Measuring Concentration of Distances—An Effective and Efficient Empirical Index

Sushma Kumari and Balasubramaniam Jayaram, *Member, IEEE*

Abstract—High dimensional data analysis gives rise to many challenges. One such that has come to gain a lot of attention recently is the concentration of distances (CoD) phenomenon, which is the inability of distance functions to distinguish points well in high dimensions. CoD affects almost every machine learning and data analysis algorithm in high dimensions. In this work, we present a novel efficient and effective empirical index that not only illustrates whether a distance function tends to concentrate for a given data set, but also enables us to measure the rate of concentration and allows us to compare different distance functions *vis-á-vis* their rate of concentration. As opposed to existing empirical indices, the proposed empirical measure uses only the internal characteristics of a given data set and hence is applicable on real data sets, which was hitherto not possible.

Index Terms—Dimensionality curse, concentration of distances, concentration function, dispersion function

1 Introduction

The term 'Big Data' has come to be related to any data whose characterics can be classified among one of the following five **V**'s: **V**olume, **V**ariety, **V**elocity, **V**eracity and **V**alue. While the first of the five V's, namely, Volume is largely taken to refer to the amount or size of data, yet another aspect related to volume is that of *Dimensionality* [39].

Data sets grow in their complexity not only due to their size but also due to the addition of more features or dimensions to the data. Given a data set $\mathcal{X}$, while Volume refers to the cardinality $\sharp\mathcal{X}$ of the data set $\mathcal{X}$, dimensionality refers to the space in which $\mathcal{X}$ itself is embedded.

Evolution of new data types such as images, videos, audio, gene expression data, etc., lead us to work with data in high dimension, thus forcing us to deal with the so called *Dimensionality Curse* (DC), a term that has come to refer to some non-intuitive phenomena that occur while dealing with data in high dimensions.

1.1 The Dimensionality Curse

The term *curse of dimensionality*, was first introduced by Bellman [5] while discussing optimisation problems involving high dimensions. However, recently, this term has come to denote or refer to many, often counterintuitive, challenges faced in high dimensions.

There are many aspects of the DC and their effects are still only being explored and is currently a hot topic of research. This is also clear from the many papers that continue to appear, see for instance, [2], [8], [23], [31]. Two of the well-known aspects of the DC are:

(i) *Combinatorial explosion in Search Space*, where the search space grows exponentially due to the increase in the number of variables [5].

(ii) *Hughes Phenomenon*—which refers to the need for at least a sub-exponential growth in the number of data points as dimension increases for many of the data analysis algorithms to be consistent, see for instance, [19], [27].

However, recently, many other aspects of the DC have also been discovered and are being investigated. For instance, the *Hubness Phenomenon*, which was first reported by [4] and later on investigated by Radovanovic et al. [28], [29], [36], which refers to the formation of hubs, i.e., a subset of data points which are more popular as nearest neighbors than other data points.

Yet another major aspect of the DC that has recently come to the fore is the *Concentration of Distances* phenomenon, which will form the main focus of this work.

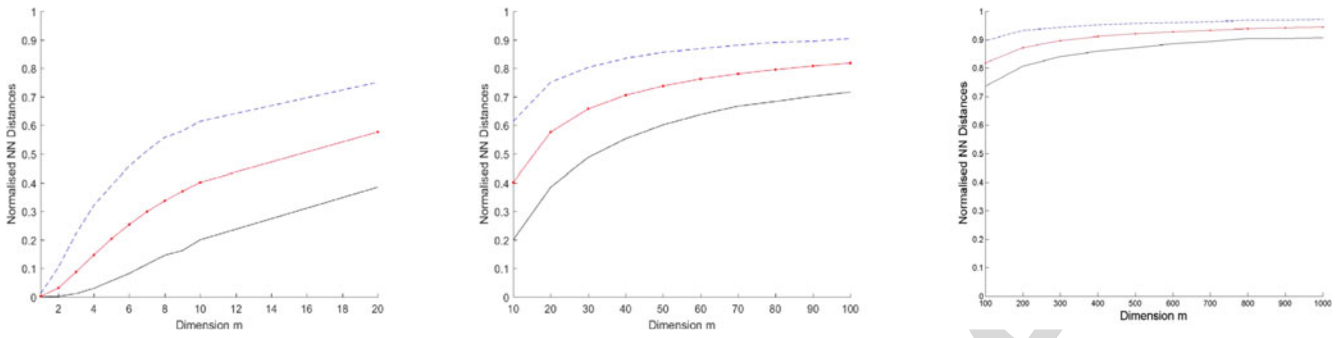
1.2 Concentration of Distances

Concentration of Distances (CoD), also referred to as *Concentration of Norms* in the literature, refers to the inability of distance functions to distinguish points well in high dimensions. To measure the closeness between any two objects/points we need the concept of a distance or its dual concept of similarity. However, as the dimension increases all the points appear to be approximately at the same distance and the distance function seems to lose its discriminative power. This phenomenon is called the concentration of distances.

Let $\mathcal{X} = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^m$ be a set of N data points from the m-dimensional Euclidean space. Let $q \in \mathbb{R}^m$ be an arbitrary but fixed query point and consider a distance function ρ to calculate the distances between points in $\mathcal{X}$—for instance, ρ could be the Euclidean distance (see (2) in Section 2). Let x^- and x^+ be the nearest and farthest points to q, i.e., $x^- = \arg\min_{x_i \in X} \rho(x_i, q),\ x^+ = \arg\max_{x_i \in X} \rho(x_i, q)$. As the dimension $m \to \infty$, one finds that $\rho(q, x^-) \approx \rho(q, x^+)$, which means that the distance of a query to the farthest point approaches the distance of the query to its nearest

- *S. Kumari is with the Department of Mathematics, Kyoto University, Kyoto, Prefecture 606-8501, Japan. E-mail: ma13m1011@iith.ac.in.*
- *B. Jayaram is with the Department of Mathematics, Indian Institute of Technology Hyderabad, Andhra Pradesh 502 205, India. E-mail: jbala@iith.ac.in.*

Manuscript received 7 Mar. 2016; revised 11 Oct. 2016; accepted 15 Oct. 2016. Date of publication 0 . 0000; date of current version 0 . 0000. Recommended for acceptance by H.T. Shen. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TKDE.2016.2622270



(a) $N = 1000,\ m = 1, \ldots, 20$ (b) $N = 1000,\ m = 10, \ldots, 100$ (c) $N = 1000,\ m = 100, \ldots, 1000$

Fig. 1. Concentration exhibited by the Euclidean distance when moving from low to high dimensions—k_M ($--$), k_m ($-$), k_A ($- \cdot -$).

point. Since $\rho(q, x^-) \leq \rho(q, x_i) \leq \rho(q, x^+)$ for $1 \leq i \leq N$, all distances to q begin to concentrate and are confined to a small domain. In other words, we can say that all the points in $\mathcal{X}$ are *almost* at the same distance to q. Thus the distances become less discriminative as the dimension grows and the distances between any two points begin to converge.

1.3 An Empiricial Illustration

Let us compare the Nearest Neighbour (NN) distances of an arbitrary query point with the average of pairwise distances of a given data set. Let $\mathcal{X} = \{x_i\}_{i=1}^N$ be N data points. Let $Y_{\max}, Y_{\min}, Y_{\text{avg}}$ denote the maximum, minimum and average of the Nearest Neighbour Euclidean distances. For instance, if z_i denotes the NN distance of x_i for each $i = 1, \ldots, N$, then $Y_{\max} = \max\{z_i : 1 \leq i \leq N\}$. We calculate $Y_{\min}, Y_{\text{avg}}$, similarly. Let $Y_{\mathcal{X}}$ denote the average of all pairwise distances in $\mathcal{X}$.

Let $k_M = \frac{Y_{\max}}{Y_{\mathcal{X}}}$ denote the normalised maximum NN distance w.r.t. the average of all pairwise distances. Similarly, let k_m and k_A denote the normalised minimum and average NN distances w.r.t. the average of all pairwise distances.

In Figs. 1a , 1b, and 1c, we plot the above three indices for $N = 1{,}000$ data points generated from Uniform distribution, viz., $\mathcal{X} \sim \mathcal{U}([-1, 1]^m)$, for varying dimensions, $m = 1, \ldots,$ 1,000. The plots allow us to make the following observations:

- In low dimensions, we see that $k_m \ll 1$ and there is enough separation between k_m and k_M, i.e., there is sufficient contrast present and hence points are well separated, see Fig. 1a.
- In medium dimensions, i.e., up to 100 dimensions, $0 \ll k_m < k_A < k_M$, which means that the minimum NN distances are beginning to increase and one can already see the presence of CoD, see Fig. 1b.
- However, as dimension increases, $k_m \to 1$, $k_M \sim k_A$ and $k_m \sim k_A$, i.e., the normalised maximum NN distances and the normalised minimum NN distances both converge to the normalised average NN distances. There is not much contrast present between the distances, i.e., all the distances seem to concentrate around the average value of the pairwise distances. Thus all points become *almost* equidistant to each other, see Fig. 1c.
- 3D surface plots of these indices for different values of m and N, see Fig. 2, shows that increasing N does not change the observed trend.

1.4 Why is CoD Important?

The concept of distance, or its dual notion of similarity, is all-pervasive and plays a central role in almost every algorithm or method in data analysis, from classification to clustering to similarity searches to pattern recognition. In many of these applications and algorithms, the distance functions which are useful in low dimensions are no longer effective in high dimensions, largely due to the affecting role of the CoD phenomenon. There are many domains where data are high dimensional and, thus, CoD poses an immediate and serious threat to their applicability to real world scenarios.

Let us consider searching, which is one of the most fundamental tasks used in every stream. The basic aim of similarity searching is to find an object or a set of objects similar to the given query object. For instance, in face recognition, one needs to search for a picture that is similar to the given query face in a database of images. A picture is made up of hundreds of thousands/millions of pixels and hence is a high dimensional object. Similarity searching methods, typically employ some kind of a distance function to measure the closeness between two objects. However, as shown above, due to the high dimensionality of the data, all pairwise distances can converge and hence our search might return a lot of candidates similar to our query object. This clearly puts a question mark on the usefulness of distance functions in high dimensions, see also [1], [18], [21].

Many nearest neighbour searching algorithms become computationally quite expensive in high dimensions [6], [7],

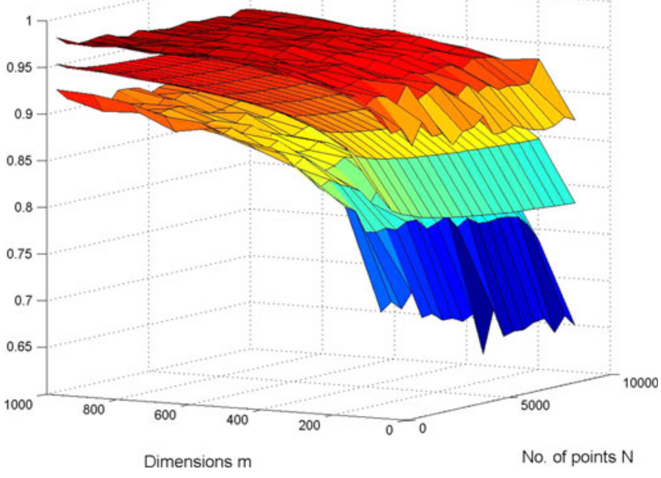


Fig. 2. Plots of the surfaces of k_m (bottom most), k_A, k_M (top most) for different values of m and N.

[10], [16], [20], [32]. However, it is made even more difficult by CoD. In fact, CoD raises the issue of whether or not the nearest neighbour is meaningful [9] in high dimension. Thus, in high dimensions, not only the efficiency of an algorithm is at stake, but also its effectiveness.

