Distances in High Dimension

Sushma Kumari

A Thesis Submitted to
Indian Institute of Technology Hyderabad
In Partial Fulfillment of the Requirements for
The Degree of Master of Science

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Mathematics

May 2015

Declaration

I hereby declare that the submission entitled 'Distances in High Dimension' is submitted by me in the partial fulfillment of the requirement for the award of the degree of MASTER OF SCIENCE in MATHEMATICS. The study was conducted at IIT Hyderabad.

The matter embodied here represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. This project represent my original work and have not been presented earlier in this manner.

(Signature)

SUSHMA KUMARI
(Student Name)

MA13M1011
(Roll No.)

Approval Sheet

This Thesis entitled Distances in High Dimensions by Sushma Kumari is approved for the degree of Master of Science from IIT Hyderabad

(Dr. Balasubramaniam Jayaram)
Supervisor
Deptarment of Mathematics
IITH

Acknowledgement

This project would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

I am highly indebted to Dr. Balasubramaniam Jayaram for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project. His enthusiasm, encouragement and faith in me throughout have been extremely helpful.

I would like to express my special gratitude towards my beloved parents for their love, encouragement and support financially and mentally that made the completion of this project possible.

My thanks and appreciations also go to my colleague, seniors and juniors who has willingly helped me out with their abilities. I also thank Department of Mathematics, IITH for their support and encouragement.

Lastly, I would like to thank every single person that has been involved directly or indirectly in the completion of this project.

Contents

0.1 Curse of dimensionality . . . 3
0.2 Concentration of Norm(CoN): An Introduction . . . 4
0.2.1 What is CoN? . . . 4
0.2.2 An Empiricial Illustration . . . 4
0.2.3 Why is CoN important? . . . 6
0.3 Studies on the CoN Phenomenon . . . 8
0.3.1 Fixing the notation . . . 8
0.3.2 Existence of CoN: Theoretical Analysis . . . 9
0.4 Study of Concentration of Minkowski-type Norms : Some Indices . . . 11
0.4.1 *Minkowski Norms ($\mathcal{L}_p$ norms) :* . . . 11
0.4.2 Some Indices to Illustrate the CoN phenomenon: An Empirical Measure . . . 11
0.4.3 Some Indices to Illustrate the CoN phenomenon: A Theoretical Measure 14
0.4.4 An Index to Measure the CoN phenomenon . . . 16
0.4.5 New Distance measures to mitigate CoN . . . 18
0.5 Motivation for and the Objectives of our current work . . . 21
0.5.1 Studying the Newly Proposed Distances like $J_p, \mathcal{K}_p$ . . . 21
0.5.2 γ_p^m, ξ_p^m and α_Ω: Some Drawbacks . . . 21
0.6 Objectives of this study: . . . 22
0.7 Analysis of $\mathcal{J}_p$ and $\mathcal{K}_p$ . . . 22
0.7.1 $\mathcal{J}_p$, $\mathcal{K}_p$ and Nearest Neighbour Distances . . . 22
0.7.2 $\mathcal{J}_p$, $\mathcal{K}_p$ and the Relative Contrast . . . 22
0.7.3 $\mathcal{J}_p$, $\mathcal{K}_p$ and the Relative Variance . . . 24
0.8 Need for Efficient Emperical Indices . . . 26
0.8.1 Advantages and Drawbacks of α_Ω . . . 26
0.9 Stability of Distance Functions . . . 27
0.9.1 Stability of a Query . . . 27
0.9.2 g-δ Stability Analysis . . . 27
0.9.3 g-Compactness of a Dataset . . . 28
0.9.4 Empirical results of η_g and $\beta_\mathcal{X}$ . . . 29
0.10 Some Novel Empirical Indices to Measure Concentration . . . 33
0.10.1 $C_g^*(x_i)$- Complement of the new index . . . 35
0.10.2 Nomenclature . . . 35
0.10.3 A New General Purpose Index : λ . . . 36
0.10.4 Two Specific Measures based on λ : $\widetilde{\lambda}_\mathcal{X}$ and $\widehat{\lambda}_\mathcal{X}$ . . . 37
0.10.5 $\alpha_\mathcal{X}$ vs $\widetilde{\lambda}_\mathcal{X}, \widehat{\lambda}_\mathcal{X}$ . . . 38
0.10.6 Are these indices really useful? . . . 39
0.11 Conclusion . . . 43

0.1 Curse of dimensionality

Nowadays data are getting more and more complex adding more features or dimension to the data. Evolution of new data types such as images, videos and audio force us to work with data in high dimension, thus leading us to deal with the so called *Curse of Dimensionality*, a term that has come to refer to the unnatural things happening in high dimension.

Breaking the term Curse of Dimensionality into two components, Dimensionality refers to the dimension of the data set and curse refers to the difficulty that arises when dimension increases. It is used to refer to the counter-intuitive challenges faced in high dimension. Working with data become more difficult with increasing dimensions, since we are not able to visualize the high dimensional data points as we no longer have the aid of paper and pencil. This lack of visualisability is only one aspect of the CoD.

There are many aspects of CoD and their effects are still not well explored and huge amount of research is still going on. Some of the well-known aspects of CoD are

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

(ii) *Need for greed* - which refers to the need for atleast 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, [Pestov(2013)], for more details,

(iii) *Relationship Among Dimensions*, which refer to the intrinsic and embedding dimensionalities of the data and their influence on the algorithms,

(iv) *Relevance of Dimensions*, which again refers to the presence of irrelevant features that interfere with the performance of similarity queries.

(v) *Hubness Phenomenon* [Radovanovic et al.(2010)Radovanovic, Nanopoulos, and Ivanovic], which refers to the formation of hubs i.e. points which more popular as nearest neighbors than other data points.

One major aspect of Curse of Dimensionality that has recently come to the fore is the *Concentration of Norms* phenomenon, which will form the main focus of this Master's thesis.

0.2 Concentration of Norm(CoN): An Introduction

0.2.1 What is CoN?

Concentration of Norms(CoN) refers to the inability of distances in high dimensions to distinguish points well. To measure the closeness between any two objects/points we need some distance function, but as the dimension increases all the points appear to be approximately at the same distance, hence the distance function loses its discriminative power. This phenomenon is called Concentration of distances.

Let $X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_n\} \subset \mathbb{R}^m$ be a set of n data points from the m-dimensional Euclidean space. Let $\bar{q} \in \mathbb{R}^m$ be a query point and consider a distance function ρ to calculate the distances between points in X - for instance, ρ could be the Euclidean distance. Let $\bar{x}^-$ and $\bar{x}^+$ be the nearest and farthest point to $\bar{q}$, i.e.,

$$\bar{x}^- = \arg\min_{\bar{x}_i \in X} \rho(\bar{x}_i, \bar{q}) \ ,$$
$$\bar{x}^+ = \arg\max_{\bar{x}_i \in X} \rho(\bar{x}_i, \bar{q}) \ .$$

As a consequence of concentration of distances, as the dimension $m \to \infty$, one finds that $\rho(\bar{q}, x^-) \approx \rho(\bar{q}, x^+)$, which means that the distance of a query to the farthest point approaches the distance of the query to its nearest point.

Since $\rho(\bar{q}, \bar{x}^-) \leq \rho(\bar{q}, \bar{x}_i) \leq \rho(\bar{q}, \bar{x}^+)$ for $1 \leq i \leq n$, all distances to $\bar{q}$ are concentrating and confined to a small domain. In other words, we can say that all the points in X are *almost* at the same distance to $\bar{q}$. Thus the distances become less precise as the dimension grows because the distance between any two points converges.

0.2.2 An Empiricial Illustration

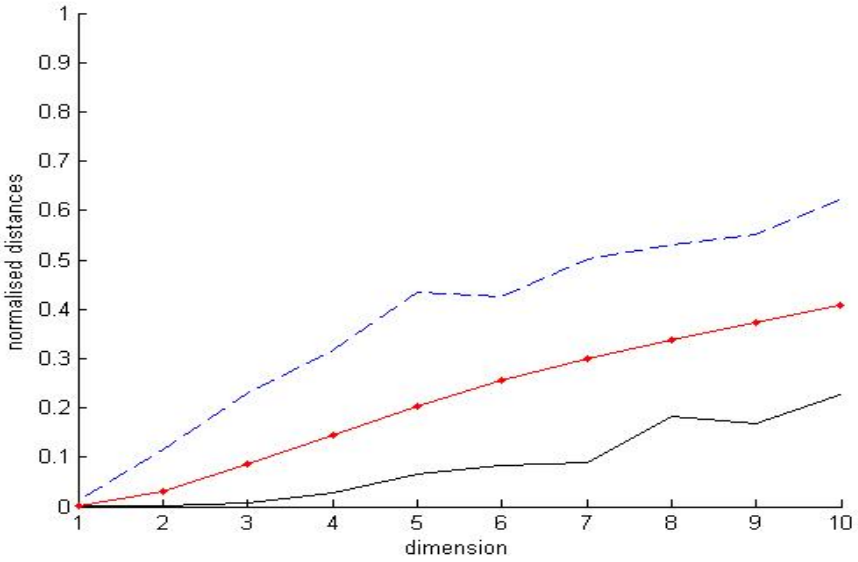
Since CoN is a counter-intuitive phenomenon, to have a practical insight into CoN, we will explain some results and graphically show the CoN effect for the Euclidean distance function. The following experiment will help in understanding the CoN phenomenon in high dimension.

What we want to do? On an average, we want to compare the Nearest Neighbour(NN) distance of a query point with the average of other pairwise distances.

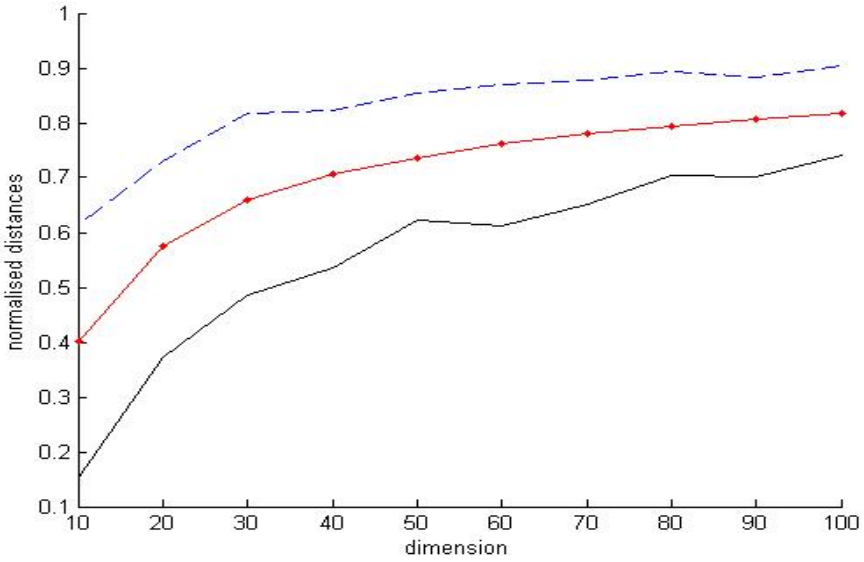
How do we do it? Let us generate N data points say $X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_N\}$ uniformly from an m-dimensional bounded domain, say $X \subsetneq [-1, 1]^m$. Now we calculate four parameters denoted as follows:

- α_i denotes the Nearest Neighbor (NN) distance of x_i for each $i = 1$ to N.
- $Y_{\max} = \max\{\alpha_i : 1 \leq i \leq N\}$, denotes the maximum of the NN distances.
- $Y_{\min} = \min\{\alpha_i : 1 \leq i \leq N\}$, denotes the minimum of the NN distances.
- $Y_{\text{avg}} = \dfrac{1}{N} \displaystyle\sum_{1 \leq i \leq N} \alpha_i$, denotes the average of the NN distances.
- $Y_X = \dfrac{2}{N(N-1)} \displaystyle\sum_{1 \leq i < j \leq N} \rho(\bar{x}_i, \bar{x}_j)$, denotes the average of all pairwise distances in X.

Let us now consider the following indices, which are denoted and defined as follows:



(a) The indices $k_M(--), k_m(-), k_A(-\cdot-)$ for the Euclidean norm in dimensions $m = 1, \ldots, 10$



(b) The indices $k_M(--), k_m(-), k_A(-\cdot-)$ for the Euclidean norm in dimensions $m = 10, \ldots, 100$

Figure 1: Concentration phenomenon exhibited by Euclidean norm when moving from low to high dimensions

- $k_M = \dfrac{Y_{\max}}{Y_X}$ denotes the normalised maximum NN distance w.r.to the average of all pairwise distances.
- $k_m = \dfrac{Y_{\min}}{Y_X}$ denotes the normalised minimum NN distance w.r.to the average of all pairwise distances.
- $k_A = \dfrac{Y_{\text{avg}}}{Y_X}$ denotes the normalised average NN distance w.r.to the average of all pairwise distances.

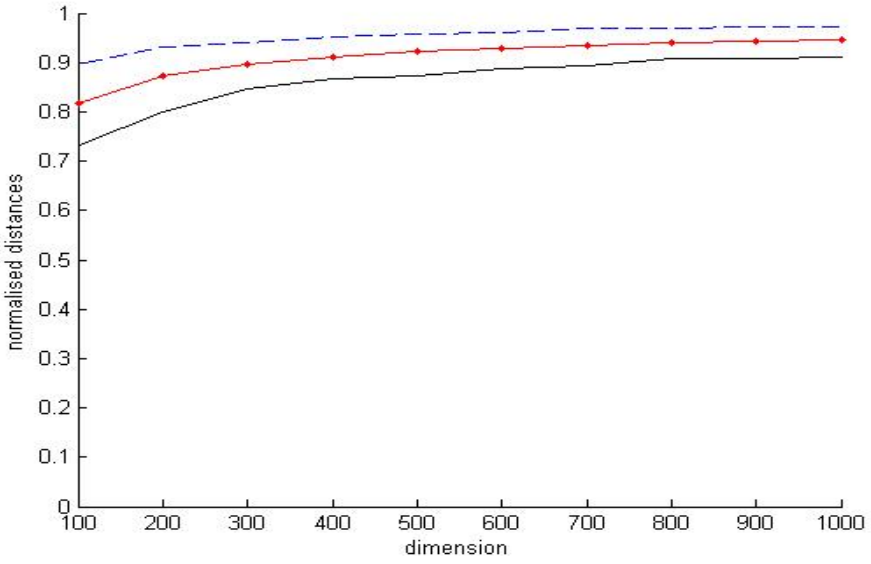


Figure 2: Concentration phenomenon clearly exhibited by Euclidean norm in high dimensions, where $m = 100, \ldots, 1000$

In Figures 1 and 2 we plot the above three indices for $N = 1000$ data points generated for varying dimensions, $m = 1, \ldots, 1000$. 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., we can say that enough contrast is present between the points and hence points are well separated (see Fig. 1(a)).
- 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 the CoN phenomenon (see Fig. 1(b)).
- However, as dimension increases, $k_m \to 1$, $k_M \sim k_A$ and $k_m \sim k_A$, i.e., average maximum NN distances and average minimum NN distances both converge to the average NN distances. There is not much dispersion or contrast present between the distances, i.e., all the distances are concentrating around the average of the distances. Hence all points become *almost* equidistant to each other (see Fig. 2).

0.2.3 Why is CoN important?

In this section let us look into some applications where CoN plays an affecting role. It turns out that in many applications, the distance functions which are useful in low dimensions are no longer relevant in high dimensions. There are many domains where data are high dimensional and CoN poses a serious threat to their applicability to real life.

CoN and Similarity Searches

One of the main areas affected by the CoN phenomenon is the searching algorithms in computer science. The goal in these type of applications is either

- to find objects whose features are similar to the query object, or

- to find objects whose feature values lie in a particular range of values nearer that of the query object.

The basic aim of search is to find an object or a set of objects similar to the given query object. Searching is the most fundamental task used in every stream.

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 thousand 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.

Is NN query meaningful?

Clearly, CoN has a serious effect on similarity searching in high dimensions. Consider yet another appliication domain that Graphical Information Systems (GIS), where we need to find the nearest city closest to one's location. It is same the as asking for the nearest neighbor to a query.

Nearest neighbor searching can already be quite computationally inefficient in high dimensions. However, it is made even more difficult by CoN. In fact, CoN raises the issue of whether or not the nearest neighbor is meaningful in high dimension!!

0.3 Studies on the CoN Phenomenon

Effective solution to a problem requires a deep and thorough understanding of the problem. The research studies done on CoN, so far, can be broadly classified into the following three types:

(i) Studies that have theoretically proven the existence of CoN,

(ii) Studies that have proposed indices or functions to illustrate or measure the CoN in specific settings,

(iii) Studies that attempt to proposing new distance functions to defeat / mitigate the CoN phenomenon.

0.3.1 Fixing the notation

Before, we begin to review the current literature on CoN, we first establish certain notations and definitions which will be required in the sequel.

- 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 Ω.

- Further, with the measure μ, we associate a distribution $\mathcal{R}$ which will be used to obtain a finite sample of n-points $X = \{x_1, x_2, \ldots, x_n\} \subset \Omega$. We will then write both $X \sim \mathcal{R}$ and $x \sim \mathcal{R}$, interchangeably, to denote that the data set $X \subset \Omega$ has been generated using the distribution $\mathcal{R}$. Often the quadruple (Ω, X, ρ, μ) is termed as a *Similarity Workload*.

- Let $\mathcal{I} \subset \mathbb{N}$ be a, possibly countably infinite, index set. If $\{(\Omega_i, \rho_i, \mu_i)\}$ is a sequence of measurable metric spaces for $i \in \mathcal{I}$, then for a finite fixed $m \in \mathbb{N}$, one can obtain an m-dimensional measurable metric space $(\Omega^m, \rho^m, \mu^m)$ as follows:

  $\Omega^m = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_m$, the Cartesian product of the domains Ω_i, ρ^m and μ^m are the product metric and product measure on Ω^m.

- Similarly, if $\mathcal{R}_i$ are the distributions associated with μ_i for $i \in \mathcal{I}$, then $X_i \sim \mathcal{R}_i$ and $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$.

- Further, one can obtain the m-dimensional data set $X^m = \{\bar{x}_1^m, \bar{x}_2^m, \ldots, \bar{x}_n^m\}$, where each $\bar{x}_j^m,\ j = 1, 2, \ldots, n$ is an m-dimensional vector such that $\bar{x}_j^m = (x_{1j}, x_{2j}, \ldots, x_{mj})$. Thus we would also write $X^m \sim \mathcal{R}^m$ or $\bar{x}^m \sim \mathcal{R}^m$.

  For example, let $m = 3$ and $\Omega^3 = \Omega_1 \times \Omega_2 \times \Omega_3$ such that $\Omega_1 = [0, 1], \Omega_2 = [-1, 1], \Omega_3 = \mathbb{R}$. Similarly $X^3 = \{\bar{x}_1^3, \bar{x}_2^3, \ldots, \bar{x}_{10}^3\}$, is a finite sample of $n = 10$ points, where $\bar{x}_i^3 = (\bar{x}_{i1}, \bar{x}_{i2}, \bar{x}_{i3})$, and the first component $\bar{x}_{i1}$ of each of the points is distributed as $\bar{x}_{i1} \sim \mathcal{U}[0, 1]$, and similarly, $\bar{x}_{i2} \sim \mathcal{U}[-1, 1], \bar{x}_{i3} \sim \mathcal{N}(0, 1)$ for every $i = 1, \ldots, n$.

- When no confusion is possible, and when the context makes it clear, we use the simpler notation Ω instead of the cumbersome Ω^m to still denote the domain of dimension m. Accordingly, ρ, μ are the metric and measure on the corresponding spaces. In fact, we will employ the vector representation for the elements of Ω, since from now on we will implicitly assume that Ω is a multi-dimensional space as explained above.

