On the Necessary and Sufficient Conditions of a Meaningful Distance Function for High Dimensional Data Space

Chih-Ming Hsu * Ming-Syan Chen †

Abstract

The use of effective distance functions has been explored for many data mining problems including clustering, nearest neighbor search, and indexing. Recent research results show that if the Pearson variation of the distance distribution converges to zero with increasing dimensionality, the distance function will become unstable (or meaningless) in high dimensional space even with the commonly used L_p metric on the Euclidean space. This result has spawned many subsequent studies. We first comment that although the prior work provided the sufficient condition for the unstability of a distance function, the corresponding proof has some defects. Also, the necessary condition for unstability (i.e., the negation of the sufficient condition for the stability) of a distance function, which is required for function design, remains unknown. Consequently, we first provide in this paper a general proof for the sufficient condition of unstability. More importantly, we go further to prove that the rapid degradation of Pearson variation for a distance distribution is in fact a necessary condition of the resulting unstability. With the result, we will then have the necessary and the sufficient conditions for unstability, which in turn imply the sufficient and necessary conditions for stability. This theoretical result derived leads to a powerful means to design a meaningful distance function. Explicitly, in light of our results, we design in this paper a meaningful distance function, called Shrinkage-Divergence Proximity (abbreviated as SDP), based on a given distance function. It is empirically shown that the SDP significantly outperforms prior measures for its being stable in high dimensional data space and robust to noise, and is thus deemed more suitable for distance-based clustering applications than the priorly used metric.

1 Introduction

The curse of dimensionality has recently been studied extensively on several data mining problems such as clustering, nearest neighbor search, and indexing. The curse of high dimensionality is critical not only with regards to the performance issue but also to the quality issue. Specifically, on the quality issue, the design of effective distance functions has been deemed a very important and challenging issue. Recent research results showed that in high dimensional space, the concept of distance (or proximity) may not even be qualitatively [1][2][3][5][6][11]. Explicitly, the theorem in [6] showed that under a broad set of conditions, in high dimensional space, the distance to the nearest data point approaches the distance to the farthest data point of a given query point with increasing dimensionality. For example, under the independent and identically distributed dimensions assumption, the commonly used L_p metrics will encounter problems in high dimensionality. This theorem has spawned many subsequent studies along the same line [1][2][3][11][13].

The scenario is shown in Figure 1 where ϵ denotes a very small number. From the query point, the ratio of the distance to the nearest neighbor to that to the farthest neighbor is almost 1. This phenomenon is called the unstable phenomenon [6] because there is poor discrimination between the nearest and farthest neighbors for proximity query. As such, the nearest neighbor problem becomes poorly defined. Moreover, most indexing structures will have a rapid degradation with increasing dimensionality which leads to an access to the entire database for any query [3]. Similar issues are encountered by distance-based clustering algorithms and classification algorithms to model the proximity for grouping data points into meaningful subclasses. In this paper, a distance function which will result in this unstable phenomenon is referred to as a meaningless function in high dimensional space, and is called meaningful otherwise. The result in [6] suggested that the design of a meaningful distance function in high dimensional space is a very important problem and has significant impact to a wide variety of data mining applications.

In [6], it is proved that the *sufficient* condition of unstability is that "the Pearson variation of the corresponding distance distribution degrades to 0 with increasing dimensionality". For example, if "under the independent and identically distributed dimensions assumption and the use an L_p metric, the Pearson

*Electrical Engineering Department, National Taiwan University, Taipei, Taiwan. Email: ming@arbor.ee.ntu.edu.tw.

†Electrical Engineering Department, National Taiwan University, Taipei, Taiwan. Email: mschen@cc.ee.ntu.edu.tw.

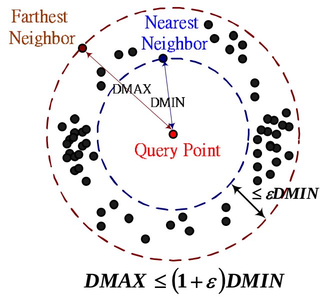


Figure 1: An example of unstable phenomenon.

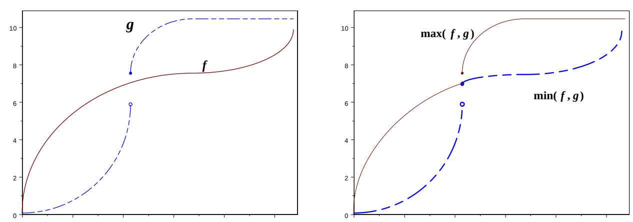


Figure 2: Example minimum and maximum of functions which are not continuous.

variation of distance distribution rapidly degrades to 0 with increasing dimensionality in high dimensional space," then the unstable phenomenon occurs. (This distance function is hence called meaningless.) Note that in light of the equivalence between "$p \rightarrow q$" and "$\neg q \rightarrow \neg p$", the negative of above sufficient condition for unstability is equivalent to the necessary condition of stability (where we have a meaningful distance function). However, the sufficient condition for stability remains unknown.

In fact, the important issue is how to design a meaningful distance (or proximity) function for high dimensional data space. The authors in [1] provided some practical desiderata for constructing a meaningful distance function, including (1) contrasting, (2) statistically sensitive, (3) skew magnification, and (4) compactness. These properties are in essence design guidelines for a meaningful distance function. However, we have no guarantee that a distance function which satisfies those needs will avoid the unstable phenomenon (since these properties are not sufficient condition for stability). Consequently, neither the result in [6] nor that in [1] provides the necessary condition for unstability which is required for us to design a meaningful distance function in high dimensional space. The design of a meaningful distance (or proximity) function hence remains as an open problem. This is the issue we shall solve in this paper.

We first comment that although the work in [6] provided the sufficient condition for unstability, the corresponding proof has some defects. [6] used the property that the minimum and maximum are continuous functions to deduce the sufficient condition for unstability, which, however, does not hold always. For example, consider the scenario of two functions f and g, as shown in Figure 2 where both the minimum and the maximum of functions f and g are discontinuous. In general, the minimum and maximum functions of random sequence are not continuous, especially for discontinuous random variables [8]. To remedy this, we will first provide in this paper a general proof for the sufficient condition of unstability. More importantly, we go further to prove that the rapid degradation of Pearson variation for a distance distribution is a necessary condition of the resulting unstability. With the result, we will then have the necessary and sufficient conditions for unstability, whose negatives in turn imply the sufficient and necessary conditions for stability. Note that with the sufficient condition for stability which was unsolved in prior works and is first derived in this paper, one will then be able to design a meaningful distance function for high dimensional space. Explicitly, this new result means that a distance function for which the degradation of Pearson variation does not approach zero rapidly will be guaranteed to be meaningful (i.e., stable) in high dimensional space. The estimation of variation is a guideline for testing the distance function is unstable or not. As such, this theoretical analysis leads a powerful means to design a meaningful distance function.

Explicitly, in light our results, we design in this paper a meaningful distance function, called Shrinkage-Divergence Proximity (abbreviated as SDP), based on a given distance function. Specifically, the SDP defines an adaptive proximity function of two data points on individual attributes separately. The proximity of two points is the aggregation of each attributive proximity. SDP magnifies the variation of the distance to detect and avoid the unstable phenomenon attribute by attribute. For each attribute of two data points, we will shrink the proximity of this attribute to zero if the projected attribute of two data points falls into a small interval. If they are more similar to each other than to others on an attribute, we are then not able to significantly discern among them statistically. On the other hand, if some projected attributes of two data points are apart from one to another for a long original distance, then they are viewed dissimilar to each other. Therefore, we will be able to spread them out to increase the degree of discrimination. Note that since we define the proximity between two data points separately on individual attributes, the noise effects of some attributes will be mitigated by other attributes. This accounts for the reason that SDP is robust to noise in our experiments.