1.5 Motivation for This Work

Since the seminal paper of Beyer et al. [9] on the concentration of distances, studies that deal with this phenomenon typically employ an index to illustrate their point. There exist three major indices that are often used in such works, viz., the relative contrast ξ, relative variance γ and the concentration function α. While all these indices are excellent illustrators of whether a distance function ρ exhibits concentration or not, they do have their own merits and drawbacks. On the one hand, while both ξ, γ are empirical indices, and are hence easy to calculate, the relative contrast ξ is not amenable for theoretical analysis, while both ξ, γ are not conducive to measure the level or rate of concentration. On the other hand, the concentration function α overcomes some of these drawbacks, but is extremely computationally intensive to calculate.

Given a data set $\mathcal{X}$ and a set of distance functions ρ_i, there does not exist any index, so far, that is able to compare them and suggest or indicate their suitability w.r.t. their level of concentration on $\mathcal{X}$. Thus there is a need for an efficient empirical index that orders a given set of distance functions w.r.t. their suitability, *vis-á-vis* their level of concentration. This forms the main motivation of this submission.

1.6 Main Contributions of This Work

The main contributions of this work are twofold. First, we propose a novel efficient and empirical index function, called the *Dispersion Function* λ_ρ, that not only illustrates whether a distance function ρ exhibits concentration or not but also enables us to measure the rate at which it concentrates. This differs from the concentration function α_ρ in two aspects, viz.,

(i) λ_ρ is far less computationally intensive than α_ρ,

(ii) λ_ρ being an empirical measure can be calculated on any given data set $\mathcal{X}$. Further, λ_ρ only makes use of the internal characteristics of the given data set $\mathcal{X}$ and hence, unlike α_ρ, λ_ρ can be calculated even if the underlying distribution of $\mathcal{X}$ is unknown.

Second, based on the dispersion function λ_ρ, we have proposed an index τ_ρ which enables us to compare distances and indicate their suitability w.r.t. their level of concentration on $\mathcal{X}$.

2 Indices to Illustrate and/or Measure CoD

As discussed in Section 1, distances do tend to concentrate in higher dimensions. The illustration in Section 1.3 was done for Eucidean distances, i.e., distances calculated based on what is often referred to as an $\mathcal{L}_2$ distance, since it is a member of the family of Minkowski's $\mathcal{L}_p$-distances for $p \in [1, \infty)$, given as follows for a pair of vectors $x = (x_1, x_2, \ldots, x_m), y = (y_1, y_2, \ldots, y_m) \in \mathbb{R}^m$:

$$\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p \right)^{\frac{1}{p}}, \tag{1}$$

TABLE 1
Some Important Notations

Ω	Non-empty Domain
ρ	Distance Function
$\mathcal{X} \subset \Omega$	Data set
ξ_ρ	Relative Contrast w.r.t. ρ
γ_ρ	Relative Variance w.r.t. ρ
α_ρ	Concentration Function w.r.t. ρ
λ_ρ	Dispersion Function w.r.t. ρ
$\mathcal{L}_p$	pth Minkowski norm $p \geq 1$
$\mathcal{F}_p$	Fractional norm—$\mathcal{L}_p$ with $p \in (0, 1)$
μ^*	Counting Measure
$\lvert\mathcal{X}\rvert$	Cardinality of the dataset $\mathcal{X}$
$\mu^*_{\mathcal{X}}$	Normalised Counting Measure w.r.t. $\lvert\mathcal{X}\rvert$

$$\mathcal{L}_p(x, y) = \|x - y\|_p = \left(\sum_{i=1}^{m} |x_i - y_i|^p \right)^{\frac{1}{p}}. \tag{2}$$

Note that $\mathcal{L}_p$ is a metric [11].

It can be easily shown that the above indices, viz., k_M, k_m, k_A, behave similarly when we use a different $\mathcal{L}_p$-distance, i.e., they all still do converge. However, the following questions arise:

(i) Even if all $\mathcal{L}_p$-distances concentrate, do they all concentrate in the same manner? In which case, can one talk about the rate of concentration?

(ii) Are there indices that allow comparison between different distance functions w.r.t. their concentration? Are they calculable empirically? Are they also amenable for theoretical studies?

Since the seminal paper on this topic by Beyer et al. [9], there have been many studies dealing with the above posers. In this section, we give a brief yet substantive review of these works and the indices proposed therein, highlight their advantages and indicate the contexts in which they are not readily applicable, thus leading up to the motivation behind this work.

2.1 Fixing the Notation

We introduce some notations and concepts which will be used in the rest of the paper. For a concise summary of them, please refer to Table 1.

- The triple (Ω, ρ, μ) will denote a measurable metric space, where Ω is the domain, ρ is the metric on Ω and μ is a probability measure on Ω.
- Further, the measure μ we consider will always be absolutely continuous and hence we can associate a distribution $\mathcal{R}$ which will be used to obtain a finite sample of N-points $\mathcal{X} = \{x_1, x_2, \ldots, x_N\} \subset \Omega$. We will then write $\mathcal{X} \sim \mathcal{R}$ to denote that the data set $\mathcal{X} \subset \Omega$ has been generated using the distribution $\mathcal{R}$. Often the quadruple $(\Omega, \mathcal{X}, \rho, \mu)$ is termed as a *Similarity Workload*, see [25], [26].
- We assume that there always exist a $\mathbf{0} \in \Omega$ designated as the origin of the domain Ω. This is almost always true since usually $\Omega \subseteq \mathbb{R}$.
- μ^* will denote the counting measure, i.e., if $\mathcal{X}$ is finite, $\mu^*(\mathcal{X}) = \sharp\mathcal{X}$, the cardinality of $\mathcal{X}$.

- Note that if $\Omega \subseteq \mathbb{R}^m$, then we will also use the notation $\Omega^m, \mathcal{X}^m$ for added emphasis. In such cases, the other quantities like $\mu^m, \rho^m, \mathcal{R}^m$, etc., are appropriately defined.
- By $\|\cdot\|$ we denote a real valued function on Ω, i.e, $\|\cdot\| : \Omega \to \mathbb{R}$, which is taken to measure the distance of an $x \in \Omega$ to the origin $\mathbf{0} \in \Omega$, i.e., $\|x\| = \rho(x, \mathbf{0})$. To remain consistent with earlier works, we term $\|x\|$ to be the *norm* of the vector x, even when $\|\cdot\|$ does not satisfy all the properties of a norm, as is common in the literature.
- $D^m_{\max}$ ($D^m_{\min}$) denotes the maximum (minimum, resp.) of the norms in a given data set $\mathcal{X}^m$ of N points, i.e., the distance of the farthest (nearest, resp.) point in $\mathcal{X}^m$ to the origin w.r.t. the metric ρ:

$$D^m_{\max} = \max\{\|x^m_i\| = \rho(x^m_i, \mathbf{0}) \ : \ x^m_i \in \mathcal{X}^m\},$$
$$D^m_{\min} = \min\{\|x^m_i\| = \rho(x^m_i, \mathbf{0}) \ : \ x^m_i \in \mathcal{X}^m\}.$$

- $E[Z]$ and $var[Z]$ will denote the expectation and variance of a random variable Z.

2.2 Existence of CoD: Theoretical Analysis

Towards discussing the questions raised in Section 2 above, we begin by recalling the seminal result of Beyer et al. [9], wherein they discussed the existence of meaningful nearest neighbours in high dimension. Their result showed that under some reasonable assumptions on the data distribution $\mathcal{R}$, distance functions do concentrate.

Theorem 2.1 ([9], Theorem 1). *Let $(\Omega^m, \rho^m, \mu^m)$ be an m-dimensional measurable metric space, let $\mathcal{X}^m = \{x^m_1, x^m_2, \ldots, x^m_N\}$ be a finite sample of N points such that $x^m \sim \mathcal{R}^m$ and $D^m_{\max}, D^m_{\min}$ are as defined above. Further, let $E[\|x^m\|]$ and $var[\|x^m\|]$ be finite and $E[\|x^m\|] \neq 0$. If*

$$\lim_{m\to\infty} var\left(\frac{\|x^m\|}{E\|x^m\|}\right) = 0, \tag{3}$$

then for all $\varepsilon > 0$,

$$\lim_{m\to\infty} P[D^m_{\max} \leq (1+\varepsilon) D^m_{\min}] = 1. \tag{4}$$

The above result points out that nearest neighbor searching is not meaningful when the variance of the ratio of the distance between any two random points, drawn from the data distribution, and the expected distance between them converges to zero as dimension goes to infinity. This, in essence, means that almost all points are equidistant to the query point.

Theorem 2.1 clearly discusses only a sufficient condition for concentration, i.e., the distance to the nearest neighbor and the distance to the farthest neighbor tend to converge, in a probabilistic sense, as the dimension m increases. In other words, we get a poor contrast if the spread between the points tends towards 0. However, the question of whether this condition is also necessary was not known. Almost after a decade after the work of Beyer et al., the converse of Theorem 2.1 was proved by Durrant and Kabán, see [12], Theorem 2, p. 387.

2.3 Some Indices to Illustrate CoD

Based on the theoretical results of Beyer et al. [9] proving the existence of CoD in high dimensions, two indices have been proposed to study the tendency of concentration among different distances.

2.3.1 Relative Contrast—An Index to Illustrate CoD

The first of them is the Relative Contrast proposed by Aggarwal et al. [3].

Definition 2.2 ([3], p. 422). *Let us consider an m-dimensional similarity workload, $(\Omega^m, \mathcal{X}^m, \rho^m, \mu^m)$. The Relative Contrast (RC), w.r.t. ρ, is defined as*

$$\xi_\rho(m) = \frac{D^m_{\max} - D^m_{\min}}{D^m_{\min}} \ . \tag{RC}$$

Defining relative contrast thus, Aggarwal et al. [3] showed that when ρ is any of the Minkowski norms $\mathcal{L}_p$ for $p \in [1, \infty)$, $\xi_{\mathcal{L}_p}(m) \to 0$ as $m \to \infty$. Interestingly, based on the bounds obtained for $\xi_{\mathcal{L}_p}(m)$ they argued that if the exponent $p \in (0, 1)$ in (1) then such *p-norms*, which they called *fractional norms* and were denoted by $\mathcal{F}_p$, were better than Minkowski distances $\mathcal{L}_p$. It should be mentioned that when $p \in (0, 1)$ then the fractional distances $\mathcal{F}_p$ are not norms, or even a metric, since they do not satisfy the triangle inequality. Thus in the sequel, we refer to all such functions with the more general term *distance functions*.

2.3.2 Relative Variance—Another Index to Illustrate CoD

While Aggarwal et al. [3] took their motivation from (4) of Theorem 2.1 to propose ξ_ρ, François et al. [15] proposed yet another index, but this time taking their cue from (3) of Theorem 2.1, to demonstrate if a distance function suffers from the concentration phenomenon or not.

Definition 2.3 ([15], p. 877). *Given an m-dimensional similarity workload, $(\Omega^m, \mathcal{X}^m, \rho^m, \mu^m)$, the relative variance of the distance function ρ is defined as*

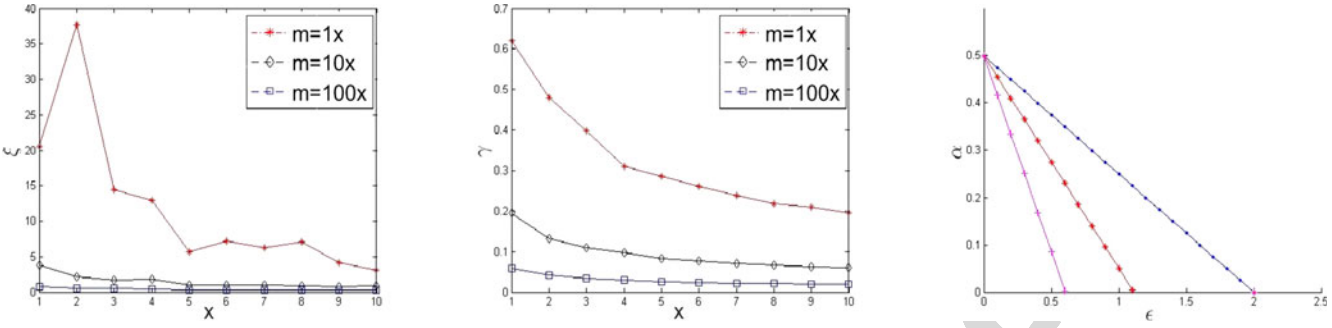
$$\gamma_\rho(m) = \frac{\sqrt{\ Var\ (\|x^m\|)}}{E(\|x^m\|)} \ ,$$

where, as usual, $\|x\| = \rho(x, \mathbf{0})$.