- We assume that there always exist a $\bar{0} \in \Omega$ designated as the origin of the domain Ω.

- 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 $\bar{x} \in \Omega$ to the origin $\bar{0} \in \Omega$, i.e., $\|\bar{x}\| = \rho(\bar{x}, \bar{0})$.

- The parameter $0 < p < \infty$ is an arbitrary constant and plays the role of an exponent in the considered distance functions.

- By $D^m_{\max}$ we denote the maximum of the norms in a given data set X^m, i.e., the distance of the farthest point in X^m to the origin w.r.to the metric ρ.

$$D^m_{\max} = \max\{\|\bar{x}^m_i\| = \rho(\bar{x}^m_i, \bar{0}) \ : 1 \leq i \leq n, \ \bar{x}^m_i \in X^m\} \ .$$

- Similarly, by $D^m_{\min}$ we denote the minimum of the norms in a given data set X^m, i.e., the distance of the farthest point in X^m to the origin w.r.to the metric ρ.

$$D^m_{\min} = \min\{\|\bar{x}^m_i\| = \rho(\bar{x}^m_i, \bar{0}) \ : 1 \leq i \leq n, \ \bar{x}^m_i \in X^m\} \ .$$

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

Definition 0.3.1 (Convergence in Probability)**.** *A sequence of random variables $\{A_n\}$ converges in probability to random variable A, if for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} P[|A_n - A| \leq \varepsilon] = 1 \ .$$

It is denoted as $A_n \xrightarrow{P} A$.

0.3.2 Existence of CoN: Theoretical Analysis

Distance functions are known to be sensitive to the dimension of data and hence reduces the efficiency of the search. While searching for the nearest neighbour the obvious approach is to search the database and compute the distance of every data to our query data and then to compare the distances. Not only is this naive approach computationally expensive with very large databases, the CoN phenomenon now adds another level of discomfort, since almost all points become equidistant to the query point, i.e., almost all points appear to be the nearest neighbours to the query data, thus questioning the very existence of meaningful nearest neighbours in high dimension.

Beyer *et. al.* were the first to point out that nearest neighbor searching may not always be meaningful when the ratio of the variance of the distance between any two random points, drawn from the data and query distributions, and the expected distance between them converges to zero as dimension goes to infinity by proving the following result.

Theorem 0.3.2 (**Beyer et. al.,** [Beyer et al.(1999)Beyer, Goldstein, Ramakrishnan, and Shaft])**.** *Let $(\Omega^m, \rho^m, \mu^m)$ be an m-dimensional measurable metric space, let $X^m = \{\bar{x}^m_1, \bar{x}^m_2, \ldots, \bar{x}^m_n\}$ be a finite sample of n points such that $\bar{x}^m \sim \mathcal{R}^m$ and $D^m_{\max}, D^m_{\min}$ are as explained above. Further, let $E[||\bar{x}^m||^p]$ and $var[||\bar{x}^m)||^p]$ be finite and $E[||\bar{x}^m||^p] \neq 0$. If*

$$\lim_{m\to\infty} \ var\left(\frac{\|\bar{x}^m\|^p}{E\|\bar{x}^m\|^p}\right) = 0 \ , \tag{1}$$

then for all $\varepsilon > 0$,

$$\lim_{m\to\infty} P[D^m_{\max} \leq (1+\varepsilon) D^m_{\min}] = 1 \ . \tag{2}$$

Thus, this result shows that under some pre-conditions on the data distribution and distance function the difference between the maximum and minimum distances become very small compared to the minimum distance in high dimension. This means all points are almost equidistant to the query point. Thus all the dimensionality issues can be traced back to the lack of contrast between the points.

Theorem 0.3.2 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 0.3.2 was proved by Durrant and Kabán in 2009.

Theorem 0.3.3 (**Durrant and Kabán**, [Durrant and Kabán(2009)]). *Let* $(\Omega^m, \rho^m, \mu^m)$ *be an* m*-dimensional measurable metric space, let* $X^m = \{\bar{x}_1^m, \bar{x}_2^m, \ldots, \bar{x}_n^m\}$ *be a finite sample of* n *points such that* $\bar{x}^m \sim \mathcal{R}^m$ *and* $D^m_{\max}, D^m_{\min}$ *are as explained above. Let the number of points* n *be large enough such that*

$$E[||\bar{x}^m||^p] \in \ [(D^m_{\min})^p, (D^m_{\max})^p \ ] \ .$$

If for any $\varepsilon > 0$,

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

then

$$\lim_{m\to\infty} \ var\left(\frac{\|\bar{x}^m\|^p}{E\|\bar{x}^m\|^p}\right) = 0 \ .$$

This result, in a sense, tries to answer the question when is nearest neighbour meaningful in high dimensions.

0.4 Study of Concentration of Minkowski-type Norms : Some Indices

Theorem 0.3.2 and Theorem 0.3.3 provided a necessary and sufficient condition on a general distance function to suffer from concentration in high dimensions. Thus, subsequently, researchers began investigating some indices, which were derived out of these results, for different types of distance functions. The most common among them being the classical Euclidean metric and its generalisations.

0.4.1 *Minkowski Norms ($\mathcal{L}_p$ norms) :*

Minkowski Norms are the family of p-norms parametrized by exponent $p \in (0, \infty)$ which are defined as follows, for an $\bar{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$:

$$\|\bar{x}\|_p = \left(\sum |x_i|^p\right)^{\frac{1}{p}}$$

.

- For $p = 1$, it is called the Manhattan norm and is denoted as $\mathcal{L}_1$ norm.
- For $p = 2$, it corresponds to the Euclidean norm and is denoted as $\mathcal{L}_2$ norm.
- In the limiting case, as $p \to \infty$, it becomes the $\mathcal{L}_\infty$-norm or the sup-norm or the Chebyshev metric.
- For $0 < p < 1$, triangle inequality does not hold for $\mathcal{L}_p$. Hence they are not norms but are called prenorms. An $\mathcal{L}_p$-norm, with $0 < p < 1$, is called a Fractional norm and is denoted by $\mathcal{F}_p$.

0.4.2 Some Indices to Illustrate the CoN phenomenon: An Empirical Measure

Theorem 0.3.2 led to researchers proposing two indices to illustrate the presence of concentration. The first of them is given in the following definition.

Definition 0.4.1. *Let us consider a similarity workload, (Ω, X, ρ, μ). The Relative Contrast with exponent p is defined as*

$$\xi_p^m = \frac{D_{\max}^m - D_{\min}^m}{D_{\min}^m} \ ,$$

where $D_{\max}^m$ and $D_{\min}^m$ are as defined earlier.

While Beyer *et al.* studied the CoN phenomenon for arbitrary norms, the first result for concentration of norms was studied for the Euclidean norms by Demartines in his doctoral thesis, who presented the following imporant theorem.

Theorem 0.4.2 (**Demartines, 1994**, [Demartines(1994)]). *let $X \subsetneq \mathbb{R}^m$ be an m-dimensional data set, where each dimension is distributed in an* i.i.d. *fashion, i.e., each $X_i \sim \mathcal{R}$ and ρ is the Euclidean $\mathcal{L}_2$ norm. Then,*

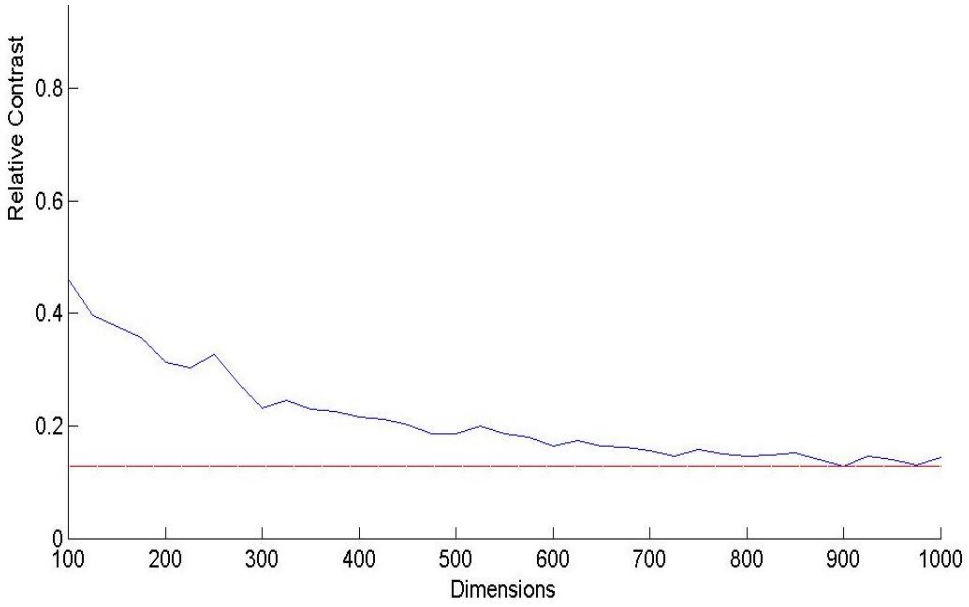
$$E(\rho(\bar{x}, \bar{0})) = E\left(\|\bar{x}\|\right) = \sqrt{am - b} + O\left(\frac{1}{m}\right),$$

$$Var(\rho(\bar{x}, \bar{0})) = Var\left(\|\bar{x}\|\right) = b + O\left(\frac{1}{\sqrt{m}}\right) \ ,$$

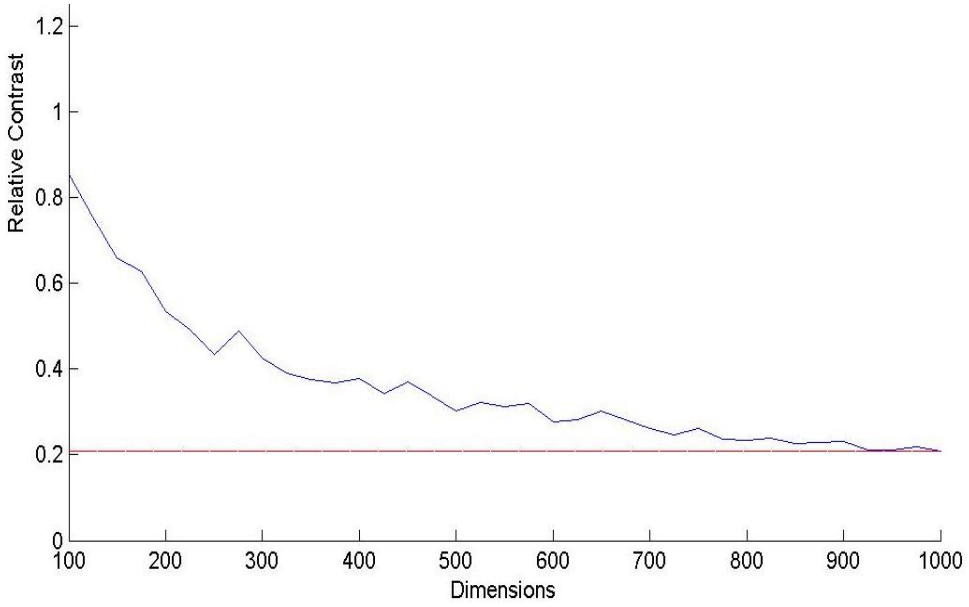
where a and b are some constants independent of the dimension m.

This theorem shows that expectation of the distances to the origin increases as dimension increases, but the variance remains a constant. Thus, when the dimension is very large, the variance will still be small as compared to the expected distance, hence the points will be closely packed.

The above can be seen from Fig. 3, where we plot the relative contrast ξ_2^m for the Euclidean distance metric $\mathcal{L}_2$. It is clear that no matter what distribution data follow, either uniform (as in Fig. 3(a)) or Gaussian (as in Fig. 3(b)), the Euclidean distances quickly concentrate in high dimensions.



(a) Relative contrast for the Euclidean norm for data obtained from Uniform distribution



(b) Relative contrast for the Euclidean norm for data obtained from Gaussian distribution

Figure 3: Relative contrast for the Euclidean norm where data are generated from Uniform and Gaussian distributions, respectively, showing the degrading separation between points with increase in dimensions.

The result of Demartines was generalised to any $\mathcal{L}_p$ norm by Hinneburg *et al.*.

Theorem 0.4.3 (**Hinneburg *et. al.***, [Hinneburg et al.(2000)Hinneburg, Aggarwal, and Keim]). *Let $X = \{\bar{x}_1^m, \bar{x}_2^m, ..., \bar{x}_n^m\}$ be n m-dimensional i.i.d. random vectors, ρ be any of the Minkowski norms $\mathcal{L}_p$ with exponent p. Then there exists a constant C_p, independent of the underlying distribution $\mathcal{R}$ of $\bar{x}_i^m$, such that*

$$C_p \leq \lim_{m\to\infty} E\left(\frac{D_{\max}^m - D_{\min}^m}{m^{\frac{1}{p}-\frac{1}{2}}}\right) \leq (n-1)C_p \ . \tag{3}$$

Theorem 0.4.3 says that the ratio of contrast to $m^{\frac{1}{p}-\frac{1}{2}}$ is bounded by C_p that depends on the exponent p. Based on (3) Hinneburg *et al.* have made the following observations on the exponent p:

- For $\mathcal{L}_p$ norm ($p \geq 3$), the relative contrast rapidly goes to 0 as m increases. It means that the distance function has lost its discriminative power for $p \geq 3$ in high dimensions.
- For the Euclidean $\mathcal{L}_2$ norm ($p = 2$), contrast remains constant.
- For the Manhattan $\mathcal{L}_1$ norm ($p = 1$), contrast increases as $\sqrt{m}$ increases.
- This tends to imply that the $\mathcal{L}_1$ norm is more preferable than the $\mathcal{L}_2$ norm for high dimensional data as it provides a better contrast than $\mathcal{L}_2$ norm.

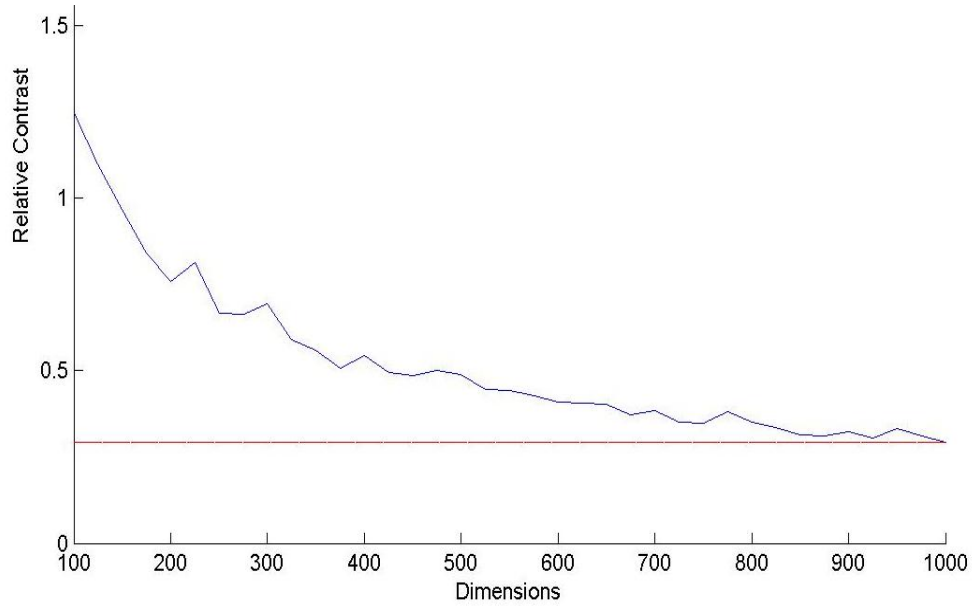
This result motivated some researchers to consider the Minkowski norms where the exponent $p \in (0, 1)$, i.e., the Fractional norms $\mathcal{F}_p$. Aggarwal *et al.* further extended Theorem 0.4.3 to study the concentration of Fractional norms.

Theorem 0.4.4 (**Aggarwal *et al.***, [Aggarwal et al.(2001)Aggarwal, Hinneburg, and Keim]). *$X = \{\bar{x}_1^m, \bar{x}_2^m, ..., \bar{x}_n^m\}$ be n m-dimensional i.i.d. random vectors uniformly distributed over $[0, 1]^m$. Then there exists a constant C, independent of p and m, such that*

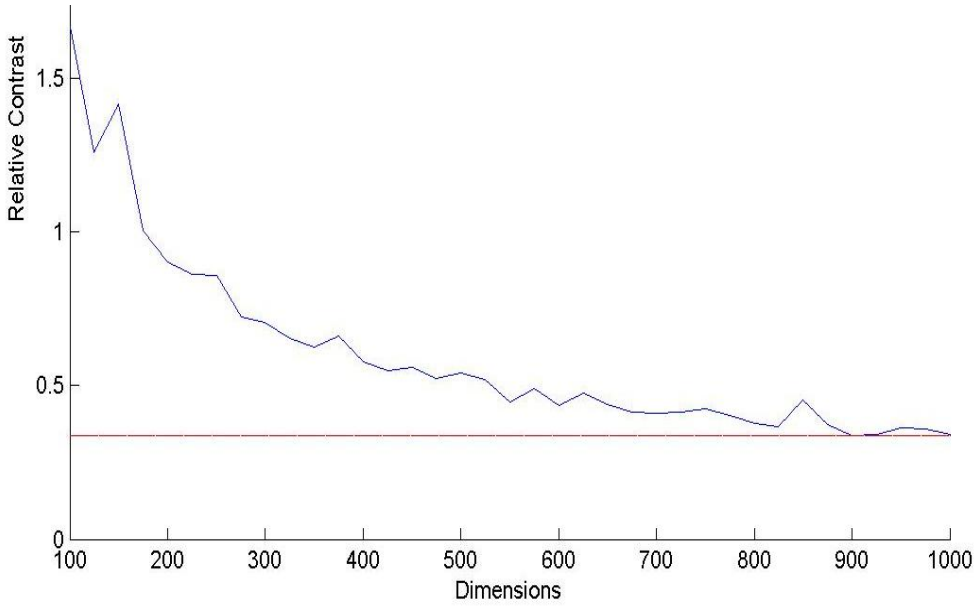
$$C\sqrt{\frac{1}{2p+1}} \leq \lim_{m\to\infty} E\left(\frac{D_{\max}^m - D_{\min}^m}{D_{\min}^m}\right).\sqrt{m} \leq (n-1).C\sqrt{\frac{1}{2p+1}} \ . \tag{4}$$

From (3), it is clear that the constant C may be independent of p but the bounds for relative contrast depend largely on $\sqrt{\frac{1}{2p+1}}$. Hence, they concluded that on an average fractional norms provide better contrast then Minkowski norms. Fig. 4 does seem to confirm the suspicions of Aggarwal *et al.* The relative contrast for the $\mathcal{F}_{0.04}$ norm shown in Figs. 4(a) and (b), compared to the relative contrast for $\mathcal{L}_2$ norm in Figs.3(a) and (b) is better. This indicates that fractional norms can provide much wider separation between points than the Euclidean norm.

Note, however, that as m increases the bounds on either side of the relative contrast tend to zero and hence $\mathcal{F}_p$ norms will also concentrate with the increasing dimensionality of the data space. This can be seen from Fig. 4(a). Note that, while **Theorem 0.4.4** is proven only for uniformly distributed data, one finds that even when the data are not uniformly distributed, the conclusions of the above result still seem to be true, see, for instance, Fig. 4(b).



(a) Relative contrast for Fractional norm from Uniform distribution



(b) Relative contrast for Fractional norm from Gaussian distribution

Figure 4: Relative contrast for $\mathcal{F}_{0.04}$ norm where data are generated from Uniform and Gaussian distributions, respectively, showing the degrading separation between points with increasing dimensions.

