(\mathcal{F}_{\mathbf{X}}, \mathcal{Y}, p) < 1$. The more extreme the value of $\mathrm{NH}(\mathcal{F}_{\mathbf{X}}, \mathcal{Y}, p)$, the greater the degree to which $\mathcal{F}_{\mathbf{X}}$ may be considered a hub or anti-hub, as the case may be.

To see how Definition 5 relates to the established notion of the hubness of a data set, consider the point sample $\{\mathbf{y}_i | \, 1 \leq i \leq n\}$ of size $n \geq 2$ drawn from $\mathbb{R}^m$ according to the distribution $\mathcal{F}$. Let us assume that we are interested in the hubness of a distinguished point $\mathbf{x} \in \mathbb{R}^m$ for a given choice of neighborhood size k, where $1 \leq k \leq n-1$. Let the probability threshold p be chosen so that $p = k/n$. For each sample point $\mathbf{y}_i$, we denote by $\mathcal{F}_{\mathbf{Y}_i}$ the distance distribution induced by $\mathcal{F}$ with respect to $\mathbf{y}_i$. The collection of all such distributions will be denoted by $\mathcal{Y}$. Similarly, we denote by $\mathcal{F}_{\mathbf{X}}$ the distance distribution induced by $\mathcal{F}$ with respect to $\mathbf{x}$.

Consider now the relationships between $\mathbf{x}$ and each of the sample points $\mathbf{y}_i$. For any sample of n points drawn according to $\mathcal{F}$, if we assume that the cumulative distribution function $F_{\mathbf{Y}_i}$ achieves the probability p for a unique distance value $r_i = F_{\mathbf{Y}_i}^{-1}(p)$, then r_i is the neighborhood distance threshold at which one would expect k members of the sample to lie within the neighborhood. Thus, if the sample point $\mathbf{y}_i$ were to lie within distance r_i of $\mathbf{x}$, the hubness of $\mathbf{x}$ would be expected to increase by one. Since the probability of a sample point falling within distance r_i is given by $F_{\mathbf{X}}(r_i)$, the total hubness score that one would expect is given by

$$\sum_{i=1}^{n} F_{\mathbf{X}}(r_i) = \sum_{i=1}^{n} F_{\mathbf{X}}\left(F_{\mathbf{Y}_i}^{-1}(p)\right) .$$

After normalization by k, the expression is seen to fit the form of Definition 5:

$$\begin{aligned} \frac{1}{k} \sum_{i=1}^{n} F_{\mathbf{X}}(r_i) &= \frac{1}{pn} \sum_{i=1}^{n} F_{\mathbf{X}}\left(F_{\mathbf{Y}_i}^{-1}(p)\right) \\ &= \mathrm{NH}(\mathcal{F}_{\mathbf{X}}, \mathcal{Y} \backslash \{\mathcal{F}_{\mathbf{X}}\}, k/n) . \end{aligned}$$

5.2 Hubness and Continuous ID

Let $\mathbf{X}$ and $\mathbf{Y}$ be any two absolutely continuous random distance variables with support $[0, \infty)$, and let $\mathcal{F}_\mathbf{X}$ and $\mathcal{F}_\mathbf{Y}$ be their respective distributions. By solving for the quotient r/w in the intrinsic dimensional representations of $F_\mathbf{X}$ and $F_\mathbf{Y}$, Theorem 6 implies that

$$\begin{aligned}\frac{r}{w} &= \left(\frac{F_\mathbf{X}(r)}{F_\mathbf{X}(w) \cdot G_{F_\mathbf{X},0,w}(r)}\right)^{1/\mathrm{ID}_{F_\mathbf{X}}(0)} \\ &= \left(\frac{F_\mathbf{Y}(r)}{F_\mathbf{Y}(w) \cdot G_{F_\mathbf{Y},0,w}(r)}\right)^{1/\mathrm{ID}_{F_\mathbf{Y}}(0)} .\end{aligned}$$

These quantities, being unitless, allow the growth characteristics of distributions to be compared across different domains.