The relative variance γ_ρ illustrates the concentration of distances by comparing the spread with the expectation of the distances. If γ_ρ has a small value then it indicates that distances are concentrated and a large value for γ_ρ denotes a good amount of spread between the distances. In some sense it is similar to relative contrast, as ξ_ρ also compares the measure of spread to the measure of location.

In fact, Theorem 2.1 and its converse can be restated as follows based on the above indices: *If the relative variance is not tending to zero then the relative contrast will also not converge to zero and therefore one does obtain a good separation between points.*

Remark 2.4. Following are some of the merits and demerits of the indices ξ_ρ and γ_ρ:



(a) $\xi_{\mathcal{L}_2}$, $N = 100m$ (b) $\gamma_{\mathcal{L}_2}$, $N = 100m$ (c) $\alpha_{\mathcal{L}_1}^{\Omega_i}$

Fig. 3. (a) Relative Contrast and (b) Relative Variance of the Euclidean distances as dimensions increase, see Remark 2.4. (c) Concentration functions $\alpha_{\mathcal{L}_1}^{\Omega_1}(-\bullet-)$,$\alpha_{\mathcal{L}_1}^{\Omega_2}$ $(-*-)$, $\alpha_{\mathcal{L}_1}^{\Omega_3}(-+-)$ of Example 2.6.

- ⊕ ξ_ρ does illustrate the concentration of distances well. Fig. 3a plots the relative contrast of the Euclidean distance $\mathcal{L}_2$ on the set $\mathcal{X}$ consisting of N data points, where $\mathcal{X} \sim \mathcal{U}([-1,1]^m)$ and $N = 100m$. It is clear from the plots that in low dimensions, viz., $m = 1, \ldots, 10$, $\xi_{\mathcal{L}_2}$ is quite high, while even in medium dimensions, viz., $m = 10, \ldots, 100$, $\xi_{\mathcal{L}_2}$ starts to fall drastically and in high dimensions, viz., $m = 100, \ldots, 1{,}000$, $\xi_{\mathcal{L}_2}$ is almost zero. γ_ρ, like ξ_ρ, does illustrate the concentration of distances well. Once again, Fig. 3b illustrates that in low dimensions $\gamma_{\mathcal{L}_2}$ is far away from zero, but in medium to large dimensions $\gamma_{\mathcal{L}_2}$ starts to approach zero faster.
- ⊕ Both ξ_ρ and γ_ρ are empirically calculable and hence are applicable on any distance function.
- ⊖ ξ_ρ is not amenable for theoretical analysis, as finding the distributions of minimum and maximum pairwise distances in a data set with given probability distribution for most distances is extremely complicated. Even in the case of $\mathcal{L}_p$ and $\mathcal{F}_p$ norms only some loose bounds have been obtained and when N is finite. Where γ_ρ overtakes ξ_ρ is in its amenability for theoretical analysis, as is clearly demonstrated by François et al. [15]. Note, however, that theoretical analysis can become difficult with arbitrary distance functions.
- ⊖ Both ξ_ρ and γ_ρ illustrate the concentration of a particular distance in the asymptotic case as $m \to \infty$. However, given two distances, say ρ_1, ρ_2, and a specific data set $\mathcal{X}^m$ (thus the dimensionality m is fixed), it is not clear if the values $\xi_{\rho_1}(m)$ and $\xi_{\rho_2}(m)$ allow us to compare the distances ρ_1, ρ_2 vis-á-vis their concentration. Note that ξ_ρ, γ_ρ are not *strictly* decreasing functions of m.

2.4 A Theoretical Index to Measure CoD

While ξ_ρ and γ_ρ illustrate the concentration phenomenon well, they do not give any information on the rate at which a distance function concentrates. Recent studies, see Pestov [25], have started considering a more general mathematical function to measure concentration.

Definition 2.5 (cf. [24], [25], [37]). *Let us be given a measurable metric space (Ω, ρ, μ). The concentration function α_ρ : $\mathbb{R}^{\geq 0} \to [0, \frac{1}{2}]$ is defined as follows:*

$$\alpha_\rho(\varepsilon) = \begin{cases} 1 - \inf\{\mu(A_\varepsilon) : A \subseteq \Omega \ \& \ \mu(A) \geq 1/2\}, & \varepsilon > 0, \\ \frac{1}{2}, & \varepsilon = 0, \end{cases}$$

where $A_\varepsilon = \{x \in \Omega : \rho(x, a) < \varepsilon$ for some $a \in A\}$.

The value $\alpha_\rho(\varepsilon)$ gives an upper bound on the measure of the complement to the ε-neighborhood A_ε of every subset A of measure greater than or equal to $\frac{1}{2}$. It can be easily seen that α_ρ is a decreasing function. Thus, the rate of concentration of a distance function ρ, in the considered workload, is measured based on the rate at which α_ρ decreases.

If a distance function ρ concentrates, the concentration function α_ρ approaches zero faster. The smaller the value of ε at which $\alpha_\rho(\varepsilon) = 0$, the faster the distance function concentrates. In fact, the rate at which α_ρ decreases is illustrative of the fact that the pairwise distances, as measured by ρ, concentrate near their mean/median value.

In Example 2.6, we consider some simple measurable metric spaces (Ω, ρ, μ) and plot their respective concentration functions in Fig. 3c, which shows that α_ρ does measure the rate of concentration, i.e., how fast a given distance ρ concentrates in a domain of interest Ω with respect to the data distribution obtained from the measure μ.

Example 2.6.

(i) Let us consider the space (Ω_1, ρ, μ), where $\Omega_1 = [0,1] \cup [2,3]$, ρ is the usual metric on $\mathbb{R}$, viz., the $\mathcal{L}_1$ metric and μ is the Lebesgue measure. The corresponding concentration function $\alpha_{\mathcal{L}_1}^{\Omega_1}$ $(-\bullet-)$ is plotted in Fig. 3c.

(ii) Let us now consider the domains $\Omega_2 = [0,1] \cup [1.1, 2.1]$ and $\Omega_3 = [-0.6, -0.1] \cup [0,1] \cup [1.1, 1.6]$, while ρ, μ remain the same. The corresponding concentration functions $\alpha_{\mathcal{L}_1}^{\Omega_2}$ $(-*-)$ and $\alpha_{\mathcal{L}_1}^{\Omega_3}$ $(-+-)$, respectively, are plotted in Fig. 3c.

Following are some of the merits and demerits of the concentration function α_ρ:

- ⊕ Not only does α_ρ illustrate the concentration of distances, it also allows us to measure it. In Example 2.6, we saw that the $\mathcal{L}_1$ distance behaves differently for different domains, even though the measure of the underlying domains $\mu(\Omega_i)$ was the same.
- ⊕ The concentration function has been used in the analysis of many a measurable metric space to obtain

TABLE 2
Comparison between the Indices ξ_ρ, γ_ρ and α_ρ : (**EC**)—Empirical Calculations, (**TA**)—Theoretical Analysis, (**MC**)—Measuring Concentration, and (**CD**)—Comparing Distances

Index	Suitable for			
	(EC)	(TA)	(MC)	(CD)
ξ_ρ	✓	×	×	×
γ_ρ	✓	✓	×	×
α_ρ	×	✓	✓	×

theoretical bounds, see, for, e.g., the monograph of Ledoux [24].

- ⊕ Given a measurable space, (Ω, μ) of an arbitrary but fixed dimension m, one can visually compare different distance functions ρ_i based on the rate at which the corresponding α_{ρ_i} tends to zero.
- ⊖ However, given two distances, say ρ_1, ρ_2, it is not clear if the functions α_{ρ_1} and α_{ρ_2} can be strictly ordered, based on the usual point-wise ordering of functions. Thus even if $\alpha_{\rho_1}(\varepsilon) < \alpha_{\rho_2}(\varepsilon)$, there can exist an $\varepsilon' \neq \varepsilon$ such that $\alpha_{\rho_1}(\varepsilon') > \alpha_{\rho_2}(\varepsilon')$. Thus other than the visual cues, purely based on α_{ρ_i}'s it is not apparent how to compare distances w.r.t. their concentration.
- ⊖ Further, if we do not know the underlying distribution of a particular dataset *a priori*, i.e., if μ is unknown, we cannot determine α_ρ theoretically.
- ⊖ While it is amenable for certain kind of theoretical analysis, as an empirical index α_ρ is highly computationally expensive. For large sets calculating α_ρ is very cumbersome as we need to find every subset of Ω with measure at least half. In fact, using the counting measure $\mu^*_{\mathcal{X}}$, given a set with cardinality N, the number of subsets with measure greater than $\frac{1}{2}$ is equal to $\sum_{k=\frac{N}{2}}^{N} {}^{N}C_k = 2^{N-1}$.

2.5 Motivation for This Work—Need for an Efficient Empirical Index to Measure CoD

It is clear from the discussions so far that, on the one hand, ξ_ρ, γ_ρ are best suited in empirical settings and are efficiently calculable, but α_ρ does not enjoy these properties. On the other hand, α_ρ can be useful in comparing distances w.r.t. their concentration, while neither ξ_ρ nor γ_ρ gives us that information. Table 2 summarises the properties of the above indices against these parameters, from whence we see the need for an efficient empirical index, *á la* ξ_ρ, γ_ρ, and that which would measure the concentration and hence allow us to compare between different distance functions, *á la* α_ρ. Our approach towards defining this index stems from the concept of stability of queries. In the next sections, we discuss these in detail and come up with an empirical index that upper bounds α_ρ, which is also comparatively easier to calculate than α_ρ.

3 A Novel Efficient Empirical Index to Measure CoD

In this section, we recall the concept of stability of range queries and discuss the stability of workloads. Based on these discussions, we present our novel index that not only illustrates and measures concentration, but with the help of which we will also be able to compare distances vis-á-vis their concentration.

3.1 Stability of a Query

Let $(\Omega, \mathcal{X}, \rho, \mu)$ be a given similarity workload. Let a query $q \in \Omega$ and an $\varepsilon \in \overline{\mathbb{R}}_+ = [0, \infty)$ be given. By a *range-query* problem we refer to the determination of the set of all points in $\mathcal{X}$ that are within ε units away from q, i.e., we need to find the ε-neighbourhood of q in $\mathcal{X}$

$$S = \mathsf{N}(q, \varepsilon) = \{x' \in \mathcal{X} : \rho(x', q) \leq \varepsilon\}.$$

In [9], the authors discuss when a range-query is stable by defining the stability of a range-query as follows:

Definition 3.1 ([25], p. 48, cf. [9], Definition 1). *Given a query point $q \in \Omega$ and an $\varepsilon \in \overline{\mathbb{R}}_+$, a range-query is said to be ε-unstable if*

$$\mu^*_{\mathcal{X}}(\mathsf{N}(q, (1+\epsilon) * \delta)) \geq \frac{\mu^*_{\mathcal{X}}(X)}{2},$$

where, $\delta = \min\{\rho(q, x) : x \in \mathcal{X}\}$, the NN-distance of q.