The contributions of this paper are twofold. First, as a theoretical foundation, we provided and proved the necessary and sufficient conditions of unstable phenomenon in high dimensional space. Note that the negative of necessary condition of unstability is in essence the sufficient condition of stability, which provides an innovative and effective guideline for the design of a meaningful (i.e., dimensionality resistant) distance function in high dimensional space. Second, in light of the theoretical results derived, we developed a new dimensionality resistant proximity function SDP. It is empirically shown that the SDP significantly outperforms prior measures for its being stable in high dimensional data space and robust to noise, and is thus deemed more suitable for distance-based clustering applications than the commonly used L_p metric.

The rest of the paper is organized as follows. Section 2 describes related works. Section 3 provides theoretical results for our work where the necessary and sufficient conditions for unstable phenomenon are derived. In Section 4, we devise a meaningful distance function SDP. Experimental results are presented in Section 5. This paper concludes with Section 6.

2 Related Works

The use of effective distance functions has been explored in several data mining problems, including nearest neighbor search, indexing, and so on. As described earlier, the work in [6] showed that under a broad set of conditions the neighbor queries become unstable in high dimensional spaces. That is, from a given query point, the distance to the nearest data point will approach that to the farthest data point in high dimensional space. For example, under the commonly used assumption that each dimension is independent, the L_p metric will be unstable for many high dimensional data spaces. For constant p ($p \geq 1$), the L_p metric for two m-dimensional data points $\overrightarrow{x} = (x_1, x_2, \cdots, x_m)$ and $\overrightarrow{y} = (y_1, y_2, \cdots, y_m)$ is defined as $L_p(\overrightarrow{x}, \overrightarrow{y}) = \sum_{i=1}^{m}(|x_i - y_i|^p)^{1/p}$. The result in [6] has spawned many subsequent studies along this direction. [11] and [2] specifically examined the behavior of L_p metric and showed that the problem of stability (i.e., meaningfulness) in high dimensionality is sensitive to the value of p. A property of L_p presented in [11] is that the value of extremal difference $|Dmax_m - Dmin_m|$ grows as $m^{1/p-1/2}$ with increasing dimensionality m, where $Dmax_m$ and $Dmin_m$ are the distances to the farthest point and that to the nearest point from the origin, respectively. As a result, the L_1 metric is the only metric of the L_p family for which the absolute difference between nearest and farthest neighbor increases with the dimensionality. For the L_2 metric, $|Dmax_m - Dmin_m|$ converges to a constant, and for distance metrics L_k for $k \geq 3$, $|Dmax_m - Dmin_m|$ converges to zero with dimensionality m. This means that the L_1 metric is more preferable than the L_2 for high dimensional data mining applications. In [2], the authors also extended the notion of a L_p metric to a fractional distance function where a fractional distance function $dist_m^l$ for $l \in (0,1)$ is defined as:

$$dist_m^l(\overrightarrow{x}, \overrightarrow{y}) = \left(\sum_{i=1}^{m}(x_i - y_i)^l\right)^{1/l}.$$

In [3], the authors proposed the IGrid-index which is a method for similarity indexing. The IGrid-index used grid-based approach to redesign the similarity function from L_p. In order to perform the proximity thresholds, the IGrid-index method discretize the data space into k_d equidepth ranges. Specifically, $\mathcal{R}[i,j]$ denotes the jth range for dimension i. For dimension i, if both x_i and y_i belong to the same range $\mathcal{R}[i,j]$, then the two data points are said to be in proximity on dimension i. Let $\mathcal{S}[\overrightarrow{x}, \overrightarrow{y}, k_d]$ be the proximity set for two data points $\overrightarrow{x}$ and $\overrightarrow{y}$ with a given level of discretization k_d, then the similarity between $\overrightarrow{x}$ and $\overrightarrow{y}$ is given by:

$$PIDist(\overrightarrow{x}, \overrightarrow{y}, k_d) = \left[\sum_{i \in \mathcal{S}[\overrightarrow{x}, \overrightarrow{y}, k_d]} \left(1 - \frac{|x_i - y_i|}{m_i - n_i}\right)^p\right]^{1/p},$$

where m_i and n_i are the upper and lower bounds for the corresponding range in the dimension i in which the data points $\overrightarrow{x}$ and $\overrightarrow{y}$ are in proximity to one another. Note that these results while being valuable from various perspectives do not provide the sufficient condition for a meaningful distance function that can be used for the distance function design in high dimensional data space.

3 On a Meaningful Distance Function

In this section, we shall derive theoretical properties for a meaningful distance function. Preliminaries are given in Section 3.1. In Section 3.2, we first provide a new proof for the sufficient condition of unstability in Theorem 1 (since, as pointed out earlier, the proof in [6] has some defects). Then, we derived the necessary condition for unstability (i.e., the negative of the sufficient condition for stability) in Theorem 2. We state the complete results (the necessary and sufficient conditions for unstability) in Theorem 3. For better readability, we put the proofs of all theorems in Section 3.3. Remarks on the valid indices for a distance function to be meaningful are made in Section 3.4.

3.1 Definitions We now introduce several terms to facilitate our presentation. Assume that d_m is a real-value distance function [1] defined on a certain m-dimensional space. d_m is well defined as m increases. For example, the L_p metric defined on the m-dimensional Euclidean space is well defined as m increases for any p in $(0, \infty)$. Let $P_{m,i}$ $i = 1, 2, \cdots, N$ be N independent data points which are sampled from some m-variate distribution F_m. F_m is also well defined on some sample space as m increases. (Note that we do not assume that the attributes of data space are independent. Essentially, many of the attributes are correlated with one another [1].) If Q_m is an arbitrary (m-dimensional) query point chosen independently from all $P_{m,i}$. Let $DMAX_m = max\{d_m(P_{m,i}, Q_m) | 1 \leq i \leq N\}$ and $DMIN_m = min\{d_m(P_{m,i}, Q_m) | 1 \leq i \leq N\}$. Hence $DMAX_m$ and $DMIN_m$ are random variables for any m.

Definition 3.1. *A family of well defined distance functions* $\{d_m | m = 1, 2, \cdots\}$ *is δ-unstable (or δ-meaningless) if* $\delta = sup\{\delta^* | \lim_{m\to\infty} P\{DMAX_m \leq (1+\epsilon)DMIN_m\} \geq \delta^* \quad \text{for any } \epsilon > 0\}$.

For ease of exposition, we also refer to 1-unstable as *unstable* (or with a meaningless distance function). As δ approaches 1, a small relative change in any query point in a direction away from the nearest neighbor could change the point into the farthest neighbor. For the purpose of this study, we shall explore the extremal case: unstable phenomenon (i.e., $\delta = 1$). A distance function is called *stable* if it is not unstable (i.e., $\delta < 1$). A list of symbols used is given in Table 1.

3.2 Theoretical Properties of Unstability From the probability theory, the unstable phenomenon is equivalent to the case that $\frac{DMAX_m}{DMIN_m}$ converges in probability to one as m increases. The work in [6] proved that under the condition that *Pearson variation* $var\left(\frac{d_m(P_{m,1},Q_m)}{E[d_m(P_{m,1},Q_m)]}\right)$ of distance distribution converges to 0 with increasing dimensionality, the *extremal ratio* $\frac{DMAX_m}{DMIN_m}$ will also converge to one with increasing dimensionality. Formally, we have the following theorem.

