Can unbounded distance measures mitigate the curse of dimensionality?

Balasubramaniam Jayaram

Department of Mathematics
Indian Institute of Technology Hyderabad
Yeddumailaram - 502 205, India
E-mail: jbala@iith.ac.in

Frank Klawonn*

Department of Computer Science
Ostfalia University of Applied Sciences
D-38302 Wolfenbuettel, Germany
E-mail: f.klawonn@ostfalia.de
and
Bioinformatics and Statistics
Helmholtz Centre for Infection Research
D-38124 Braunschweig, Germany
E-mail: frank.klawonn@helmholtz-hzi.de
*Corresponding author

Abstract: In this work, we revisit the Curse of Dimensionality, especially the concentration of the norm phenomenon which is the inability of distance functions to separate points well in high dimensions. We study the influence of the different properties of a distance measure, viz., triangle inequality, boundedness and translation invariance, on this phenomenon. Our studies indicate that unbounded distance measures whose expectations do not exist are to be preferred. We propose some new distance measures based on our studies and present many experimental results which seem to confirm our analysis. In particular, we study these distance measures w.r.t. indices like Relative Variance and Relative Contrast and further compare and contrast these measures in the setting of nearest neighbour/proximity searches and hierarchical clustering.

Keywords: Curse of dimensionality; nearest neighbour classifier; cluster analysis.

Reference A full version of this work can be found at doi: 10.1504/ijdmmm.2012.049883.

1 Introduction

The concept of "nearness" is all pervasive and important in every field of knowledge and sphere of human activity. Be it to define a mathematical concept like continuity or to match faces in an identity parade or to express the feeling of liking between human beings or to determine the closeness between points in a cluster or to find an approximate match to a given query in a database. Central to any evaluation of the nearness of two objects lies a similarity, or equivalently, its dual distance measure. While a norm or even a metric does the job admirably in mathematical settings, one has to resort to more subjective measures in other areas. However, one area which has attracted a lot of attention recently because of the difficulties in measuring this concept of "nearness" is the setting of high dimensional spaces.

1.1 The curse of dimensionality

Recently many works have dealt with what has now come to be called the "*Curse of Dimensionality*" (CoD). The term presently connotes two different phenomena whose effects are typically seen when one deals with high dimensional spaces. This term was firstly introduced by Bellmann [1961] to refer to the combinatorial explosion in the number of variables in optimisation problems involving high dimensions. Recently, this term has also been used to refer to the degration in the effectiveness of methods employed in similarity searches, clustering and indexing in high dimensional spaces – typically the dimension is in the order of 100s. In this work, we deal with CoD in the context of the latter interpretation.

Research on this topic over the last decade and more have attributed this effect largely to the following:

(i) The *intrinsic dimension* of the data which can lie on a manifold whose dimension is far less than the space in which the data reside. For instance, the data from a 10-dimensional space can all lie on a straight line and hence its intrinsic dimension is just 1. For more details, see the works of Pestov [2000, 2007, 2008] and the references therein.

(ii) The inability of the *distance functions* to separate points well in high dimensions. This inability of a distance measure manifests itself, rather unpleasantly, in nearest neighbourhood algorithms, clustering schemes, query searching in data bases with high dimensionality and often leads to 'instabilities' or convergence to sub-optimal solutions.

The scope of this work is restricted to dealing with the second of the above two factors.

1.2 Distance measures and their concentration

The "*Concentration of the Norm*" (CoN) phenomenon, in some sense, refers to the concentration of points and hence their distances, which is to say, that as the dimension n increases the distances between a query point and its nearest and farthest neighbours are no more significantly different.

Studies on the influence of distance functions on CoD can be broadly classified into those that:

(i) determine the conditions on the data distribution and the properties of the distance measures which lead to unstable situations (see, for instance, Demartines [1994], Beyer et al. [1999], Durrant and Kabán [2009]),

(ii) analyze existing and/or new distance measures with respect to some indices Hinneburg et al. [2000], Aggarwal et al. [2001], François et al. [2007], Doherty et al. [2004], Hsu and Chen [2009]).

This work can be considered to fall in the second category. Of course, needless to state, the existing results from the first category have to be complied to.

In this work, we study the influence of the different properties of a distance measure, viz., triangle inequality, boundedness and translation invariance, on this phenomenon. Our studies indicate that unbounded distance measures whose expectations do not exist are to be preferred. We propose some new distance measures based on our studies and present many experimental results which seem to confirm our analysis. In particular, we study these distance measures w.r.t. indices like Relative Variance and Relative Contrast and further compare and contrast these measures in the setting of nearest neighbour/proximity searches and hierarchical clustering.

1.3 Outline of the work

In Section 2, after giving some preliminaries that fix the definitions and notations, we formally introduce the CoN phenomenon and discuss in brief the different works related to it. Section 3 contains the study of the influence of the different mathematical properties of a distance measure, viz., triangle inequality, boundedness and translation invariance, on this phenomenon. Based on this analysis, in Section 4 we discuss the desirability of the above properties and whether they can co-exist. In Section 5 we propose some new distance measures that conform to the analysis in the previous sections. Following this, we study these new and also some existing distance measures w.r.t. indices like Relative Variance and Relative Contrast. Further, we compare and contrast these measures in the setting of nearest neighbour/proximity searches and hierarchical clustering on both real and synthetic data sets. It can be seen that the experimental results seem to confirm our analysis. In Section 6 some concluding remarks are given.

2 Concentration of norms in high-dimensional spaces

2.1 Some preliminaries

Let $\bar{X}, \bar{Y} \in \mathcal{U} \subset \mathbb{R}^n$ for some $n \in \mathbb{N}$, i.e., $\bar{X} = (x_1, \ldots, x_n)$, $\bar{Y} = (y_1, \ldots, y_n)$ are n-dimensional real vectors. We say $\bar{X} \preceq \bar{Y}$ if $x_i \leq y_i$ for all $i = 1, 2, \ldots, n$. A set $\mathcal{V} \subset \mathcal{U}$ is called a chain if any two elements in it are comparable w.r.t. the order given by $\preceq$ above.

A mapping $d \colon \mathcal{U} \times \mathcal{U} \to [0, \infty]$ is called a **distance measure** if

(i) $d(\bar{X}, \bar{Y}) = 0 \iff \bar{X} = \bar{Y}$,

(ii) it is symmetric, i.e., $d(\bar{X}, \bar{Y}) = d(\bar{Y}, \bar{X})$

(iii) it is monotonic on any chain $\mathcal{V} \subset \mathcal{U}$, i.e., if $\bar{X}, \bar{Y}, \bar{Z} \in \mathcal{V}$ such that $\bar{X} \preceq \bar{Y} \preceq \bar{Z}$ then $d(\bar{X}, \bar{Y}) \leq d(\bar{X}, \bar{Z})$.

A distance measure d is called a **metric** if it further satisfies, for any $\bar{X}, \bar{Y}, \bar{Z} \in \mathcal{U}$,

(iv) the triangle inequality, i.e., $d(\bar{X}, \bar{Z}) \leq d(\bar{X}, \bar{Y}) + d(\bar{Y}, \bar{Z})$.

Given a normed vector space $(\mathbb{R}^n, +, \cdot, \|.\|)$ and any $\bar{X}, \bar{Y} \in \mathbb{R}^n$, the **norm** – usually denoted as $\|.\|$ – is a function from $\mathbb{R}^n \to [0, \infty]$ such that

(i) $\left\|\bar{X}\right\| = 0 \iff \bar{X} = \bar{0}$,

(ii) $\left\|a\bar{X}\right\| = |a| \left\|\bar{X}\right\|$ for any scalar a, (**Linearity**)

(iii) $\left\|\bar{X} + \bar{Y}\right\| \leq \left\|\bar{X}\right\| + \left\|\bar{Y}\right\|$. (**Triangle Inequality**)

It is well known that one can get a metric from a norm as $d(\bar{X}, \bar{Y}) = \left\|\bar{X} - \bar{Y}\right\|$, though the converse is not always possible. For instance, let $\bar{X}, \bar{Y} \in \mathbb{R}^n$. Consider the function

$$\left\|\bar{X}\right\| = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}. \tag{1}$$

For $p \geq 1$, it can be easily shown that (1) is a norm – usually denoted as $\|\cdot\|_p$ – and one obtains the following distance function which is a metric from it as follows:

$$d(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}. \tag{2}$$

If $1 \leq p \in \mathbb{N}$ then d is usually called the Minkowski norm/metric and is denoted as $\mathcal{L}_p$ in this work. Setting $p = 2$ gives the Euclidean metric. If $p \in (0, 1)$ then d is called the Fractional norm/metric (see Aggarwal et al. [2001]) and we denote it by $\mathcal{F}_p$. Note also that $\mathcal{F}_p$ is actually not a metric since the triangle inequality does not hold.