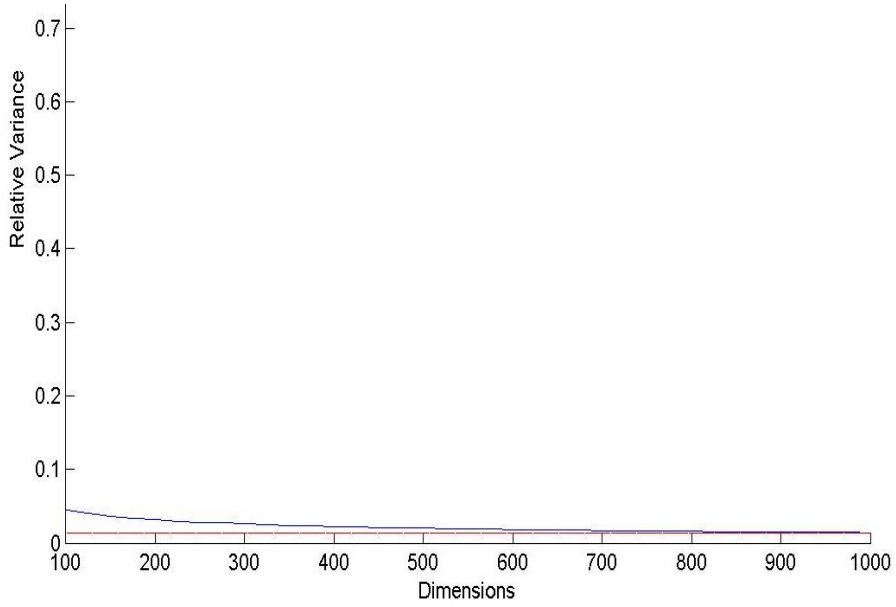
0.4.3 Some Indices to Illustrate the CoN phenomenon: A Theoretical Measure

While ξ_p^m is a good empirical measure to illustrate whether a norm concentrates or not, it is not amenable to theoretical analysis. This motivated François *et al.* **error** [François et al.(2007)François, to introduce a more theoretical index to measure the concentration in a similarity workload (Ω, X, ρ, μ). Note that this index is also derived from the result of Beyer *et al.*, **Theorem 0.3.2**.

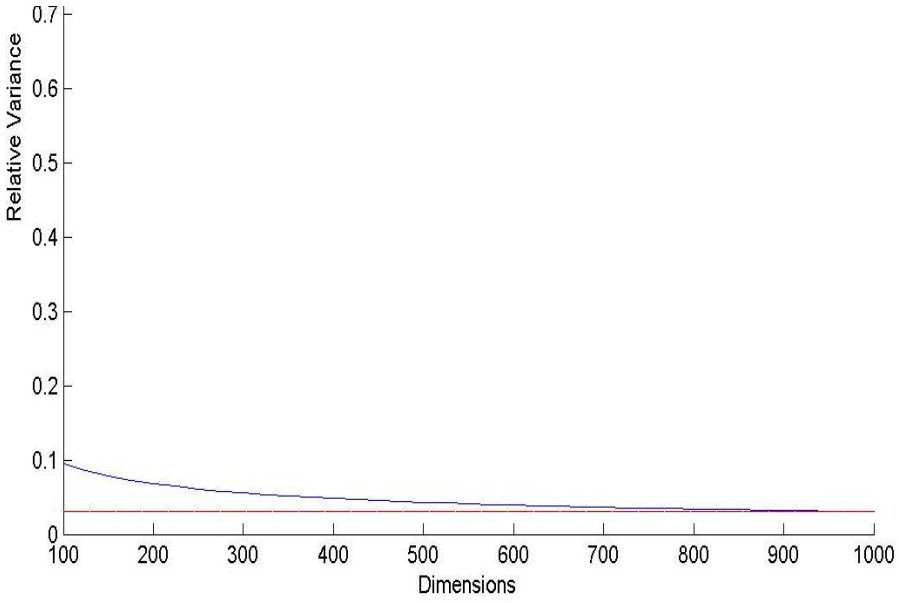
Definition 0.4.5 (François *et al.* [François et al.(2007)François, Wertz, and Verleysen], pg.

877). *Given a similarity workload,* (Ω, X, ρ, μ)*, where* Ω *is* m*- dimensional, the relative variance of* $\rho(\bar{x}, \bar{0}) = \|\bar{x}\|$ *is defined as:*

$$\gamma_p^m = \frac{\sqrt{Var\ (\|\bar{x}^m\|^p)}}{E\left(\|\bar{x}^m\|^p\right)}\ .$$



(a) Relative Variance for Euclidean norm from Uniform distribution



(b) Relative variance for Fractional norm from Uniform distribution

Figure 5: Relative variance for Euclidean norm and the Fractional Norm with $p = .04$, where data are from Uniform distribution. It is clear that both of them tend to zero in high dimensions, however the rate of convergence to zero does vary.

The relative variance γ_p^m illustrates the concentration of distances by comparing the spread of points with the expectation. If γ_p^m has small value then it indicates that norms are concentrated and a large value for γ_p^m denotes a good amount of spread between the

points. In some sense it is similar to ξ_p^m as ξ_p^m also compares the measure of spread to measure of location.

In fact, Theorems 0.3.2 and 0.3.3 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.*

For a fixed but large dimension m, François *et al.* also determined the explicit relation between γ_p^m and p as follows (see [François et al.(2007)François, Wertz, and Verleysen], **Theorem 6**):

$$\gamma_p^m = \frac{\sqrt{Var\|\bar{x}^m\|^p}}{E\|\bar{x}^m\|^p} \approx \frac{1}{p}\left(\frac{\sigma p}{\nu_p}\right) \ , \tag{5}$$

where $\nu_p = E(|X_i|^p)$ and $\sigma_p = Var(|X_i|^p)$.

The above relation (5) shows that for a fixed large m, as p decreases the relative variance γ_p^m increases and thus explains why an $\mathcal{F}_p$ norm $(0 < p < 1)$ gives better contrast than other $\mathcal{L}_p$ norms where $p \geq 1$. This can also be seen by comparing the relative contrast for $\mathcal{F}_{0.04}$ in Figs. 4(a) and (b) to those of $\mathcal{L}_2$ in Figs. 3(a) and (b).

However, François *et al.* also showed that for *any* fixed $p \in (0, \infty)$, as $m \to \infty$, $\gamma_p^m \to 0$. In fact, using relative variance as an index to measure concentration, François *et al.* proved that all Minkowski-type norms concentrate (see [François et al.(2007)François, Wertz, and Verleysen], **Theorem 5**) and showed that concentration is indeed an intrinsic property of Minkowski-type norms, though the rate of concentration may vary depending on the exponent p, as illustrated in Figs. 5(a) and (b).

For yet another illustration of the same, let us consider the indices k_m, k_A, k_M as discussed in Section 0.2.2 for the same similarity workload. Comparing Fig. 6(a) with Fig. 1(b) we see that all the indices k_m, k_A, k_M grow moderately in medium dimensions for the $\mathcal{F}_{0.04}$ norm as against those for the $\mathcal{L}_2$ norm. However, comparisons between Fig. 6(b) and Fig. 2 show that in high dimensions, all Minkowski-type norms do concentrate.

0.4.4 An Index to Measure the CoN phenomenon

While ξ_p^m and γ_p^m illustrate the concentration phenomenon well, they do not give any information on the rate at which a norm concentrates. Recently, Pestov [Pestov(2000)] introduced a more general mathematical function to measure concentration.

Definition 0.4.6 (Pestov, [Pestov(2000)])**.** *Let us be given a measurable metric space* (Ω, ρ, μ)*. The concentration function* $\alpha_\Omega : \mathbb{R}^{\geq 0} \to [\frac{1}{2}, 1]$ *is defined as follows:*

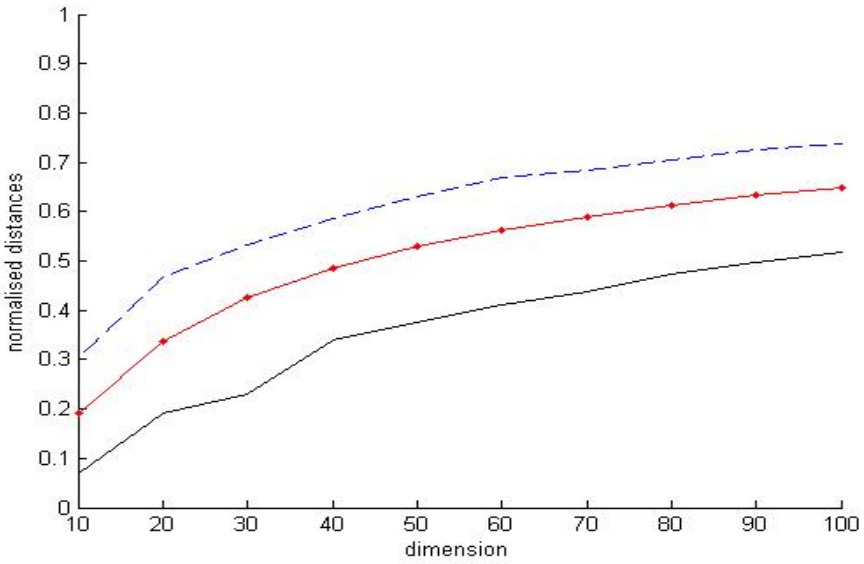
$$\alpha_\Omega(\varepsilon) = \begin{cases} 1 - \inf\{\mu(O_\varepsilon(A)) : A \subseteq \Omega \text{ is } \textit{Borel} \ \& \ \mu(A) \geq 1/2\} \ , & \text{if } \varepsilon > 0 \ , \\ \frac{1}{2}, & \text{if } \varepsilon = 0 \ , \end{cases}$$

where

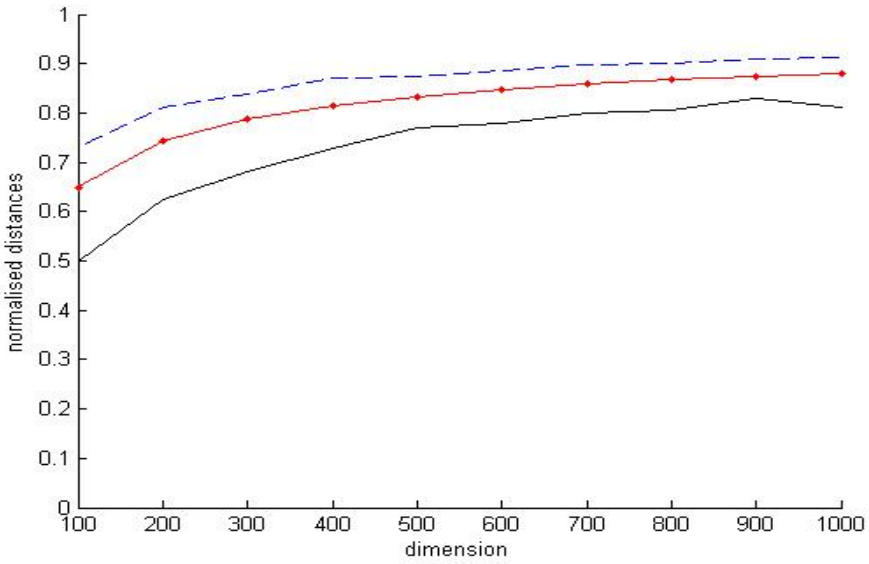
$$O_\varepsilon(A) = \{x \in \Omega : \textit{for some } a \in A, \rho(x, a) < \varepsilon\} \ .$$

The value $\alpha_\Omega(\varepsilon)$ 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 $\frac{1}{2}$.

To gain a better understanding, let us calculate and plot the concentration function for some measurable metric spaces (Ω, ρ, μ) and show that α_Ω does, indeed, 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.



(a) The indices $k_M(--), k_m(-), k_A(-\cdot-)$ for the Fractional Norm with $p = .04$ in dimensions $m = 10, \ldots, 100$



(b) The indices $k_M(--), k_m(-), k_A(-\cdot-)$ for the Fractional Norm with $p = .04$ in dimensions $m = 100, \ldots, 1000$

Figure 6: Concentration phenomenon exhibited by Fractional norms when moving from low to high dimensions

Example 0.4.7. *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* $\mathcal{L}_1$ *norm and* μ *is the counting measure. Table 1 shows the steps involved in the calculation of* α_{Ω_1} *for a few values of* $\varepsilon = 0.1, 0.5, 1.5$*. The final concentration function* α_{Ω_1} *(– • –) is plotted in Fig. 7.*

Example 0.4.8. *Let us consider the same space as in Example 0.4.7, but with the domain* $\Omega_2 = [0, 1] \cup [1.1, 2.1]$*, while* ρ, μ *remain the same. Once again, Table 0.4.4 shows the steps involved in the calculation of* α_{Ω_2} *for a few values of* $\varepsilon = 0.1, 0.5, 1.1$ *and the final concentration function* α_{Ω_2} *(– –) is plotted in Fig. 7.*

S.No.	ε	A	$\mu(A)$	$O_\varepsilon(A)$	$\mu(O_\varepsilon(A))$	$\alpha_\Omega(\varepsilon)$
1		[0,1]	0.5	[0,1]	0.5	
2	0.1	[2,3]	0.5	[2,3]	0.5	0.5
3		[0,0.6] ∪[2,2.8]	0.7	[0,0.7] ∪ [2,2.9]	0.7	
4		Ω_1	1	Ω_1	1	
1		[0,1]	0.5	[0,1]	0.5	
2	0.5	[2,3]	0.5	[2,3]	0.5	0.5
3		[0,0.6]∪[2,2.8]	0.7	[0,1]∪ [2,3]	1	
4		Ω_1	1	Ω_1	1	
1		[0,1]	0.5	[0,1] ∪ [2,2.5]	0.75	
2	1.5	[2,3]	0.5	[0.5,1] ∪ [2,3]	0.75	0.25
3		[0,0.6]∪[2,2.8]	0.7	[0,1] ∪ [2,2.1] ∪ [2,3]	1	
4		Ω_1	1	Ω_1	1	

Table 1: Calculating concentration function for Ω_1

S.No.	ε	A	$\mu(A)$	$O_\varepsilon(A)$	$\mu(O_\varepsilon(A))$	$\alpha_\Omega(\varepsilon)$
1		[0,1]	0.5	[0,1]	0.5	
2	0.1	[1.1,2.1]	0.5	[1.1,2.1]	0.5	0.5
3		[0,0.6]∪[1.1,1.8]	0.65	[0,0.7]∪[1.1,1.9]	0.75	
4		Ω_2	1	Ω_2	1	
1		[0,1]	0.5	[0,1]∪[1.1,1.5]	0.7	
2	0.5	[1.1,2.1]	0.5	[0.6,1]∪[1.1,2.1]	0.7	0.3
3		[0,0.6]∪[1.1,1.8]	0.65	[0,1]∪ [1.1,2.1]	1	
4		Ω_2	1	Ω_2	1	
1		[0,1]	0.5	[0,1] ∪ [1.1,2.1]	1	
2	1.1	[1.1,2.1]	0.5	[0,1] ∪ [1.1,2.1]	1	0
3		[0,0.6]∪[1.1,1.8]	0.65	[0,1] ∪ [2,2.1]	1	
4		Ω_2	1	Ω_2	1	

Table 2: Calculating concentration function for Ω_2

Example 0.4.9. *As a final example, let us consider the same space as in Example 0.4.7, except now the domain* $\Omega_3 = [-0.6, -0.1] \cup [0, 1] \cup [1.1, 1.6]$, *while* ρ, μ *remain the same. Once again, Table 0.4.4 shows the steps involved in the calculation of* α_{Ω_2} *for a few values of* $\varepsilon = 0.1, 0.2, 0.6$ *and the concentration function* α_{Ω_3} *(– + –) is plotted in Fig. 7.*

From Tables 1, 0.4.4 and 0.4.4 and Fig. 7 we see that α_Ω is a decreasing function of ε. The smaller the value of ε the faster the norm concentrates. In fact, the rate at which α_Ω decreases is illustrative of the fact that the pairwise distances, as measured by ρ, concentrate near their mean/median value.

0.4.5 New Distance measures to mitigate CoN

Let $(\Omega, \leq)$ be a poset with a special element $\bar{0} \in \Omega$. A $\rho : \Omega \times \Omega \to \mathbb{R}^{\geq 0}$ is called a *distance function* if it satisfies the following :

- $\rho(\bar{x}, \bar{y}) = \rho(\bar{y}, \bar{x})$, for all $\bar{x}, \bar{y} \in \Omega$,
- $\rho(\bar{x}, \bar{y}) = \Leftrightarrow \bar{x} = \bar{y}$, for all $\bar{x}, \bar{y} \in \Omega$,

S.No.	ε	A	$\mu(A)$	$O_\varepsilon(A)$	$\mu(O_\varepsilon(A))$	$\alpha_\Omega(\varepsilon)$
1		[0,1]	0.5	[0,1]	0.5	
2	0.1	[-0.6,-0.1]∪ [1.1,1.6]	0.5	[-0.6,-0.1]∪ [1.1,1.6]	0.5	0.5
3		[0,1]∪[1.1,1.2]	0.55	[0,1]∪[1.1,1.3]	0.6	
4		Ω_3	1	Ω_3	1	
1		[0,1]	0.5	[-0.2,-0.1]∪[0,1]∪[1.1,1.2]	0.6	
2	0.2	[-0.6,-0.1]∪ [1.1,1.6]	0.5	[-0.6,-0.1]∪[0,0.1]∪[0.9,1]∪ [1.1,1.6]	0.6	0.4
3		[0,1]∪[1.1,1.2]	0.55	[-0.2,-0.1]∪[0,1]∪[1.1,1.3]	0.8	
4		Ω_3	1	Ω_3	1	
1		[0,1]	0.5	[-0.6,0.1]∪ [0,1] ∪ [1.1,1.6]	1	
2	0.6	[-0.6,-0.1]∪ [1.1,1.6]	0.5	[-0.6,0.1]∪ [0,1] ∪ [1.1,1.6]	1	0
3		[0,1]∪[1.1,1.2]	0.55	[-0.6,0.1]∪ [0,1] ∪ [1.1,1.6]	1	
4		Ω_3	1	Ω_3	1	

Table 3: Calculating concentration function for Ω_3

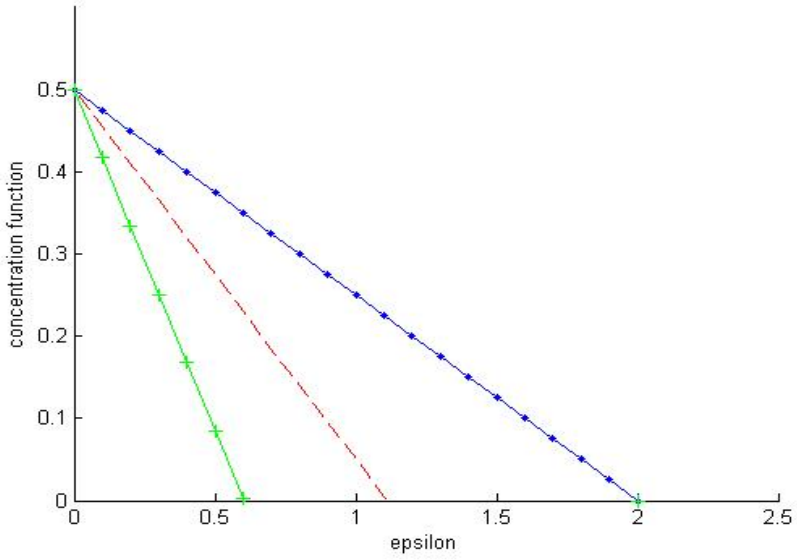


Figure 7: Concentration functions $\alpha_{\Omega_1}(--\bullet-)$, $\alpha_{\Omega_2}(---)$, $\alpha_{\Omega_3}(-+-)$ vs ε

- if it is monotonic on a chain i.e. ,

$$\bar{x} \leq \bar{y} \leq \bar{z} \Longrightarrow \rho(\bar{x}, \bar{y}) \leq \rho(\bar{x}, \bar{z}) \text{ , for all } \bar{x}, \bar{y}, \bar{z} \in \Omega \text{ .}$$

Further, a distance function ρ is said to be

- a *metric* if it satisfies the triangle inequality, i.e.,

$$\rho(\bar{x}, \bar{y}) \leq \rho(\bar{x}, \bar{z}) + \rho(\bar{y}, \bar{z}) \text{ , for all } \bar{x}, \bar{y}, \bar{z} \in \Omega \text{ .}$$

- *unbounded* on a bounded domain Ω if there exists an $\bar{x}_0 \in \Omega$ such that $\lim_{\bar{x} \to \bar{x}_0} \|\bar{x}\| = \infty$, where $\|\bar{x}\| = \rho(\bar{x}, \bar{0})$.