[1]In this paper, the distance function need not be a metric. A nonnegative function $d : \mathcal{X} \times \mathcal{X} \to R$ is a *metric* for data space $\mathcal{X}$ if it satisfies the following properties: (1) $d(x, y) \geq 0 \ \ \forall x, y \in \mathcal{X}$, (2) $d(x, y) = 0$ if and only if $x = y$, (3) $d(x, y) = d(y, x) \ \ \forall x, y \in \mathcal{X}$, and (4) $d(x, z) + d(z, y) \geq d(x, y) \ \ \forall x, y, z \in \mathcal{X}$

Table 1: List of Symbols

Notation	Definition
m	Dimensionality of the data space
N	Number of data points
$P\{e\}$	Probability of an event e
X, Y, X_m, Y_m	Random variables defined on some probability space
$E[X]$	Expectation of a random variable X
$var(X)$	Variance of a random variable X
iid	Independent and identically distribution
d_m	A distance function of a m-dimensional data space $m = 1, 2, \cdots$
$x \sim F$	x is a random sample point from the distribution F
$X_n \xrightarrow{P} X$	A sequence of random variables X_1, $X_2 \cdots$ converges in probability to a random variable X if $\forall \epsilon \ \ lim_{n\to\infty} P\{\|X_n - X\| \leq \epsilon\} = 1$

Recall that this theorem was rendered in [6] and we mainly provide a correct and more general proof in this paper.

Theorem 3.1. *(Sufficient condition of unstability [6]) Let p be a constant* $(0 < p < \infty)$.
If $\lim_{m\to\infty} var\left(\frac{d_m(P_{m,1},Q_m)^p}{E[d_m(P_{m,1},Q_m)^p]}\right) = 0$, *then for every* $\epsilon > 0$

$$\lim_{m\to\infty} P\{DMAX_m \leq (1+\epsilon)DMIN_m\} = 1.$$

From Theorem 3.1, a distance function is unstable in high dimensional space if its Pearson variation of distance distribution approaches 0 with increasing dimensionality. This phenomenon leads poor discrimination between the nearest and farthest neighbor for proximity query in high dimensional space. Note that as mentioned in [6], the condition of Theorem 3.1 is applicable to a variety of data mining applications.

Example 1. Suppose that we have the data whose distribution and query are iid from some distribution with finite fourth moments in all dimensions [6]. If P_{m,i_j} and Q_{m_j} are, respectively, the jth attribute of ith data point and the jth attribute of the query point. Hence, $(P_{m,i_j} - Q_{m_j})^2$, $j = 1, 2, \cdots, m$ are iid with some expectation μ and some nonnegative variance σ^2 that are the same regardless of the values of i and m. If we use the L_2 metric for proximity query, then $d_m(P_{m,i}, Q_m) = (\sum_{j=1}^{m}(P_{m,i_j} - Q_{m_j})^2)^{1/2}$. Under the iid assumptions, we have $E[d_m(P_{m,i}, Q_m)^2] = m\mu$ and $var\left(d_m(P_{m,i}, Q_m)^2\right) = m\sigma^2$. Therefore, the Pearson variation $var(\frac{d_m(P_{m,1},Q_m)^2}{E[d_m(P_{m,1},Q_m)^2]}) = \sigma^2/m\mu^2$ converges to 0 with increasing dimensionality. Hence, the corresponding L_2 is meaningless under such a data space. In general, if the data distribution and query are iid from some distribution with finite $2p$th moments in all dimensions, the L_p metric is meaningless for all $1 \leq p < \infty$[6]. □

We next prove in Theorem 3.2 that "the condition for the Pearson variation of the distance distribution to any given target to converge to 0 with increasing dimensionality" is not only the sufficient condition (as stated in Theorem 3.1) but also the necessary condition of unstability of a distance function in high dimensional space.

Theorem 3.2. *(Necessary condition of unstability)* *If* $\lim_{m\to\infty} P\{DMAX_m \leq (1+\epsilon)DMIN_m\} = 1$ *for every* $\epsilon > 0$*, then*

$$\lim_{m\to\infty} var(\frac{d_m(P_{m,1},Q_m)^p}{E[d_m(P_{m,1},Q_m)^p]}) = 0 \ (0 < p < \infty).$$

With Theorem 3.2, we will then have the necessary and the sufficient conditions for unstability, whose negatives in turn imply the sufficient and necessary conditions for stability. The statement that the Pearson variation of the distance distribution to any given target should converge to 0 with increasing dimensionality is equivalent to the unstable phenomenon. Following Theorems 3.1 and 3.2, we reach Theorem 3.3 below which provides a theoretical guideline for designing dimensionality resistant distance functions.

Theorem 3.3. *(Main Theorem)*
Let p *be a constant* $(0 < p < \infty)$.
For every $\epsilon > 0$,

$$\lim_{m\to\infty} P\{DMAX_m \leq (1+\epsilon)DMIN_m\} = 1$$

if and only if

$$\lim_{m\to\infty} var(\frac{d_m(P_{m,1},Q_m)^p}{E[d_m(P_{m,1},Q_m)^p]}) = 0.$$

Example 2. Theorem 3.3 shows that we have to increase the variation of distance distribution for redesigning a dimensionality resistant distance function. Assume that the data points and query are iid from some distribution in all dimension, and $E[d_m(P_{m,i},Q_m)^p] = m\mu$, $var\,(D_m(P_{m,1},Q_m)^p) = m\sigma^2$ for some constant μ and $\sigma (> 0)$, as in Example 1. Since the exponential function e^x (for $x \geq 0$) is strictly convex, hence the Jensen's inequality [4] [15] suggests that the variation of distance distribution can be magnified by applying the exponential function. Let $f(x) = e^x$ for $x \geq 0$. We obtain some interesting results by applying f to some well-known distance distributions, as shown in Table 2. □

The above results show that the transformations of distance functions by the exponential function can remedy the meaningless behaviors on some high dimensional data spaces, especially for those cases that distance distributions have long tails. However, in such case, the stable property of distance functions may not be useful for application needs. In addition, *the large variation makes data sparsely, and then the concept of proximity will also be difficult to visualize.* From our experimental results, it is shown that the Pearson variations, whose distance functions are translated from L_1 metric or L_2 metric by exponential function f, will start to diverge even when encountering only 10 dimensions on many data spaces. We make some remarks on the valid indices for a distance function to be meaningful in Section 3.4.

3.3 Proof of Main Theorem In order to prove our main theorem, we need to drive some properties of unstable phenomenon. For interest of space, those properties and some important theorems of probability are presented in Appendix A.

Proof for Theorem 3.1:
Let $V_{m,i} = \frac{d_m(P_{m,i},Q_m)^p}{E[d_m(P_{m,i},Q_m)^p]}$ and $\overrightarrow{X_m} = (V_{m,1}, V_{m,2}, \cdots, V_{m,N})$.
Hence, $V_{m,i}$ $i = 1,2,\cdots,N$ are non-negative iid random variables for all m and $V_{m,i} \xrightarrow{P} 1$ as $m \to \infty$.
Since for any $\epsilon > 0$,
$lim_{m\to\infty} P\{|\max(\overrightarrow{X_m}) - 1| \geq \epsilon\}$
$= lim_{m\to\infty} P\{\max(\overrightarrow{X_m}) \geq (1+\epsilon)\}$
$+ lim_{m\to\infty} P\{0 \leq \max(\overrightarrow{X_m}) \leq (1-\epsilon)\}$
$= \lim_{m\to\infty} (1 - P\{V_{m,j} < (1+\epsilon) \quad \forall j = 1,2,\cdots,N\})$
$+ \lim_{m\to\infty} P\{V_{m,j} \leq (1-\epsilon) \ \ \forall j = 1,2,\cdots,N\}$
(by Theorem A.4)
$= \lim_{m\to\infty} \left(1 - \prod_{j=1}^{N} P\{V_{m,j} < (1+\epsilon)\}\right)$
$+ \lim_{m\to\infty} \prod_{j=1}^{N} P\{V_{m,j} \leq (1-\epsilon)\}$
(by $V_{m,j}$ $j = 1,2,\cdots,N$ are iid and Theorem A.4)
$= 0$ (by $V_{m,j} \xrightarrow{P} 1$ as $m \to \infty$), and
$lim_{m\to\infty} P\{|min(\overrightarrow{X_m}) - 1| \geq \epsilon\}$
$= lim_{m\to\infty} P\{min(\overrightarrow{X_m}) \geq (1+\epsilon)\}$
$+ lim_{m\to\infty} P\{0 \leq min(\overrightarrow{X_m}) \leq (1-\epsilon)\}$
$= \lim_{m\to\infty} P\{V_{m,j} \geq (1+\epsilon) \ \ \forall j = 1,2,\cdots,N\}$
$+\lim_{m\to\infty} (1 - P\{V_{m,j} > (1-\epsilon) \ \ \forall j = 1,2,\cdots,N\})$
(by Theorem A.4)
$= \lim_{m\to\infty} \prod_{j=1}^{N} P\{V_{m,j} \geq (1+\epsilon)\}$
$+ \lim_{m\to\infty} \left(1 - \prod_{j=1}^{N} P\{V_{m,j} > (1-\epsilon)\}\right)$
(by $V_{m,j}$ $j = 1,2,\cdots,N$ are iid and Theorem A.4)