In this work, we will denote the distance of a vector $\bar{X}$ from the origin w.r.t. d by $\left\|\bar{X}\right\|_d$, as the "norm" of the vector, i.e., $\left\|\bar{X}\right\|_d = d(\bar{X}, \bar{0}) = \left\|\bar{X} - \bar{0}\right\|_d$ even if d

is not a metric and only a distance measure. Of course, this is true if d is a metric obtained from a norm.

In the sequel, we deal only within the framework of $[0,1]^n$, which means it is implicitly assumed that all data can be normalised to fall within this domain. This is done to maintain uniformity and consistency in the presentation of arguments and results. There are many forms of normalisation each with its own effectiveness and efficiency, the presentation of which is quite outside of the scope of the present work. For more on the benefits and effects of data normalisation see Sneath and Sokal, Milligan and Cooper [1985]. We only assume that the data are reasonably and satisfactorily normalised. As is usual, let us take the query point to be situated at the origin. Note that this can be assumed without loss of generality only if the distance measure used is translation invariant.

2.2 Relative contrast of a metric

Consider a finite data set $\mathcal{U} \subset [0,1]^n$. Given a distance measure d on $[0,1]^n$, let us denote by D_M, D_m the distance from our query point $\bar{Q} = \bar{0}$ (the origin here) to those members $\bar{X}, \bar{Y} \in \mathcal{U}$ that are the farthest and nearest to $\bar{Q}$, respectively, w.r.t. d. Consider the following quotient called the "relative contrast" (RC):

$$\rho_d^n = \frac{D_M - D_m}{D_m}. \tag{3}$$

The concentration of the norm phenomenon, in some sense, refers to the concentration of points and hence their distances, which is to say, that as the dimension n increases ρ_d^n goes to zero, i.e. $\lim_{n\to\infty} \rho_d^n = 0$. In other words, as the dimension increases most of the data seem to be distributed closer to the corners and hence the difference between the "farthest" and the nearest neighbours, as determined by the distance measure d, becomes indistinguishable.

2.3 Studies on the concentration of norms

Here we give a brief summary of only those works that have a direct bearing on our studies and refer the reader to the excellent article of François et al. [2007] and the references therein for further details.

Demartines [1994] is credited to have been the first to determine the bounds on ρ_d^n for any arbitrary but i.i.distributed data but only for the Euclidean norm. Later, independently, Beyer et al. [1999] proposed rather mild conditions on the data distribution by discussing the ratio between the variance and the expectation of the distance distribution under which the CoN phenomenon occurs (see Beyer et al. [1999], Theorem **1**). Recently, the authors in Durrant and Kabán [2009], Hsu and Chen [2009] have shown that the converse also holds when the number of points is "large".

Following this Aggarwal et al. [2001] (see also Hinneburg et al. [2000]) discussed the CoN for uniformly distributed data consisting of finite points for both the $\mathcal{L}_p$ norm and made a strong case for $\mathcal{F}_p$ by showing that relative contrast was better with decreasing p, i.e., as $p \to 0$. These results were further generalised by François et al. [2007] for any arbitrary distributions, not necessarily uniform or i.i.d. However, they also showed that the fractional metric too concentrated, i.e.,

the relative contrast approached zero but just that the rate at which it decreased was slow. Further they also present data sets (real and synthetic) where it is observed that $\mathcal{L}_p$ norms fare better than the fractional metrics. Doherty et al. [2004] studied these norms in the setting of NN classifiers and their results seem to further confirm the earlier studies.

2.4 A case of treating only the symptoms?

So far studies on the CoN phenomenon w.r.t. some specific distance measures have largely restricted themselves with the $\mathcal{L}_p$ and $\mathcal{F}_p$ measures.

It should also be highlighted that the effectiveness of distance measures is very much contextual and depends on the domain of the application, the distribution of the data and to a certain extent even on the range of the data. Studies have shown that different distance functions behave differently on normalised data, see for example Doherty et al. [2004, 2007] (see also Remarks 5.1 & 5.2). In fact, Aggarwal [2001, 2003] makes a case for a distance function to be user centric by stating " *The most important aspect of distance function design is that since a human is the end-user for any application, the design must satisfy the user requirements with regard to effectiveness.*"

Thus both emperical studies related to measuring the different indices like the relative contrast or the relative variance and also the effectiveness of these distance measures in different applications have been studied in Aggarwal et al. [2001], François et al. [2007], Doherty et al. [2007].

However, the following questions still remain: What happens to the concentration of the norm if p is fixed a priori and $n \to \infty$? On the other hand, if the data come from a fixed but arbitrarily large dimension $n = n_0 >> 1$, how should the value of p be chosen so that $\rho_d^n >> 0$? Does there exist any relation between n, p so as to ensure that $\rho_d^n >> 0$? Note that in the works of Aggarwal et al. [2001] and François et al. [2007] one needs to vary simultaneously both the dimension n and the exponent power p of the metric.

In other words, though the relative contrast ρ_d^n reflects well the inability of the distance function to distinguish points in higher dimensions, does proposing metrics to ensure the slowness of its concentration in itself address the root of the problem?

In the following section we revisit the CoN phenomenon and show a likely cause for it and propose some metrics that overcome this drawback. Towards this end, we firstly discuss the desirable properties of a distance measure and show the interplay between these properties among themselves and w.r.t. the CoN.

3 Properties of a distance measure and the concentration of norm phenomenon

In this section we take a look at the properties of typical distance measures, which are usually based on a norm or a metric, and discuss their desirability vis-á-vis the CoN phenomenon.

3.1 Can bounded measures ever do the job?

A distance measure d on $[0,1]^n$ is called **unbounded** if $\lim_{\bar{X}\to\bar{1}} \left\|\bar{X}\right\|_d = \infty$.

Let us consider the formula for relative contrast (3). Once again let $\bar{Q} = \bar{0}$ be the query point. Clearly, $\rho_d^n \to 0$ either if the nearest neighbour $\bar{Y}$ of $\bar{Q}$ goes closer to the farthest neighbour $\bar{X}$ of $\bar{Q}$ and/or if the distance of $\bar{Y}$ itself is large (w.r.t. the origin, here $\bar{Q}$). Let us consider a bounded distance measure d. Note that $D_M = d(\bar{Q},\bar{X}) = \left\|\bar{X}\right\|_d$ and $D_m = d(\bar{Q},\bar{Y}) = \left\|\bar{Y}\right\|_d$.

Let the data come from a fixed but arbitrarily large n dimensional space, $[0,1]^n$. Then the theoretical farthest neighbour for our query point is $\bar{1} = (1,\ldots,1)$ at a distance $D_M = K$, for some $K \in \mathbb{R}$. Letting $\bar{X} = \bar{1}$ we have that as $\bar{Y} \to \bar{X}$, $\left\|\bar{Y}\right\|_d \to K$ and the numerator of (3), viz., $\left(D_M - \left\|\bar{Y}\right\|\right) \to 0$ while its denominator $\left\|\bar{Y}\right\|_d$ is increasing. Clearly, however large K may be this phenomenon is waiting to happen, especially if the number of data points are fixed but the dimension n is allowed to increase.

The above discussion seems to call for unbounded metrics d, i.e., $\lim_{\bar{X}\to\bar{1}} \left\|\bar{X}\right\|_d = \infty$. Clearly, then d cannot be linear either. For, if d is linear, then consider some $c \in (0,1)$. With $a = \frac{1}{c} > 0$, $\bar{C} = (c,\ldots,c)$, we have that $\left\|\bar{C}\right\|_d = K < \infty$ which implies that $|a|\left\|\bar{C}\right\|_d = \left\|a\bar{C}\right\|_d = \|\bar{1}\|_d = |a|K < \infty$, contradicting the fact that d is unbounded.