- *translation invariant* on a domain with well-defined addition of elements, denoted $+$, if $\bar{x}, \bar{y} \in \Omega$ and for any $\bar{z} \in \Omega$ such that $\bar{x} + \bar{z}, \bar{y} + \bar{z} \in \Omega$ the following equality holds:

$$\rho(\bar{x} + \bar{z}, \bar{y} + \bar{z}) = \rho(\bar{x}, \bar{y}) \text{ .}$$

In [Jayaram and Klawonn(2012)] the authors did a rigorous math analysis on the factors in a distance function that lead to their concentration. Their study indicated that unbounded measures seem more preferable and that, while triangle inequality and translation variance are desirable properties for a distance function ρ, they also contribute towards its concentration. Further, they proved the following result:

Theorem 0.4.10 (Jayaram & Klawonn, [Jayaram and Klawonn(2012)])**.** *Given a bounded metric space (Ω, ρ), with a suitable ordering $\leq$ and a well-defined addition $+$, a distance function ρ can have atmost two of the following properties:*

- *Unboundedness*
- *Translation invariance*
- *Triangle Inequality*

Towards illustrating that such unbounded distance functions which have the desirable properties to fight the concentration phenomenon do exist, they have introduced two new distance functions as defined below:

Definition 0.4.11. *Consider the poset $(\Omega, \leq)$, where $\Omega = [-1, 1]^m$ and with componentwise ordering and let $\bar{x} = (x_1, x_2, \ldots, x_m), \bar{y} = (y_1, y_2, \ldots, y_m) \in \Omega$. For any $p > 1$, the following functions $\mathcal{J}_p, \mathcal{K}_p : \Omega \times \Omega \to [0, \infty]$ are valid distance functions:*

$$\mathcal{J}_p(\bar{x}, \bar{y}) = \left(\sum_{i=1}^{m} \left| \frac{|x_i|}{1 - |x_i|} - \frac{|y_i|}{1 - |y_i|} \right|^p \right)^{\frac{1}{p}} , \tag{6}$$

$$\mathcal{K}_p(\bar{x}, \bar{y}) = \left(\sum_{i=1}^{m} \left| \frac{x_i - y_i}{1 - |x_i - y_i|} \right|^p \right)^{\frac{1}{p}} . \tag{7}$$

Note that a complete and thorough investigation of the stability of the above distance functions $\mathcal{J}_p, \mathcal{K}_p$ and their suitability in applications is yet to be done, though some partial studies are available in [Jayaram and Klawonn(2012)].

0.5 Motivation for and the Objectives of our current work

0.5.1 Studying the Newly Proposed Distances like $J_p, \mathcal{K}_p$

As noted already in **Section 0.4.5**, a complete and thorough investigation of the stability of the above distance functions $\mathcal{J}_p, \mathcal{K}_p$ and their suitability in applications is yet to be done. Theoretical analysis and results in [Jayaram and Klawonn(2012)] indicate that these functions are better prepared to fight concentration compared to, say the Euclidean norm. Hence, we would like to take up deeper investiagations of the same and study their performance vis-á-vis the existing distance functions both w.r.to the different indices, like γ_p^m and ξ_p^m, and also in typical applications.

0.5.2 γ_p^m, ξ_p^m and α_Ω: Some Drawbacks

From the above, it is clear that there exist three main and strong indices to catch hold of concentration, namely, ξ_p^m, γ_p^m and $\alpha_\Omega(\varepsilon)$. However, these indices do have some drawbacks.

- We know γ_p^m and ξ_p^m can deal with concentration in more general way but however, it is not always easy to find variance and expectation of norms for arbitrary distribution and distances.
- Given an m-dimensional data set of size N, drawn i.i.d. from $\mathcal{R}^m$, given a metric ρ and an $\varepsilon > 0$, what is $P[D^m_{\max} \leq (1+\varepsilon)D^m_{min}]$? In other words, how likely it is that in an arbitrary sample of size N the largest distance would be no more than $1+\varepsilon$ distance away from the smallest one?
- While one could get an estimate of γ_p^m from the data set, how *small* should that be to conclude that the above probability is *large*?
- Although γ_p^m and ξ_p^m are strong indices to measure concentration, in some workloads, they may be very time consuming and difficult to calculate. For example, computing the Euclidean distance between two points in a high-dimensional space, say m, requires m multiplication operations and $m-1$ addition operations, as well as a square root operation.
- Similarly while the concentration function α_Ω is a general and theoretical concept, often it is difficult to determine the concentration function for given metric and distribution.
- Calculating concentration function empirically and applying to synthetic data sets can be very dry. It includes taking the ε-neighbourhood of a set and then calculating its infimum, which can be done theoretically for simple sets but empirically may be too complicated.

0.6 Objectives of this study:

Based on the above discussions and observations on the CoN phenomenon, our objectives for this study are as follows:

Objective 1: Study the stability of norms, especially the newly proposed $\mathcal{J}_p, \mathcal{K}_p$ with respect to existing indices and also in applications.

Objective 2: Propose simpler indices that would (i) measure the rate of concentration and (ii) allow being applied in emprical settings.

0.7 Analysis of $\mathcal{J}_p$ and $\mathcal{K}_p$

Since $\mathcal{J}_p$ and $\mathcal{K}_p$ are newly introduced distance functions, not much have been explored about them neither theoretically nor empirically. In this section, we present some empirical results done on $\mathcal{J}_p$ and $\mathcal{K}_p$ norms for $p = 2$ to show how they behave in higher dimension. We document our findings in the following sections.

0.7.1 $\mathcal{J}_p$, $\mathcal{K}_p$ and Nearest Neighbour Distances

The concentration phenomenon for the Euclidean $\mathcal{L}_2$ norm was shown in Section 0.2.2 by discussing the normalised minimum / average / maximum NN distance w.r.to the average of all pairwise distances. A similar study was conducted by us on exactly the same dataset used to analyse the indices k_m, k_A, k_M for the $\mathcal{L}_2$ and $\mathcal{F}_{0.04}$ distance functions. We present the results in Fig. 8.

From Fig. 8(a) we see that, for the $\mathcal{J}_2$ distance function, the index k_m is very close to 0 even in high dimensions as it should be, while $k_A \approx 0.5$, indicating that even average NN distances are much below the average pairwise distances, once again even in high dimensions. The values of k_M were far greater than 1, in every repeated trial, and hence is not plotted here to retain a sense of proportion. This indicates that the maximum NN distances far exceeded the average pairwise distances, which augurs very well when it comes to distinguishing points in high dimensions.

From Fig. 8(b) we see that, for the $\mathcal{K}_2$ distance function, the indices $k_m < k_A < k_M \ll 1$ showing that *all* of the NN distances are much below the average pairwise distances, thus ensuring exceptional contrast between the points.

0.7.2 $\mathcal{J}_p$, $\mathcal{K}_p$ and the Relative Contrast

Considering $\Omega = [-1, 1]^m$ we generated two data sets each containing $N = 100,000$ points which were distributed as follows, for each of the dimensions $m = 100, \ldots, 1000$:

(i) $X^m \sim \mathcal{U}\left((-1, 1)^m\right)$,

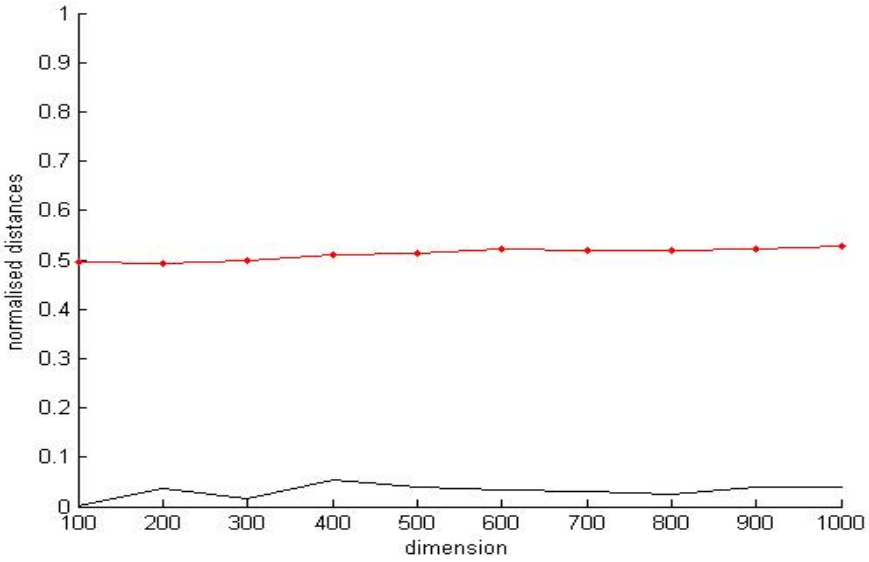
(ii) $X^m \sim \mathcal{N}\left((0, 0.3)^m\right)$.

We then plotted the relative contrasts ξ_2^m of the $\mathcal{J}_2$ norm for these data sets with query point taken to be the origin of Ω, i.e., $\bar{q} = \bar{0}$, in which case

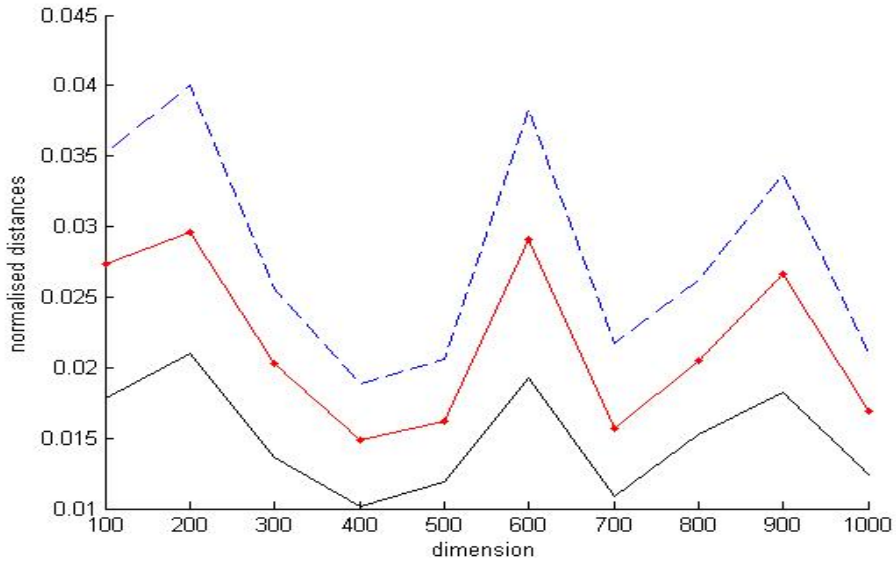
$$\mathcal{J}_2\left(\bar{x}, \bar{q}\right) = \mathcal{J}_2\left(\bar{x}, \bar{0}\right) = \|\bar{x}\|_{\mathcal{J}} = \mathcal{K}_2\left(\bar{x}, \bar{0}\right) = \|\bar{x}\|_{\mathcal{K}} \ .$$

Thus the plots of relative contrast for the $\mathcal{J}_2$ and $\mathcal{K}_2$ norms coincide.

From Figs. 9(a) and (b), one can make the following two observations:



(a) The indices $k_m(-), k_A(-\cdot-)$ for the $\mathcal{J}_2$ norm



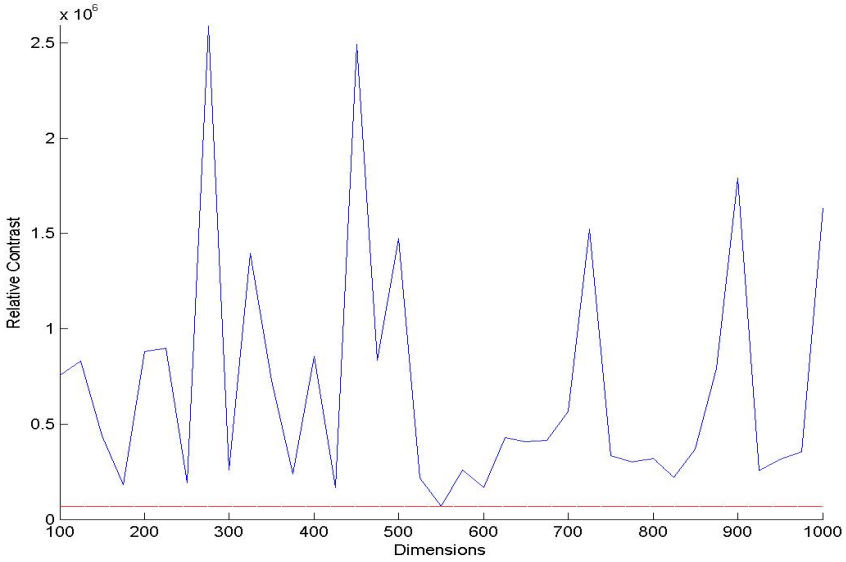
(b) The indices $k_M(--), k_m(-), k_A(-\cdot-)$ for the $\mathcal{K}_2$ norm

Figure 8: The indices k_M, k_m, k_A for the $\mathcal{J}_2$ and $\mathcal{K}_2$ norms for data generated from uniform distributions on $[-1, 1]^m$ for dimensions $m = 100, \ldots, 1000$

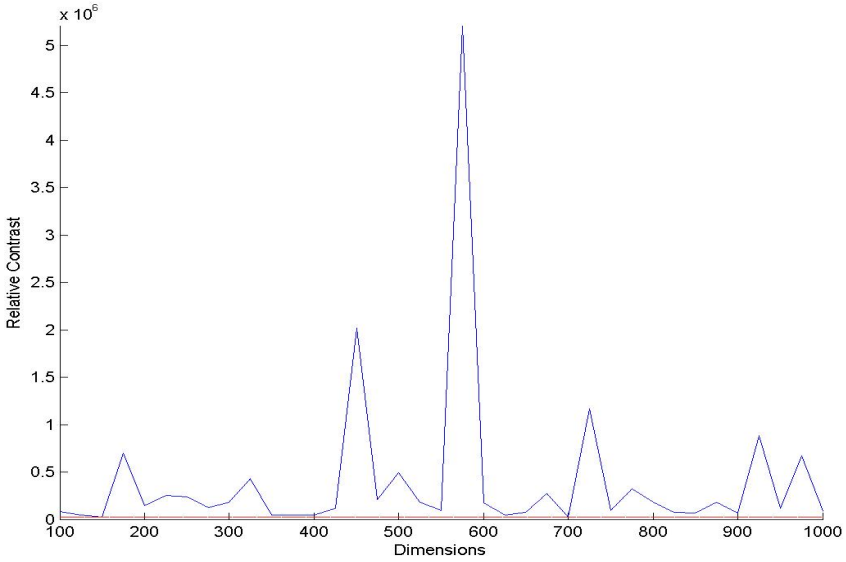
(i) The plots for ξ_2^m show that there is no tendency to decrease to a particular value. Thus we see that the $\mathcal{J}_2$ and $\mathcal{K}_2$ norms buck the trend shown by $\mathcal{L}_p$ and $\mathcal{F}_p$ norms.

(ii) Further, at first glance, it does appear that the values of ξ_2^m are too close to zero for comfort. However, a closer inspection shows that the scale is of the order 10^5 and hence the separation between points is excellent.

Note that these were the identical data sets used to calculate the relative contrasts of $\mathcal{L}_2$ and $\mathcal{F}_{0.04}$ norms in Figs. 3 and 4.



(a) Relative contrast for $\mathcal{J}_2$ norm from Uniform distribution



(b) Relative contrast for $\mathcal{J}_2$ norm from Gaussian distribution

Figure 9: Relative contrast for the $\mathcal{J}_2$ norm where $\Omega = [-1, 1]^m$ for dimensions $m = 100, \ldots, 1000$. (a) $X^m \sim \mathcal{U}\left((-1, 1)^m\right)$ (b) $X^m \sim \mathcal{N}\left((0, 0.3)^m\right)$

0.7.3 $\mathcal{J}_p$, $\mathcal{K}_p$ and the Relative Variance

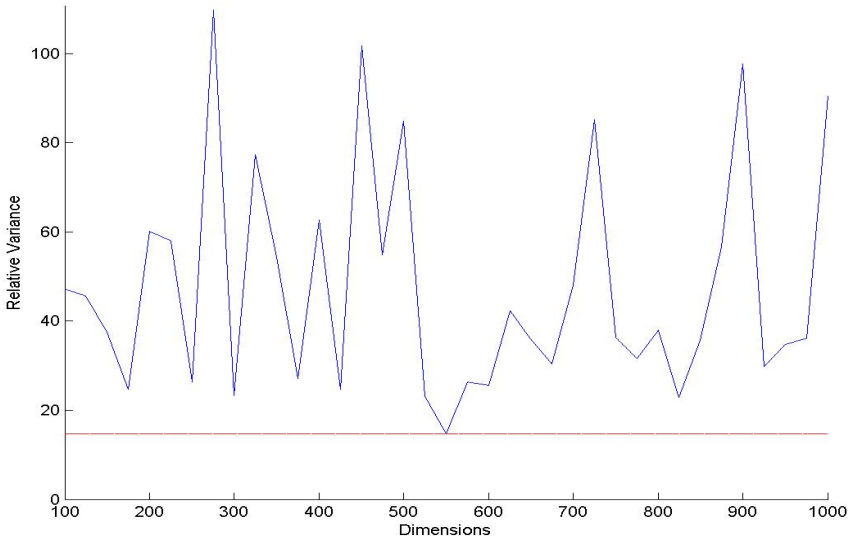
Once again, for the same datasets as presented above, we calculated the relative variance γ_2^m of the $\mathcal{J}_2$ norm, which, once again, is equal to that of the relative variance of the $\mathcal{K}_2$ norm.

Figs. 10(a) and (b), once again show that the relative variance γ_2^m of the $\mathcal{J}_2$ norm

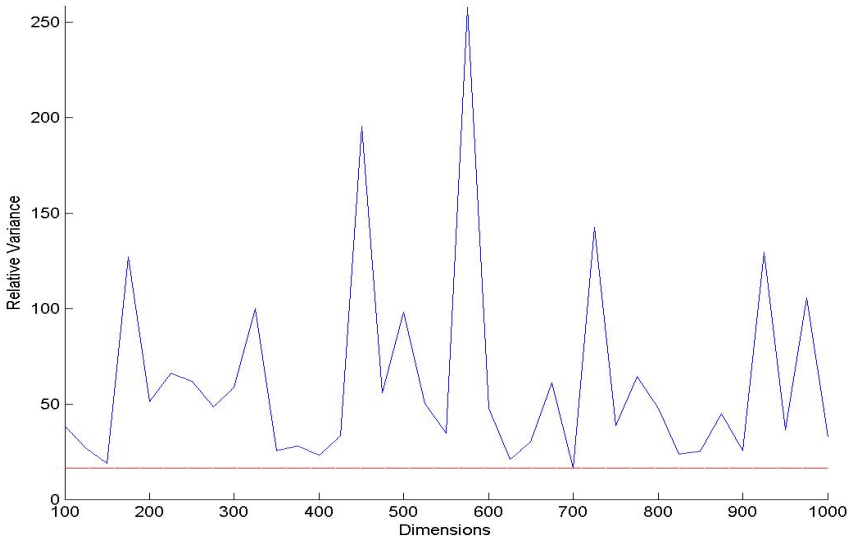
(i) bucks the decreasing tendency shown by the Minkowski-type norms, and

(ii) the values are far greater than zero,

thus providing confidence on the separation powers of the newly proposed $\mathcal{J}_p$ and $\mathcal{K}_p$ norms.