In other words, as formulated initially in [9], a range-query is said to be *unstable* if *most* of the data set is covered within the ε-δ sphere of the query q. The subsequent quantification to half (*from most*) of the data set in Definition 3.1 was done by Pestov [25].

Taking a cue from Definition 3.1, we discuss the stability of a particular workload and propose an index that will help us in achieving our goals.

3.2 The g-δ-Count

Let $\mathbb{N}_N = \{1, 2, \ldots, N\}$. Let $\mathcal{X}$ be a given data set whose cardinality is N, i.e., $\sharp\mathcal{X} = N$.

Consider an $x \in \mathcal{X}$ and let δ denote the NN distance of x. For a $g \in \overline{\mathbb{R}}_+$, let us define the g-δ count of the point x as

$$C(x, g\delta) = \mu^*_{\mathcal{X}}(\mathsf{N}(x, g\delta)).$$

Clearly, $C(x, g\delta)$ gives the fraction of the number of data points in the g-δ neighborhood of x. Note that $\mu^*_{\mathcal{X}}$ is the normalised counting measure, i.e., for an $A \subset X$ and $|X| < \infty$ we have $\mu^*_{\mathcal{X}}(A) = \frac{|A|}{|X|}$.

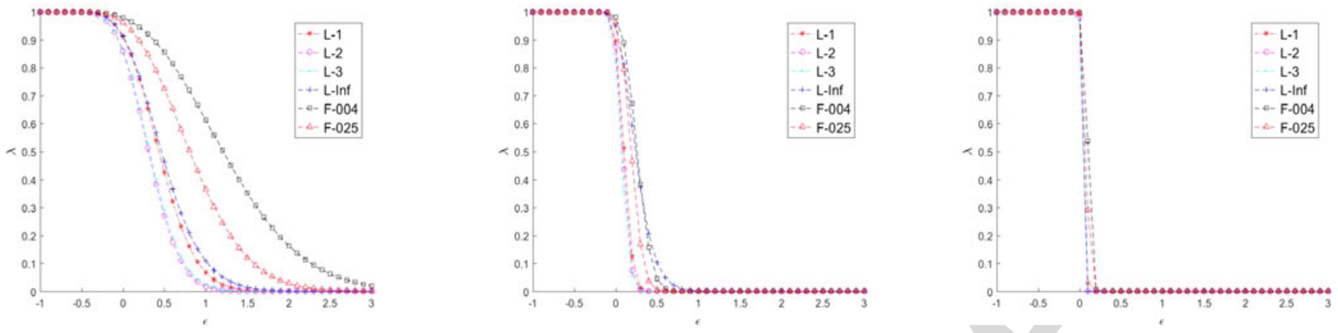
For small values of g, if the $C(x, g\delta)$ values of most of the $x \in \mathcal{X}$ are high, then one surmises that more points lie in the dilated g-δ neighborhood of each $x \in \mathcal{X}$ and hence the data are distributed very close to each other and the relative distances between the data points will be small. Thus, $C(\cdot, g\delta)$ does keep track of the concentration of points. Specifically, given a dataset, even without the information of the distribution of the dataset, $C(\cdot, g\delta)$ is computable and hence further analysis is possible.

Now, we define C^*, the complement of C as follows:

$$C^*(x, g\delta) = 1 - C(x, g\delta).$$

$C^*(x, g\delta)$ gives the fraction of the data set that the point x is not able to arrest through its dilated g-δ neighborhood.

Clearly, if $C(x, g\delta)$ is large for a point x then $C^*(x, g\delta)$ will be small. Thus, when g is relatively small, small values of



(a) Gaussian, $N = 10\text{K}$, $m = 10$

(b) Gaussian, $N = 10\text{K}$, $m = 100$

(c) Uniform, $N = 10\text{K}$, $m = 1000$

Fig. 4. Plot of λ_{ρ_i} for different workloads—see Section 4.1.

$C^*(\cdot, g\delta)$ for most of the data set would indicate that distances are concentrating and vice versa. Note, however, that the above observation is valid only if a large number of points have small values of $C^*(\cdot, g\delta)$. For instance, if $C^*(\cdot, g\delta)$ values are very large for only a minority of the data points, it does not mean that the distances are not concentrating. It may happen that these points are outliers and rest of the data points that are not captured by these outliers are closely packed. Therefore, we need to check the overall behavior of all the data points. Taking cue from this observation, we propose a novel index that will help us to accomplish our objective in the next section.

3.3 The Dispersion Function—λ_ρ

Consider the similarity workload $(\Omega, \mathcal{X}, \rho, \mu^*_{\mathcal{X}})$. If δ_i is the NN distance of an $x_i \in \mathcal{X}$, let δ_0 denote the maximum of the NN distances in $\mathcal{X}$, i.e.,

$$\delta_0 = \max_{x_i \in \mathcal{X}}\{\delta_i\} = \max_{x_i \in \mathcal{X}}\left\{\min_{x_j \in \mathcal{X}} \rho(x_i, x_j)\right\}. \tag{5}$$

Definition 3.2. *Let us consider the similarity workload* $(\Omega, \mathcal{X}, \mu^*_{\mathcal{X}}, \rho)$ *and let* δ_0 *be as defined in (5). The* dispersion function $\lambda_\rho : [-1, \infty) \to [0, 1]$ *is defined as follows:*

$$\lambda_\rho(\varepsilon) = \underset{x_i \in \mathcal{X}}{\mathbf{avg}}\{C^*(x_i, (1+\varepsilon)\delta_0)\}, \tag{6}$$

*where **avg** is the the usual statistical average of the values.*

What does the dispersion function λ_ρ really indicate? For a given $\varepsilon > 0$, λ_ρ returns the average of the fraction of the data set that is not captured by a data point in its dilated $(1+\varepsilon)\delta_0$ neighborhood. Thus, when N is large, high values of λ_ρ indicate that a large part of the data set are such that most of the data are lying at a distance greater than $(1+\varepsilon)\delta_0$ to each of them. If we take ε to be small, data are at least δ_0 distance away and so data will still be well separated. Thus, essentially, λ_ρ can be considered as a statistical measure of the dispersion as measured by the distance function ρ. As will be shown in the next section, where we discuss its properties, the dispersion function λ_ρ does have many desirable properties, one of which is that it forms an upper bound for the concentration function α_ρ when the distance function under consideration is known to concentrate; thus if λ_ρ decreases at a faster rate, then so does α_ρ.

From the definition of λ_ρ and the discussion above, the following remarks are readily verifiable:

⊕ λ_ρ is an empirical function and illustrates the concentration of distances well.

⊕ λ_ρ is calculable for arbitrary distances.

⊕ λ_ρ, like α_ρ, is calculable for any data set with fixed dimensionality.

⊕ Even if we do not know the underlying distribution of a particular dataset *a priori*, i.e., even if μ is unknown, we can still determine λ_ρ.

However, it is not immediately clear whether λ_ρ, like α_ρ, measures the rate of concentration or allows us to compare distance functions w.r.t. their concentration. We take this up in detail in the next section.

4 Empirical and Theoretical Analysis of λ_ρ

In this section, we discuss the properties and characteristics of the dispersion function λ_ρ. We begin by empirically plotting λ_ρ for different distance functions ρ, on both synthetic and some real data sets and make some evidential observations based on them, some of which are also validated theoretically later on. Following this, we discuss the behaviour of the dispersion function on distance functions and show how λ_ρ can help in distinguishing distance functions based on their concentration. Finally, we do a comparative study between λ_ρ and α_ρ.

4.1 Studies on Synthetic and Real Datasets

Let $\Omega = [-1, 1]^m$ be the m-dimensional unit hyper cube. We consider two data sets $\mathcal{X}_1, \mathcal{X}_2$ of cardinality N, where

- $\mathcal{X}_1$ is a set of N uniformly distributed points in Ω, i.e., $\mathcal{X}_1 \sim \mathcal{U}([-1, 1]^m)$,
- $\mathcal{X}_2$ is a set of N points generated using a normal distribution with mean 0 and variance 0.09, on each of the m dimensions, so as to ensure that $\mathcal{X}_2 \subset \Omega$, i.e., $\mathcal{X}_2 \sim \mathcal{N}(0, 0.09)$.

For the distance function ρ, we consider the Minkowski distance functions $\mathcal{L}_p$ for $p = 1, 2, 3, \infty$ and the fractional distances $\mathcal{F}_p$ for $p = 0.04, 0.25$. We consider the following two sets of synthetic workloads $\mathcal{W}_1 = (\Omega, \mathcal{X}_1, \rho_i)$ and $\mathcal{W}_2 = (\Omega, \mathcal{X}_2, \rho_i)$, where ρ_i is one of the six distance functions listed above. We typically took the number of data points $N = 10$ K, restricted largely due to the computational power available, and plotted the λ_{ρ_i} values for $m = 10, 100$ and 1,000. In Figs. 4a, 4b, and 4c we plot the graph of different λ_{ρ_i} as ε varies from -1 to 3 in steps of 0.1.

One of the main advantages of λ_ρ over α_ρ is that it could be applied to real data sets, where usually there is no *a priori*

TABLE 3
Real Data Sets

Dataset $\mathcal{X}_i$	Dimension m	$\sharp$ of datapoints N
Splice	60	1,000
Protein	357	6,621
Colon Cancer	2,000	62
Gisette	5,000	6,000
Duke	7,129	44
Dexter	20,000	300

knowledge about the underlying distribution. We consider the UCI data sets given in Table 3, i.e., the sets of workloads $(\mathcal{X}_i, \rho_j, \mu^*_{\mathcal{X}})$, where $\rho_j,\ j = 1, 2, \ldots, 6$ are the same set of six Minkowski distance functions considered above. The superimposed plots of λ_{ρ_j} for four of the workloads, with $m \geq 2{,}000$, is given in Figs. 5a, 5b, 5c, and 5d.

From Figs. 4a, 4b, and 4c and Figs. 5a, 5b, 5c, and 5d we see that λ_ρ is a decreasing function of ε. The following result demonstrates theoretically the above empirical observation.

Theorem 4.1. *Given a similarity workload* $(\Omega, \mathcal{X}, \rho, \mu)$, λ_ρ *is a decreasing function, i.e.,* $\varepsilon_1 \leq \varepsilon_2 \longrightarrow \lambda_\rho(\varepsilon_1) \geq \lambda_\rho(\varepsilon_2)$.