3.2 The effect of triangle inequality

The triangle inequality property of a norm automatically fixes an upper bound for ρ_d^n:

$$\frac{\left\|\bar{X}-\bar{Y}\right\|_d}{\left\|\bar{Y}\right\|_d} \geq \frac{\left\|\bar{X}\right\|_d - \left\|\bar{Y}\right\|_d}{\left\|\bar{Y}\right\|_d} = \rho_d^n \,. \tag{4}$$

Once again, as $\bar{Y} \to \bar{X}$ we have that $(\bar{X}-\bar{Y}) \to \bar{0}$ and hence $\left\|\bar{X}-\bar{Y}\right\|_d \to 0$ and so does ρ_d^n.

Thus even if d is unbounded but satisfies the triangle inequality the CoN phenomenon is certain to manifest sooner or later. This also suggests that, for ρ_d^n to not to go to zero, not only should d continue to increase for points far away from the origin, the distances "closer to the origin" should also be "relatively large".

3.3 Translation invariance and the curse of dimensionality

From the above it is clear that one needs to ensure that $\left\|\bar{X}-\bar{Y}\right\|_d$ remains large even when $(\bar{X}-\bar{Y}) \to \bar{0}$. One way to achieve this is to make the distance between 2 points – which is essentially a measure of their relative position – to somehow depend on their "absolute positions" also. For instance, consider $\bar{X},\bar{Y},\bar{Z} \in [0,1]^n$ such that

$$\|\bar{0}\|_d < \left\|\bar{X}\right\|_d, \left\|\bar{Y}\right\|_d << \left\|\bar{Z}\right\|_d < \|\bar{1}\|_d \,.$$

Let $\bar{C} \in [0,1]^n$ be such that $\bar{X}+\bar{C}, \bar{Y}+\bar{C} \in [0,1]^n$ and

$$\|\bar{0}\|_d << \left\|\bar{X}+\bar{C}\right\|_d, \left\|\bar{Y}+\bar{C}\right\|_d < \|\bar{1}\|_d \,.$$

To ensure that the numerator of the upper bound in (4) is also relatively large compared to its denominator, we need that

$$d\left(\bar{X}+\bar{C},\bar{Y}+\bar{C}\right) > d\left(\bar{X},\bar{Y}\right).$$

In short, d should not be translation invariant!

Perhaps the above property is captured in some sense in another index firstly proposed by François et al. [2007] (and later generalised in Durrant and Kabán [2009]) to investigate the CoN phenomenon. The relative variance of a given data distribution is given as:

$$RV_d = \frac{\sqrt{Var\left(\left\|\bar{X}\right\|_d\right)}}{E\left(\left\|\bar{X}\right\|_d\right)}, \tag{5}$$

where $\|.\|_d$ is any distance measure. Once again, a small value for RV_d reflects the concentration of d. As already stated by the authors, "*$RV_{\mathcal{F}_p}$ measures the concentration by relating a measure of spread (variance) to a measure of location (expectation)*".

4 Desirable properties of a distance measure

It is clear from the above discussion that we need distance measures d that are unbounded, and hence non-linear, nevertheless satisfying both the triangle inequality and translation invariance.

4.1 Can a distance measure d have the above three properties?

We show below that there cannot exist distance measures d that possess all the above 3 properties.

Theorem 4.1: *Let d be a distance measure on $[0,1]^n$. Then d can have at most two of the following three properties:*

(i) Unboundedness,

(ii) Translation Invariance,

(iii) Triangle Inequality.

Proof.

(i) & (ii) $\implies$ not (iii):

Let d be unbounded with $d(\bar{0},\overline{0.5}) = K < \infty$. By the translation invariance of d, we have $d(\bar{0},\overline{0.5}) = d(\overline{0.5},\bar{1})$. If d were to also have the triangle inequality property, then

$$d(\bar{0},\bar{1}) \leq d(\bar{0},\overline{0.5}) + d(\overline{0.5},\bar{1}) = 2K < \infty, \tag{6}$$

contradicting the fact that d is unbounded.

(ii) & (iii) $\Longrightarrow$ not (i) and

(iii) & (i) $\Longrightarrow$ not (ii)

follow easily from the inequality (6) above.

4.2 Can translation invariance be given the slip?

While it is clear that we need a measure that evaluates distances between points not only based on their relative separation but also somehow taking into account their absolute positions also, translation invariance is an essential property in many applications, for instance, in clustering applications. Deviating from translation invariance would imply that the distance or similarity differs depending on the range in which data fall. This might be desirable in certain applications, but could only be justified by domain-specific knowledge that suggests how to measure distance in different ranges.

In fact, note that the usual assumption of considering the origin as a query point in any analysis is valid only if the distance measures employed are translation invariant.

4.3 Is triangle inequality always desirable?

The triangle inequality property of a norm states that the straight line is the shortest path between any two points. How valid is this intuitive property of Euclidean spaces in higher dimensions? More importantly, is it even desirable? Even in the cases where the distance measures satisfy this property, one can get some counter-intuitive results.

In fact, the triangle inequality is important so that the corresponding similarity measure $S\colon \mathcal{U} \times \mathcal{U} \to [0,1]$ induced by the distance measure d satisfies the transitivity property. In this context, it is enough to find a well-defined operation $\oplus$ such that

$$d(\bar{X}, \bar{Z}) \leq d(\bar{X}, \bar{Y}) \oplus d(\bar{Y}, \bar{Z}).$$

Then one can suitably find a $*$ (usually a conjunctive operator) such that

$$S(\bar{X}, \bar{Z}) \geq S(\bar{X}, \bar{Y}) * S(\bar{Y}, \bar{Z})$$

holds.

5 Analysis of distance measures and some experimental results

In this section, we analyse the two commonly used and investigated distance measures, viz., the Minkowski's $\mathcal{L}_p$ and Fractional norms $\mathcal{F}_p$. Moreover based on our analysis in the previous sections we also propose two new unbounded measures and investigate all these measures w.r.t. the two indices that have now become commonly accepted as measures of concentration, viz., the relative contrast ρ_d^n and the relative variance RV_d.

Table 1 Properties of different d with fixed n, p. K is the bound of d, i.e., $K = \|\bar{1}\|_d$. See Section 5.1 for more details.

d	p parameter	Bounded	K	Triangle Inequality	Translation Invariance
$\mathcal{L}_p$	$p \geq 1$	✓	$n^{\frac{1}{p}}$	✓	✓
$\mathcal{F}_p$	$0 < p < 1$	✓	$n^{\frac{1}{p}}$	×	✓
$\mathcal{J}_p$	$p \geq 1$	×	∞	✓	×
$\mathcal{J}_p$	$0 < p < 1$	×	∞	×	×
$\mathcal{J}'_p$	$p \geq 1$	×	∞	×	✓
$\mathcal{J}'_p$	$0 < p < 1$	×	∞	×	✓

5.1 Bounded and unbounded distance measures

Let us once again consider the data to come from $[0,1]^n$ for a fixed but arbitrarily large $n \in \mathbb{N}$ and assume that our query point is placed at the origin $\bar{0} \in [0,1]^n$.

5.1.1 The Minkowski norms $\mathcal{L}_p,\ p \in \mathbb{N}$

From Eqn.(1) it is clear that for a fixed p it is bounded above by $K = \|\bar{1}\|_d = \mathcal{L}_p(\bar{1}) = n^{\frac{1}{p}}$. Also, as p increases this bound decreases. Being a norm it does satisfy both the triangle inequality and translation invariance properties.

5.1.2 The fractional norms $\mathcal{F}_p,\ p \in (0,1)$

Once again, from Eqn.(1) with a fixed $p \in (0,1)$ we see that the fractional norms are also bounded above by $K = \|\bar{1}\|_d = \mathcal{F}_p(\bar{1}) = n^{\frac{1}{p}}$. Also, as p decreases this bound increases, since as $p \to 0$ we have that $\frac{1}{p} \to \infty$. Thus even though it is bounded we see that as $n \to \infty$ the bound increases but for a fixed n, p it is very much bounded. Once again, it is well known that $\mathcal{F}_p$ is translation invariant but does not satisfy the triangle inequality.

Note that the above analysis also clarifies the seemingly contradictory results found in Aggarwal et al. [2001] and Doherty et al. [2004], François et al. [2007]. In Aggarwal et al. [2001], the authors prove that as $p \to 0$ the constant for the upper bound increases non-linearly, but the moment p is fixed this constant is also fixed, thus it only slows down the rate of concentration without completely eliminating it as the results in Doherty et al. [2004], François et al. [2007] show.