Theorem 10 *Let $\mathcal{Y}$ be a non-empty collection of distance distributions, and let $\mathcal{F}_\mathbf{X}$ be a distinguished distance distribution, with the assumption that all are absolutely continuous. Let p be a probability threshold $(0 < p < 1)$ such that for all $\mathcal{F}_\mathbf{Y} \in \mathcal{Y}$, the cumulative distribution function $F_\mathbf{Y}$ of $\mathcal{F}_\mathbf{Y}$ achieves p uniquely at the distance value $r_y > 0$, and the cumulative distribution function $F_\mathbf{X}$ of $\mathcal{F}_\mathbf{X}$ achieves p uniquely at the distance value $w > 0$. If in addition for all $\mathcal{F}_\mathbf{Y} \in \mathcal{Y}$, $\mathrm{ID}_{F_\mathbf{X}}(0)$ and $\mathrm{ID}_{F_\mathbf{Y}}(0)$ exist, and $F_\mathbf{X}$ and $F_\mathbf{Y}$ are positive everywhere except at distance zero, then the hubness of $\mathcal{F}_\mathbf{X}$ relative to $\mathcal{Y}$ and p equals*

$$\mathrm{NH}(\mathcal{F}_\mathbf{X}, \mathcal{Y}, p) = \frac{1}{|\mathcal{Y}|} \sum_{\mathcal{F}_\mathbf{Y} \in \mathcal{Y}} G_{F_\mathbf{X},0,w}(r_y) \left(\frac{p/F_\mathbf{Y}(w)}{G_{F_\mathbf{Y},0,w}(r_y)}\right)^{\frac{\mathrm{ID}_{F_\mathbf{X}}(0)}{\mathrm{ID}_{F_\mathbf{Y}}(0)}} .$$

Proof We may apply Theorem 6 to the definition of hubness of distributions, obtaining

$$\begin{aligned}\mathrm{NH}(\mathcal{F}_\mathbf{X}, \mathcal{Y}, p) &= \frac{1}{p \cdot |\mathcal{Y}|} \sum_{\mathcal{F}_\mathbf{Y} \in \mathcal{Y}} F_\mathbf{X}\left(F_\mathbf{Y}^{-1}(p)\right) \\ &= \frac{1}{p \cdot |\mathcal{Y}|} \sum_{\mathcal{F}_\mathbf{Y} \in \mathcal{Y}} F_\mathbf{X}(r_y) \\ &= \frac{F_\mathbf{X}(w)}{p \cdot |\mathcal{Y}|} \sum_{\mathcal{F}_\mathbf{Y} \in \mathcal{Y}} G_{F_\mathbf{X},0,w}(r_y) \cdot \left(\frac{r_y}{w}\right)^{\mathrm{ID}_{F_\mathbf{X}}(0)} \\ &= \frac{1}{|\mathcal{Y}|} \sum_{\mathcal{F}_\mathbf{Y} \in \mathcal{Y}} G_{F_\mathbf{X},0,w}(r_y) \cdot \left(\frac{F_\mathbf{Y}(r_y)}{F_\mathbf{Y}(w) \cdot G_{F_\mathbf{Y},0,w}(r_y)}\right)^{\frac{\mathrm{ID}_{F_\mathbf{X}}(0)}{\mathrm{ID}_{F_\mathbf{Y}}(0)}} \\ &= \frac{1}{|\mathcal{Y}|} \sum_{\mathcal{F}_\mathbf{Y} \in \mathcal{Y}} G_{F_\mathbf{X},0,w}(r_y) \cdot \left(\frac{p/F_\mathbf{Y}(w)}{G_{F_\mathbf{Y},0,w}(r_y)}\right)^{\frac{\mathrm{ID}_{F_\mathbf{X}}(0)}{\mathrm{ID}_{F_\mathbf{Y}}(0)}}\end{aligned}$$

as required. □

In the established notion of the hubness of a data set, as the data sample size n grows, the proportion of points k/n in the neighborhoods tends to zero. For the hubness of distributions, this is analogous to the tail probability p tending to zero. Asymptotically, then, as $p \to 0$, Theorem 7 implies that the hubness of $\mathcal{F}_{\mathbf{X}}$ relative to $\mathcal{Y}$ and p converges as:

$$\lim_{p\to 0} \mathrm{NH}(\mathcal{F}_{\mathbf{X}}, \mathcal{Y}, p) = \lim_{p\to 0} \frac{1}{|\mathcal{Y}|} \sum_{\mathcal{F}_{\mathbf{Y}} \in \mathcal{Y}} \left(\frac{p}{F_{\mathbf{Y}}\left(F_{\mathbf{X}}^{-1}(p)\right)} \right)^{\frac{\mathrm{ID}_{F_{\mathbf{X}}}(0)}{\mathrm{ID}_{F_{\mathbf{Y}}}(0)}} .$$