Table 2: The Pearson variation of some translated distance functions.

Distance distribution	Binomial	Uniform	Normal	Gamma	Exponential
$\lim_{m\to\infty} var(\frac{f(d_m(P_{m,1},Q_m)^p)}{E[f(d_m(P_{m,1},Q_m)^p)]})$	0	0	∞	∞	∞

$= 0$ (by $V_{m,j} \xrightarrow{P} 1$ as $m \to \infty$),
then $max(\overrightarrow{X_m})$ and $min(\overrightarrow{X_m})$ converge in probability to 1 as $m \to \infty$.
Further, by Slutsky's theorem, proposition 2 of Theorem A.1, and

$$\frac{DMAX_m}{DMIN_m} = \left(\frac{E[(d_m(P_{m,i},Q_m)^p]max(\overrightarrow{X_m})}{E[(d_m(P_{m,i},Q_m)^p]min(\overrightarrow{X_m})}\right)^{1/p} = \left(\frac{max(\overrightarrow{X_m})}{min(\overrightarrow{X_m})}\right)^{1/p},$$

then $\frac{DMAX_m}{DMIN_m} \xrightarrow{P} 1$ as $m \to \infty$. **Q.E.D.**

Proof for Theorem 3.2:
Let $W_{m,i} = d_m(P_{m,i},Q_m)^p$.
By third property of Lemma A.1 and the second property of Theorem A.1, we have $\frac{W_{m,i}}{W_{m,j}} \xrightarrow{P} 1$ for any i, j.
Hence, an application of Theorem A.3 and the forth property of Lemma A.1, we have

$$var\left(\frac{W_{m,i}}{W_{m,j}}\right) = E\left[var\left(\frac{W_{m,i}}{W_{m,j}}|W_{m,j}\right)\right] + var\left(E\left[\frac{W_{m,i}}{W_{m,j}}|W_{m,j}\right]\right)$$

and

$$\lim_{m\to\infty} var\left(\frac{W_{m,i}}{W_{m,j}}\right) = 0,$$

respectively.
Furthermore, since $E\left[var\left(\frac{W_{m,i}}{W_{m,j}}|W_{m,j}\right)\right]$ and $var\left(E\left[\frac{W_{m,i}}{W_{m,j}}|W_{m,j}\right]\right)$ are nonnegative for all m, hence

$$\lim_{m\to\infty} E\left[var\left(\frac{W_{m,i}}{W_{m,j}}|W_{m,j}\right)\right] = 0 \tag{3.1}$$

and

$$\lim_{m\to\infty} var\left(E\left[\frac{W_{m,i}}{W_{m,j}}|W_{m,j}\right]\right) = 0. \tag{3.2}$$

Also, $var\left(\frac{W_{m,i}}{W_{m,j}}|W_{m,j}=x\right) \geq 0$ for all x and for all m, therefore Equation (3.1) implies that the probability of this set $\{x|var\left(\frac{W_{m,i}}{W_{m,j}}|W_{m,j}=x\right) > 0\}$ must be 0 or

$$var\left(\frac{W_{m,i}}{W_{m,j}}|W_{m,j}=x\right) = 0 \text{ for all } x,$$

as $m \to \infty$.
Set $x = E[d_m(P_{m,i},Q_m)^p]$, hence we have

$$\lim_{m\to\infty} var(\frac{d_m(P_{m,i},Q_m)^p}{E[d_m(P_{m,i},Q_m)^p]}) = 0 \text{ for any } i.$$

Then

$$\lim_{m\to\infty} var(\frac{d_m(P_{m,1},Q_m)^p}{E[d_m(P_{m,1},Q_m)^p]}) = 0.$$

(Note that $W_{m,i}$ $i = 1, 2, \cdots$ are iid for all m.) **Q.E.D.**

The main theorem, i.e., Theorem 3.3, thus follows from Theorem 3.1 and 3.2.

3.4 Remarks on Indices to Test Meaningful Functions

Many researchers used the extremal ratio $\frac{DMIN_m}{DMAX_m}$ [6] or the relative contrast $\frac{DMAX_m - DMIN_m}{DMAX_m}$ [2] to test the stable phenomenon of distance functions. However, as shown by our experimental results, these indices are sensitive to outliers and found inconsistent for many cases. Here, we will comment on the reason that Pearson variation is a valid index to evaluate the stable phenomenon of distance functions.

In light of Theorem 3.3 devised, we have the relationship shown in Figure 3, which leads to the following four conceivable indices to test the stable (or unstable) phenomenon:

Index 1. (*Extremal ratio*) $\frac{DMIN_m}{DMAX_m}$;
Index 2. (*Pearson variation*) $var\left(\frac{d_m(P_{m,1},Q_m)}{E[d_m(P_{m,1},Q_m)]}\right)$;
Index 3. (*Mean of extremal ratio*) $E\left[\frac{DMIN_m}{DMAX_m}\right]$;
Index 4. (*Variance of extremal ratio*) $var\left(\frac{DMIN_m}{DMAX_m}\right)$.

A robust meaningful distance function needs to satisfy $\lim_{m\to\infty} var(\frac{d_m(P_{m,1},Q_m)}{E[d_m(P_{m,1},Q_m)]}) > 0$ independently of the distribution of data set. Suppose that we use those indices to evaluate the meaningless behavior, for some given distance functions, through all possible data sets. If $\lim_{m\to\infty} \frac{DMIN_m}{DMAX_m} = 1$[2] or $\lim_{m\to\infty} var\left(\frac{d_m(P_{m,1},Q_m)}{E[d_m(P_{m,1},Q_m)]}\right) = 0$, we can conclude that this distance function is meaningless. On the other hand, we can say that it is stable if $\lim_{m\to\infty} var\left(\frac{d_m(P_{m,1},Q_m)}{E[d_m(P_{m,1},Q_m)]}\right) >$

[2]Note that $DMAX_m$ and $DMIN_m$ are random variables. Therefore, this convergence is much stronger than convergence in probability [4][7][14][15].

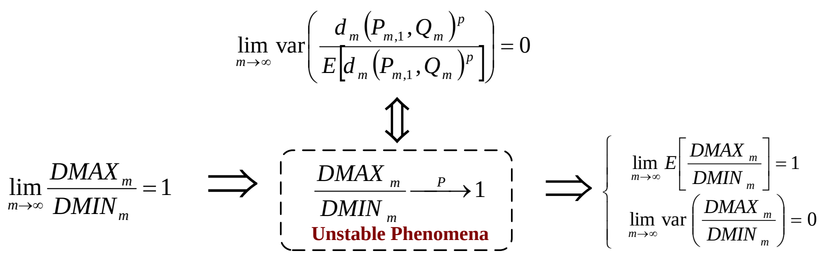


Figure 3: The convergent relationships of extremal ratio.