5.1.3 Some new unbounded measures

Consider the following two functions on $[0,1)^n$ with $p > 1$:

$$\mathcal{J}_p(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^{n} \left| \frac{x_i}{1 - x_i} - \frac{y_i}{1 - y_i} \right|^p \right)^{\frac{1}{p}} \tag{7}$$

$$\mathcal{J}'_p(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^{n} \left| \frac{x_i - y_i}{1 - |x_i - y_i|} \right|^p \right)^{\frac{1}{p}} \tag{8}$$

It can be easily verified that both $\mathcal{J}_p, \mathcal{J}'_p$ satisfy the basic conditions of a distance measure, viz., symmetry and monotonicity on any chain in $[0,1)^n$. Moreover, both of them are unbounded measures, i.e.,

$$\lim_{\bar{X}\to\bar{1}} \mathcal{J}_p(\bar{0},\bar{X}) = \left\|\bar{X}\right\|_{\mathcal{J}_p} = \left(\sum_{i=1}^{n}\left|\frac{x_i}{1-x_i}\right|^p\right)^{\frac{1}{p}} = \left\|\bar{X}\right\|_{\mathcal{J}'_p} = \lim_{\bar{X}\to\bar{1}} \mathcal{J}'_p(\bar{0},\bar{X}) = \infty\ ,$$

since $\bar{X}\to\bar{1}$ implies that $x_i \to 1$ for all $i = 1,\ldots,n$. (Note that $\|\cdot\|_{\mathcal{J}'_p} = \|\cdot\|_{\mathcal{J}_p}$, since the distance of a point from the origin remains the same under both the measures.) However, while $\mathcal{J}_p$ satisfies the triangle inequality (this can be easily proven using the usual Minkowski's inequality) it is not translation invariant. In contrast, $\mathcal{J}'_p$ is translation invariant but does not satisfy the triangle inequality.

Of course, if we let $p \in (0,1)$ it only affects the triangle inequality property of $\mathcal{J}_p$.

The distance measure $\mathcal{J}_p$ is in the line of kernel functions as they are used for support vector machines. It is based on the nonlinear transformation

$$(x_1,\ldots,x_n) \mapsto \left(\frac{1}{1-x_1},\ldots,\frac{1}{1-x_n}\right)$$

and then computing the distance with the corresponding $\mathcal{L}_p$ norm. The distance measure $\mathcal{J}'_p$ can be seen as a modification of $\mathcal{J}_p$, since it can be rewritten in the form

$$\mathcal{J}'_p(\bar{X},\bar{Y}) = \mathcal{J}_p(\bar{0},\bar{X}-\bar{Y}).$$

The distance $\mathcal{J}'_p$ defines the distance between to vectors $\bar{X}$ and $\bar{Y}$ as the $\mathcal{J}_p$ distance of the difference between $\bar{X}$ and $\bar{Y}$ to the zero vector.

In François et al. [2007] the authors have considered a sample of $100,000$ uniformly sampled points from $[0,1]^n$ for n varying from $1-100$. Considering the set of distances $A = \{\left\|\bar{X}^i\right\|_d, i = 1,\ldots,100,000\}$, they plot the minimum, maximum, average and the variance of the set A for dimensions $1-100$ with the Euclidean metric for d. We do a similar study for the measures $\mathcal{J}_p, \mathcal{J}'_p$.

In Figures 1–3 we plot the above indices for 3 data sets each containing $25,000; 50,000$ and $100,000$ vectors. Each set contained roughly the same number of points that were Gaussian and uniformly distributed over $[0,1)^n$. Not only is the relative contrast quite high – the difference between minimum and maximum distances is in the order of 10^6, as is expected it tends to increase with the increase in the number of points for the same dimension. This is further confirmed by the relative variance index which is plotted in Figure 4 which was measured on the same data sets.

5.2 K-NN search on some normalised UCI data sets

In Aggarwal et al. [2001] a K-nearest neighbor (K-NN) search was done on some UCI data sets to test the quality of the $\mathcal{L}_p$ and $\mathcal{F}_p$ distance measures for different parameter values. The test consisted of stripping off the class variable data from

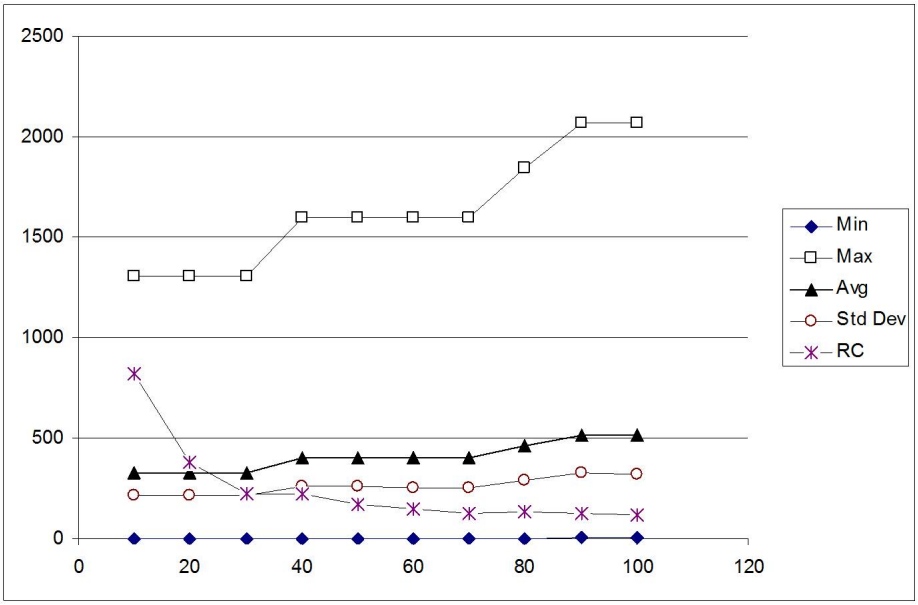


Figure 1 The minimum, maximum, average, standard deviation and relative contrast of $\mathcal{J}_p$ (and $\mathcal{J}'_p$) with $p = 2$ for distances on 25,000 points distributed over $[0, 1)^n$ for dimensions $n = 10 - 100$.

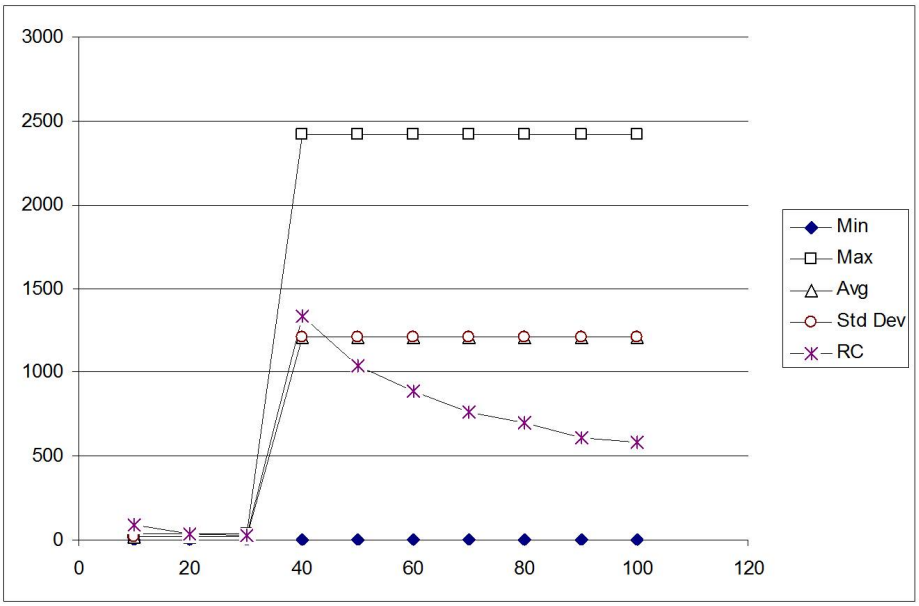


Figure 2 The minimum, maximum, average, standard deviation and relative contrast of $\mathcal{J}_p$ (and $\mathcal{J}'_p$) with $p = 2$ for distances on 50,000 points distributed over $[0, 1)^n$ for dimensions $n = 10 - 100$.