The contribution to the hubness due to distribution $\mathcal{F}_{\mathbf{Y}} \in \mathcal{Y}$ thus depends on both the relative intrinsic dimensionality with respect to $\mathcal{F}_{\mathbf{X}}$, and the relative density with respect to $\mathcal{F}_{\mathbf{X}}$. If $\mathcal{F}_{\mathbf{X}}$ simultaneously has a higher ID than $\mathcal{F}_{\mathbf{Y}}$ (that is, if $\mathrm{ID}_{F_{\mathbf{X}}}(0) > \mathrm{ID}_{F_{\mathbf{Y}}}(0)$), and a higher local density within distance $w = F_{\mathbf{X}}^{-1}(p)$ (that is, if $p > F_{\mathbf{Y}}(w)$), then the contribution of $\mathcal{F}_{\mathbf{Y}}$ to the hubness score is greater than 1. On the other hand, if the ID and local density of $\mathcal{F}_{\mathbf{X}}$ are both smaller than that of $\mathcal{F}_{\mathbf{Y}}$, the contribution is less than 1. Distributions $\mathcal{F}_{\mathbf{X}}$ of high hubness thus tend to have an intrinsic dimensionality and a local neighborhood density that is higher than is typical for the distributions in $\mathcal{Y}$. This theoretical explanation is in accordance with the empirical assessment of [63], who observed that hubness of data tends to increase with the intrinsic dimensionality of the subspace containing data clusters, and with proximity to cluster centers.

6 Conclusion

The theory presented in this paper constitutes a step towards the development of an overall theory of data mining: under the statistical framework of extreme value theory, it formally unites the notions of similarity measure, data density, data discriminability, intrinsic dimensionality, local inlierness (cluster membership) and outlierness, and hubness.

To realize the full practical potential of this theory of intrinsic dimensionality, efficient and accurate estimators are needed. Although existing estimators for the first-order EVT index can be used as is for first-order ID [4,32,42], no estimators seem to have yet been developed for the second-order inlierness/outlierness measure proposed in Section 3.3.

One quite promising direction for future work is that of density-based approaches to data mining, particularly for those applications in which subspaces are explicitly or implicitly explored (such as clustering and outlier detection, or subspace identification). The ID model presented in this paper can be regarded as a normalization of density information by (intrinsic) dimensional information, allowing for both influences to be accounted for in a natural way. Future work in this area should involve the design and testing of ID-based criteria for density-based data mining applications.

Another important direction is that of feature selection and metric learning. The ID 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.