Proof. Let $\varepsilon_1 \leq \varepsilon_2$ for $\varepsilon_1, \varepsilon_2 \in [-1, \infty)$ and $\delta_0 > 0$ be as defined in (5). Then,

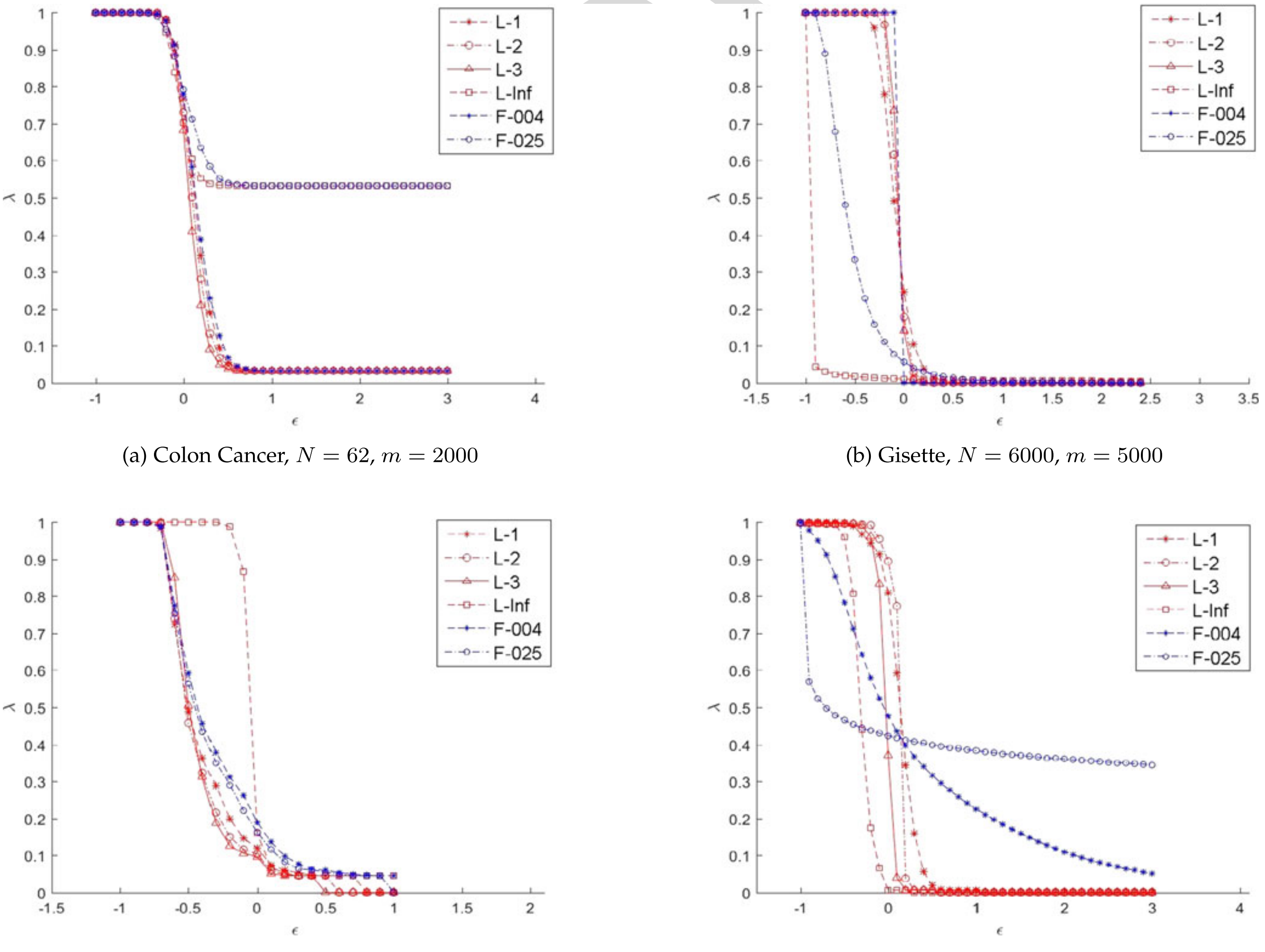
$$\begin{aligned}
&(1+\varepsilon_1)\delta_0 \leq (1+\varepsilon_2)\delta_0 \\
\Longrightarrow\ & \mathsf{N}(x_i, (1+\varepsilon_1)\delta_0) \subset \mathsf{N}(x_i, (1+\varepsilon_2)\delta_0) \qquad (\forall i) \\
\Longrightarrow\ & \mu^*_{\mathcal{X}}(\mathsf{N}(x_i, (1+\varepsilon_1)\delta_0)) \leq \mu^*_{\mathcal{X}}(\mathsf{N}(x_i, (1+\varepsilon_2)\delta_0))\ (\forall i) \\
\Longrightarrow\ & C(x_i, (1+\varepsilon_1)\delta_0) \leq C(x_i, (1+\varepsilon_2)\delta_0) \qquad (\forall i) \\
\Longrightarrow\ & 1 - C(x_i, (1+\varepsilon_1)\delta_0) \geq 1 - C(x_i, (1+\varepsilon_2)\delta_0)\ (\forall i) \\
\Longrightarrow\ & C^*(x_i, (1+\varepsilon_1)\delta_0) \geq C^*(x_i, (1+\varepsilon_2)\delta_0)\ , \qquad (\forall i)
\end{aligned}$$

from whence, we obtain $\lambda_{\mathcal{X}}(\varepsilon_1) \geq \lambda_{\mathcal{X}}(\varepsilon_2)$, since **avg** is a monotonic operation. □

4.2 Suitability of Distance Functions Based on λ_ρ

Once again, from Figs. 4a, 4b, and 4c, the rate of descent of λ_ρ does indicate the rate of concentration. The faster it falls, the more is the concentration. For instance, from the above plots, it does appear that Fractional distances ($\mathcal{F}_{.04}$) concentrate at a much slower rate than the other distance functions considered.

However, before comparing distance functions, perhaps an even more fundamental question that needs to be addressed is the following: *Given a workload, what is a suitable*



(a) Colon Cancer, $N = 62$, $m = 2000$

(b) Gisette, $N = 6000$, $m = 5000$

(c) Duke, $N = 44$, $m = 7129$

(d) Dexter, $N = 300$, $m = 20000$

Fig. 5. Plot of λ_{ρ_i} for some UCI data sets listed in Table 3.

TABLE 4
Suitability of Distance Functions on the Basis of λ_ρ—Synthetic Workloads

Dataset	m	N	Indices	$\mathcal{F}_{0.04}$	$\mathcal{F}_{0.25}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_\infty$
Gaussian	10	1,000	ε_ρ^+	– 0.7	– 0.8	– 0.7	– 0.8	– 0.8	– 0.7
			ε_ρ^-	2.9	1.9	1.2	1.	1.	1.3
			$[\varepsilon_\rho^+, \varepsilon_\rho^-]$	3.6	2.7	1.9	1.8	1.8	2.0
Gaussian	100	10,000	ε_ρ^+	– 0.4	– 0.4	– 0.4	– 0.3	– 0.3	– 0.4
			ε_ρ^-	0.5	0.4	0.3	0.2	0.2	0.7
			$[\varepsilon_\rho^+, \varepsilon_\rho^-]$	0.9	0.8	0.7	0.5	0.5	1.1
Uniform	1,000	10,000	ε_ρ^+	– 0.1	– 0.1	– 0.1	– 0.1	– 0.1	– 0.1
			ε_ρ^-	0.1	0.1	0.1	0.0	0.0	0.0
			$[\varepsilon_\rho^+, \varepsilon_\rho^-]$	0.2	0.2	0.2	0.1	0.1	0.1

distance function and how does λ_ρ enable us to identify it? It is clear that, in the context of this work, a distance function ρ is suitable if it does not concentrate *much*. In other words, a ρ is suitable if all pairwise distances do not converge around a single value and provide good contrast. Since λ_ρ measures this in terms of nearest neighbour distances, what is expected of a suitable distance function is that the NN distances of a large number of data points should be far less compared to that of the largest NN distance.[1] In the following, we attempt to place this intuitive idea in a more formal setting.

Definition 4.2. *Let $(\Omega, \mathcal{X}, \rho, \mu_{\mathcal{X}}^*)$ be a given similarity workload and λ_ρ be the corresponding dispersion function. We define $\varepsilon_\rho^+, \varepsilon_\rho^-$ as follows:*

$$\varepsilon_\rho^+ = \sup\{\varepsilon \in [-1, \infty) \mid \lambda_\rho(\varepsilon) = 1\} \ , \tag{7}$$

$$\varepsilon_\rho^- = \inf\{\varepsilon \in [-1, \infty) \mid \lambda_\rho(\varepsilon) = 0\} \ . \tag{8}$$

In other words, $\varepsilon_\rho^+, \varepsilon_\rho^-$ are the values at which the dispersion function λ_ρ begins and completes its descent. As we show below, both the point of decrease and the interval length play a role in classifying a distance function as suitable or not for a given workload.

Based on $\varepsilon_\rho^+, \varepsilon_\rho^-$ the following observations can be made:

- Clearly, $\varepsilon_\rho^+ \leq 0$. To see this, let $x_0 \in \mathcal{X}$ be the data point that has the maximum NN distance, i.e., its nearest neighbour is at a distance δ_0. Hence, for any $\varepsilon > 0$, we have $C(x_0, (1+\varepsilon)\delta_0) > 1$, since x_0 contains (at least) its nearest neighbour and hence $\lambda_\rho(\varepsilon) \neq 1$.
- If ρ is a suitable distance function, i.e., ρ does not concentrate much for the given workload, then clearly $\delta_0 \gg \delta_i$ for a *large* portion of the data set $\mathcal{X}$. This would imply that even for $\varepsilon \in [-1, 0)$, $(1+\varepsilon)\delta_0 > \delta_i$ for a large portion of the data set and hence $\varepsilon_\rho^+ \ll 0$.
- Let λ_ρ be a slowly decreasing function and hence the interval $[\varepsilon^+, \varepsilon^-]$ is large. This would mean that even for $\varepsilon \gg 0$ a large number of data points do not capture *much* of the rest of the data set in their $(1+\varepsilon)$-δ_0 neighbourhood. Or equivalently, that their $(1+\varepsilon)$-δ_0 count is much lesser than N indicating that ρ is still able to provide a good contrast for the considered workload and hence ρ is a suitable distance function.
- Let us consider a workload in high dimensions with a distance function that is known to concentrate. Then $\delta_0 \approx \delta_i$ for *almost* every data point x_i and hence

$$\begin{aligned} C(x_i, (1+\varepsilon)\delta_0) &= \frac{1}{N} \approx C(x_i, (1+\varepsilon)\delta_i), \text{ and} \\ C^*(x_i, (1+\varepsilon)\delta_0) &\approx C^*(x_i, (1+\varepsilon)\delta_i) = \frac{N-1}{N} \approx 1. \end{aligned} \tag{9}$$

  Thus, λ_ρ is almost a constant at 1 for $\varepsilon \in [-1, 0)$, i.e., $\varepsilon_\rho^+ \approx 0$. Further, even for small values of $\varepsilon > 0$, $(1+\varepsilon)\delta_0 > \delta_i$ for a large number of data points x_i and hence the rate of decrease of λ_ρ is very steep.

Thus, given a similarity workload, a suitable distance function ρ is such that either $\varepsilon_\rho^+ \ll 0$ or the interval $[\varepsilon_\rho^+, \varepsilon_\rho^-]$ is large, while if both $\varepsilon_\rho^+ \approx 0$ and the interval $[\varepsilon_\rho^+, \varepsilon_\rho^-]$ is small, it shows that the ρ concentrates for $\mathcal{W}$ and is not so suitable.

Note that both these phenomena are noticeable from Fig. 4, where at low dimensions λ_ρ does begin to decrease when $\varepsilon \in [-1, 0)$ but at a slower rate (see Figs. 4a and 4b), while at higher dimensions λ_ρ is almost a constant at 1 and dips steeply (see Fig. 4c). Further, from Table 4, the following can be observed:

(i) At low dimensions ($m = 10$), the ε_ρ^+ values are far lesser than zero, in fact, ε_ρ^+ values are closer to -1. Also the intervals $[\varepsilon_\rho^+, \varepsilon_\rho^-]$ are large, indicating that all the considered distance functions seem suitable.

(ii) At medium dimensions ($m = 100$), we see a gradual shift in the ε_ρ^+ values, which are now closer to zero than -1. We also see a shrinking in the lengths of the corresponding $[\varepsilon_\rho^+, \varepsilon_\rho^-]$ intervals.

(iii) At high dimensions ($m = 1{,}000$), the behaviour of λ_ρ is way different. For all the distance functions, both the $\varepsilon_\rho^+ \approx 0$ and the lengths of the corresponding $[\varepsilon_\rho^+, \varepsilon_\rho^-]$ intervals are small. Clearly, this indicates that all the six distance functions seem not so suitable for the considered workloads.