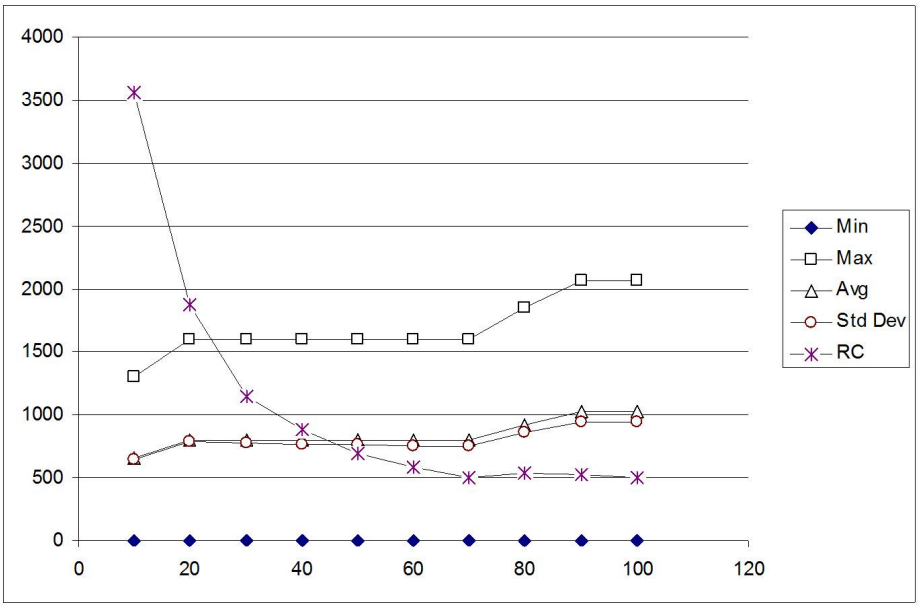


Figure 3 The minimum, maximum, average, standard deviation and relative contrast of $\mathcal{J}_p$ (and $\mathcal{J}'_p$) with $p = 2$ for distances on 100,000 points distributed over $[0, 1)^n$ for dimensions $n = 10 - 100$.

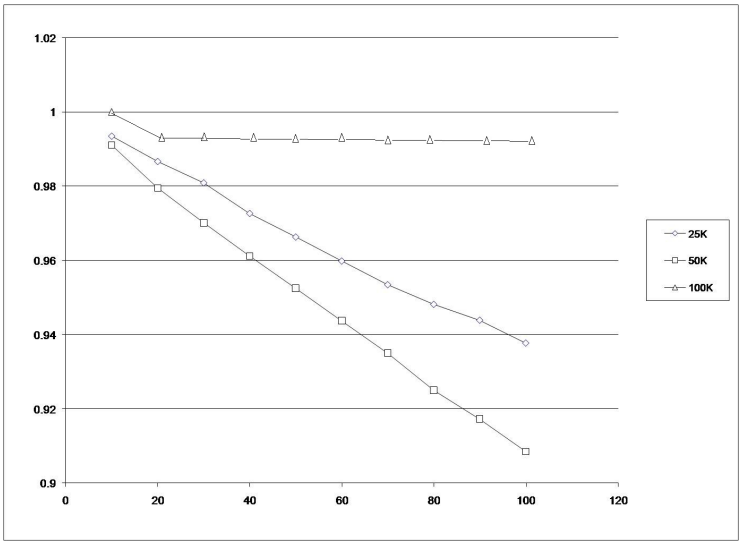


Figure 4 The relative variance of $\mathcal{J}_p$ (and $\mathcal{J}'_p$) with $p = 2$ for distances on points distributed over $[0, 1)^n$ for dimensions $n = 10 - 100$.

the data sets and using only the feature variables. Picking each data member from the data set as the query point, the K-nearest neighbours were determined using a distance d (here $\mathcal{L}_p$ or $\mathcal{F}_p$). This set of K-nearest neighbours was then checked for class variable accuracy, i.e., how many among the K-nearest neighbours actually belonged to the same class as the query point. Though this is primarily a measure that is evidential in nature, their study still indicated that the class variable accuracy increased with decreasing values of p with $p \in (0, 1)$, thus suggesting the use of fractional norms.

Later on Doherty et al. [2004] performed a similar experiment on the following data sets from the UCI repository: Ionosphere, Wisconsin Breast Cancer Diagonostic (WBCD) and Image Segmentation training data. When the tests were performed on data that were normalised employing the usual formula

$$y^i = \frac{x^i - x^i_m}{x^i_M - x^i_m} \ , \tag{9}$$

along each of the feature variables, where x^i_m, x^i_M were the minimum and maximum values of the i-th feature variable x^i, their results show that there was no clear relation between the values of the parameter p and the class variable accuracy.

We performed the same experiment on the above three data sets using the normalisation formula (9) for the four distance measures, viz., $\mathcal{L}_2, \mathcal{F}_{0.04}, \mathcal{J}'_2, \mathcal{J}_2$. To ensure that the distance measures $\mathcal{J}'_2, \mathcal{J}_2$ did not saturate, when a particular feature variable had the value $y^i = 1$ after normalisation, we reassigned the value to $y^i = 0.9999$. The results are presented in Table 2. (Though we present here results of $\mathcal{L}_2, \mathcal{F}_{0.04}$ for such a data set, the results were not significantly different on the normalised data without this modification as can be readily verified for the case of $\mathcal{L}_2$ from Table 4 in Doherty et al. [2004]. Also note that $p = 0.1$ is the smallest value for which the results are tabulated in Aggarwal et al. [2001] and Doherty et al. [2004].) In the first column of the table indicates the corresponding data set and the number of instances contained in the data sets. The number K in the second column specifies the number of nearest neighbours that are considered. For each instance we compute how many of the K nearest neighbours with respect to the mentioned distance measure belong to the same the same class as the instance itself. The average number and the average percentage of these K nearest neighbours from the same class is given in the remaining columns.

The results show that while the fractional norm with $p = 0.04$ was the best for the Ionosphere data set, this honour belonged to the Euclidean norm for the other two data sets. They also show that $\mathcal{J}'_2, \mathcal{J}_2$ do perform consistently well with one of them being the second best in every scenario.

Although the four distance measures yield very different values for the distances, their performance in Table 2 differs not so drastically as might have been expected. The reason might be that for the nearest neighbour search, only a few nearest neighbours are considered. These nearest neighbours are in most cases really close to the reference point, no matter which distance measured is used. The situation changes when cluster analysis is carried out. In this case, the number of data objects in a cluster is usually much larger than the number of considered nearest neighbours in classification. As will be discussed in Section 5.4, the results

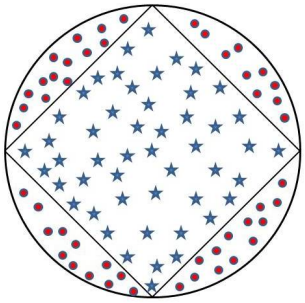


Figure 5 Distance measures: For classification or clustering!

for the four distance measures differ much more for clustering than for nearest neighbour search.

Remark 5.1: As was noted in Aggarwal et al. [2001] any commentary on the quality of a distance measure based on such experiments is largely evidential in nature. Moreover, it is not clear whether distance measures which are naturally more suited for clustering of spatial data should be expected to or can also perform well in classification problems, where the data members of respective classes may not share any intrinsic spatial similarity. For instance, the Manhattan metric can effectively classify the two simple data sets in Figure 5, while the Euclidean metric is unlikely to classify it accurately.

5.3 RC and RV on a real data set

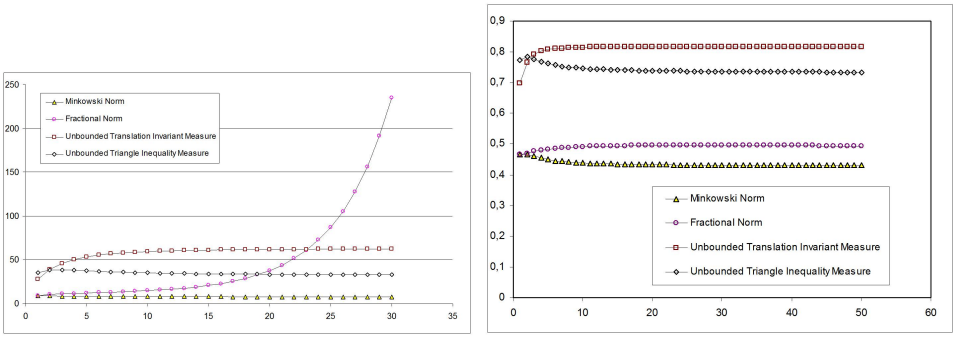
Now we present some results related to the relative contrasts and relative variances of the above distance measures on a real data set. In François et al. [2007] they have plotted the relative contrasts for some real data sets from the UCI depository, specifically the Wisconsin Breast Cancer Diagonostic (WBCD) data and the Image Segmentation data set. Their study shows that the Minkowski norms $\mathcal{L}_p$ perform better than fractional norms $\mathcal{F}_p$ and, in fact, their relative contrast increases with increasing p with $p \geq 1$.