References

1. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: OPTICS: Ordering points to identify the clustering structure. In: Proc. 1999 ACM SIGMOD Int. Conf. on Management of Data, pp. 49–60 (1999)
2. Balkema, A.A., de Haan, L.: Residual life time at great age. The Annals of Probability **2**, 792–804 (1974)
3. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: Statistics of Extremes: Theory and Applications. John Wiley & Sons (2004)
4. Beirlant, J., Vynckier, P., Teugels, J.: Excess function and estimation of the extreme-value index. Bernoulli **2**, 293–318 (1996)
5. Bernecker, T., Houle, M.E., Kriegel, H., Kröger, P., Renz, M., Schubert, E., Zimek, A.: Quality of similarity rankings in time series. In: Advances in Spatial and Temporal Databases – 12th International Symposium, SSTD 2011, Minneapolis, MN, USA, August 24-26, 2011, Proceedings, pp. 422–440 (2011)
6. Beyer, K.S., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is 'nearest neighbor' meaningful? In: Proceedings of the 7th International Conference on Database Theory (ICDT), pp. 217–235. Springer (1999)
7. Beygelzimer, A., Kakade, S., Langford, J.: Cover trees for nearest neighbors. In: International Conference on Machine Learning (ICML), pp. 97–104 (2006)
8. Bingham, N., Goldie, C., Teugels, J.: Regular Variation, vol. 27. Cambridge University Press (1989)
9. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. SIGMOD Rec. **29**(2), 93–104 (2000)
10. Bruske, J., Sommer, G.: Intrinsic dimensionality estimation with optimally topology preserving maps. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(5), 572–575 (1998)
11. Camastra, F., Vinciarelli, A.: Estimating the intrinsic dimension of data with a fractal-based method. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(10), 1404–1407 (2002)
12. Clarkson, K.L.: Nearest neighbor queries in metric spaces. Discrete & Comput. Geom. **22**, 63–93 (1999)
13. Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer (2001)
14. Dahan, E., Mendelson, H.: An extreme-value model of concept testing. Management Science **47**, 102–116 (2001)
15. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statistical Society, Series B **39**(1), 1–38 (1977)
16. Dunne, J., Ghanbari, M.: Efficient extreme value prediction for nonlinear beam vibrations using measured random response histories. Nonlinear Dynamics **24**, 71–101 (2001)
17. Falconer, K.: Fractal Geometry: Mathematical Foundations and Applications. John Wiley & Sons (2003)
18. Faloutsos, C., Kamel, I.: Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In: Proceedings of the 13th ACM Symposium on Principles of Database Systems (PODS), pp. 4–13 (1994)
19. Flexer, A., Schnitzer, D., Schlüter, J.: A MIREX meta-analysis of hubness in audio music similarity. In: Proc. 13th Int. Soc. for Music Information Retrieval Conf. (ISMIR), pp. 175–180 (2012)
20. Fukunaga, K., Olsen, D.R.: An algorithm for finding intrinsic dimensionality of data. Transactions on Computers **C-20**(2), 176–183 (1971)
21. Furon, T., Jégou, H.: Using Extreme Value Theory for Image Detection. Research Report RR-8244, INRIA (2013)
22. 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)