0, $lim_{m\to\infty} E\left[\frac{DMIN_m}{DMAX_m}\right] \neq 1$, or $lim_{m\to\infty} var\left(\frac{DMIN_m}{DMAX_m}\right) > 0$.

If we decide to apply some distance-based data mining algorithms to explore a given high dimensional data set. We first need to select a meaningful distance function depending on our applications. Therefore, we may compute those indices for each candidate of distance function, using the whole data set or a sampling subset, to evaluate its meaningful behavior. However, both the mean of extremal ratio (Index 3) and the variance of extremal ratio (Index 4) are invalid to estimate in this case. Also, though one can deduce that the distance is meaningless if its $\frac{DMIN_m}{DMAX_m}$ value is very close to one, it is not decided whether the function is meaningful or not if its value of extremal ratio (Index 1) is apart from one. In addition, the extremal ratio (Index 1) is sensitive to outliers. On other hand, if we apply some resampling techniques, such as bootstrap [10], to test the stable property of distance functions for some given high dimensional data set. The extremal ratio (Index 1) could be inconsistent for many sampling subsets. Consequently, in light of the theoretical results derived and also as will be validated, Pearson variation (Index 2) emerges as the index to use to evaluate the meaningfulness of distance function.

4 Shrinkage-Divergence Proximity

As deduced before, the unstable phenomenon is rooted in the variation of the distance distribution. In light of Theorem 3.3 devised, we will next design a new proximity function, called SDP (**S**hrinkage-**D**ivergence **P**roximity), based on some well defined family of distance functions $\{d_m | m = 1, 2, \cdots\}$.

4.1 Definition of SDP Let f be a non-negative real function defined on the set of non-negative real numbers such that

$$f_{a,b}(x) = \begin{cases} 0 & \text{if } 0 \le x < a, \\ x & \text{if } a \le x < b, \\ e^x & \text{otherwise.} \end{cases}$$

For any m-dimensional data points $\overrightarrow{x} = (x_1, x_2, \cdots, x_m)$ and $\overrightarrow{y} = (y_1, y_2, \cdots, y_m)$, we define the SDP function as

$$SDP^G(\overrightarrow{x}, \overrightarrow{y}) = \sum_{i=1}^{m} w_i f_{s_{i1}, s_{i2}}(d_1(x_i, y_i)).$$

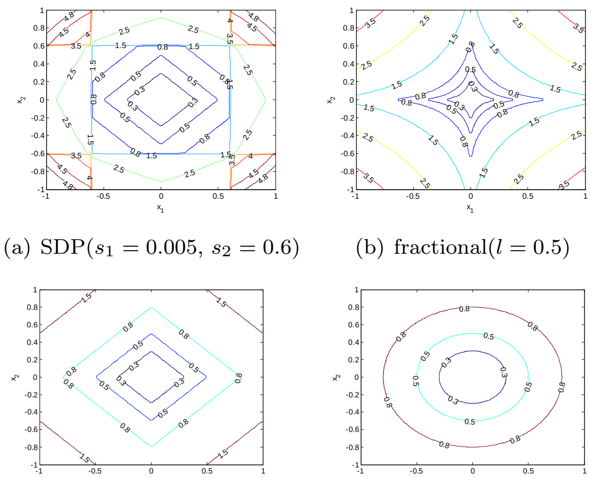
The general form of SDP, denoted by SDP^G defines a distance function $f_{s_{i1}, s_{i2}}$ between $\overrightarrow{x}$ and $\overrightarrow{y}$ on each individual attribute. Parameters w_i $i = 1, 2, \cdots, m$ are determined by the domain knowledge subject to the importance of attribute i for application needs. Also, parameters s_{i1} and s_{i2} for attribute i are dependent on the distribution of the data points projected on ith dimension.

In many situations, we have no prior knowledge on the weights of importance among attributes, and do not know the distribution of each attribute either. In such a case, the general SDP function, i.e., SDP^G, will be degenerated to,

$$SDP_{s_1, s_2}(\overrightarrow{x}, \overrightarrow{y}) = \sum_{i=1}^{m} f_{s_1, s_2}(d_1(x_i, y_i)).$$

In this paper, we only discuss the properties and applications on this SDP. The parameters s_1 and s_2 are, respectively, called as *shrinkage threshold* and *divergence threshold*. For illustrative purposes, we present in Figure 4 some 2-dimensional balls of center $(0, 0)$ with different radius for SDP, fractional function, L_1 metric, and L_2 metric.

4.2 Properties of SDP We next discuss the properties of SDP for similarity search and data clustering problems.



(a) SDP($s_1 = 0.005$, $s_2 = 0.6$) (b) fractional($l = 0.5$)

(c) L_1 metric (d) L_2 metric

Figure 4: The 2-dimensional balls of center $(0, 0)$ for different radius.

Proposition of SDP 1.
1. If d_m is the L_1 metric defined on Euclidean space, then SDP_{s_1,s_2} is equivalent to L_1 as $s_1 \to 0$ and $s_2 \to \infty$.
2. $SDP_{s_1,s_2}(\overrightarrow{x}, \overrightarrow{y}) = 0$ if and only if $0 \leq d_1(x_i, y_i) < s_1$ for all i.
3. $SDP_{s_1,s_2}(\overrightarrow{x}, \overrightarrow{y}) \geq me^{s_2}$ if $d_1(x_i, y_i) \geq s_2$ for all i.

The first property shows that SDP is a general form of L_1 metric. The second property means that the SDP is similar to grid approaches [3] in that all data points within a small rectangle are more similar to each other than to others. Thus, we cannot significantly discern among them statistically. Therefore, it is reasonable to shrink the proximity of them to zero. In order to construct a noise insensitive proximity function, and to avoid over magnifying the distance variation, the SDP defines an adaptive proximity of two data points on individual attributes. For two data points, if values of any attribute are projected into the same small interval, we will shrink the proximity of this attribute to zero. On the other hand, if all projected attributes of two data points are apart from one to another for a long original distance, then they are dissimilar to each other. As such, we are able to spread them out to increase discrimination. The SDP can remedy the edge effects problem of grid approach [3] caused by two adjacent grids which may contain data points very close to one another. It is worth mentioning that same as the fractional function [2] and *PIDist* function of the IGrid-index [3], the SDP function is in essence not a metric with triangle inequality. It can be verified that $SDP_{s_1,s_2}(\overrightarrow{x}, \overrightarrow{y}) = 0$ does not imply $\overrightarrow{x} = \overrightarrow{y}$ for $s_1 > 0$ and the triangle inequality does not hold in general. However, the influence of triangle inequality is usually insignificant in many clustering applications [9][12], in particular for high dimensional space.

Statistical View. Here, we examine the perspectives of SDP for clustering applications statistically. Assume that the data distributions are independent in all dimensions of attributes. Given a small nonnegative number ϵ, let s_{i1} be the maximum value (or supremum) such that $P\{d_1(x_i, y_i) \leq s_{i1}\} \leq \epsilon$ if $\overrightarrow{x}$ and $\overrightarrow{y}$ belong to distinct clusters. Similarly, let s_{i2} be the minimum value (or infimum) such that $P\{d_1(x_i, y_i) \geq s_{i2}\} \leq \epsilon$ if $\overrightarrow{x}$ and $\overrightarrow{y}$ belong to the same cluster. Set $s_1 = min\{s_{i1} | i = 1, 2, \cdots, m\}$ and $s_2 = max\{s_{i2} | i = 1, 2, \cdots, m\}$. Then, both

$$P\{SDP_{s_1,s_2}(\overrightarrow{x}, \overrightarrow{y}) = 0 | \overrightarrow{x} \text{ and } \overrightarrow{y} \text{ belong to distinct clusters}\} \leq (\epsilon)^m$$

and

$$P\{SDP_{s_1,s_2}(\overrightarrow{x}, \overrightarrow{y}) \geq me^{s_2} | \overrightarrow{x} \text{ and } \overrightarrow{y} \text{ belong to the same cluster}\} \leq (\epsilon)^m$$

will approach 0 as m is large.