We consider the WBCD data consisting of 569 vectors which, after stripping down the categorical variables of Patient ID and the type of cancer, contains 30 dimensions. Once again, we normalise the data as explained in Section 5.2. We plot the relative contrasts and relative variances of the above distance measures on this data set for different values of p, typically $p = 1, \ldots, 50$. Note that for the fractional metric we used the reciprocal of the parameter p. As can be seen in Figure 6 the comparison among the indices tell quite an interesting story.

In Figure 6(a) the relative contrast is plotted for $p = 1, \ldots, 36$, which for the fractional metric means p assumes the values $1, \frac{1}{2}, \ldots, \frac{1}{36}$ (we do not present the results for $p > 36$ just to retain the overall scale of the plot). The relative contrasts for the Minkowski norms $\mathcal{L}_p$ and the unbounded measures $\mathcal{J}_p$ decrease with increasing p, however they seem to stabilise at not a very far value $(7.81; 35.26)$ from where they begin $(9.21; 32.66)$. In contrast, the relative contrast for the unbounded measure $\mathcal{J}'_p$ initially increases with increasing p but seems to stabilise, but with a value (63.21) far away from what it was for $p = 1$ (27.79). The fractional

Table 2 K-NN Search on some *normalised* UCI data sets

Data Set (no. of objects)	K	$\mathcal{L}_2$ Euclidean	$\mathcal{F}_{\frac{1}{25}}$ Fractional	$\mathcal{J}_2'$ Translation Invariant	$\mathcal{J}_2$ Triangle Inequality
Ionosphere (351)	3	2.56 (85%)	2.71 (90%)	2.59 (86%)	2.36 (79%)
	5	4.20 (84%)	4.38 (88%)	4.24 (85%)	3.88 (77%)
	9	7.42 (80%)	7.72 (86%)	7.57 (84%)	6.83 (76%)
Segmentation (210)	3	2.51 (84%)	2.03 (68%)	2.46 (82%)	2.27 (76%)
	5	4.12 (82%)	3.17 (63%)	4.00 (80%)	3.58 (72%)
	9	7.00 (78%)	5.29 (66%)	6.71 (75%)	5.83 (65%)
WBCD (569)	3	2.86 (95%)	2.71 (90%)	2.80 (93%)	2.85 (95%)
	5	4.74 (95%)	4.55 (91%)	4.74 (95%)	4.63 (93%)
	9	8.48 (94%)	8.14 (90%)	8.44 (94%)	8.28 (92%)



(a) Relative Contrast ρ_d^n (b) Relative Variance RV_d

Figure 6 Plots of the (a) Relative Contrast (b) Relative Variance of the above distance measures for the Wisconsin Breast Cancer Diagonostic Data

norm $\mathcal{F}_p$ shows a totally different behaviour and continues its increasing trend even after $p = 36$.

In Figure 6(b) the relative contrast is plotted for $p = 1, \ldots, 50$ (for the fractional metric $p = 1, \frac{1}{2}, \ldots, \frac{1}{50}$). Once again, the relative variances for the Minkowski norms $\mathcal{L}_p$ and the unbounded measure $\mathcal{J}_p$ decrease with increasing p, while those for the unbounded measure $\mathcal{J}_p$ and the fractional norm $\mathcal{F}_p$ increase with increasing p. However, all of them soon saturate to an almost stable value.

What is quite revealing here is that the fractional norm that increased almost exponentially with respect to the relative contrast not only does its relative variance stabilise but to a value that is far below those of the unbounded measures. Note that while the relative contrast of $\mathcal{F}_p$ is far superior to those of the other measures considered, the values for $\mathcal{J}_p, \mathcal{J}'_p$ are also quite on the higher side even without increasing the parameter value beyond 2. In fact, this is one of the nice properties of these unbounded measures: they do not need the help of an extra parameter. Also note that through out this work we have consistently used a value of $p = 2$ for $\mathcal{J}_p, \mathcal{J}'_p$, while for $\mathcal{F}_p$ we use a rather low value of $p = 0.04$.

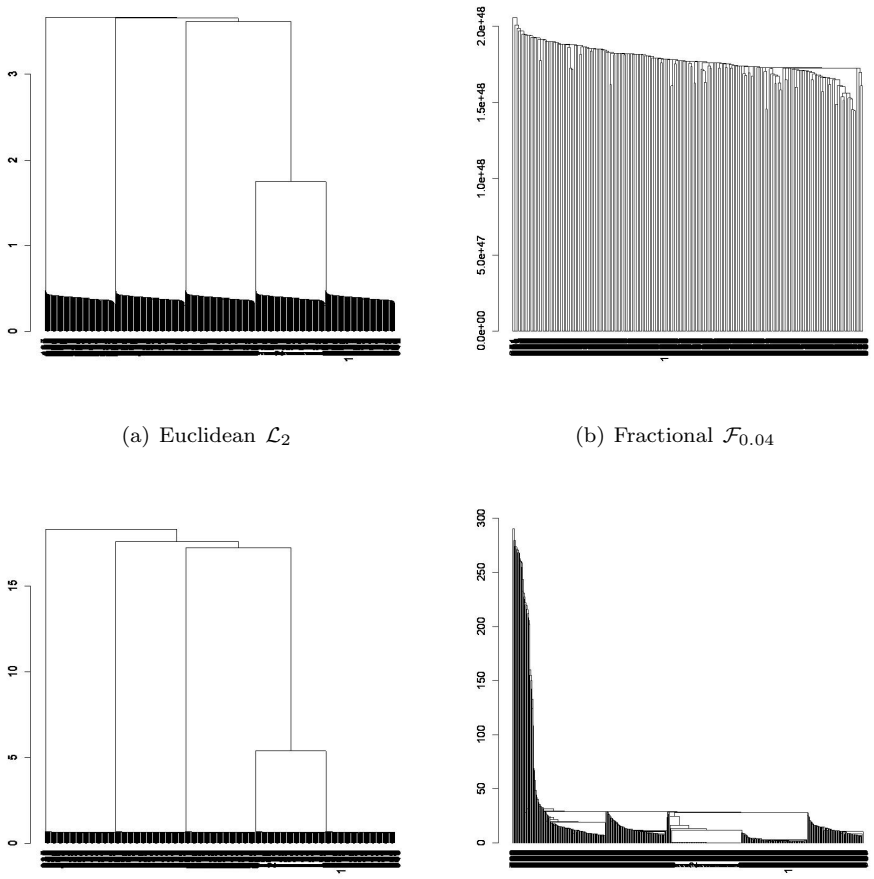
Remark 5.2: Note that we have presented our results on normalised data (Figure 6(a)) and hence we do not expect it to conform to what is reported in François et al. [2007]. This once again clearly highlights the difference between employing data which is normalised and that which is not.

5.4 Hierarchical clustering

In this section we present results of Hierarchical clustering performed with the above 4 distance measures. For this purpose we considered two data sets consisting of 1000 vectors of 100 dimensions. The first data set resembles a Gaussian mixture model. Firstly, we generated 200 data points with 100 dimensions with values lying in $[0, 0.16666]$, , i.e., $A = \{\bar{X}^i = (x^i_j) | x^i_j \in [0, 0.16666],\ i = 1, 2, \ldots, 200, j = 1, \ldots, 100\}$. Then we created another 800 points from these by adding 0.833333 to different but non-overlapping dimensions to create the other 4 clusters. For instance, we added the above constant to the dimensions $j = 10, \ldots, 28$ of every $\bar{X}^i \in A$ to get the next cluster B. Similarly, we obtained clusters C, D, E by modifiying the dimensions in the range of $[35, 55], [58, 74], [88, 98]$.

Using a similar procedure as above, for the second data set we generated uniformly distributed data in $[0, 1]^{100}$ such that there were 4 clear and distinct clusters each with 200 points. Then we generated 200 points, once again uniformly over $[0, 1]^{100}$, and added them to the data set as a fifth noise cluster.