23. Goyal, N., Lifshits, Y., Schütze, H.: Disorder inequality: a combinatorial approach to nearest neighbor search. In: WSDM, pp. 25–32 (2008). DOI http://doi.acm.org/10.1145/1341531.1341538
24. Grassberger, P., Procaccia, I.: Measuring the strangeness of strange attractors. Physica D: Nonlinear Phenomena **9**(1–2), 189–208 (1983)
25. Grimshaw, S.D.: Computing maximum likelihood estimates for the generalized Pareto distribution. Technometrics **35**(2), 185–191 (1993)
26. Gupta, A., Krauthgamer, R., Lee, J.R.: Bounded geometries, fractals, and low-distortion embeddings. In: Proceedings of the 44th IEEE Symposium on Foundations of Computer Science (FOCS), pp. 534–543. IEEE Computer Society (2003)
27. de Haan, L., Resnick, S.: Second-order regular variation and rates of convergence in extreme-value theory. The Annals of Probability **24**(1), 97–124 (1996)
28. de Haan, L., Stadtmüller, U.: Generalized regular variation of second order. J. Australian Mathematical Society (Series A) **61**(3), 381–395 (1996)
29. Harris, R.I.: The accuracy of design values predicted from extreme value analysis. Journal of Wind Engineering and Industrial Aerodynamics **89**, 153–164 (2001)
30. Hausdorff, F.: Dimension und äußeres Maß. Mathematische Annalen **79**(1–2), 157–179 (1919)
31. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: NIPS 2005. MIT Press (2005)
32. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Annals of Statistics **3**(5), 1163–1174 (1975)
33. Hinneburg, A., Gabriel, H.H.: DENCLUE 2.0: Fast clustering based on kernel density estimation. In: Proc. 7th Int. Conf. on Intelligent Data Analysis, IDA'07, pp. 70–80 (2007)
34. Houle, M.E.: Dimensionality, discriminability, density & distance distributions. In: Proceedings of the 13th IEEE International Conference on Data Mining Workshops (ICDMW), pp. 468–473 (2013)
35. Houle, M.E., Kashima, H., Nett, M.: Generalized expansion dimension. In: Proceedings of the 12th IEEE International Conference on Data Mining Workshops (ICDMW), pp. 587–594 (2012)
36. Houle, M.E., Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? In: International Conference on Scientific and Statistical Database Management (SSDBM), pp. 482–500. Springer (2010)
37. Houle, M.E., Ma, X., Nett, M., Oria, V.: Dimensional testing for multi-step similarity search. In: Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), pp. 299–308 (2012)
38. Houle, M.E., Ma, X., Oria, V., Sun, J.: Efficient algorithms for similarity search in axis-aligned subspaces. In: 7th Int. Conf. on Similarity Search and Applications (SISAP), pp. 1–12 (2014)
39. Houle, M.E., Nett, M.: Rank cover trees for nearest neighbor search. In: 6th Int. Conf. on Similarity Search and Applications (SISAP), pp. 16–29 (2013)
40. Houle, M.E., Nett, M.: Rank-based similarity search: Reducing the dimensional dependence. IEEE Trans. Pattern Analysis and Machine Intelligence **37**(1), 136–150 (2015)
41. Hui, K., Wang, C.: Clustering-based locally linear embedding. In: 19th Int. Conf. on Pattern Recognition (ICPR), pp. 1–4 (2008)
42. Huisman, R., Koedijk, K.G., Kool, C.J.M., Palm, F.: Tail-index estimates in small samples. Journal of business and economic statistics **19**(2), 208–216 (2001)
43. Karamata, J.: Sur un mode de croissance réguliere des fonctions. Mathematica (Cluj) **4**, 38–53 (1930)
44. Karamata, J.: Sur un mode de croissance réguliere, théorèmes fondamentaux. Bull. Soc. Math. France **61**, 55–62 (1933)
45. Karger, D.R., Ruhl, M.: Finding nearest neighbors in growth-restricted metrics. In: Proceedings of the 34th ACM Symposium on Theory of Computing (STOC), pp. 741–750 (2002)
46. Karhunen, J., Joutsensalo, J.: Representation and separation of signals using nonlinear PCA type learning. Neural Networks **7**(1), 113–127 (1994)