Based on Table 4 above, our analysis shows that Fractional distance functions ($\mathcal{F}_{.04}$) concentrate at a much slower rate than other Minkowski distance functions. This observation is in tune with what many studies have reported

[1] Note that, in the presence of outliers or in noisy data, δ_0 can dominate other δ_i and make λ_ρ less interpretable. However, in such scenarios, in some sense, one could contend that there is no concentration of distances. Further, taking δ_0 to be any other internal operation of the δ_i's, say for instance $\mathbf{avg}(\delta_i)$ did not seem to change the trend of the λ_ρ curves, especially in the context of comparing distances based on λ_ρ.

earlier, see, for instance, [3]: that smaller values of p in the Minkowski distance functions seem to concentrate less. In fact, as we show in Section 5.1, the values in Table 6 for some real workloads also seem to confirm their claim.

4.3 Relation between α_ρ and λ_ρ

From the above discussion, it does appear that α_ρ and λ_ρ exhibit some similarities. In fact, comparing λ_ρ and α_ρ the following commonalities can be observed:

(i) Both α_ρ and λ_ρ not only illustrate but also can measure the rate of concentration of distance functions.

(ii) Unlike ξ_ρ and γ_ρ, which indicate the presence of concentration as a function of or dependent on increase in dimensions, both α_ρ and λ_ρ make use of the *internal*, and hence fixed, characteristics of the workload to do the same.

(iii) Both α_ρ and λ_ρ are non-increasing functions and hence can offer insightful comparisons on the rate of concentration of different distance functions for an arbitrary but fixed workload under consideration.

However, α_ρ and λ_ρ are not without their share of differences as enumerated below:

(i) α_ρ is a purely theoretical index while λ_ρ is an empirical index.

(ii) Unless the underlying distribution is known, calculation of α_ρ is not possible, whereas λ_ρ can still be determined. Hence λ_ρ can be applied on real data sets to glean some useful information on the distances that could be considered when applying data analysis algorithms on them. This was already seen in Section 4.2. Also see Section 5.

(iii) In fact, the suitability of α_ρ to be employed as an empirical measure in practice is largely questionable. For instance, calculating α_ρ even for smaller data sets is extremely cumbersome. Recall that to find subsets with measure greater than $\frac{1}{2}$ requires $\sum_{k=\frac{N}{2}}^{N} C_k \approx 2^{N-1}$ computations. However, to evaluate λ_ρ one only needs to work with N *singleton* subsets. In fact, the computational complexity of determining α_ρ is of $\mathcal{O}(N2^{N-1})$, while that of λ_ρ is only of $\mathcal{O}(N^2 \log N)$. See the Appendix for more details. Thus determination of λ_ρ is computationally far less expensive than α_ρ.

(iv) However, an advantage that α_ρ enjoys that is not available to λ_ρ is the following: α_ρ—being a theoretical index—is amenable for theoretical analysis, for instance, to obtain lower and upper bounds for a given workload when the underlying measurable metric space (Ω, ρ, μ) is well defined. See [24] for some interesting existing results.

While the above discussion was based largely on empirical observations, the question that naturally arises is the following: Is there a relation between λ_ρ and α_ρ? The following result shows that, given a similarity workload where the distance ρ is known to concentrate, λ remains an upper bound of α.

Recall, from Section 4.2, that if a ρ concentrates, then both $\varepsilon_\rho^+ \approx 0$ and the interval $[\varepsilon^+, \varepsilon^-]$ is narrow. Further, since, $\delta_0 \approx \delta_i$ for *most* of the data points x_i, from (9)

$$\max_{x_i \in \mathcal{X}} C^*(x_i, (1+\varepsilon)\delta_0) \approx \mathop{\mathbf{avg}}_{x_i \in \mathcal{X}} C^*(x_i, (1+\varepsilon)\delta_0) \ . \tag{10}$$

Theorem 4.3. *Let $(\Omega, \mathcal{X}, \rho, \mu^*_{\mathcal{X}})$ be a given similarity workload, where $\mathcal{X} \subset \Omega$ is finite and $\mu^*_{\mathcal{X}}$ is the normalised counting measure and the distance function ρ is known to concentrate for this workload. Let $\varepsilon \in [-1, \infty)$ and δ_0 be as defined in (5). Let us denote by $r = (1+\varepsilon)\delta_0$. Then,*

(i) $\alpha_\rho(r) \leq \lambda_\rho\left(\frac{r}{\delta_0} - 1\right)$.

(ii) $\alpha_\rho(r) = 0$, *for any* $r \ > \ r^-$, *where* $r^- = (1+\varepsilon_\rho^-)\delta_0$.

Proof. First, note that r is a function of ε and hence as ε varies from $[-1, \infty)$, we have that r varies over $[0, \infty) = \overline{\mathbb{R}}_+$ and hence $\alpha_\rho(r)$ is well-defined.

(i): Let $\varepsilon \in [-1, \infty)$ be arbitrary but fixed and r be as defined above. Let $\mathcal{A}$ be the collection of all the subsets of $\mathcal{X}$ having measure greater than half, i.e.,

$$\mathcal{A} = \left\{ A \subset \mathcal{X} : \mu^*_{\mathcal{X}}(A) \geq \frac{1}{2} \right\} .$$

Let $A_r = \{x \in \mathcal{X} : \rho(x, a) \leq r \text{ for any } a \in A\}$, be the r-neighborhood of A for $r \geq 0$. Since $(1+\varepsilon)\delta_i \leq (1+\varepsilon)\delta_0$ for every $\varepsilon \in [-1, \infty)$ and $i = 1, 2, \ldots, n$, for any arbitrary but fixed $A \in \mathcal{A}$ and for every $x_i \in A$, we have

$$\begin{aligned}
&\mathsf{N}(x_i, (1+\varepsilon)\delta_i) \subset \mathsf{N}(x_i, (1+\varepsilon)\delta_0) \subset A_r \\
&\Longrightarrow \mu^*_{\mathcal{X}}(\mathsf{N}(x_i, (1+\varepsilon)\delta_i)) \leq \mu^*_{\mathcal{X}}(\mathsf{N}(x_i, (1+\varepsilon)\delta_0)) \leq \mu^*_{\mathcal{X}}(A_r) \\
&\Longrightarrow C(x_i, (1+\varepsilon)\delta_i) \leq C(x_i, (1+\varepsilon)\delta_0) \leq \mu^*_{\mathcal{X}}(A_r) \quad (\forall i) \\
&\Longrightarrow \mathcal{C}_A = \min_{x_i \in A} C(x_i, (1+\varepsilon)\delta_0) \leq \mu^*_{\mathcal{X}}(A_r) \ .
\end{aligned}$$

Now, since

$$\begin{aligned}
\inf_{A \in \mathcal{A}} \mathcal{C}_A &= \inf_{A \in \mathcal{A}} \left\{ \min\nolimits_{x_i \in A} C(x_i, (1+\varepsilon)\delta_0) \right\} \\
&= \min_{x_i \in \mathcal{X}} C(x_i, (1+\varepsilon)\delta_0) \ ,
\end{aligned}$$

and from (10), we have the following implications:

$$\begin{aligned}
&\min_{x_i \in \mathcal{X}} C(x_i, (1+\varepsilon)\delta_0) \approx \mathop{\mathbf{avg}}_{x_i \in \mathcal{X}} C(x_i, (1+\varepsilon)\delta_0) \leq \inf_{A \in \mathcal{A}} \ \mu^*_{\mathcal{X}}(A_r) \\
&\quad \Longrightarrow 1 - \left(\min_{x_i \in \mathcal{X}} C(x_i, (1+\varepsilon)\delta_0) \right) \geq 1 - \inf_{A \in \mathcal{A}} \{ \mu^*_{\mathcal{X}}(A_r) \} \\
&\quad \Longrightarrow \max_{x_i \in \mathcal{X}} (1 - C(x_i, (1+\varepsilon)\delta_0)) \geq \sup_{A \in \mathcal{A}} \{ \mu^*_{\mathcal{X}}(A_r^c) \} \\
&\quad \Longrightarrow \max_{x_i \in \mathcal{X}} \{ C^*(x_i, (1+\varepsilon)\delta_0) \} \geq \sup_{A \in \mathcal{A}} \{ \mu^*_{\mathcal{X}}(A_r^c) \} \\
&\quad \Longrightarrow \max_{x_i \in \mathcal{X}} \{ C^*(x_i, r) \} \geq \sup_{A \in \mathcal{A}} \{ \mu^*_{\mathcal{X}}(A_r^c) \} \\
&\Longrightarrow \lambda_\rho(\varepsilon) = \lambda_\rho\left(\frac{r}{\delta_0} - 1\right) \geq \alpha_\rho(r).
\end{aligned}$$

(ii): Follows from part (i). □

Remark 4.4.

(i) From Theorem 4.3, clearly λ_ρ forms an upper bound for α_ρ. Hence, for a given workload, if λ_ρ itself falls very steeply to 0, then α_ρ will fall faster, which indicates that the rate of concentration of the distance function under consideration will be very high.

(ii) It should also be noted that λ_ρ is a tighter *tail* bound for α_ρ and is largely a loose bound for smaller values of ε, i.e., as r, equivalently ε increases, λ_ρ

TABLE 5
Comparing Distance Functions on the Basis of τ_ρ—Synthetic Workloads

Dataset	m	N	$\mathcal{F}_{0.04}$	$\mathcal{F}_{0.25}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_\infty$
Gaussian	10	1,000	18.04	14.92	11.02	11.20	11.00	10.76
Gaussian	100	10,000	7.12	6.42	5.54	4.38	4.22	7.10
Uniform	1,000	10,000	2.54	2.89	2.02	1.99	1.97	2.00

bounds α_ρ tighter. In fact, if ρ is known to concentrate, then α_ρ falls steeply between $r^+ = (1+\varepsilon_\rho^+)\delta_0$ and $r^- = (1+\varepsilon_\rho^-)\delta_0$.

(iii) Also note that, if λ_ρ falls at a slower rate, i.e., the interval $[\varepsilon_\rho^+, \varepsilon_\rho^-]$ is large, no conclusions can be made on α_ρ and hence on whether the distance function under consideration concentrates or not.

5 Comparing Distance Functions Using λ_ρ

In Section 4.2, we discussed how to recognise the suitability of a distance function ρ for a given workload based on its dispersion function λ_ρ. However, let us be given a workload and two distance functions ρ_1, ρ_2 whose dispersion functions, viz., $\lambda_{\rho_1}, \lambda_{\rho_2}$, show similar trends, say for instance, $\varepsilon_{\rho_1}^+ \approx \varepsilon_{\rho_2}^+$ and/or the lengths of the intervals $[\varepsilon_{\rho_1}^+, \varepsilon_{\rho_1}^-]$ and $[\varepsilon_{\rho_2}^+, \varepsilon_{\rho_2}^-]$ could be the same. It is not yet clear how to compare them based on these parameters.

Further, one may not always be able to give an ordering (say, the usual point-wise ordering of functions) between λ_{ρ_1} and λ_{ρ_2}, see for instance, Figs. 5a and 5d. Earlier, from Figs. 4a, 4b, and 4c, based on the rate of descent of λ_ρ we surmised that it does appear that Fractional distances ($\mathcal{F}_{.04}$) concentrate at a much slower rate than the other distance functions considered, which also coincided with many earlier studies that reported that smaller values of p in the Minkowski distances seem to concentrate less. However, note that the above observation is purely based on visual illustrations. Also, we do not yet have a satisfactory result, *á la* if $0 < p < q < \infty$ then $\lambda_{\mathcal{L}_p} < \lambda_{\mathcal{L}_q}$, either empirically or theoretically. In addition, the rate of decrease of λ_ρ can be different over different intervals.

Nevertheless, our motivation is to propose an empirical index that would not only visually illustrate and measure the rate of concentration, but also allow us to compare between different distance functions on a given workload.

Towards this end, given a workload $(\Omega, \mathcal{X}, \rho, \mu^*_{\mathcal{X}})$, we propose another empirical index τ_ρ that assigns a real value to every distance function ρ based on λ_ρ, with the help of the parameters $\varepsilon_\rho^+, \varepsilon_\rho^-$ proposed in Definition 4.2.