We used two of the hierarchical clustering methods available with R statistical package, viz., the Single-linkage and the Ward's method. In Figures 7–10 we give the plots of the dendrograms and the heat maps for the clustering obtained with the above distance measures. The result shows that the Euclidean metric $\mathcal{L}_2$ and the unbounded measures $\mathcal{J}_2, \mathcal{J}'_2$ perform consistently well, while the fractional metric ($\mathcal{F}_{0.04}$) is found wanting. Though we have presented the results only for the above two methods and the specified data sets, our experiments with other types of data distributions and different hierarchical clustering methods showed a similar trend.



(a) Euclidean $\mathcal{L}_2$

(b) Fractional $\mathcal{F}_{0.04}$

(c) Translation Invariant $\mathcal{J}_2'$

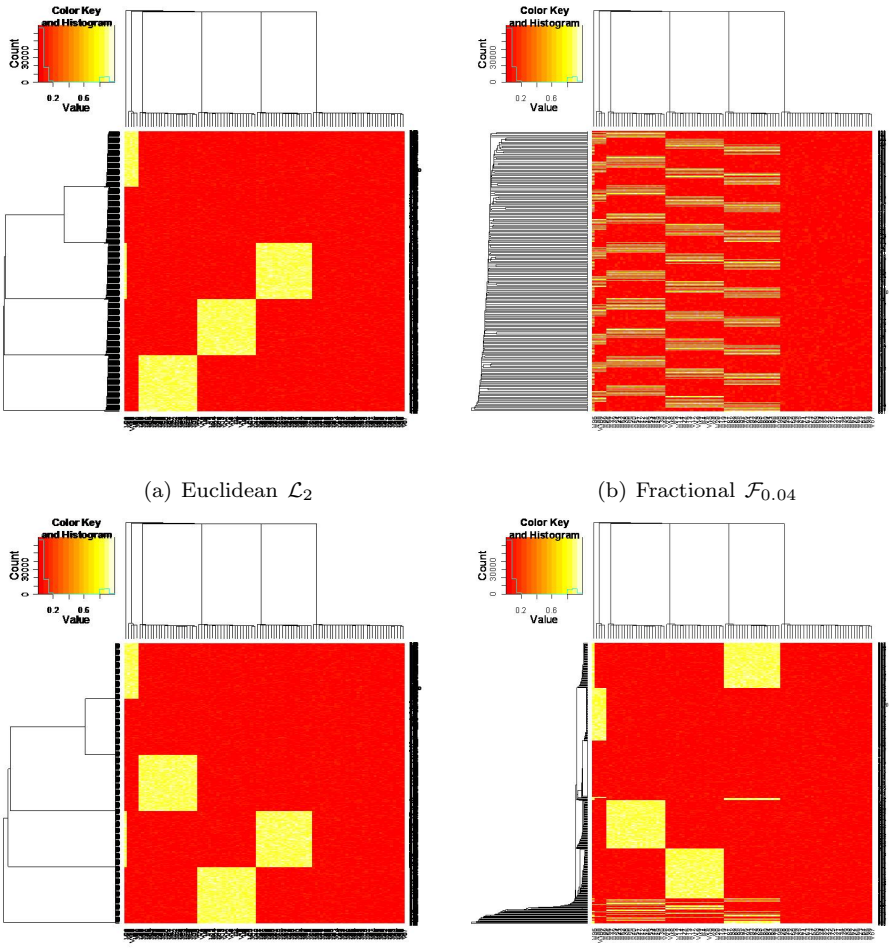
(d) Trianlge Inequality $\mathcal{J}_2$

Figure 7 Dendrograms of the single linkage hierarchical clustering for 1000 points Gaussian distributed with 5 distinct clusters

It should be noted that the data clusters are more or less well-separated. When clusters tend to overlap more, the situation will become more difficult for the distance measures. The Euclidean metric $\mathcal{L}_2$ and our distance measure $\mathcal{J}_2$ can both discover the the five clusters, these clusters are even more visible for the distance measure $\mathcal{J}_2$ in the dendrogram than for the Euclidean metric. The "lawn" of smaller clusters is much shorter for the distance measure $\mathcal{J}_2$.

5.5 High expectations from an expectationless distance measure

In Hsu and Chen [2009] the authors state "*Our theoretical results show that all distance functions should be meaningless in high-dimensional space, except that it can resist the rapid degradation of distance variation with increasing dimensionality.*" With reference to this statement, how does one view the results



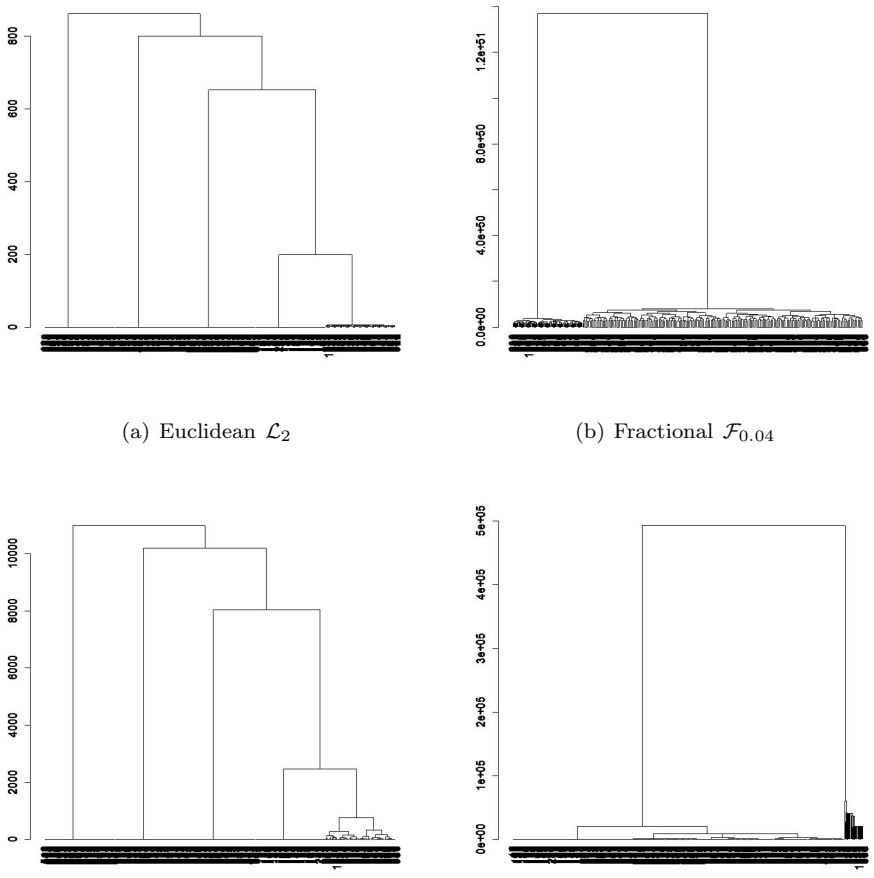
(a) Euclidean $\mathcal{L}_2$

(b) Fractional $\mathcal{F}_{0.04}$

(c) Translation Invariant $\mathcal{J}_2'$

(d) Trianlge Inequality $\mathcal{J}_2$

Figure 8 Heat Maps of the Single-linkage Hierarchical Clustering for 1000 points Gaussian distributed with 5 distinct clusters