47. Karydis, I., Radovanović, M., Nanopoulos, A., Ivanović, M.: Looking through the "glass ceiling": A conceptual framework for the problems of spectral similarity. In: Proc. 11th Int. Soc. for Music Information Retrieval Conf. (ISMIR), pp. 267–272 (2010)
48. Knees, P., Schnitzer, D., Flexer, A.: Improving neighborhood-based collaborative filtering by reducing hubness. In: Proc. 4th ACM Int. Conf. on Multimedia Retrieval (ICMR), pp. 161–168 (2014)
49. Krauthgamer, R., Lee, J.R.: Navigating nets: simple algorithms for proximity search. In: SODA '04: Proc. of 15th Annual ACM-SIAM Symp. on Discrete Algorithms, pp. 798–807 (2004)
50. Larrañaga, P., Lozano, J.A.: Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation, vol. 2. Springer (2002)
51. Lavenda, B.H., Cipollone, E.: Extreme value statistics and thermodynamics of earthquakes: Aftershock sequences. The Annals of Geophysics **43**, 967–982 (2000)
52. Levina, E., Bickel, P.J.: Maximum likelihood estimation of intrinsic dimension. In: Advances in Neural Information Processing Systems (NIPS) (2004)
53. Lifshits, Y., Zhang, S.: Combinatorial algorithms for nearest neighbors, near-duplicates and small-world design. In: SODA, pp. 318–326 (2009)
54. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Springer (1998)
55. Liu, H., Zhao, Z.: Spectral feature selection for supervised and unsupervised learning. ICML 2007 (2007)
56. Lloyd, S.P.: Least squares quantization in PCM. IEEE Trans. Information Theory **28**(2), 129–137 (1982)
57. McNulty, P.J., Scheick, L.Z., Roth, D.R., Davis, M.G., Tortora, M.R.S.: First failure predictions for EPROMs of the type flown on the MPTB satellite. Transactions on Nuclear Science **47**, 2237–2243 (2000)
58. Nanopoulos, A., Radovanović, M., Ivanović, M.: How does high dimensionality affect collaborative filtering? In: Proc. 3rd ACM Conf. on Recommender Systems (RecSys), pp. 293–296 (2009)
59. Pestov, V.: On the geometry of similarity search: Dimensionality curse and concentration of measure. Information Processing Letters **73**(1–2), 47–51 (2000)
60. Pettis, E., Bailey, T., Jain, A., Dubes, R.: An intrinsic dimensionality estimator from nearest-neighbor information. IEEE Transactions on Pattern Analysis and Machine Intelligence **1**, 25–37 (1979)
61. Pickands, J.: Statistical inference using extreme order statistics. The Annals of Statistics **3**, 119–131 (1975)
62. Radovanović, M., Nanopoulos, A., Ivanović, M.: Nearest neighbors in high-dimensional data: The emergence and influence of hubs. In: Proc. 26th Annual International Conference on Machine Learning, ICML '09, pp. 865–872 (2009)
63. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research **11** (2010)
64. Radovanović, M., Nanopoulos, A., Ivanović, M.: On the existence of obstinate results in vector space models. In: Proc. 33rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 186–193 (2010)
65. Radovanović, M., Nanopoulos, A., Ivanović, M.: Time-series classification in many intrinsic dimensions. In: Proc. 10th SIAM Int. Conf. on Data Mining (SDM), pp. 677–688 (2010)
66. Roberts, S.J.: Extreme value statistics for novelty detection in biomedical data processing. Proceedings of Science, Measurement and Technology **147**, 363–367 (2000)
67. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
68. Rozza, A., Lombardi, G., Ceruti, C., Casiraghi, E., Campadelli, P.: Novel high intrinsic dimensionality estimators. Machine Learning Journal **89**(1-2), 37–65 (2012)
69. Sander, J., Ester, M., Kriegel, H.P., Xu, X.: Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. Data Mining and Knowledge Discovery **2**(2), 169–194 (1998)
70. Schölkopf, B., Smola, A.J., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation **10**(5), 1299–1319 (1998)

71. Tenenbaum, J., Silva, V.D., Langford, J.: A global geometric framework for non linear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
72. Tomašev, N., Pracner, D., Brehar, R., Radovanović, M., Mladenić, D., Ivanović, M., Nedevschi, S.: Object recognition in WIKImage data based on local invariant image features. In: Proc. 9th Int. Conf. on Intelligent Computer Communication and Processing, ICCP '13, pp. 139–146 (2013)
73. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: The role of hubness in clustering high-dimensional data. IEEE Trans. Knowledge and Data Engineering **26**(3), 739–751 (2014)
74. Tryon, R.G., Cruse, T.A.: Probabilistic mesomechanics for high cycle fatigue life prediction. Journal of Engineering Materials and Technology **122**, 209–214 (2000)
75. Venna, J., Kaski, S.: Local multidimensional scaling. Neural Networks **19**(6–7), 889–899 (2006)
76. Verveer, P., Duin, R.: An evaluation of intrinsic dimensionality estimators. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(1), 81–86 (1995)
77. 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)
78. Wang, F., Sun, J.: Survey on distance metric learning and dimensionality reduction in data mining. Data Mining and Knowledge Discovery **29**(2), 534–564 (2015)
79. Wang, T., Guan, S.U., Liu, F.: Feature discriminability for pattern classification based on neural incremental attribute learning. In: Foundations of Intelligent Systems, *Advances in Intelligent and Soft Computing*, vol. 122, pp. 275–280. Springer (2012)
80. Weber, R., Schek, H.J., Blott, S.: A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: Proceedings of the 24rd International Conference on Very Large Data Bases (VLDB), pp. 194–205 (1998)
81. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.J.: Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems (NIPS), pp. 505–512 (2002)
82. Young, L.S.: Dimension, entropy and Lyapunov exponents. Ergodic Theory and Dynamical Systems **2**, 109–129 (1982)
83. Zimek, A., Schubert, E., Kriegel, H.P.: A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining **5**(5), 363–387 (2012)