(a) Relative variance for $\mathcal{J}_2$ norm from Uniform distribution



(b) Relative variance for $\mathcal{J}_2$ norm from Gaussian distribution

Figure 10: Relative variance for the $\mathcal{J}_2$ norm where $\Omega = [-1, 1]^m$ for dimensions $m = 100, \ldots, 1000$. (a) $X^m \sim \mathcal{U}((-1, 1)^m)$ (b) $X^m \sim \mathcal{N}((0, 0.3)^m)$

From the above empirical results it is clear that new distance functions $\mathcal{J}_p$ and $\mathcal{K}_p$ behave better than the Minkowski-type norms in high dimensions.

0.8 Need for Efficient Emperical Indices

There were three main indices proposed by the researchers to measure concentration viz. Relative contrast, Relative variance and Concentration function. But these indices comes with the problem that some are not comfortable with the empirical settings while some are not suitable for theoretical study. For example, Concentration function (α_Ω) is a theoretical index. It can be studied theoretically and as well as measures the rate of concentration but it is very difficult to compute α_Ω for a empirical settings. Where as Relative Contrast is an empirical index but it cannot be studied theoretically. Since Relative Contrast and Relative variance have been studied in detail in the previous section, so the studies done ahead mainly focuses on α_Ω.

0.8.1 Advantages and Drawbacks of α_Ω

Recalling from **Section 0.5.2**, we see that despite the fact that α_Ω is a pure theoretical tool and is not so difficult to calculate for smaller set with pen and paper, it does have its drawbacks. For instance,

1. What if we do not know the underlying distribution of a particular dataset *a priori*? Then we do not know μ and hence cannot determine α_Ω.

2. Also, for large sets calculating α_Ω is very cumbersome as we need to find every subset of Ω with measure at least half. In other words, given a set with cardinality n, the number of subsets with measure greater than $\frac{1}{2}$ is equal to

$$\sum_{k=\frac{n}{2}}^{n} {}^nC_k = 2^{n-1} \ .$$

   So, α_Ω may prove to be computationally inefficient if we move to empirical settings.

The above questions poses the problem of stability of workloads and the usefulness of α_Ω in empirical settings. Given a similarity workload, we want to know whether it is stable or not? In the next sections, we discuss these in detail and come up with an empirical index that upper bounds α_Ω and is also comparitively easier to calculate than α_Ω.

0.9 Stability of Distance Functions

This section mainly discusses the stability of a range queries and establishes the setting in which we can discuss the stability of distance functions.

0.9.1 Stability of a Query

Let $(\Omega, \mathcal{X}, \rho, \mu)$ be a given similarity workload. Given a query $q \in \Omega$ and an $\varepsilon \in \bar{\mathbb{R}}_+$ we need to find the set of all points in $\mathcal{X}$ that are within ε units away from q, i.e., we need to find the following subset of $\mathcal{X}$:

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

Note that $S = N(q, \varepsilon) \subset \mathcal{X}$ and hence the problem of finding S is also known as the *range-query*.

In [Beyer et al.(1999)Beyer, Goldstein, Ramakrishnan, and Shaft], the authors discuss when a range-query is stable by defining the stability of a range-query as follows:

Definition 0.9.1. *Given a query point $q \in \Omega$, a range-query is said to be ε-unstable if*

$$\sharp\{x \in X : \rho(q, x) \leq (1 + \epsilon) * \delta\} \geq \frac{\sharp X}{2}$$

where, $\delta = \min\{d(q, x) : x \in \mathcal{X}\}$, the nearest neighbor distance of the query point q.

In other words, if half of the data set is covered within the ε-δ sphere of the query q, then the range-query is said to be *unstable*.

Taking a cue from Definition 0.9.1, we define the stability of a particular workload and propose an index that can overcome the drawbacks of α_Ω. We term the analysis done along these lines as the *g-δ Stability Analysis* .

0.9.2 *g*-δ Stability Analysis

Let $x_i \in \mathcal{X}$ and let δ_i denote the nearest neighbor distance of x_i, i.e. $\delta_i = \min\{\rho(x, x_i) : x \in \mathcal{X}\}$. For any $g \in \bar{\mathbb{R}}_+$, the *g*-δ_i neighborhood of x_i is defined as :

$$N_{g\delta_i}(x_i) = N_g(x_i, \delta_i) = \{x \in \mathcal{X} : d(x, x_i) \leq g * \delta_i\} \ .$$

By $\mathbb{N}_n$ we denote the first n natural numbers, i.e., $\mathbb{N}_n = \{1, 2, \ldots, n\}$. Let us define a function $C_g : \mathcal{X} \to \mathbb{N}_n$ such that

$$C_g(x_i) = \sharp N_g(x_i) \ .$$

It counts the number of data points in the g-δ_i neighborhood of x_i. In some sense it tells us how closely a data set is distributed.

Consider a point x_i and take its δ_i-neighborhood. Now dilate the δ_i neighborhood with radius $g * \delta_i$. So counting the number of points lying in the dilated sphere will give the C_g count for point x_i. For example, let $C_g(x_i) = 5$. This means that the point x_i has 5 other data points in its dilated *g*-δ sphere. If the C_g values of most of the $x \in \mathcal{X}$ is high, then more points are lying 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. So, in a way C_g does keep track of the concentration of points. Specifically, given a dataset without the information of the distribution of the dataset, C_g is easily computable and further analysis can be done easily.

0.9.3 g-Compactness of a Dataset

Given a similarity workload, we want to look on the C_g values of the dataset and therefore we discuss about the density function for C_g. As a result we have yet another definition.

Definition 0.9.2. *Let* $\eta_g : \mathbb{N}_n \to \mathbb{N}_n$ *be a function such that*

$$\eta_g(k) = \sharp\{C_g^{-1}(k)\}$$

where $C_g^{-1} : \mathbb{N}_n \to \mathbb{P}(\mathcal{X})$.

η_g expresses the cardinal number of data points in $\mathcal{X}$ that have their C_g values as k. For instance, if $\eta_g(k) = \ell$ then it means 'ℓ' points in $\mathcal{X}$ have 'k' other data points in their dilated $g - \delta$ sphere. This afresh introduced index will be known as **g-compactness** of the point x_i such that $C_g(x_i) = k$ for a given k. Therefore, $\eta_g(k)$ is just the density function for C_g with $C_g(x_i) = k$ for some $x_i \in \mathcal{X}$.
Normalized probability mass function of $C_g(x_i)$ is given as :

$$\tilde{\eta_g} = \frac{\eta_g(k)}{N}$$

Properties of η_g

η_g is an indicator of the flow of the densities of C_g for different data points. If η_g is large for large values of k, this means more number of points have more other members in their dilated $g - \delta$ sphere leading to the concentrating of points. This condition is not desirable. So k and $\eta_g(k)$ should be inversely proportional to each other and hence with increasing values of k, $\eta_g(k)$ should be a decreasing function.

We abstract out the properties of η_g as follows:

(i) In other notation, $\eta_g(k) = \sum_{i=1}^{n} \mathcal{I}\{C_g(x_i) = k\}$

(ii) η_g is a decreasing function i.e. given $n_1, n_2 \in \mathbb{N}_n$ such that $n_1 \leq n_2$ then, $\eta_g(n_1) \leq \eta_g(n_2)$.

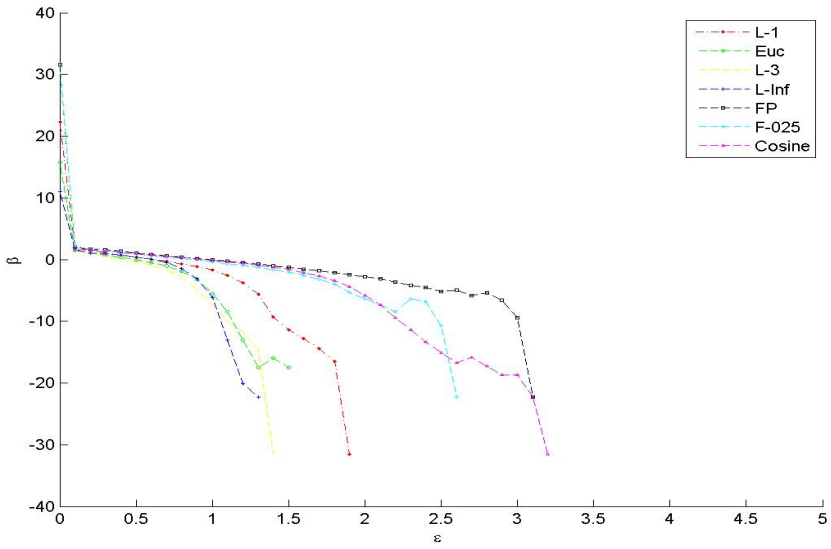
What can be more appealing is to look at the graph of $\widetilde{\eta}_g$ for different k since we are more interested in the density of C_g.

Definition 0.9.3. *Let* $\beta_{\mathcal{X}} : [0, 1] \to (-\infty, \infty)$ *be a function defined as:*

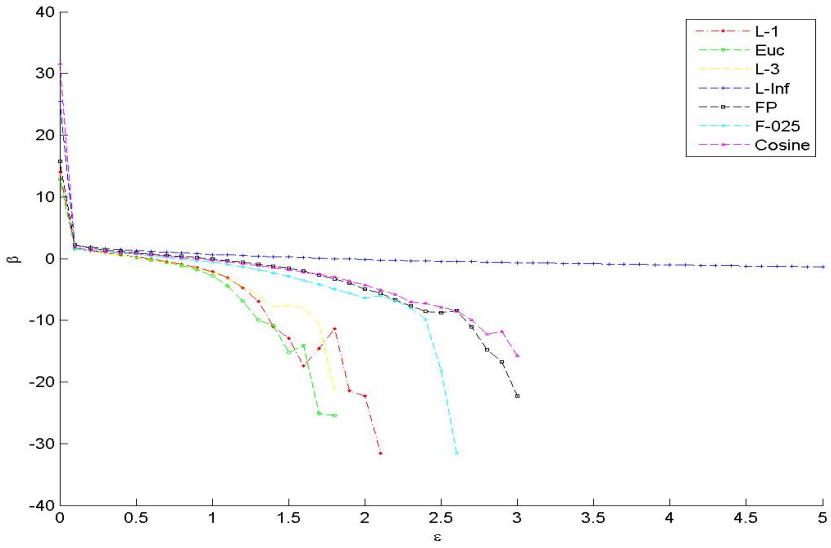
$$\beta_{\mathcal{X}}(\varepsilon) = S_{1+\varepsilon}(\mathcal{X}) = S_g(\mathcal{X})$$

where $S_g(\mathcal{X}) = \frac{E(\widetilde{\eta}_g - \mu_{\widetilde{\eta}_g})^3}{{\sigma_{\widetilde{\eta}_g}}^3}$.
The function $\beta_{\mathcal{X}}$ *is called the skewness of* $\widetilde{\eta}_g$.

From the Demartines results, we can see that $\beta_{\mathcal{X}}$ will always be a decreasing function as expectation depends on dimension whereas variance is independent of dimension. We want the graph of $\widetilde{\eta}_g$ to fall steeply with increasing k to have less concentration. Talking in terms of density, density should be more on the left so that k is less and $\widetilde{\eta}_g$ is large. Therefore, the density of $\widetilde{\eta}_g$ should be more on the left side and thus the graph of $\beta_{\mathcal{X}}$ should be positively



(a) Uniform Distribution with $n = m = 1000$



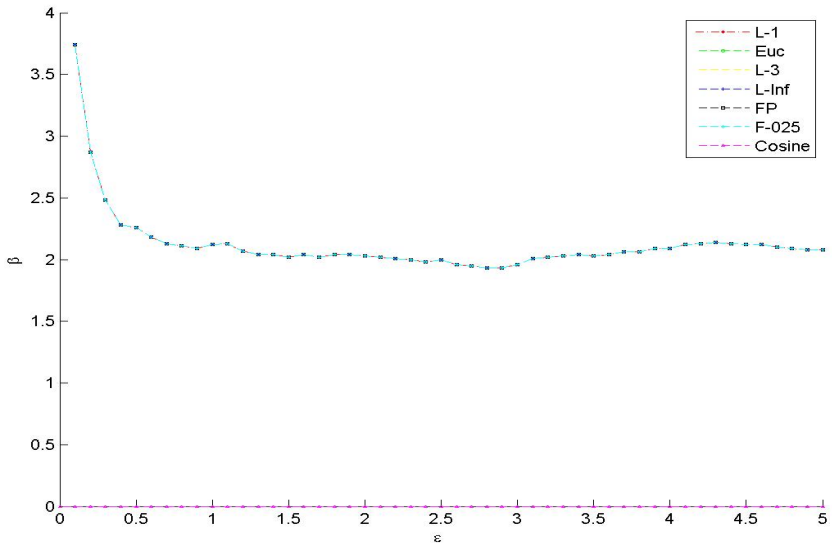
(b) Gaussian Distribution with $n = m = 1000$

Figure 11: Skewness Plots for Synthetic Data sets for different Distance Functions

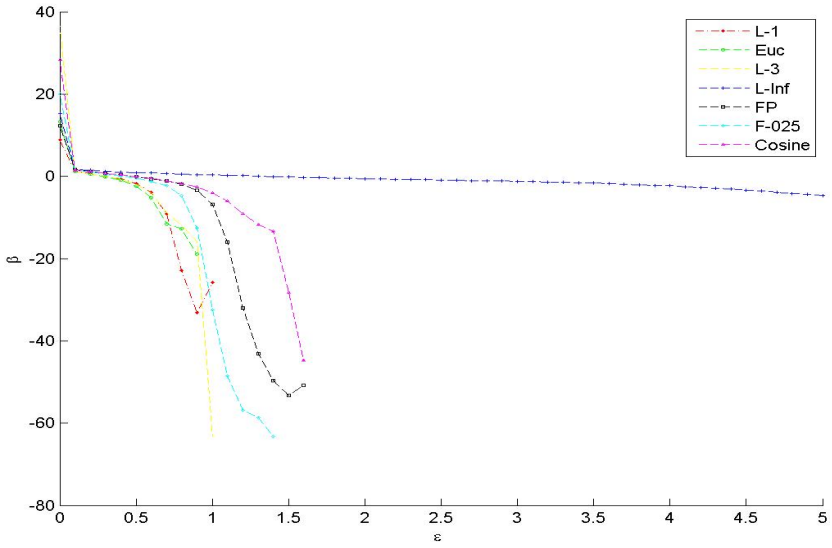
skewed. As a result, we yearn $\beta_{\mathcal{X}} \geq 0$ even if $\varepsilon \gg 1$. The quicker the $\beta_{\mathcal{X}}$ falls the quicker the efficiency of distance functions to distinguish points well narrows. The rate of falling of $\beta_{\mathcal{X}}$ clearly demonstrates the degrading power of distance functions.
Now we can define the analogy between the stability of a point x with respect to the C_g values for x . If $C_g(x) \geq \frac{\sharp\mathcal{X}}{2}$ i.e. x has more than half of the total number of points in its g-δ neighborhood then x is said to be ε- unstable.

0.9.4 Empirical results of η_g and $\beta_{\mathcal{X}}$

So far we have seen that, it is very easy to understand all the terms C_g,$\widetilde{\eta_g}$ and $\beta_{\mathcal{X}}$ theoretically. Now in this section we want to justify our interpretation. Summarizing all the efforts



(a) Uniform Distribution with $n = m = 4000$



(b) Gaussian Distribution with $n = m = 4000$

Figure 12: Skewness Plots for Synthetic Data sets for different Distance Functions

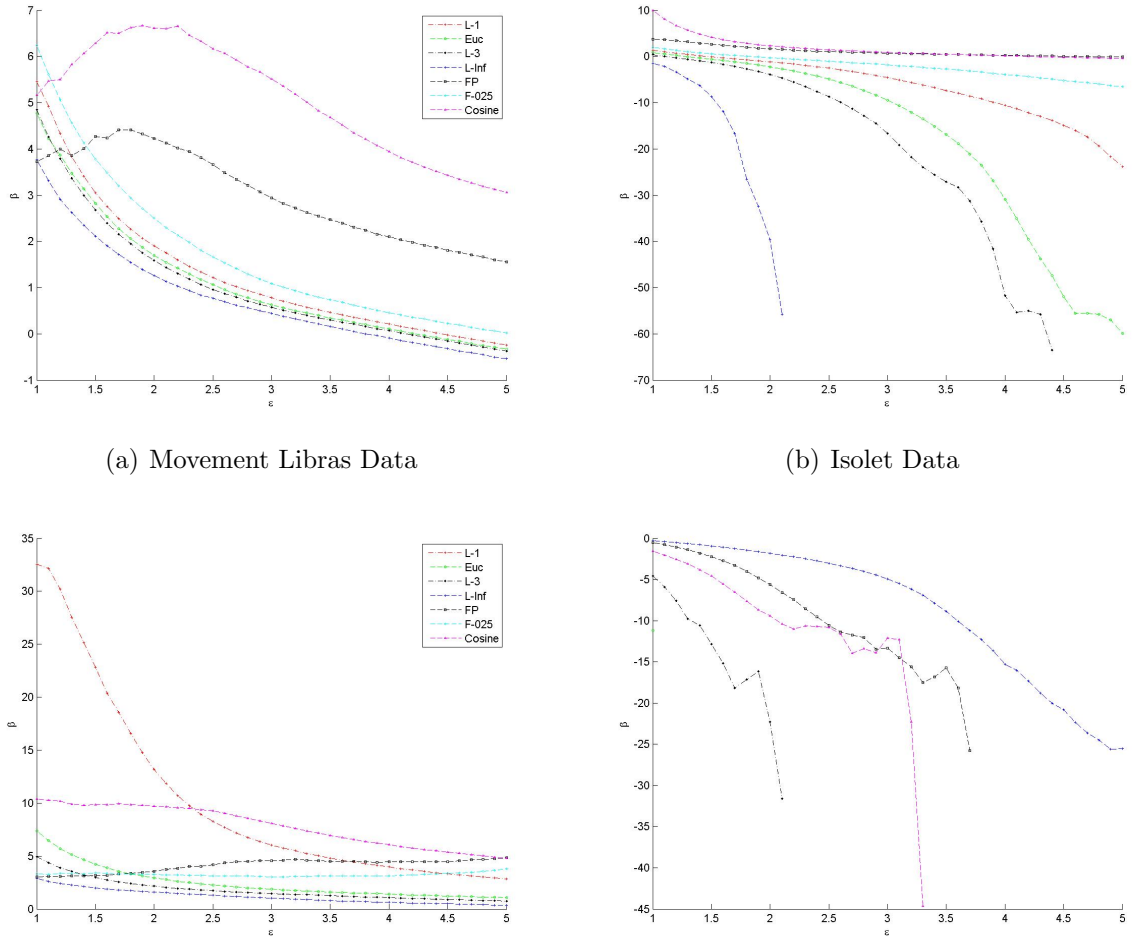
done up to now, tells that given a similarity workload, we need to check the density of η_g values and measure the concentration.

What do we want to do?

Basically we want to measure the density of η_g, in turn we want the skewness of η_g. Density lying more on left side is desirable for less concentration. It follows that well separated data set will have positive skewness and perform better.