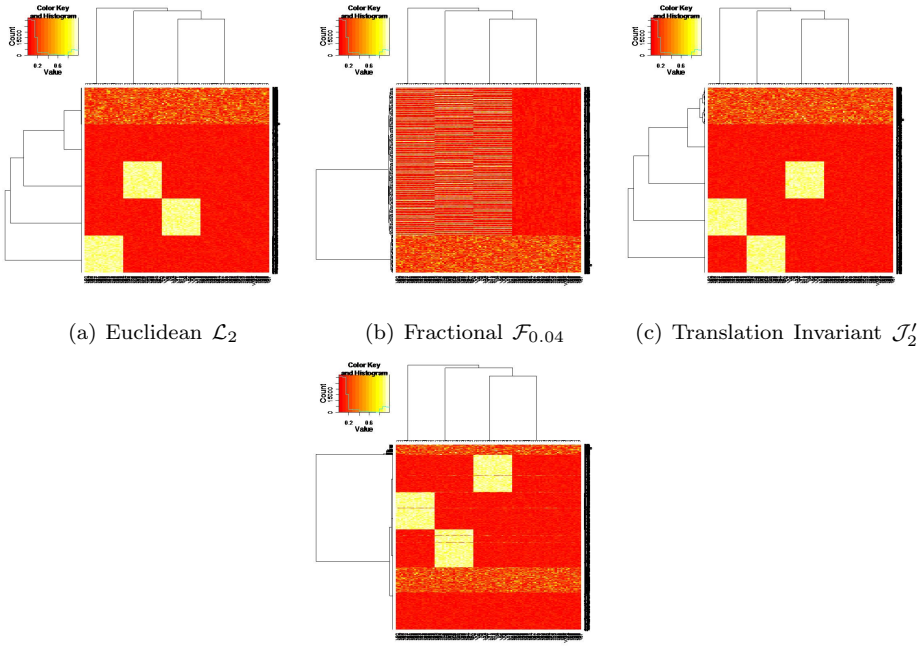
(a) Euclidean $\mathcal{L}_2$

(b) Fractional $\mathcal{F}_{0.04}$

(c) Translation Invariant $\mathcal{J}_2'$

(d) Trianlge Inequality $\mathcal{J}_2$

Figure 9 Dendrograms of Ward's Hierarchical Clustering for 1000 points uniformly distributed with 4 distinct clusters and a noise cluster



(a) Euclidean $\mathcal{L}_2$ (b) Fractional $\mathcal{F}_{0.04}$ (c) Translation Invariant $\mathcal{J}'_2$

(d) Trianlge Inequality $\mathcal{J}_2$

Figure 10 Heat Maps of the Ward's Hierarchical Clustering for 1000 points uniformly distributed with 4 distinct clusters and a noise cluster

and analysis presented so far? Are the new functions $\mathcal{J}_p, \mathcal{J}'_p$ prone to concentration but only much slower? Is this true of any general unbounded measure too?

The answer perhaps lies in the fact that the known theoretical results are valid for only those distance measures whose expectation is finite. Let us consider the uniformly distributed random variable $X \sim U(0,1)$ on a single dimension. Then the expectation of the distance from the origin for X w.r.t. the measures $\mathcal{J}_p, \mathcal{J}'_p$ is given as

$$E[\|X\|] = \int_0^1 \frac{x}{1-x}\, dx\,,$$

which clearly does not exist.

So far, we have dealt with only $[0,1]^n$ assuming the underlying data is normalised. However, outliers in data can upset the normalisation and hence most of the actual data may not belong as close to the vertices as assumed so far. In other words, even if d is an unbounded metric the data distribution may be such that the distances may not 'saturate' and hence the expectation of d may very well exist. In the following we show that the above metrics can be easily adapted to this situation.

Let us assume that even though the normalised data occupy the $[0,1]^n$ space, they are effectively concentrated in some 'sub-space' $[a,b]^n \subset [0,1]^n$. Since $b \in (0,1)$ we have that $\lim_{t\to\infty} b^{\frac{1}{t}} = 1$. From the denseness of reals we can easily find a

$q \in (0,1)$ such that $|1-b^q| < \epsilon$ for any arbitrarily small $\epsilon > 0$. Now consider the modified translation invariant metric $\mathcal{J}_q^*,\ q \in (0,1)$ given by:

$$\mathcal{J}_q^*(\bar{X},\bar{Y}) = \left(\sum_{i=1}^{n} \frac{|x_i - y_i|^q}{1-|x_i-y_i|^q}\right)^{\frac{1}{q}} . \tag{10}$$

Clearly, for a uniformly distributed random variable $X \sim U(0,1)$

$$E[\|X\|] = \int_0^1 \frac{x^q}{1-x^q}\, dx\ ,$$

does not exist.

Measuring the distance as in Equation (10) is related to the concept of power transform Box and Cox [1964], Carroll and Ruppert [1981], a well known concept from statistics where it is normally applied directly to the data, whereas it is applied to the distances here.

6 Concluding remarks

In this work, we have analysed distance measures to cope with the curse of dimensionality. Our main observation is that unbounded distance measures can help to overcome certain problems caused by the curse of dimensionality. We have also given examples for such unbounded distance measures and have evaluated them based on some characteristic indices and real data sets.

However, we would like to reiterate that we do not proclaim the superiority of the unbounded measures proposed in this work. Firstly, they are only for illustrative purposes and more such measures along these lines can easily be proposed. Secondly, as mentioned repeatedly in this work, the effectiveness of a distance measure is very much contextual and hence distance measures that perform consistently well in all areas and aspects can be quite hard to come by. In any case, it is better to stick to the ideas of intelligent data analysis Berthold and Hand [2009], Berthold et al. [2010] and to involve as much domain knowledge into the data analysis and handling as possible. So if a domain-specific distance measure is known, this should be preferred over other general purpose measures. However, especially when dealing with high-dimensional data, a canonical domain-specific distance measure is not known or very difficult to define. In such cases, it is useful to try out different general purpose distance measures that reduce the effects of the curse of dimensionality.

This work can and should be seen as yet another honest effort in understanding the curse of dimensionality and an attempt at mitigating its effects.

Acknowledgements

This work was done during the visit of the first author to Department of Computer Science, Ostfalia University of Applied Sciences under the fellowship provided by the Alexander von Humboldt Foundation.

References

Charu C. Aggarwal. Re-designing distance functions and distance-based applications for high dimensional data. *SIGMOD Record*, 30(1):13–18, 2001.

Charu C. Aggarwal. Towards systematic design of distance functions for data mining applications. In Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, *KDD*, pages 9–18. ACM, 2003. ISBN 1-58113-737-0.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In Jan Van den Bussche and Victor Vianu, editors, *ICDT*, volume 1973 of *Lecture Notes in Computer Science*, pages 420–434. Springer, 2001. ISBN 3-540-41456-8.

R. Bellmann. *Adaptive Control Processes: A Guided Tour*. Princeton Univ. Press, 1961.

M.R. Berthold and D. Hand, editors. *Intelligent Data Analysis*. Springer, Berlin, 2nd edition, 2009.

M.R. Berthold, C. Borgelt, F. Höppner, and F. Klawonn. *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*. Springer, London, 2010.

Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In Catriel Beeri and Peter Buneman, editors, *ICDT*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer, 1999. ISBN 3-540-65452-6.

G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2):211–252, 1964.

R.J. Carroll and D. Ruppert. On prediction and the power transformation family. *Biometrika*, 68:609–615, 1981.

P. Demartines. *Analyse de Données par Réseaux de Neurons Auto-Organisés*. PhD dissertation, Institut Nat'l Polytechnique de Grenoble, Grenoble, France, 1994. (in French).

Kevin Doherty, Rod Adams, and Neil Davey. Non-euclidean norms and data normalisation. In *ESANN*, pages 181–186, 2004.

Kevin Doherty, Rod Adams, and Neil Davey. Unsupervised learning with normalised data and non-euclidean norms. *Appl. Soft Comput.*, 7(1):203–210, 2007.

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

Damien François, Vincent Wertz, and Michel Verleysen. The concentration of fractional distances. *IEEE Trans. Knowl. Data Eng.*, 19(7):873–886, 2007.

Alexander Hinneburg, Charu C. Aggarwal, and Daniel A. Keim. What is the nearest neighbor in high dimensional spaces? In Amr El Abbadi, Michael L. Brodie, Sharma Chakravarthy, Umeshwar Dayal, Nabil Kamel, Gunter Schlageter, and Kyu-Young Whang, editors, *VLDB*, pages 506–515. Morgan Kaufmann, 2000. ISBN 1-55860-715-3.

Chih-Ming Hsu and Ming-Syan Chen. On the design and applicability of distance functions in high-dimensional data space. *IEEE Trans. Knowl. Data Eng.*, 21(4): 523–536, 2009.

G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159179, 1985.

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

Vladimir Pestov. Intrinsic dimension of a dataset: what properties does one expect? In *IJCNN*, pages 2959–2964. IEEE, 2007.

Vladimir Pestov. An axiomatic approach to intrinsic dimension of a dataset. *Neural Networks*, 21(2-3):204–213, 2008.

P.H. Sneath and R.R. Sokal. *Numerical Taxonomy – The Priciples and Practice of Numerical Classification*. W.H. Freeman and Company, San Francisco.