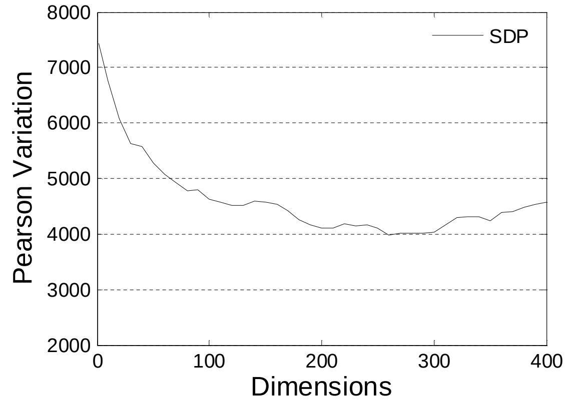
Furthermore, SDP is insensitive to noise in high dimensional space, because SDP disposes individual attribute separately. The noise effects of some attributes will be mitigated by other attributes. Also, the SDP is able to avoid spreading data points too sparsely for discrimination. Overall, SDP has better discrimination power than the original distance function, and is hence more proper for distance-based clustering algorithms.

5 Experimental Evaluation

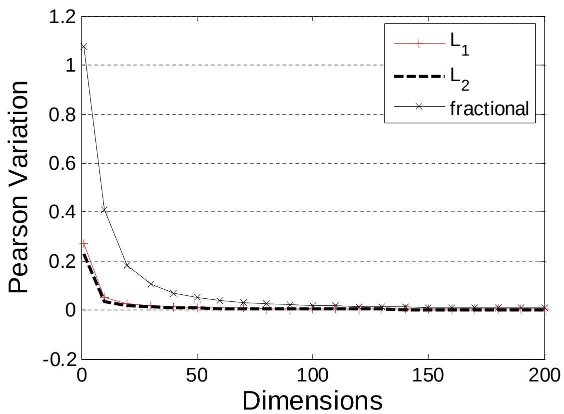
To assess the stableness and the performance of SDP function, we have conducted a series of experiments. We compare in our study the stable behaviors and the performances for distance-based clustering of SDP with several well-known distance functions, including L_1 metric, L_2 metric, and fractional distance function dis^l_m.

Meaningful Behaviors. First we compared the stable behavior of SDP, which is based on L_P metric, with several widely used of distance functions. We use the Pearson variation (Index 2): $var\left(\frac{d_m(P_{m,1}, Q_m)}{E[d_m(P_{m,1}, Q_m)]}\right)$, the mean of the extremal ratio (Index 3): $E\left[\frac{DMIN_m}{DMAX_m}\right]$, and the variance of the extremal ratio (Index 4): $var\left(\frac{DMIN_m}{DMAX_m}\right)$

to evaluate the stability of those distance functions. Recall that comments on these indices and their use are given in Section 3.4.



(a) SDP



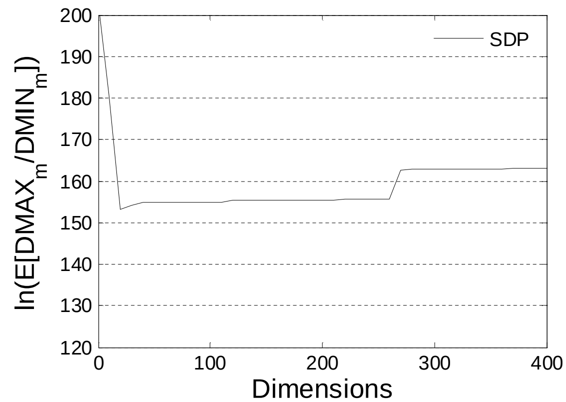
(b) L_1, L_2, and fractional function

Figure 5: The average value of Pearson variation.

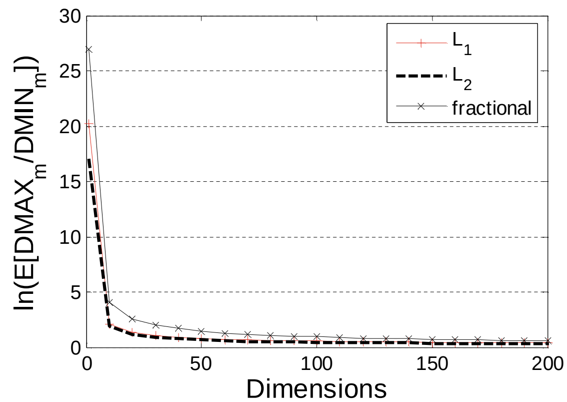
The synthetic m-dimensional sample data sets and query points of our experiments were generated as follows. Each data set includes 10 thousand independent data points. For each data set, the jth entry x_{ij} of ith data point $\overrightarrow{x_i} = (x_{i1}, x_{i2}, \cdots, x_{i,m})$ was randomly sampled from uniform distribution $U(a, b+a)$, or exponential distribution $Exp(\lambda)$, or normal distribution $N(\mu, \sigma)$. In each generation for x_{ij}, the parameters a, b, λ, μ, and σ are randomly sampled from uniform distributions with range $(0, 100)$, $(0, 100)$, $(0.1, 2)$, $(0, 100)$, and (0.1,10), respectively. Formally, for each data set, for any $i = 1, 2, \cdots, 10000$, for any $j = 1, 2, \cdots, m$,

$$x_{ij} \sim \begin{cases} U(a, a+b) & \text{with probability} 1/3, \\ Exp(\lambda) & \text{with probability} 1/3, \\ N(\mu, \sigma) & \text{with probability} 1/3, \end{cases}$$

where a, b, $\mu \sim U(0, 100)$, $\lambda \sim U(0.1, 2)$, and $\sigma \sim U(0.1, 10)$. The query points were also generated by the same manner. We repeated such process to generate 100 data sets for each dimensionality m. Dimensionality m varied from one to 1000. The shrinkage threshold and divergence threshold of SDP are s_1=0.005 and s_2=0.6, respectively. The parameter l for the fractional distance function dis^l_m was set as 0.5. The estimations of $E\left[\frac{DMIN_m}{DMAX_m}\right]$, $var\left(\frac{DMIN_m}{DMAX_m}\right)$ and the average value of $var\left(\frac{d_m(P_{m,1}, Q_m)}{E[d_m(P_{m,1}, Q_m)]}\right)$ were computed in natural logarithm scale to measure the meaningful behavior.