How do we do?

We pick a data set and calculate the skewness for different distance functions and do the required analysis.



(a) Movement Libras Data

(b) Isolet Data

(c) Sensor Readings 24 Data

(d) Madelon Training Data

Figure 13: Skewness Plots for Real Data sets as in Table 4 for different Distance Functions

Empirical results on Synthetic datasets

We generated some data sets from Uniform or Gaussian Distribution. Then computed skewness values for various distance functions and plotted it all in one graph. The different parameters used for the computation of skewness of η_g is enlisted as:

- Data from a single distribution either Uniform or Gaussian Distribution. Note that in this study, we are not considering data from mixed distribution.
- We kept number of points to be same as number of dimension i.e., $n = m$, e.g. $n = 1000, 4000 = m$.
- Different Distance Functions, mainly :
  1. Minkowski distance function for $p = 0.04, 0.25, 1, 2, 3$.
  2. Cosine distance function.

Some of the plots is shown in the Figure 11 and 12 for different data sets and for all the distance functions in one plot for a single distribution, as mentioned above.

Some observations made from Figure11 :

Dataset	Dimension	Number of datapoints
Movement Libras Data	90	360
Isolet	618	7797
Gas Sensor Data	24	5456
Madelon Training Data	2000	500
Madelon Valid Data	600	500
Multiple Features with correlation coefficients	216	2000
Multiple Features with Fourier coefficients	76	2000
Multiple Features with Karhumen-Love coefficients	64	2000
Multiple Features with Morphological Features	6	2000
Multiple Features with Pixel Features	240	2000
Multiple Features with Zernike moments Features	47	2000

Table 4: Real Data sets

(i) We see that skewness was initially positive for very small value of ε and starts decreasing at a very rapid rate and soon hits 0 as the ε increases slowly. Almost every distance function used for the experiment is affected in the same way.

(ii) Also, note that for $\varepsilon \approx 0.01$ itself, the skewness has hit 0 for all the distance functions.

(iii) This clearly shows that these distance functions easily succumb to the concentration effect. This effect is happening for both the distribution, Gaussian and Uniform.

Remark : $\alpha_{\mathcal{X}}$ just been presented as an illustrative index for the concentration effect of the distance functions. It has not been identified as a comparative index for concentration effect.

Empirical results on Real data sets

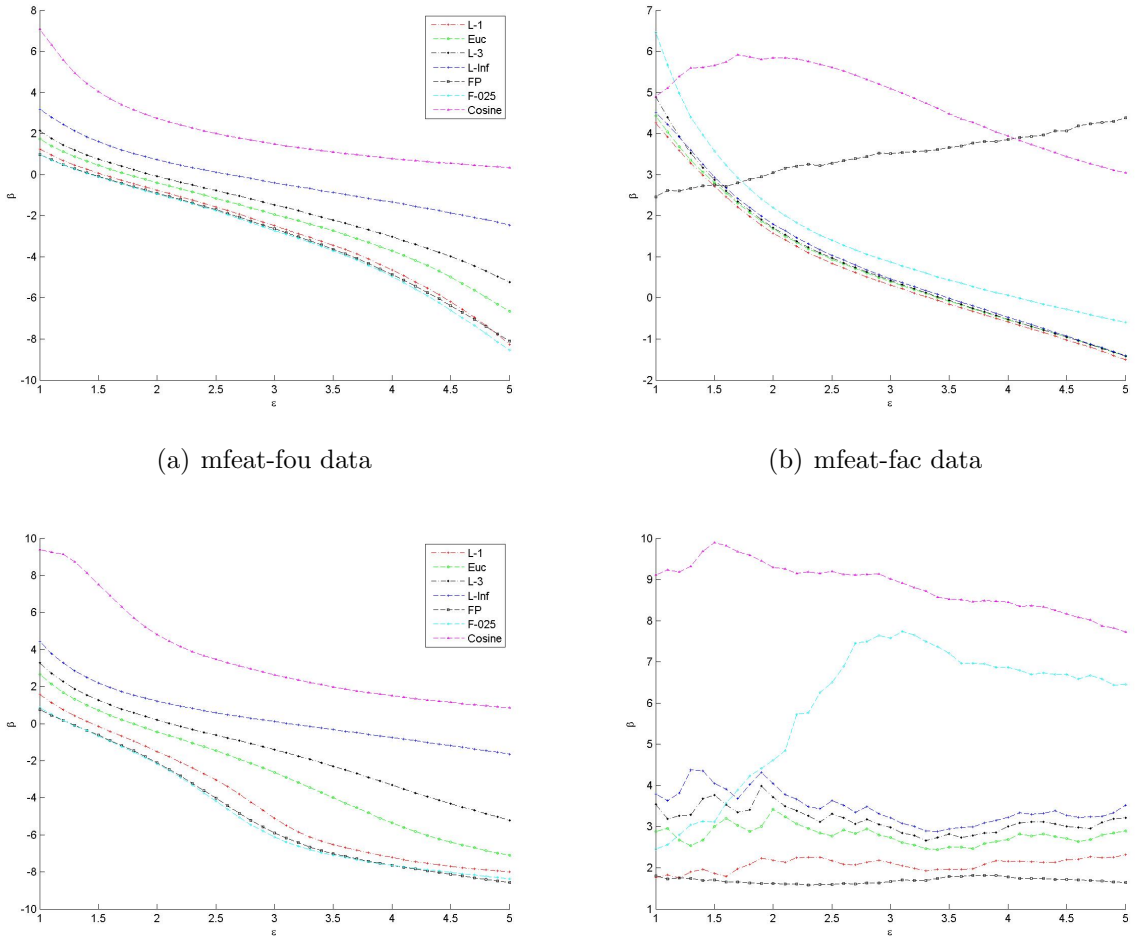
In this section, we show the different plots for $\beta_{\mathcal{X}}$, try to make conclusion for $\beta\mathcal{X}$. Figures 13,14 and 15, shows the different graphs for different real data sets picked from UCI Database as shown in Table 4. .

Some observations made from Figure 13,14 , 15 and 16 :

(i) We see that the skewness function decreases slowly with the increase in ε for almost all data sets considered for the experiment.

(ii) Observe that $\beta_{\mathcal{X}}$ was still positive when $\varepsilon \approx 5$ for some data sets, while $\beta_{\mathcal{X}}$ becomes negative for some data sets as soon as $\varepsilon \approx 1$.

So, $\beta_{\mathcal{X}}$ help us in showing the concentration effect for some particular settings. It does not measure the rate of concentration, this aspire us to define another index that can be more indicative.



(a) mfeat-fou data

(b) mfeat-fac data

(c) mfeat-kar data

(d) mfeat-mor data

Figure 14: Skewness Plot for Real Data sets as in Table 4 for different Distance Functions

0.10 Some Novel Empirical Indices to Measure Concentration

Through this section we want to define two new indices that are easy to compute, illustrates the rate of concentration and finally make us able to relate it to concentration function α_Ω. We will also discuss the properties of these indices and see the empirical results for different synthetic and real data sets.

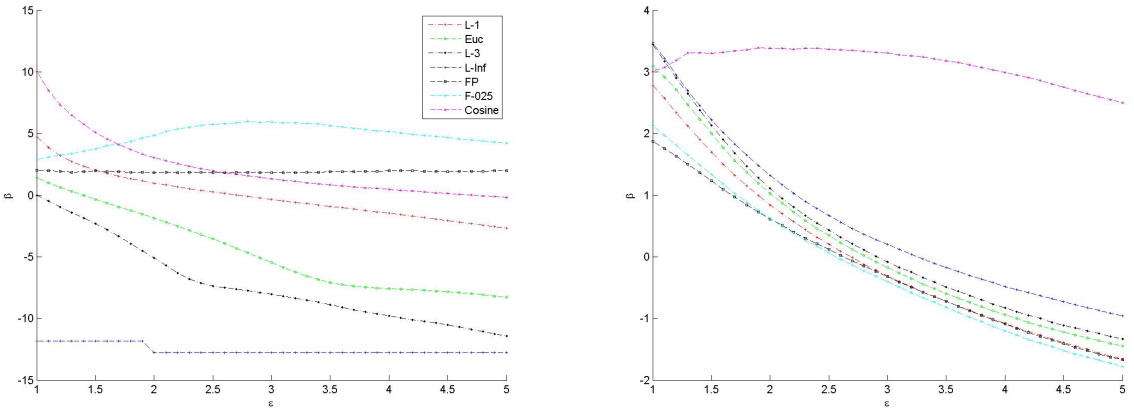
When we work with synthetic data sets, we have some underlying assumptions about the distribution of the datas that varies greatly when we change our domain from synthetic data to real data sets. Some of the assumptions are :

(i) Its been presumed that data are coming from an independent distribution.

(ii) Data points do not have much interaction among themselves i.e. the correlation between the points is almost negligible.

(iii) The intrinsic dimension of the data is not insignificant in front of the embedded dimension.

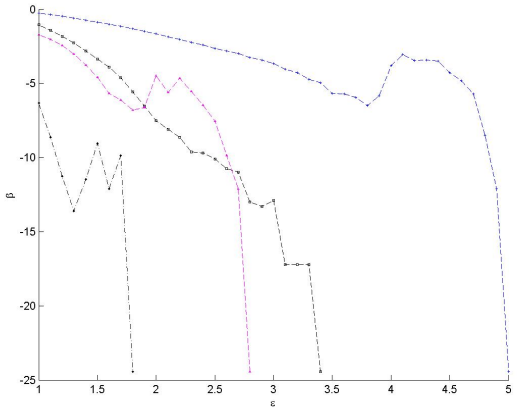
But all these assumptions fails to hold for many real data sets. As a result, $\beta_{\mathcal{X}}$ which appears to be an excellent index to measure concentration fails to follow the trend for real



(a) mfeat-pix data

(b) mfeat-zer data

Figure 15: Skewness Plot for Real Data sets as in Table 4 for different Distance Functions



(a) Madelon Valid Data

Figure 16: Skewness Plot for Real Data sets as in Table 4 for different Distance Functions

data sets. This poses the problem to another level to find or modify the index. This pushes us to find another index that can be applied to different data sets in empirical settings.

0.10.1 $C_g^*(x_i)$- Complement of the new index

In the previous section we were talking about the number of data points captured by the dilated $g-\delta_i$ sphere of a point x_i, but we could not come up with the results we wished for. *Whether discussing the other way round for $C_g(x_i)$ will help us?*

Consider $\mathcal{X}$ be the data set. Define $C_g^*(x_i)$ be the average number of points that the point x_i is not able to arrest through its $g-\delta$ neighborhood i.e.

$$C_g^*(x_i) = \frac{\sharp\mathcal{X} - C_g(x_i)}{\sharp\mathcal{X}} = 1 - \frac{C_g(x_i)}{n}$$

. We observe that if $C_g(x_i)$ is large for a point x_i then $C_g^*(x_i)$ will be small. It implies that vaguely we can say that $C_g^*(x_i)$ is inversely proportional to the concentration of points. Hence, small values of $C_g^*(x_i)$ for almost every data point means more concentrating of the points and vice versa.

One important thing to note that the above observation should be true for large number of points then only it will hold. Lets say C_g^* values is very large for 2 or 3 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 and thus we move to the better index λ .

Based on the previous section, we have a more successful index that will help us to accomplish our objective. C_g^* is an indicator related to only to a single point x_i, so we generalize it to all the data points and the indicator of the overall effect is what we call λ .

0.10.2 Nomenclature

Let $(\Omega, \mathcal{D}, \rho, \mu)$ be our similarity workload where $\mathcal{D} = (x_1, \ldots, x_n)$. Let $n = \sharp\mathcal{D}$, the number of data points in $\mathcal{D}$. Let us denote by

- $\bar{\mathbb{R}}_+$ the set of all non-negative reals, i.e., $\bar{\mathbb{R}}_+ = \mathbb{R}^+ \cup \{0\}$.
- μ_c is the counting measure, i.e., if $A \neq \emptyset$ then $\mu_c(A) = \sharp A$.
- If $\delta \in \bar{\mathbb{R}}_+$, then the δ neighbourhood of a point $x \in \mathcal{D}$ is given by

$$N_\delta(x) = N(x, \delta) = \{y \in \mathcal{X} : \rho(x, y) < \delta\}.$$

- $C(x, \delta) = \sharp\{N(x, \delta)\} = \mu_c(N(x, \delta)) = \mu_c(N_\delta(x))$.
- $C^*(x, \delta) = 1 - \frac{C(x,\delta)}{n} = 1 - \frac{\sharp\{N(x,\delta)\}}{n} = \frac{\mu_c(N(x,\delta))}{n} = \frac{\mu_c(N_\delta(x))}{n}$.
- An n-dimensional vector $\bar{\delta} \in \bar{\mathbb{R}}_+^n$ in terms of its components will be written as $\bar{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$, where $\bar{\mathbb{R}}_+^n$ denotes the n-dimensional Cartesian product of $\bar{\mathbb{R}}_+$.
- Let $\bar{\delta}, \bar{\gamma} \in \bar{\mathbb{R}}_+^n$. We say that $\bar{\delta} \leq \bar{\gamma}$ if $\delta_i \leq \gamma_i$ for all $i = 1, 2, \ldots, n$.
- If $\delta \in \bar{\mathbb{R}}_+$, then by $\widehat{\delta} = (\delta, \ldots, \delta) \in \bar{\mathbb{R}}_+^n$, we denote the n-dimensional vector with all identical components.

0.10.3 A New General Purpose Index : λ

Definition 0.10.1. *Let $\mathcal{D} = \{x_1, \ldots, x_n\}$ be the data set and μ_c the counting measure. We define a function $\lambda_{\mathcal{D}} : [-1, \infty) \times \bar{\mathbb{R}}^n_+ \to [0, 1]$ as follows:*

$$\lambda_{\mathcal{D}}(\varepsilon, \bar{\delta}) = \max_{x_i \in \mathcal{D}} \{C^*(x_i, (1+\varepsilon)\delta_i)\} \ , \tag{8}$$

where $\varepsilon \in [-1, \infty)$ and $\bar{\delta} = (\delta_1, \delta_2, \ldots, \delta_n) \in \bar{\mathbb{R}}^n_+$.

The following properties of λ are immediate:

Lemma 0.10.2. *Let $\lambda_{\mathcal{D}}$ be as defined in (8) of Definition 0.10.1.*

(i) Let $\bar{\delta}, \bar{\gamma} \in \bar{\mathbb{R}}^n_+$ be such that $\bar{\delta} \leq \bar{\gamma}$. Then $\lambda_{\mathcal{D}}(\varepsilon, \bar{\delta}) \geq \lambda_{\mathcal{D}}(\varepsilon, \bar{\gamma})$.

(ii) Let $\varepsilon, \varepsilon' \in [-1, \infty)$ such that $\varepsilon \leq \varepsilon'$. Then, $\lambda_{\mathcal{D}}(\varepsilon, \bar{\delta}) \geq \lambda_{\mathcal{D}}(\varepsilon', \bar{\delta})$.

In other words, $\lambda_{\mathcal{D}}$ is decreasing in both the variables.

Proof. (i) Let $x \in \mathcal{D}$ and $\varepsilon \in [-1, \infty)$ such that $(1+\varepsilon) > 0$.
Since $\bar{\delta} \leq \bar{\gamma}$, we have that $\delta_i \leq \gamma_i$, for $i = 1, 2, \ldots, n$. Hence, we have that

$$\begin{aligned}
(1+\varepsilon)\delta_i &\leq (1+\varepsilon)\gamma_i && (\forall i)\\
&\implies N(x, (1+\varepsilon)\delta_i) \subset N(x, (1+\varepsilon)\gamma_i) && (\forall i)\\
&\implies \sharp N(x, (1+\varepsilon)\delta_i) \leq \sharp N(x, (1+\varepsilon)\gamma_i) && (\forall i)\\
&\implies C(x, \delta_i) \leq C(x, \gamma_i) && (\forall i)\\
&\implies 1 - \frac{C(x, \delta_i)}{n} \geq 1 - \frac{C(x, \gamma_i)}{n} && (\forall i)\\
&\implies C^*(x, \delta_i) \geq C^*(x, \gamma_i) && (\forall i)\\
&\implies \max_{x \in \mathcal{D}} \{C^*(x, \delta_i)\} \geq \max_{x \in \mathcal{D}} \{C^*(x, \gamma_i)\}\\
&\implies \lambda_{\mathcal{D}}(\varepsilon, \bar{\delta}) \geq \lambda_{\mathcal{D}}(\varepsilon, \bar{\gamma})
\end{aligned}$$

(ii) Let $\bar{\delta} \in \bar{\mathbb{R}}^n_+$ and $x \in \mathcal{D}$.
Since $\varepsilon \leq \varepsilon'$

$$\begin{aligned}
&\implies (1+\varepsilon)\delta_i \leq (1+\varepsilon')\delta_i && (\forall i)\\
&\implies N(x, (1+\varepsilon)\delta i) \leq N(x, (1+\varepsilon')\delta_i) && (\forall i)\\
&\implies C(x, 1+\varepsilon)\delta_i) \leq C(x, (1+\varepsilon')\delta_i) && (\forall i)\\
&\implies 1 - \frac{C(x, (1+\varepsilon)\delta_i)}{n} \geq 1 - \frac{C((1+\varepsilon')\delta_i)}{n} && (\forall i)\\
&\implies C^*(x, (1+\varepsilon)\delta_i) \geq C^*(x, (1+\varepsilon')\delta_i) && (\forall i)\\
&\implies \max_{x \in \mathcal{D}} \{C^*(x, (1+\varepsilon)\delta_i)\} \geq \max_{x \in \mathcal{D}} \{C^*(x, (1+\varepsilon')\delta_i)\}\\
&\implies \lambda(\varepsilon, \bar{\delta}) \geq \lambda(\varepsilon', \bar{\delta})
\end{aligned}$$

Hence Proved. □

0.10.4 Two Specific Measures based on λ : $\widetilde{\lambda}_{\mathcal{X}}$ and $\widehat{\lambda}_{\mathcal{X}}$

We introduce two new restricted functions based on λ and show how they help us in measuring the concentration efficiently.

As soon as we fix $\mathcal{X}$ to be our data set, the distance of one point to other is fixed and so λ become a function in one variable.

Definition 0.10.3. *We define the nearest neighbor distance vector as :*

(i) $\widetilde{\delta} = (\delta_1, \ldots, \delta_n) \in \bar{\mathcal{R}}_+^n$, *where* δ_i *is the distance of point* x_i *to the point that is closest to it.*

(ii) Let $\delta_0 = \max\limits_{x_i \in \mathcal{X}}\{\delta_i\}$. *Then,* $\widehat{\delta_0} = (\delta_0, \ldots, \delta_0)$.

On the basis of above nearest neighbor distance vector, we introduced two new functions by restricting λ on $\mathcal{X}$.

Definition 0.10.4. *Defining two functions* $\widetilde{\lambda}_{\mathcal{X}}, \widehat{\lambda}_{\mathcal{X}} : [-1, \infty) \to [0, 1]$ *as follows:*

$$\widetilde{\lambda}_{\mathcal{X}}(r_i) = \lambda(\varepsilon, \widetilde{\delta}) = \max_{x_i \in \mathcal{X}}\{C^*(x_i, (1+\varepsilon)\delta_i\} \tag{9}$$

$$\widehat{\lambda}_{\mathcal{X}}(r) = \lambda(\varepsilon, \widehat{\delta_0}) = \max_{x_i \in \mathcal{X}}\{C^*(x_i, (1+\varepsilon)\delta_0\} \ , \tag{10}$$

where $r = (1+\varepsilon)\delta_0$ *and* $r_i = (1+\varepsilon)\delta_i$.

Note that:

(i) For a fixed $\varepsilon \in [-1, \infty)$, $\widetilde{\lambda}_{\mathcal{X}} = \widehat{\lambda}_{\mathcal{X}}$ if $\widetilde{\delta} = \widehat{\delta}$.