Definition 5.1. *Given a similarity workload* $(\Omega, \mathcal{X}, \rho, \mu^*_{\mathcal{X}})$ *and the corresponding dispersion function* λ_ρ*, let us define the index* τ_ρ *as follows:*

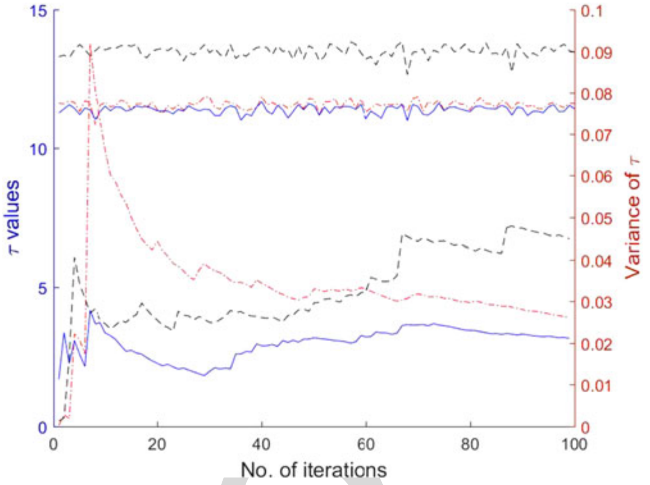


Fig. 6. τ_ρ values and its variance for the following distance functions, viz., $\mathcal{L}_1$(− · −), $\mathcal{L}_2$ (–) and $\mathcal{L}_\infty$ (−−).

$$\tau_\rho = \int_{\varepsilon_\rho^+}^{\varepsilon_\rho^-} \lambda_\rho(\varepsilon)\, d\varepsilon. \tag{11}$$

It is clear that τ_ρ calculates the area under λ_ρ over the interval $[\varepsilon_\rho^+, \varepsilon_\rho^-]$. In the discrete case, with uniform step size for ε, τ_ρ can be calculated as $\tau_\rho = \sum_{\varepsilon=\varepsilon_\rho^+}^{\varepsilon_\rho^-} \lambda_\rho(\varepsilon)$.

Table 5 tabulates the τ_{ρ_i} values for the synthetic workloads and distance functions ρ_i considered in Section 4.1. It is clear from the descending order of values of τ_{ρ_i} that the observation/claim made in earlier works that Fractional distances ($\mathcal{F}_{.04}$) concentrate at a much slower rate still seems to hold true.

In Fig. 6, we give the calculated τ_ρ values (at the i-the iteration) and its variance (upto the i-the iteration) for $\rho = \mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_\infty$, based on repeated sampling (100 realisations) from the same high-dimensional distribution $\mathcal{X} \sim \mathcal{U}(0,1)^m$ with $m = 100$ and $N = 10{,}000$. The relatively small and almost constant values of the variance of τ_ρ shows that concentration of distances is rather a stable phenomenon in high dimensions and that τ_ρ does present one way of determining it consistently.

While the index τ_ρ does induce an ordering on the considered distance functions, the question that arises now is this: Can it provide any more interesting or revealing information other than endorsing some general claims? What role does the ordering based on τ_ρ play in the algorithms applied on these real workloads? In the next section, we try to address these questions.

5.1 Ordering Based on Class Variable Accuracy

Aggarwal et al. [3], based on their study of the concentration of Minkowski distances $\mathcal{L}_p$, strongly advocated the use of

TABLE 6
Comparing Minkowski Distance Functions on the Basis of τ_ρ—Some UCI Data Sets Considered in [3]

Dataset	m	N	$\mathcal{F}_{0.10}$	$\mathcal{F}_{0.50}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_4$	$\mathcal{L}_{10}$	$\mathcal{L}_\infty$
Musk	167	476	85.79	41.18	34.38	25.59	19.03	15.13	15.85
Breast Cancer WDBC	30	569	11.62	10.58	6.54	6.54	6.37	6.23	6.20
Ionosphere	34	351	18.79	11.46	8.46	8.09	9.28	11.34	13.84
Segment	19	210	21.94	8.94	4.21	3.55	3.35	3.32	3.31

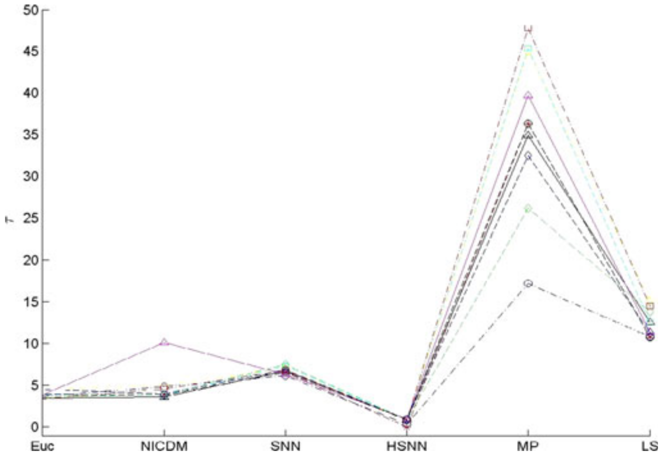


Fig. 7. Plot of τ_{ρ_i} for the 10 Synthetic Datasets considered in [35].

these distances where the parameter $p \in (0,1)$ instead of the usual $p \in [1,\infty)$.

Towards empirical validation of their claim, they considered the following UCI data sets, given in Table 6, i.e., the sets of workloads $(\mathcal{X}_i, \rho_j, \mu^*_{\mathcal{X}})$, where $\rho_j,\ j = 1, 2, \ldots, 7$ are the following Minkowski distance functions with $p = 0.1, 0.5, 1, 2, 4, 10, \infty$. They determine the *'Class Variable Accuracy'* β_{ij} for each of the above workloads as follows:

Step 1: Remove the class label information from the data.

Step 2: Fix an $x_k \in \mathcal{X}_i$ and search for ℓ nearest neighbours based on ρ_j.

Step 3: Among these ℓ neighbours count those that belong to the same class as x_k, say β^k_{ij}.

Step 4: Repeat *Step 3* for all $x_k \in \mathcal{X}_i$ and find $\beta_{ij} = \sum_k \beta^k_{ij}$.

Step 5: Repeat *Steps 2-4* for different ρ_j.

Step 6: Repeat *Steps 2-5* for different data sets $\mathcal{X}_i$.

Based on the decreasing order of β_{ij}, they ranked the above seven distance functions as follows:

$$\mathcal{F}_{0.1} \succ \mathcal{F}_{0.5} \succ \mathcal{L}_1 \succ \mathcal{L}_2 \succ \mathcal{L}_4 \succ \mathcal{L}_{10} \succ \mathcal{L}_{\infty}. \qquad (12)$$

In Table 6 we present the corresponding τ_ρ values for these work loads. Clearly from the τ_ρ values we observe that the ordering given in (12) is exactly the one obtained for descending values of τ_ρ.

5.2 Comparison of Secondary Distance Measures

So far, we have considered only the Minkowski norms. Recently there have been attempts to employ *Secondary Distance measures*, so called since these are themselves derived from other primary distance measures, typically the Minkowski metrics. These have been shown to work well, especially, in mitigating some aspects of DC in many Machine Learning problems like clustering and classification, see for instance, [13], [17], [33], [34].

We consider the following five secondary distance measures in this work, viz., the Shared Nearest Neighbour (SNN) [17], Local Scaling (LS) [38], Mutual Proximity (MP) [30], Non-iterative Contextual Dissimilarity Measure (NICDM) [22] and Hubness-aware SNN (HSNN) [35].

5.2.1 On Some Synthetic Data Sets

In this section, we consider the 10 synthetic datasets employed in [35]. These are high-dimensional Gaussian mixture data with high class overlap. Each of these data sets is 100-dimensional with 10-classes consisting of more than a thousand points each.[2] According to [35] these data sets exhibit substantial hubness and hence are very difficult for k-NN classification.

The τ_ρ values were calculated for each of the 10 data sets and all the above five secondary distance measures, with the primary distance being the Euclidean distance. The superimposed τ_ρ values for the above six distance measures is given in Fig. 7. Larger the τ_ρ value greater is the suitability of the distance function ρ. It is clear that our studies also validate the arguments of the authors in [35] that *secondary distances demonstrate great potential in correcting hubness-related problems*.

TABLE 7
Comparing Secondary Distance Functions on the Basis of τ_ρ—Real Workloads

Dataset	$\mathcal{L}_2$	NICDM	SNN	HSNN	MP	LS
Splice	6.69	2.42	9.67	0.48	**52.73**	15.46
Protein	5.43	3.08	6.69	0.0009	10.4	**20.85**
Cancer	5.77	20.51	34.75	18.05	**39.26**	25.17
Gisette	1.8	2.77	10.12	1.23	**84.117**	7.87
Duke	3.34	27.46	31.56	26.8	29.69	**43.97**
Dexter	11.3	50.1	36.23	0.47	90.37	**181.4**

5.2.2 On Real Data Sets

In this section, we consider the six UCI data sets listed in Table 3, which were chosen not only for their high dimensionality but also since a few of them, due to their intrinsic *hubness*, have been employed in many works, see for instance [14], and hence provides for meaningful comparative analysis. Further, they also provide a good mix of datasets, where for Datasets 1 & 2 we have $N \gg m$, while for Datasets 3, 5 & 6, $m \gg N$ and for Dataset 4, $m \approx N$.

The τ_ρ values for the above secondary distances on these data sets are presented in Table 7. The largest values are given in *bold*.[3]

In [14], the authors have made a comparative study of the suitability of three secondary distance functions, viz., MP, LS and SNN, on the Datasets 1, 2, 4 & 6 given in Table 7, w.r.t. their classification accuracy. It is worthy to note that the ranking of the distance functions based on τ_ρ and classification accuracy remain identical, except in the case of Dataset 1 where the ranks of MP and LS are interchanged.[4] We have used the same parameter values for all the secondary distance measures as employed in [14], i.e., $r = 10$ in SNN, $q = 10$ in LS and used $1 - \cdot$ to obtain the distances from the similarities.

6 Conclusion

In this work, we began by discussing the concentration of distances phenomenon. Our examination of the different

[2] Available at http://ailab.ijs.si/nenad_tomasev/datasets/

[3] It is interesting to note that it does appear that distances that can handle one aspect of the Dimensionality Curse, say the hubness phenomenon or the CoD, may not necessarily be good at handling other aspects.

[4] Note, however, that the classification accuracies of these two distances on Dataset 1 are almost the same at 77.2 & 77.9, respectively.

indices that either illustrate or measure this phenomenon revealed the need for an efficient empirical index that would not only illustrate but also measure the rate of concentration and enable comparison of distance functions with regards to their suitability in real workloads. With this as the motivation for this work, we have proposed a novel yardstick called the dispersion function λ_ρ which is an empirical measure.

Based on the dispersion function, we have also introduced an index τ_ρ that allows us to compare distances with regards to their suitability in real workloads, thus helping us to achieve our twin objectives. Further exploration of both λ_ρ and τ_ρ on some real workloads that have been used in existing studies seem to validate the usefulness of both the dispersion function λ_ρ and the index τ_ρ in judging the suitability of a distance function for a given workload.

So far the theory of concentration of norms has been well studied and explored but always in a non-positive way. From Sections 2.3.1 and 2.3.2, we see that Euclidean norms and other Minkowski-type distances do not behave well in high dimension. In fact, we have that all the Minkowski-type distances concentrate, but only at differing rates. The current work differs from existing studies in the following important ways:

(i) To the best of the authors' knowledge, this is the first empirical index to measure the rate of concentration of a distance function, while other empirical indices only illustrate the effect and that too only as a function of increasing dimensions.

(ii) This work not only tries to determine whether a distance function concentrates for a given workload, but also proposes an index to compare distances and suggest their suitability.