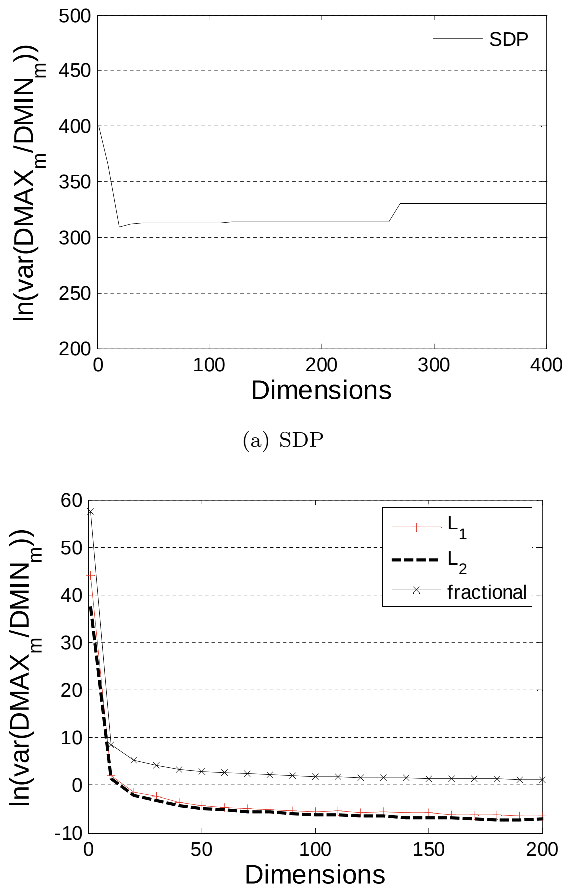
(a) SDP



(b) L_1, L_2, and fractional function

Figure 6: The estimations of the mean of the extremal ratio (in logarithmic scale).

Those results are shown from Figure 5 to Figure 7. (In order to perceive the differences among these performances easily, we only present the outputs of the first 400 (or 200) dimensions.) As shown in these figures, the SDP is an effective dimensionality resistant distance function. It is noted that L_1 and L_2 metric become unstable with as few as 20 dimensions. The fractional distance function is more effective at preserving meaningfulness of proximity than L_1 and L_2, but starts to suffer from unstability after the dimensionality exceeds 80. In contrast, the SDP remains stable even if the dimensionality is greater than 1000, showing the prominent advantage of using SDP.



(a) SDP

(b) L_1, L_2, and fractional function

Figure 7: The estimations of the variance of the extremal ratio (in logarithmic scale).

Clustering Applications. In order to compare the qualitative performances, we applied the SDP function and L_2 metric for clustering on multivariate mixture normal models. We synthesized many mixtures of two m-variate Gaussian data sets with diagonal covariance matrix. All dimensions of Cluster 1 and Cluster 2 are sampled iid from normal distribution $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. We applied matrix powering algorithm [16] on those generated data sets to compare the performances of SDP with L_2 metric. The outline of matrix powering algorithm is given in Table 3 [16].

Table 3: The matrix powering algorithm.

Algorithm: Matrix Powering Algorithm

Inputs: A(pairwise distance matrix), ϵ

Output: A partition of data points into clusters

1. Compute $A^2 = A \times A$

2. for each pair of yet unclassified points i, j

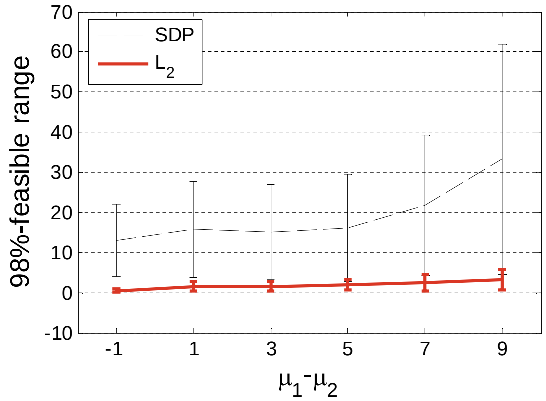
  a. If $\sqrt{(A_i^2 - A_j^2) \times (A_i^2 - A_j^2)^T} < \epsilon$, then i and j are in the same cluster.

  b. If $\sqrt{(A_i^2 - A_j^2) \times (A_i^2 - A_j^2)^T} \geq \epsilon$, then i and j are in different clusters.

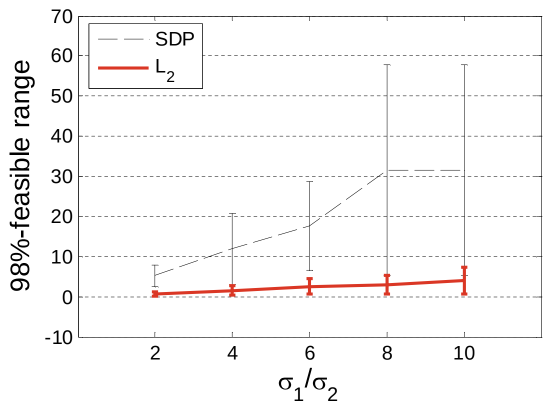
For any matrix M, we use M_j, M_{ij}, and M^T to denote, respectively, the j-th row of M, the ij-th entry of M and the transpose of M. Let the *precision ratio* of the algorithm be the percentage of the $\binom{N}{2}$ pairwise relationship (classified as same or different cluster) that it partitions correctly. Suppose that the expectation of $A_{i,j}$ is p (respectively, q) for data points i and j in the same cluster (respectively, different clusters). Also, $q > p$. Then, the optimal threshold given in [16] is $\epsilon = (q-p)^2 N^{3/2}/\sqrt{2}$. However, the knowledge of $q - p$ is usually unavailable. In addition, the scales of SDP and L_2 metric varied. We then modify the threshold as $\epsilon(k) = kmN(DMAX_m - DMIN_m)/2$ for variant k, and search the *r-feasible range* which is defined as the maximal interval of k such that the precision ratio is at least r with threshold $\epsilon(k)$.

We empirically investigated the behaviors of L_2 and SDP by using the above matrix powering algorithm. The shrinkage and divergence thresholds of SDP were set to 0.005 and $6(\sigma_1 + \sigma_2)$, respectively. First, we used 100-dimensional synthetic data sets drawn from the mixture of two normal distributions in variant clusters for mean difference $\mu_1 - \mu_2$. Each data set has 200 data points, and each cluster contains 100 data points. The results are shown in Figure 8a. We also considered the 100-dimensional multivariate mixture models with several variance ratios σ_1/σ_2, and showed the results in Figure 8b. Finally, we tested both SDP and L_2 metric for searching feasible ranges with increasing dimensionality. The empirical results are shown in Figure 9.

Note that for using matrix powering algorithm to solve our data clustering problems, we first need to choose an optimal threshold. A wider feasible range offers more adequate solution space to this problem. As shown in these figures, the feasible ranges of L_2



(a) $\sigma_1 = 1$, $\mu_2 = 3$, $\sigma_2 = 4$



(b) $\mu_1 = \mu_2 = 3$, $\sigma_2 = 1$

Figure 8: (a) The feasible ranges for SDP and L_2 with varied mean difference of two clusters, (b) the feasible ranges for SDP and L_2 with varied variance ratios of two clusters.

metric are much narrower and the boundary points of feasible ranges are very close to 0. On the other hand, the feasible ranges of SDP are much wider than L_2 even in high dimensional space. Further, the SDP obtains appropriate response to varying characters of clusters. A larger mean difference (or variance) ratio of two clusters implies a better discrimination between them. As shown in Figure 8, the feasible range of SDP becomes wider with magnifying the mean difference or the variance ratios of two clusters. Also, the width of feasible ranges for SDP increase with increasing dimensionality as in Figure 9. On the other hand, due to the unstable phenomenon, the width of feasible ranges for L_2 rapidly degrades to 0 with increasing dimensionality. From Figure 8 and Figure 9, it is shown that SDP significantly outperforms the priorly used L_2 metric.