(ii) The motivation for defining $\widehat{\lambda}_{\mathcal{X}}$ is Theorem 0.10.7.

From Lemma 0.10.2, the following result is straight forward:

Corollary 0.10.5. *If* $\widetilde{\lambda}_{\mathcal{X}}$ *and* $\widehat{\lambda}_{\mathcal{X}}$ *are the indices to measure concentration defined as above then* $\widetilde{\lambda}_{\mathcal{X}}$ *and* $\widehat{\lambda}_{\mathcal{X}}$ *are decreasing functions, i.e., given* $r_1 \le r_2$ *for* $r_1, r_2 \in [0, \infty)$, $\widetilde{\lambda}_{\mathcal{X}}(r_1) \ge \widetilde{\lambda}_{\mathcal{X}}(r_2)$ *and* $\widehat{\lambda}_{\mathcal{X}}(r_1) \ge \widehat{\lambda}_{\mathcal{X}}(r_2)$.

Lemma 0.10.6. *Let* $\widetilde{\lambda}_{\mathcal{X}}, \widehat{\lambda}_{\mathcal{X}}$ *be the indices to measure concentration, then*

$$\widetilde{\lambda}_{\mathcal{X}} \ge \widehat{\lambda}_{\mathcal{X}} \quad \text{for fixed } \varepsilon \in [-1, \infty)$$

Proof. Let $\widetilde{\delta} = (\delta_1, \ldots, \delta_n)$ then $\delta_0 = \max\{\delta_1, \ldots, \delta_n\}$ and hence $\widehat{\delta_0} = (\delta_0, \ldots, \delta_0)$.
Let $\varepsilon \ge -1$ be any real number.
As, $\delta_i \le \delta_0$ for $i = 1 \ldots n$

$$\begin{aligned}
&\implies (1+\varepsilon)\delta_i \le (1+\varepsilon)\delta_0 && (\forall i)\\
&\implies C(x, (1+\varepsilon)\delta_i) \le C(x, (1+\varepsilon)\delta_0) && (\forall i)\\
&\implies 1 - \frac{C(x, (1+\varepsilon)\delta_i)}{n} \ge 1 - \frac{C(x, (1+\varepsilon)\delta_0)}{n} && (\forall i)\\
&\implies C^*(x, (1+\varepsilon)\delta_i) \ge C^*(x, (1+\varepsilon)\delta_0) && (\forall i)\\
&\implies \min_{x_i \in \mathcal{X}}\{C^*(x, (1+\varepsilon)\delta_i)\} \ge \min_{x_i \in \mathcal{X}}\{C^*(x, (1+\varepsilon)\delta_0)\}\\
&\implies \widetilde{\lambda}_{\mathcal{X}}(r) \ge \widehat{\lambda}_{\mathcal{X}}(r)
\end{aligned}$$

Since ε is arbitrary, so $\widetilde{\lambda}_{\mathcal{X}} \ge \widehat{\lambda}_{\mathcal{X}}$. Hence proved.

□

0.10.5 $\alpha_{\mathcal{X}}$ vs $\widetilde{\lambda}_{\mathcal{X}}, \widehat{\lambda}_{\mathcal{X}}$

Comparison between $\widetilde{\lambda}_{\mathcal{X}}, \widehat{\lambda}_{\mathcal{X}}$ and $\alpha_{\mathcal{X}}$:

(i) $\alpha_{\mathcal{X}}$ is a purely theoretical index so calculating it even for a smaller set is very cumbersome. Earlier also we said that if the data set is high dimensional then $\alpha_{\mathcal{X}}$ is very improper for empirical settings. This is the advantage our index $\widetilde{\lambda}_{\mathcal{X}}$ and $\widehat{\lambda}_{\mathcal{X}}$ gives over other $\alpha_{\mathcal{X}}$. Given a small set we can easily find the $\widetilde{\lambda}_{\mathcal{X}}$ with paper and pen.

(ii) $\widetilde{\lambda}_{\mathcal{X}}$ is computationally efficient than $\alpha_{\mathcal{X}}$. Recalling 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 while to evaluate $\widetilde{\lambda}_{\mathcal{X}}$ or $\widehat{\lambda}_{\mathcal{X}}$ we just need to work with n subsets. In simpler way, to find $\widetilde{\lambda}_{\mathcal{X}}$ only n nearest neighbor distances are evaluated instead of searching for the subsets that weighs at least half the total weight.

(iii) Since min,max and median is an statistical tool so the theoretical studies can also be done smoothly by all these tools. They exihibits resembling results only.

From the above comparison, we get the impression that $\widetilde{\lambda}_{\mathcal{X}}$ and $\widehat{\lambda}_{\mathcal{X}}$ may be more indicative and sharper than $\alpha_{\mathcal{X}}$. Although $\alpha_{\mathcal{X}}$ is a strong index to work theoretically but it is very complex for experimental studies. *We were curious whether we can give any relation between* $\widetilde{\lambda}_{\mathcal{X}}$, $\widehat{\lambda}_{\mathcal{X}}$ *and* $\alpha_{\mathcal{X}}$. We are now stating a result that will show theoretically that $\widetilde{\lambda}_{\mathcal{X}}$ and $\widehat{\lambda}_{\mathcal{X}}$ are indeed more supreme to $\alpha_{\mathcal{X}}$ in experimental sense.

Theorem 0.10.7. *Let $(\Omega, \mathcal{X}, \rho, \mu)$ be a given similarity workload. Let $\varepsilon \in [-1, \infty)$ and $\widetilde{\delta}, \delta_0$ be as defined in Definition 0.10.3. Let us denote by $r = (1+\varepsilon)\delta_0$ and let $r_i = (1+\varepsilon)\delta_i$. Then,*

$$\alpha_{\mathcal{X}}(r) \leq \widehat{\lambda}_{\mathcal{X}}(r) \leq \widetilde{\lambda}_{\mathcal{X}}(r) \leq \widetilde{\lambda}_{\mathcal{X}}(r_i) \ . \tag{11}$$

Proof. We prove this theorem in three steps, proving each inequality at every step.

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 the domain of α_Ω is $\overline{\mathbb{R}}_+$ and is well-defined.

$\underline{\alpha_{\mathcal{X}} \leq \widehat{\lambda}_{\mathcal{X}}}$:

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.,

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

Also, the r-neighborhood of A for $r \geq 0$ is defined as :

$$A_r = \{x \in \mathcal{X} : \rho(x, a) \leq r \ for \ any \ a \in A\}$$

Since $(1+\varepsilon)\delta_i \leq (1+\varepsilon)\delta_0$ for every $\varepsilon \in [-1, \infty)$ and $i = 1, 2, \ldots, n$. Therefore, for any arbitrary but fixed $A \in \mathcal{A}$ and for every $x_i \in A$, we have

$$\begin{aligned} N(x_i, (1+\varepsilon)\delta_i) \subset N(x_i, (1+\varepsilon)\delta_0) \subset A_r & \\ \implies \mu_c(N(x_i, (1+\varepsilon)\delta_i)) \leq \mu_c(N(x_i, (1+\varepsilon)\delta_0)) \leq \mu_c(A_r) & \quad (\forall i) \\ \implies C(x_i, (1+\varepsilon)\delta_i) \leq C(x_i, (1+\varepsilon)\delta_0) \leq \mu_c(A_r) & \quad (\forall i) \\ \implies \mathcal{C}_A = \min_{x_i \in A} C(x_i, (1+\varepsilon)\delta_0) \leq \mu_c(A_r) \ . & \end{aligned}$$

Now, since

$$\inf_{A\in\mathcal{A}} \mathcal{C}_A = \inf_{A\in\mathcal{A}} \left\{ \min_{x_i\in A} C(x_i, (1+\varepsilon)\delta_0) \right\} = \min_{x_i\in\mathcal{X}} C(x_i, (1+\varepsilon)\delta_0) \ ,$$

we have the following implications:

$$\begin{aligned}
\min_{x_i\in\mathcal{X}} C(x_i, (1+\varepsilon)\delta_0) \leq \inf_{A\in\mathcal{A}} \ & \mu_c(A_r) \\
\implies 1 - \left(\frac{\min_{x_i\in\mathcal{X}} C(x_i, (1+\varepsilon)\delta_0)}{n} \right) & \geq 1 - \inf_{A\in\mathcal{A}} \{\mu_c(A_r)\} \\
\implies \max_{x_i\in\mathcal{X}} \left(1 - \frac{C(x_i, (1+\varepsilon)\delta_0)}{n} \right) & \geq \sup_{A\in\mathcal{A}} \{\mu_c(A_r^c)\} \\
\implies \max_{x_i\in\mathcal{X}} \{C^*(x_i, (1+\varepsilon)\delta_0)\} & \geq \sup_{A\in\mathcal{A}} \{\mu_c(A_r^c)\} \\
\implies \max_{x_i\in\mathcal{X}} \{C^*(x_i, r)\} & \geq \sup_{A\in\mathcal{A}} \{\mu_c(A_r^c)\} \\
\implies \widehat{\lambda}_{\mathcal{X}}(r) & \geq \alpha_{\mathcal{X}}(r) \ ,
\end{aligned}$$

where μ_c is the normalized measure of μ.

$\underline{\widehat{\lambda}_{\mathcal{X}} \leq \widetilde{\lambda}_{\mathcal{X}}}$**:** Second inequality is just Lemma 0.10.6.

$\underline{\widetilde{\lambda}_{\mathcal{X}}(r) \leq \widetilde{\lambda}_{\mathcal{X}}(r_i)}$**:** Since $\widetilde{\delta} \leq \widehat{\delta}_0$
$\implies (1+\epsilon)\delta_i \leq (1+\epsilon)\delta_0$ for any $\varepsilon \in [-1, \infty)$ and for every $i = 1 \ldots n$
$\implies r_i \leq r$ for any $\varepsilon \in [-1, \infty)$ and for every $i = 1 \ldots n$
Also, from Corollary 0.10.5 we know that $\widetilde{\lambda}$ is a decreasing function and hence $\widetilde{\lambda}(r) \leq \widetilde{\lambda}(r_i)$ for any $\varepsilon \in [-1, \infty)$.

□

From Thoerem 0.10.7, we have $\alpha_\Omega(r) \leq \widetilde{\lambda}_{\mathcal{X}}(r)$. As a result, $\widetilde{\lambda}_{\mathcal{X}}$ forms an upper bound for α_Ω. Hence if $\widetilde{\lambda}_{\mathcal{X}}$ is itself very small in magnitude for a given dataset then α_Ω will be small and concentration will be very large. Although, $\widetilde{\lambda}_{\mathcal{X}}$ is an upper bound for $\alpha_{\mathcal{X}}$ but if $\widetilde{\lambda}_{\mathcal{X}}$ is very large then we cannot say anything for $\alpha_{\mathcal{X}}$. Certainly a more tighter bound for $\alpha_{\mathcal{X}}$ will help us, so this was the motivation behind defining $\widehat{\lambda}_{\mathcal{X}}$ as it will be a more narrower bound than $\widetilde{\lambda}_{\mathcal{X}}$ for $\alpha_{\mathcal{X}}$ (see Thoerem 0.10.7) .

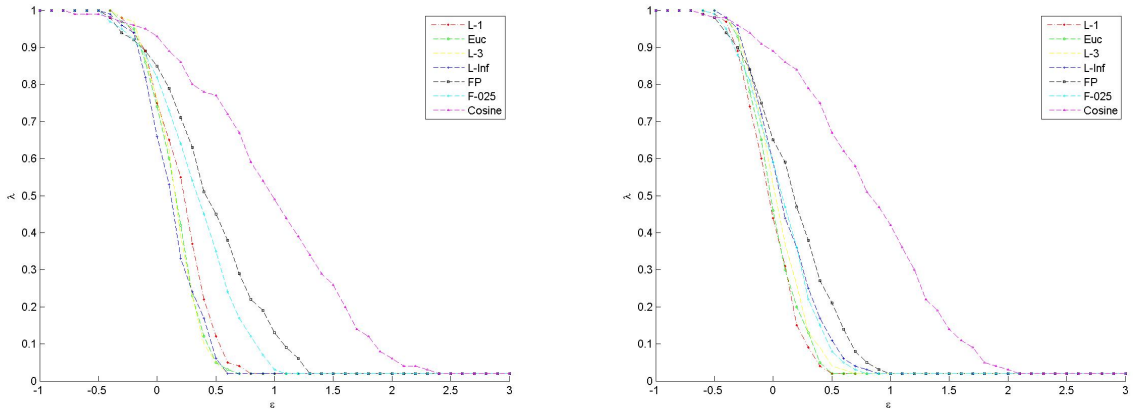
0.10.6 Are these indices really useful?

Though we introduced a bunch of indices one after other but *whether they serve our purpose?* Whether they just show concentration or really measure the rate of concentration? If yes, then whether they really measure concentration well? Will we get the freedom to classify a good distance function from a bad distance function on the basis of these indices. Note that, by a good distance function we mean a distance function that concentrates less i.e. comparatively it can distinguish points better than other distance functions. Can we talk of stability of Similarity workloads with respect to these indices? and many more...
To discuss the above questions and check their performance, we divide the studies related to these indices into two groups:

(i) Studies on Synthetic Data sets.

S.No.	Number of datapoints(N)	Dimension(m)
1	100	10
2	1000	100
3	10000	10000

Table 5: Features of the synthetic data set



(a) Uniform

(b) Gaussian

Figure 17: Plot for $\widehat{\lambda}_{\mathcal{X}}$ for Uniform and Gaussian distribution with $N = 100$ and $m = 10$

(ii) Studies on Real Data sets.

Studies on Synthetic Datasets

We generated some synthetic data sets following a particular distribution and computed the $\widehat{\lambda}_{\mathcal{X}}$ values and plotted the curves for different distance functions. Then we ran k-NN classification on the already generated Dataset and check if there is any correlation between the $\widehat{\lambda}_{\mathcal{X}}$ values and number of mismatches in k-NN. To see the behavior of $\widehat{\lambda}_{\mathcal{X}}$, we generated n points in m dimension in the interval [-1,1] such that data is either coming from Uniformly distribution or Gaussian distribution. Then we computed the pairwise distances and calculated the $\widehat{\lambda}_{\mathcal{X}}$ values for each distance functions.

The following are the parameters that we used during the computation of $\widehat{\lambda}_{\mathcal{X}}$:

(i) m- dimension of the data.

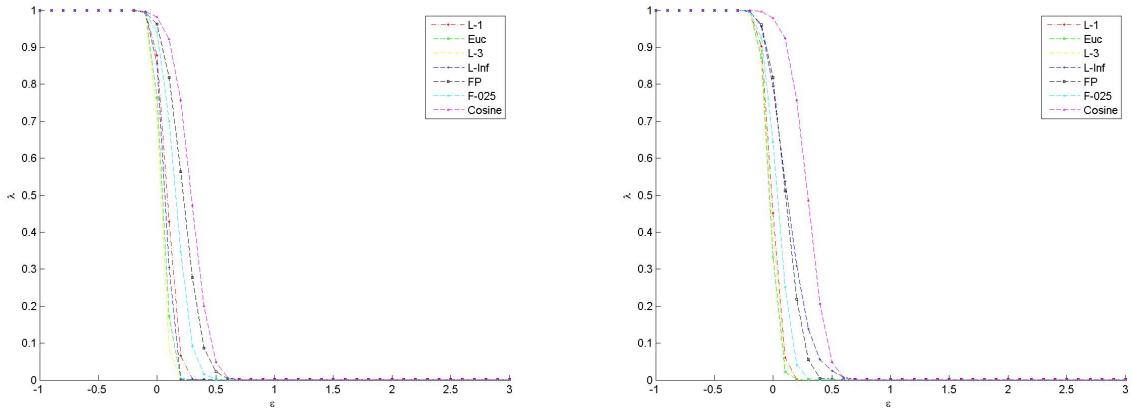
(ii) n - Number of data point. Usually $n = 10 * m$.

(iii) Different Distance Functions :

(a) Minkowski distance function for $p = 0.04, 0.25, 1, 2, 3$.

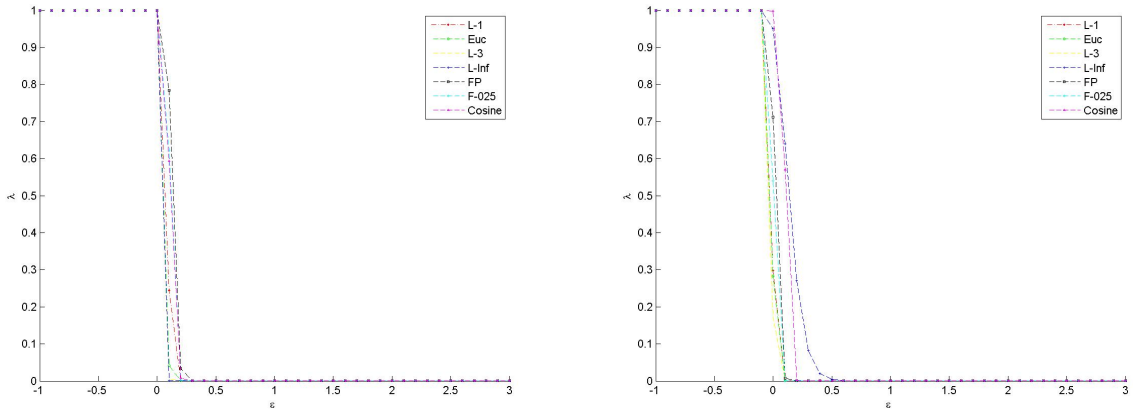
(b) Cosine distance function for $p = 2$.

(iv) Distribution of the data set : We have generated Synthetic data from both Uniform and Gaussian Distribution to find that not much deviation can be seen.



(a) Uniform (b) Gaussian

Figure 18: Plot for $\widehat{\lambda}_{\mathcal{X}}$ for Uniform and Gaussian distribution with $N = 1000$ and $m = 100$



(a) Uniform (b) Gaussian

Figure 19: Plot for $\widehat{\lambda}_{\mathcal{X}}$ for Uniform and Gaussian distribution with $N = 10000$ and $m = 1000$

Table-5 gives the abstracted form of the features of the data set that we used during our experiment.

Fig 17 , Fig 18 and Fig 19 shows the graph for different $\widehat{\lambda}_{\mathcal{X}}$ values vs ε . The following are the observations made by us from Fig 17 , Fig 18 and Fig 19.

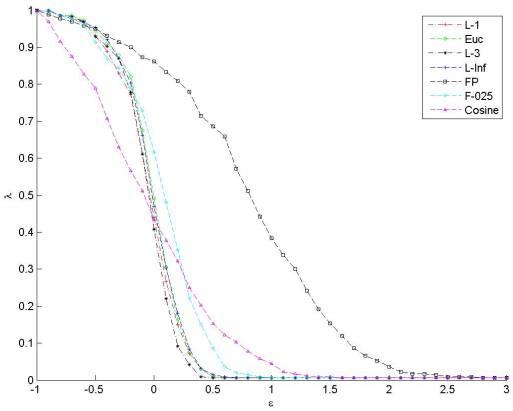
(i) $\widehat{\lambda}_{\mathcal{X}}$ is a decreasing function as expected.

(ii) $\widehat{\lambda}_{\mathcal{X}}$ starts at 1 and slowly dies off to 0 as the epsilon increases.

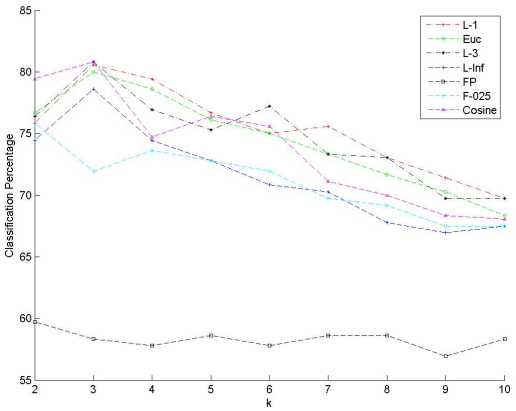
(iii) The rate of falling of $\widehat{\lambda}_{\mathcal{X}}$ does indicates the rate of concentration. The faster it falls, the more is the concentration.