This work can, and should be, seen as yet another exploratory but a positive study on this phenomenon. Clearly, a more theoretical analysis of the dispersion function is in order, which we intend to take up in the near future.

Appendix

Complexity Considerations: α_ρ versus λ_ρ

In this section we study the time complexity of calculating α_ρ versus λ_ρ for a given data set $\mathcal{X}$ and a distance function ρ.

Let $\mathcal{K} = \{A \subseteq \mathcal{X} \mid \sharp A \geq \frac{N}{2}\}$. Let us fix an $\epsilon > 0$. For an $x \in \mathcal{X}$, let $\sharp([\mathsf{N}_\epsilon(x)]^C)$ denote the number of elements that are at a distance greater than ϵ from x w.r.t. ρ and, further, for an $A \in \mathcal{K}$, let A_ϵ^C denote the complement of the ϵ-dilation of A.

To calculate $\alpha_\rho(\epsilon) = \sup_{A\in\mathcal{K}}\{\sharp(A_\epsilon^C)\}$ we proceed as follows:

Step 1: There are $\sum_{k=\frac{N}{2}}^{N} {}^N C_k = 2^{N-1}$ elements in $\mathcal{K}$ and hence finding $\mathcal{K}$ is of the order 2^{N-1}.

Step 2: We calculate $\sharp([\mathsf{N}_\epsilon(x)]^C)$ as follows:

- Sort the pairwise distance matrix, which is of $N^2 \log N$ complexity.
- For the fixed $\epsilon > 0$, for each $x \in \mathcal{X}$—whose sorted pairwise distances to other data points is a row in the distance matrix—the complexity of determining $\sharp([\mathsf{N}_\epsilon(x)]^C)$ is $N \log N$.

Step 3: Let us fix an $A \in \mathcal{K}$. Now the cardinality of ϵ-dilation of A is given by $\sharp(A_\epsilon^C) = \sum_{x\in A} \sharp([\mathsf{N}_\epsilon(x)]^C)$. Hence the complexity of finding $\sharp(A_\epsilon^C)$ for each $A \in \mathcal{K}$ is $\sum_{k=\frac{N}{2}}^{N} {}^N C_k \cdot k \approx N \cdot 2^{N-1}$.

Step 4: Finally, the complexity of finding the $\sup_{A\in\mathcal{K}}\sharp(A_\epsilon^C)$ is of the order 2^{N-1}.

Thus, on the whole, the complexity of calculating $\alpha_\rho(\epsilon)$, for a given $\epsilon > 0$ is the sum of the above, viz., $2^{N-1} + N^2 \log N + N \log N + N2^{N-1} + 2^{N-1}$ and hence is of the order $\mathcal{O}(n \cdot 2^{N-1})$.

From Definition 3.2, we have $\lambda_\rho(\varepsilon) = \mathbf{avg}_{x_i\in\mathcal{X}}\{C^*(x_i, (1+\varepsilon)\delta_0)\} = \mathbf{avg}_{x_i\in\mathcal{X}}\{\mathsf{N}_{(1+\varepsilon)\delta_0}(x)\}$. Its time complexity can be calculated as follows:

Step 1: To find δ_0, we once again sort the pairwise distance matrix and find the maximum of the column 2 which is of order $N^2 \log N + N$.

Step 2: Finding $C^*(x, (1+\varepsilon)\delta_0)$ has the same complexity as that of finding $\mathsf{N}_{(1+\varepsilon)\delta_0}(x)$, which is $N \log N$.

Step 3: Now averaging over all $x \in \mathcal{X}$ is of linear order.

Thus the total time complexity of calculating $\lambda_\rho(\epsilon)$ is of the order $\mathcal{O}(N^2 \cdot \log N)$. Note that the costliest step in this algorithm is that of calculating δ_0. However, if one were to use an approximate nearest neighbour search method, then the complexity can be brought down considerably. For instance, if one employs the LSH, keeping the width parameter '*k*' and the time taken for evaluation '*t*' are kept fixed, its complexity is $O(N^{1+p})$, with $p \ll 1$. Thus the overall complexity of calculating lambda is $O(N^{1+p})$, which is near linear complexity.

Acknowledgments

This work was done when S. Kumari was at IIT Hyderabad, India. The authors would like to acknowledge Dr. Naveen Sivadasan and Dr. N. R. Aravind for their help in the study of complexity analysis.

References

[1] C. C. Aggarwal, "Re-designing distance functions and distance-based applications for high dimensional data," *ACM SIGMOD Rec.*, vol. 30, no. 1, pp. 13–18, 2001.

[2] C. C. Aggarwal, "On the analytical properties of high-dimensional randomization," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 7, pp. 1628–1642, Jul. 2013.

[3] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in *Proc. 8th Int. Conf. Database Theory*, 2001, pp. 420–434.

[4] J. Aucouturier and F. Pachet, "A scale-free distribution of false positives for a large class of audio similarity measures," *Pattern Recognit.*, vol. 41, no. 1, pp. 272–284, 2007.

[5] R. Bellmann, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ, USA: Princeton Univ. Press, 1961.

[6] K. P. Bennett, U. M. Fayyad, and D. Geiger, "Density-based indexing for approximate nearest-neighbor queries," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 1999, pp. 233–243.

[7] S. Berchtold, B. Ertl, D. A. Keim, H. Kriegel, and T. Seidl, "Fast nearest neighbor search in high-dimensional space," in *Proc. 14th Int. Conf. Data Eng.*, 1998, pp. 209–218.

[8] S. Berchtold, D. A. Keim, H. Kriegel, and T. Seidl, "Indexing the solution space: A new technique for nearest neighbor search in high-dimensional space," *IEEE Trans. Knowl. Data Eng.*, vol. 12, no. 1, pp. 45–57, Jan./Feb. 2000.

[9] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Proc. 7th Int. Conf. Database Theory*, 1999, pp. 217–235.

[10] E. Chávez, G. Navarro, R. A. Baeza-Yates, and J. L. Marroquín, "Searching in metric spaces," *ACM Comput. Surveys*, vol. 33, no. 3, pp. 273–321, 2001.

[11] F. H. Croom, *Principles of Topology*. Boston, MA, USA: Thomson Learn. Asia, 2002.

[12] R. J. Durrant and A. Kabán, "When is 'nearest neighbour' meaningful: A converse theorem and implications," *J. Complexity*, vol. 25, no. 4, pp. 385–397, 2009.

[13] A. Flexer and D. Schnitzer, "Can shared nearest neighbors reduce hubness in high-dimensional spaces?" in *Proc. 13th IEEE Int. Conf. Data Mining Workshops*, 2013, pp. 460–467.

[14] A. Flexer and D. Schnitzer, "Choosing l^P norms in high-dimensional spaces based on hub analysis," *Neurocomputing*, vol. 169, pp. 281–287, 2015.

[15] D. François, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 7, pp. 873–886, Jul. 2007.

[16] S. Har-Peled, P. Indyk, and R. Motwani, "Approximate nearest neighbor: Towards removing the curse of dimensionality," *Theory Comput.*, vol. 8, no. 1, pp. 321–350, 2012.

[17] M. E. Houle, H. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?" in *Proc. 22nd Int. Conf. Sci. Statistical Database Manage.*, 2010, pp. 482–500.

[18] C.-M. Hsu and M.-S. Chen, "On the design and applicability of distance functions in high-dimensional data space," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 4, pp. 523–536, Apr. 2009.

[19] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

[20] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. 30th Annu. ACM Symp. Theory Comput.*, 1998, pp. 604–613.

[21] B. Jayaram and F. Klawonn, "Can unbounded distance measures mitigate the curse of dimensionality?" *Int. J. Data Mining Modelling Manage.*, vol. 4, no. 4, pp. 361–383, 2012.

[22] H. Jegou, H. Harzallah, and C. Schmid, "A contextual dissimilarity measure for accurate and efficient image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[23] F. Korn, B. Pagel, and C. Faloutsos, "On the 'Dimensionality Curse' and the 'Self-Similarity Blessing'," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 1, pp. 96–111, Jan./Feb. 2001.

[24] M. Ledoux, *The Concentration of Measure Phenomenon*. Providence, RI, USA: Amer. Math. Soc., 2001.

[25] V. Pestov, "On the geometry of similarity search: Dimensionality curse and concentration of measure," *Inf. Process. Lett.*, vol. 73, no. 1/2, pp. 47–51, 2000.

[26] V. Pestov, "Intrinsic dimension of a dataset: What properties does one expect?" in *Proc. Int. Joint Conf. Neural Netw.*, 2007, pp. 2959–2964.

[27] V. Pestov, "Is the *k*-NN classifier in high dimensions affected by the curse of dimensionality?" *Comput. Math. with Appl.*, vol. 65, no. 10, pp. 1427–1437, 2013.

[28] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Nearest neighbors in high-dimensional data: The emergence and influence of hubs," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 865–872.

[29] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space: Popular nearest neighbors in high-dimensional data," *J. Mach. Learn. Res.*, vol. 11, pp. 2487–2531, 2010.

[30] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer, "Local and global scaling reduce hubs in space," *J. Mach. Learn. Res.*, vol. 13, pp. 2871–2902, 2012.

[31] D. W. Scott, *Multivariate Density Estimation*. Hoboken, NJ, USA: Wiley, 2015.

[32] U. Shaft and R. Ramakrishnan, "When is nearest neighbors indexable?" in *Proc. 10th Int. Conf. Database Theory*, 2005, pp. 158–172.

[33] N. Tomasev, "Taming the empirical hubness risk in many dimensions," in *Proc. SIAM Int. Conf. Data Mining*, 2015, pp. 891–899.

[34] N. Tomasev and K. Buza, "Hubness-aware kNN classification of high-dimensional data in presence of label noise," *Neurocomputing*, vol. 160, pp. 157–172, 2015.

[35] N. Tomasev and D. Mladenic, "Hubness-aware shared neighbor distances for high-dimensional k-nearest neighbor classification," *Knowl. Inf. Syst.*, vol. 39, no. 1, pp. 89–122, 2014.

[36] N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "The role of hubness in clustering high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 3, pp. 739–751, Mar. 2014.

[37] G. Schechtman and V. D. Milman, *Asymptotic Theory of Finite Dimensional Normed Spaces*, B. E. A. Dold, ed. Berlin, Germany: Springer, 1986.

[38] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Advances Neural Inf. Process. Syst. 17*, 2004, pp. 1601–1608.

[39] Y. Zhai, Y. Ong, and I. W. Tsang, "The emerging 'Big Dimensionality'," *IEEE Comput. Int. Mag.*, vol. 9, no. 3, pp. 14–26, Aug. 2014.

Sushma Kumari received the BSc degree in mathematics from Delhi University, India, in 2013, and the master's degree from the Department of Mathematics, Indian Institute of Technology Hyderabad, India. Currently, she is working toward the doctoral degree in the Department of Mathematics, Kyoto University, Japan. Her research interests include data analysis and randomized matrices.

Balasubramaniam Jayaram (S'02-A'03-M'04) received the MSc and PhD degrees in mathematics from the Sri Sathya Sai Institute of Higher Learning, India, in 1999 and 2004, respectively. He has been a visiting researcher with universities in Germany, Austria, Slovakia, and Czech Republic. He is currently an associate professor in the Department of Mathematics, Indian Institute of Technology Hyderabad, India. His current research interests include fuzzy aggregation operations, approximate reasoning, and issues in high dimensional data analysis. He has co-authored a research monograph on fuzzy implications and is the author or co-author of more than 60 published papers. He is an experienced research fellow of the Alexander von Humboldt Foundation. He is an associate editor of the *IEEE Transactions on Fuzzy Systems* (2014-) and *Fuzzy Sets and Systems* (2016-). He is a member of the EUSFLAT, the IEEE, and the IEEE Computational Intelligence Societies.