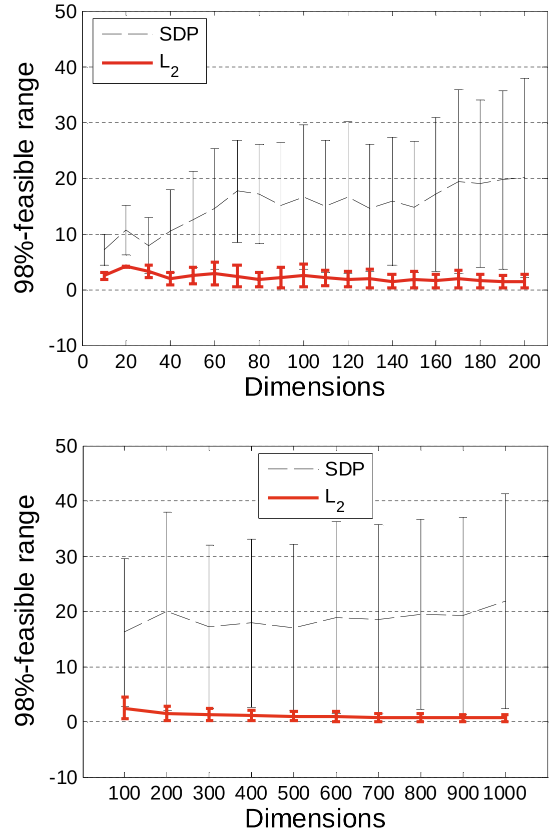


Figure 9: The feasible range for SDP and L_2 with varied dimensionality. $((\mu_1 = 8, \sigma_1 = 1)$ v.s. $(\mu_1 = 3, \sigma_1 = 4))$

6 Conclusions

In this paper, we derived the necessary and the sufficient conditions for the stability of a distance function in high dimensional space. Explicitly, we proved that the rapidly degraded Pearson variation of distance distribution with increasing dimensionality is equivalent to (i.e., being necessary and sufficient conditions of) unstable phenomenon. This theoretical result on the sufficient condition of a meaningful distance function design derived in this paper leads a powerful means to test the stability of a distance function in high dimensional data space. Explicitly, in light of our results, we have designed a meaningful distance function SDP based on a certain given distance function. It was empirically shown that the SDP significantly outperforms prior measures for its being stable in high dimensional data space and also robust to noise, and is thus deemed more suitable for distance-based clustering applications than the priorly used L_p metric.

References

[1] C. C. Aggarwal. Re-designing distance functions and distance-based applications for high dimensional data. *ACM SIGMOD Record*, Vol. 30, pp.13-18, 2001.

[2] C. C. Aggarwal, A. Hinneburg, and D. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *ICDT Conference*, 2001.

[3] C. C. Aggarwal,and P. S. Yu. The IGrid Index: Reversing the Dimensionality Curse for Similarity Indexing in High Dimensional Space. *ACM SIGKDD Conference*, 2000.

[4] P. J. Bickel, and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Vol. 1. 2nd edition. Prentice Hall, 2001.

[5] K. P. Bennett, U, . Fayyad, and D. Geiger. Density-Based Indexing for Approximate Nearest Neighbor Queries. *ACM SIGKDD Conference*, 1999.

[6] Beyer K., Goldstein J., Ramakrishnan R, and Shaft U. When is Nearest Neighbors Meaningful? *ICDT Conference Proceedings*, 1999.

[7] K. L. Chung. *A Course in Probability Theory*. 3rd edition. Academic Press, 2001.

[8] H. A. David, and H. N. Nagaraja. *Order Statistics*. 3rd edition. Wiley & Sons, 2003.

[9] L. Ertöz, M. Steinbach, and V. Kumar. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. *SDM Conference*, 2003.

[10] B. Efron, and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[11] A. Hinneburg, C. C. Aggarwal,and D. Keim. What is the nearest Neighbor in High Dimensional Spaces? *VLDB Conference*, 2000.

[12] A. K. Jain, M. N. Murty and P. J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 1999.

[13] N. Katayama, and S. Satoh. Distinctiveness-Sensitive Nearest-Neighbor Search for Efficient Similarity Retrieval of Multimedia Information. *ICDE Conference*, 2001.

[14] V. V. Petrov. *Limit Theorems of Probability Theory*. Oxford University Press, 1995

[15] V. K. Rohatgi, and A. K. Md. Ehsanes Saleh. *An Introduction to Probability and Statistics*. 2nd edition. Wiley, 2001.

[16] H. Zhou, and D. Woodruff. Clustering via Matrix Powering. *PODS Conference*, 2004.

Appendix A: Related Theorems on Probability

In order to prove our main theorem, we present some important theorems from the probability theory [15] [7] [8]. The proofs are omitted for interest of space.

Theorem A. 1. *If* $X_m \xrightarrow{P} X$, $Y_m \xrightarrow{P} Y$ *and* g *is a continuous function defined on real numbers, then we have the following properties:*
1. $X_m - X \xrightarrow{P} 0$.
2. $g(X_m) \xrightarrow{P} g(X)$.
3. $aX_m \pm Y_m \xrightarrow{P} aX \pm Y$ *for any constant* a.
4. $X_m Y_m \xrightarrow{P} XY$.
5. $X_m/Y_m \xrightarrow{P} X/a$, *provided* $Y = a(\neq 0)$ *(Slutsky's theorem).*

Theorem A. 2. *If* $X_m \xrightarrow{P} 1$ *then* $X_m^{-1} \xrightarrow{P} 1$.

Theorem A. 3. *If* $E[X^2] < \infty$ *then* $var(X) = var(E[X|Y]) + E[var(X|Y)]$.

Theorem A. 4. *If* X_j $j = 1, 2 \dots, m$ *are independent, then* $P\{max(X_1, \dots, X_m) \leq \epsilon\} = P\{X_1 \leq \epsilon,\ X_2 \leq \epsilon,\ \dots,\ X_m \leq \epsilon\} = \prod_{j=1}^{m} P\{X_j \leq \epsilon\}$.

In order to prove the necessary condition, we also need to derive the following lemma.

Lemma A. 1. *For every* $\epsilon > 0$, *if*

$$\lim_{m\to\infty} P\{DMAX_m \leq (1+\epsilon)DMIN_m\} = 1$$

then we have the following properties:
1. $\frac{DMIN_m}{DMAX_m} - 1 \xrightarrow{P} 0$.
2. $\lim_{m\to\infty} E[\frac{DMAX_m}{DMIN_m}] = 1$ *and* $\lim_{m\to\infty} var\left(\frac{DMAX_m}{DMIN_m}\right) = 0$.
3. For any i, j, $\lim_{m\to\infty} P\{d_m(P_{m,i}, Q_m) \leq (1+\epsilon) d_m(P_{m,j}, Q_m)\} = 1$.
4. For any i, j, $\lim_{m\to\infty} E[\frac{d_m(P_{m,i},Q_m)}{d_m(P_{m,j},Q_m)}] = 1$ *and* $\lim_{m\to\infty} var\left(\frac{d_m(P_{m,i},Q_m)}{d_m(P_{m,j},Q_m)}\right) = 0$.

Proof. 1. The first proposition follows from Theorem A.2.
2. From the probability theory [15] ,the following properties are all equivalent:
a. $\lim_{m\to\infty} P\{DMAX_m \leq (1+\epsilon)DMIN_m\} = 1$,
b. $\frac{DMAX_m}{DMIN_m} - 1 \xrightarrow{P} 0$,
c. $\frac{DMAX_m}{DMIN_m} - 1$ converges in distribution to the degenerate distribution $D(x)$, where $D(x) = 1$ if $x > 0$ and $D(x) = 0$ if $x \leq 0$.
Hence, we have $\lim_{m\to\infty} E[\frac{DMAX_m}{DMIN_m}] = 1$ and $\lim_{m\to\infty} var\left(\frac{DMAX_m}{DMIN_m}\right) = 0$.
3. Since $\frac{DMIN_m}{DMAX_m} \leq \frac{d_m(P_{m,i},Q_m)}{d_m(P_{m,j},Q_m)} \leq \frac{DMAX_m}{DMIN_m}$ for any i, j, hence we have $\frac{d_m(P_{m,i},Q_m)}{d_m(P_{m,j},Q_m)} \xrightarrow{P} 1$
*4.*The third proposition also leads to the fourth one.

■