(iv) We can easily identify the good distance function from the bad distance function from the figure **??**.

(v) Cosine and Fractional distance functions are clearly emerging as the better distance functions.



(a) $\widehat{\lambda}_{\mathcal{X}}$ for Movement Libras data



(b) Classification rate for Movement Libras data

Figure 20: Plots for $\widehat{\lambda}_{\mathcal{X}}$ and K-NN classification for data as described in Table 6

Dataset	m	N
Movement Libras Data	90	360
Isolet	618	7797
Gas Sensor Data	24	5456
Madelon Training Data	2000	500
Multiple Features with correlation coefficients(mfeat-fac)	216	2000
Multiple Features with Fourier coefficients(mfeat-fou)	76	2000
Multiple Features with Karhumen-Love coefficients(mfeat-kar)	64	2000
Multiple Features with Morphological Features(mfeat-mor)	6	2000
Multiple Features with Pixel Features(mfeat-pix)	240	2000
Multiple Features with Zernike moments Features(mfeat-zer)	47	2000

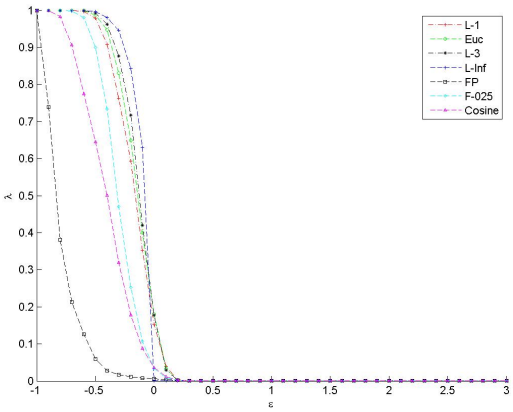
Table 6: Real Data sets

Thus, we see that $\widehat{\lambda}_{\mathcal{X}}$ is amenable to our analysis for synthetic data set. Note that, $\widehat{\lambda}_{\mathcal{X}}$ is just an upper bound for $\alpha_{\mathcal{X}}$, so we can only talk about the worst case for distance functions, saying this means if the $\widehat{\lambda}_{\mathcal{X}}$ values for a particular distance functions is very small then surely the distances are concentrated for that distance function.

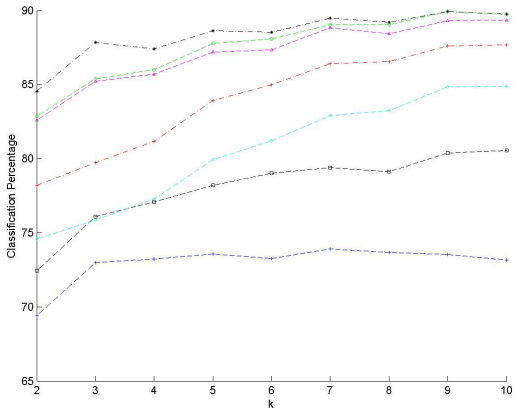
Studies on Real Datasets

In the previous section, we justified our intuition for synthetic data sets that $\widehat{\lambda}_{\mathcal{X}}$ is indeed a better index to measure the rate of concentration, where the synthetic data sets comes with two of the main amenities: firstly the distribution of the sets was known to us and secondly the data were independently generated. But when we change our domain to real data sets, these amenities are lost and we have to work on these data sets without knowing its most of the properties.
We did the same analysis here also, picking a real data set, computing its $\widehat{\lambda}_{\mathcal{X}}$ values for different distance functions and plotting the graph of the computed $\widehat{\lambda}_{\mathcal{X}}$ values and then running k-NN classification for the same data set and inspecting for any relationship between $\widehat{\lambda}_{\mathcal{X}}$ values and number of mismatches in k-NN. We further examine the classification rate for different distance functions.

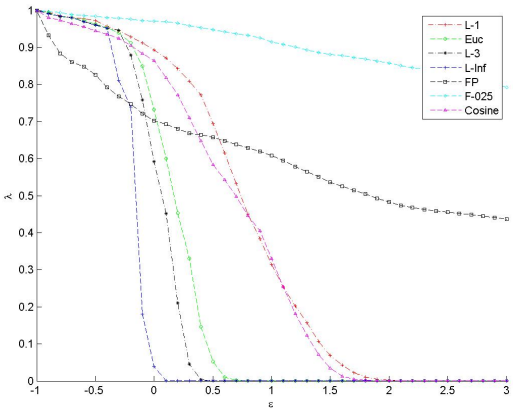
Table 6 give a brief introduction to the features of the Real Data sets used for our studies on $\widehat{\lambda}_{\mathcal{X}}$.



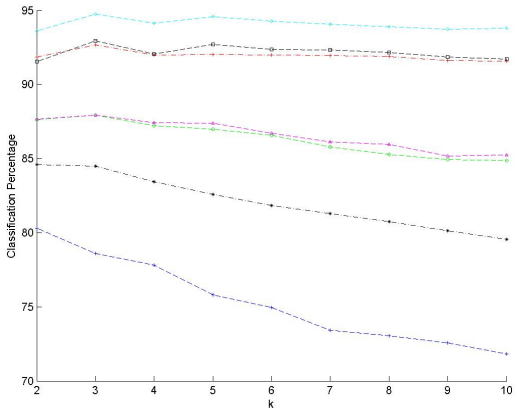
(a) $\widehat{\lambda}_{\mathcal{X}}$ for Isolet Data



(b) Classification rate for Isolet data



(c) $\widehat{\lambda}_{\mathcal{X}}$ for Sensor Readings Data



(d) Classification rate for Sensor Readings data

Figure 21: Plots for $\widehat{\lambda}_{\mathcal{X}}$ and K-NN classification for data as described in Table 6

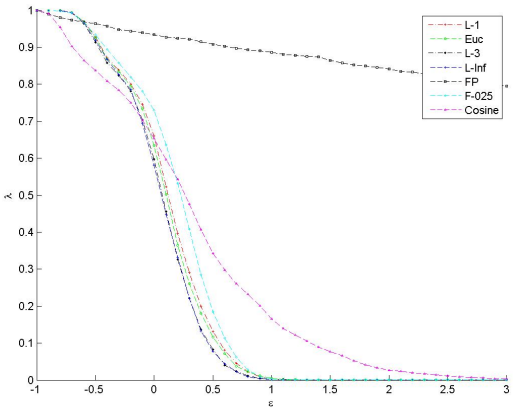
Observations made from fig 20, 21 , 22 , 23 and 24 are listed as:

(i) We observe that $\widehat{\lambda}_{\mathcal{X}}$ is a decreasing function. For most of the real data sets, $\widehat{\lambda}_{\mathcal{X}}$ goes to 0 as $\varepsilon \approx 0$.

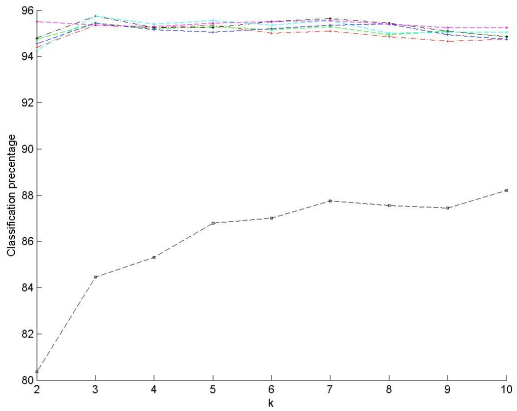
(ii) $\widehat{\lambda}_{\mathcal{X}}$ clearly measures the rate of concentration. The faster it falls the stronger is the concentration.

0.11 Conclusion

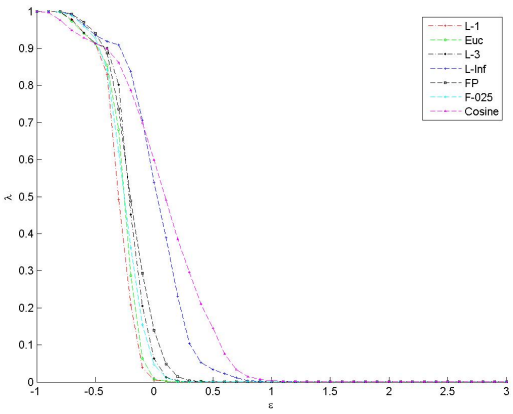
In this work, we attempted to analyze two new distance function namely, $\mathcal{J}_p$ and $\mathcal{K}_p$ with respect to existing indices Relative Contrast and Relative Variance and further we went on to find a new index called $\widehat{\lambda}_{\mathcal{X}}$ that can measure the concentration empirically and as well as theoretically. Also, we have proved it theoretically that $\widehat{\lambda}_{\mathcal{X}}$ can be used as a measure to measure concentration. Then we presented some experimental results for $\widehat{\lambda}_{\mathcal{X}}$ to validate our findings.



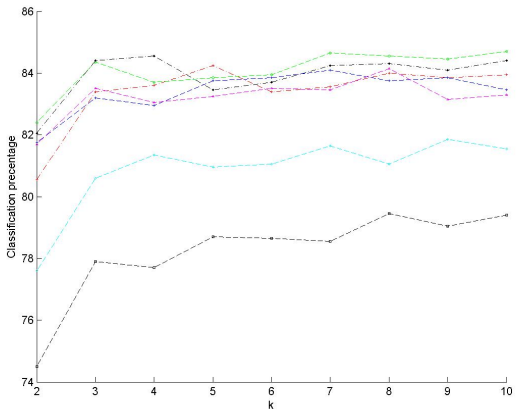
(a) $\widehat{\lambda}_{\mathcal{X}}$ for mfeat-fac Data



(b) Classification rate for mfeat-fac data



(c) $\widehat{\lambda}_{\mathcal{X}}$ for mfeat-fou Data

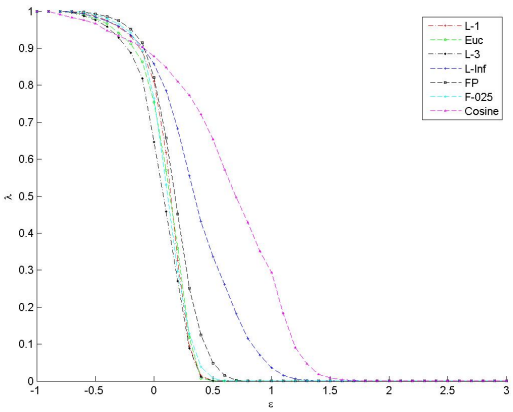


(d) Classification rate for mfeat-fou data

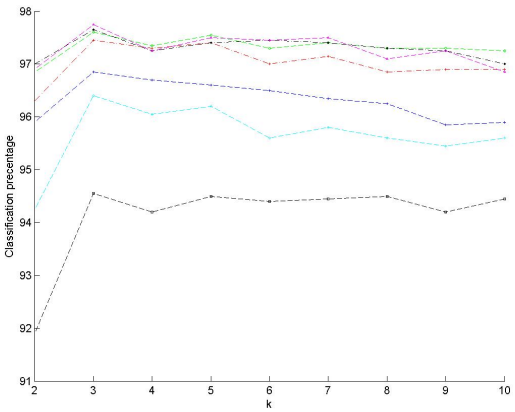
Figure 22: Plots for $\widehat{\lambda}_{\mathcal{X}}$ and K-NN classification for data as described in Table 6

So far the theory of concentration of norms has been well studied and explored but always in a non-positive way. From Sections 0.4.2 and 0.4.3, we see that Euclidean norms and other Minkowski-type norms do not behave well in high dimension. In fact, we have that all the Minkowski-type norms concentrate.

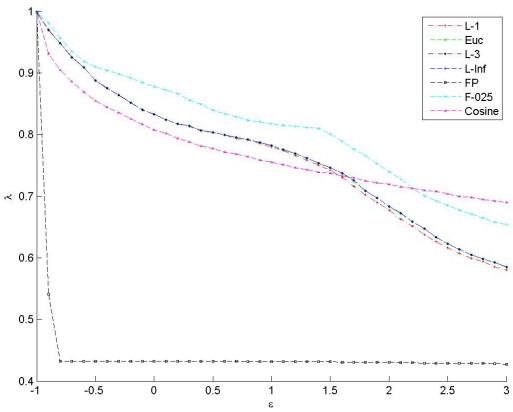
So, our future work includes studying the concentration in somewhat positive sense. Instead of discussing when and whether a distance function concentrates, we would like to investigate the stability of norms, i.e., when can we say that norms are stable even in high dimensions. In other words we would like to determine similarity workloads that are stable. Very few works have been done along such lines, see for instance, [Durrant and Kabán(2009)], [Bennett et al.(1999)Bennett, Fayyad, and Geiger], where the investigations focus on for what type of distributions the Minkowski norms do not concentrate. We would like to extend this to more general settings and applications.



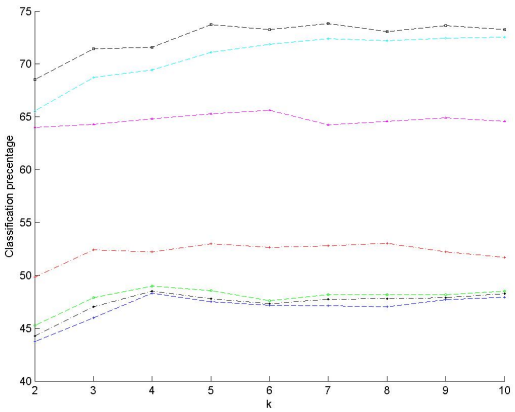
(a) $\widehat{\lambda}_{\mathcal{X}}$ for mfeat-kar Data



(b) Classification rate for mfeat-kar data

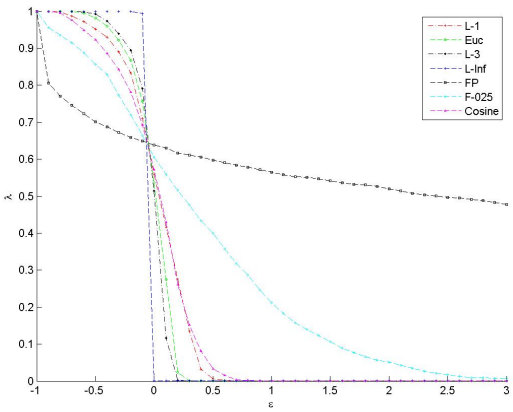


(c) $\widehat{\lambda}_{\mathcal{X}}$ for mfeat-mor Data

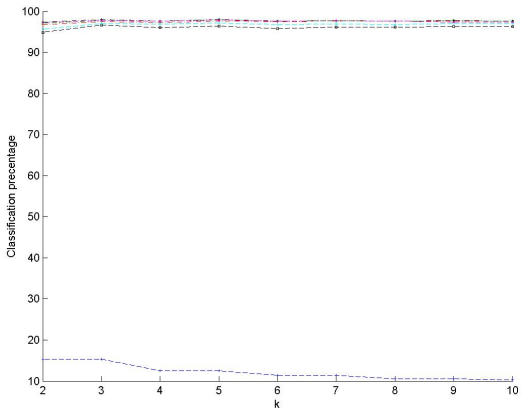


(d) Classification rate for mfeat-mor data

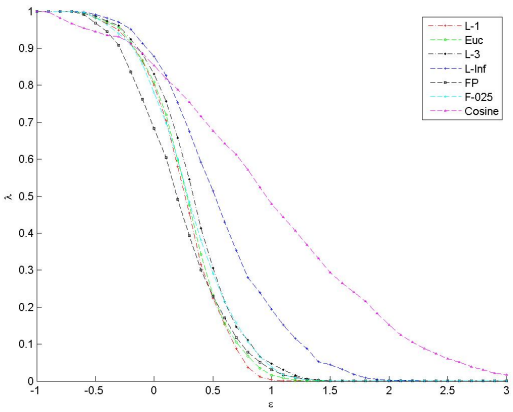
Figure 23: Plots for $\widehat{\lambda}_{\mathcal{X}}$ and K-NN classification for data as described in Table 6



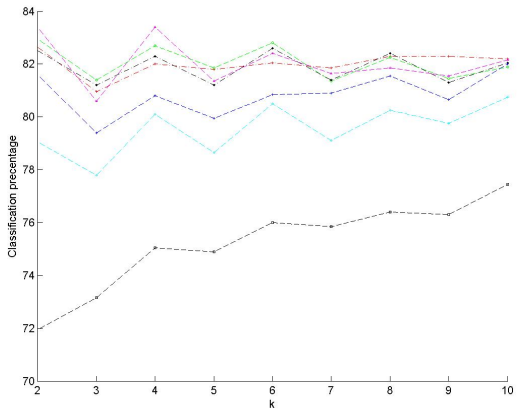
(a) $\widehat{\lambda}_{\mathcal{X}}$ for mfeat-pix Data



(b) Classification rate for mfeat-pix data



(c) $\widehat{\lambda}_{\mathcal{X}}$ for mfeat-zer Data



(d) Classification rate for mfeat-zer data

Figure 24: Plots for $\widehat{\lambda}_{\mathcal{X}}$ and K-NN classification for data as described in Table 6

Bibliography

[Aggarwal et al.(2001)Aggarwal, Hinneburg, and Keim] Aggarwal, C. C., Hinneburg, A., Keim, D. A., 2001. On the surprising behavior of distance metrics in high dimensional spaces. In: Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings. pp. 420–434.

[Bennett et al.(1999)Bennett, Fayyad, and Geiger] Bennett, K. P., Fayyad, U. M., Geiger, D., 1999. Density-based indexing for approximate nearest-neighbor queries. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 15-18, 1999. pp. 233–243.

[Beyer et al.(1999)Beyer, Goldstein, Ramakrishnan, and Shaft] Beyer, K. S., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is "nearest neighbor" meaningful? In: Database Theory - ICDT '99, 7th International Conference, Jerusalem, Israel, January 10-12, 1999, Proceedings. pp. 217–235.

[Demartines(1994)] Demartines, P., 1994. Analyse de donné,es par ré,seaux de neurones auto-organisé,s. PhD dissertation, Institut Nat',l Polytechnique de Grenoble, Grenoble, France.

[Durrant and Kabán(2009)] Durrant, R. J., Kabán, A., 2009. When is 'nearest neighbour' meaningful: A converse theorem and implications. J. Complexity 25 (4), 385–397.

[François et al.(2007)François, Wertz, and Verleysen] François, D., Wertz, V., Verleysen, M., 2007. The concentration of fractional distances. IEEE Trans. Knowl. Data Eng. 19 (7), 873–886.

[Hinneburg et al.(2000)Hinneburg, Aggarwal, and Keim] Hinneburg, A., Aggarwal, C. C., Keim, D. A., 2000. What is the nearest neighbor in high dimensional spaces? In: VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt. pp. 506–515.

[Jayaram and Klawonn(2012)] Jayaram, B., Klawonn, F., 2012. Can unbounded distance measures mitigate the curse of dimensionality? IJDMMM 4 (4), 361–383.

[Pestov(2000)] Pestov, V., 2000. On the geometry of similarity search: Dimensionality curse and concentration of measure. Inf. Process. Lett. 73 (1-2), 47–51.

[Pestov(2013)] Pestov, V., 2013. Is the kk-nn classifier in high dimensions affected by the curse of dimensionality? Computers & Mathematics with Applications 65 (10), 1427–1437.

[Radovanovic et al.(2010)Radovanovic, Nanopoulos, and Ivanovic] Radovanovic, M., Nanopoulos, A., Ivanovic, M., 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research 11, 2487–2531.