Measuring the Difficulty of Distance-Based Indexing

Matthew Skala

University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada
mskala@cs.uwaterloo.ca

Abstract. Data structures for similarity search are commonly evaluated on data in vector spaces, but distance-based data structures are also applicable to non-vector spaces with no natural concept of dimensionality. The intrinsic dimensionality statistic of Chávez and Navarro provides a way to compare the performance of similarity indexing and search algorithms across different spaces, and predict the performance of index data structures on non-vector spaces by relating them to equivalent vector spaces. We characterise its asymptotic behaviour, and give experimental results to calibrate these comparisons.

1 Introduction

Suppose we wish to index a database for similarity search. For instance, we might have a database of text documents which we query with an example document to find others close to the example. Speaking of closeness implies we must have a distance function applicable to the objects in the database. Maybe our objects are actually vectors of real numbers with a Minkowski L_p metric. Many effective data structures are known for that case, including R-trees and variants [3,11,17], SR-trees [13], and pyramid-trees [4].

But maybe the objects are not vectors; and maybe the distance function is not an L_p metric. Edit distance on strings, for instance, forms a metric space that is not a vector space. Structures for indexing general metric spaces include VP-trees [20], MVP-trees [5], GH-trees [19], and FQ-trees [2]. Such structures are called "distance-based" because they rely exclusively on the distances between the query point and other points in the space.

The problem of distance-based indexing seems to become harder in spaces with more dimensions, but we cannot easily count dimensions in a non-vector space. Even when we represent our documents as long vectors, indexing algorithms behave much differently on real document databases from the prediction for similar-length randomly generated vectors. In this work we consider how to predict indexing performance on practical spaces by comparison with random vector spaces of similar difficulty.

1.1 Intrinsic Dimensionality

Suppose we have a general space, from which we can choose objects according to a fixed probability distribution, and measure the distance between any two

M. Consens and G. Navarro (Eds.): SPIRE 2005, LNCS 3772, pp. 103–114, 2005.
© Springer-Verlag Berlin Heidelberg 2005

objects, but the objects are opaque: all we know about an object is its distance from other objects. We might wish to assume that we have a metric space, with the triangle inequality, but even that might only hold in an approximate way—for instance, only to within a constant factor, as with the "almost metrics" defined by Sahinalp and others [16]. Some functions we might like to use do not naturally obey the triangle inequality—such as relative entropy measured by compression, proposed in bioinformatics applications [10,14].

Given such a space, the only way we can describe the space or distinguish it from other general spaces is by choosing random points and considering the probability distribution of distances between them. Chávez and Navarro introduce a statistic called "intrinsic dimensionality" for describing spaces in terms of the distribution of the distance between two randomly chosen points. Where μ and σ^2 are the mean and variance of that distance, the intrinsic dimensionality ρ is defined as $\mu^2/(2\sigma^2)$ [6]. Squaring the mean puts it in the same units as the variance; and as we prove, the constant 2 makes ρ equal the number of vector dimensions for uniform random vectors with L_1 and approach it for normal random vectors with L_2.

Chávez and Navarro prove bounds on the performance of several kinds of distance-based index structures for metric spaces in terms of ρ. Spaces that are easy to index have small ρ, and the statistic increases as the spaces become harder to index. They also give an argument (using a proof of Yianilos) for why intrinsic dimensionality ought to be proportional to the number of vector components when applied to points chosen uniformly at random from vector spaces [6,21]. To calibrate the dimensionality measurement, they show experimental results for low-dimensional spaces to estimate the asymptotic constant of proportionality for ρ in terms of n, with n-component vectors having each component chosen from a uniform distribution and using L_p metrics [6].

We analyse the behaviour of ρ for vectors chosen with independent identically distributed real components, and distance measured by an L_p metric; the result is exact for L_1. We find that $\rho(n)$ is $\Theta(n)$ for L_p with finite p, but not necessarily for L_∞. We show ρ to be $\Theta(\log^2 n)$ in the case of normally-distributed random vectors with the L_∞ metric. We also present experimental results corroborating our theory. The slopes of the lines are found to be significantly greater than predicted by previous experiments, because the true asymptotic behaviour only shows itself at large n. The behaviour of the asymptotic lines as p varies is seen to be counter-intuitive, with the L_∞ metric on uniform vectors much different from the L_p metric for large but finite p.

1.2 Notation

Following the notation used by Arnold, Balakrishnan, and Nagaraja, [1] we write $X \stackrel{d}{=} Y$ if X and Y are identically distributed, $X \stackrel{d}{\to} Y$ if the distribution of $X(n)$ converges to the distribution of Y as n goes to positive infinity, and $X \stackrel{d}{\leftrightarrow} Y$ if the distributions of both X and Y depend on n and converge to each other. We also write $f(n) \to x$ if x is the limit of $f(n)$ as n goes to positive infinity, $E[X]$ and $V[X]$ for the expectation and variance of X respectively, $\log x$ for the natural

logarithm of x, and $\Gamma(x)$ for the standard gamma function (generalised factorial). Random variables that are **independent and identically distributed** are called iid, a random variable's **probability density function** is called its pdf, and its **cumulative distribution function** is called its cdf.

Let $X = Y$ be a real random variate realised as random variables X_i and Y_i. Let $\boldsymbol{x}_n = \langle X_1, X_2, \ldots, X_n \rangle$ and $\boldsymbol{y}_n = \langle Y_1, Y_2, \ldots, Y_n \rangle$ be vector random variables with n iid components each, each component drawn from the variate. Let $D_{p,n}$ be the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ under the L_p metric $\|\boldsymbol{x} - \boldsymbol{y}\|_p$, defined as $\left(\sum_{i=1}^{n} |X_i - Y_i|^p\right)^{1/p}$ for real $p > 0$ or $\max_{i=1}^{n} |X_i - Y_i|$ where $p = \infty$. We are concerned with the distribution of the random variable $D_{p,n}$, and in particular the asymptotic behaviour for large n of the intrinsic dimensionality statistic $\rho_p(n) = E[D_{p,n}]^2/2V[D_{p,n}]$ [6].

When discussing L_∞, which is defined in terms of the maximum function, it is convenient to define for any real random variate Z random variates $\max^{(k)}\{Z\}$ and $\min^{(k)}\{Z\}$ realised as random variables $\max_i^{(k)}\{Z\}$. and $\min_i^{(k)}\{Z\}$ respectively. Each $\max_i^{(k)}\{Z\}$ is the maximum, and each $\min_i^{(k)}\{Z\}$ the minimum, of k random variables from Z.

1.3 Extreme Order Statistics

Extreme order statistics of collections of random variables (the maximum, the minimum, and generalisations of them) have been thoroughly studied [1,9]. If $F(x)$ is the cdf of Z, then $F^n(x)$ is the cdf of $\max^{(n)}\{Z\}$. We say that the random variable W with non-degenerate cdf $G(x)$ is the limiting distribution of the maximum of Z if there exist sequences $\{a_n\}$ and $\{b_n > 0\}$ such that $F^n(a_n + b_n x) \to G(x)$. There are only a few possible distributions for W, if it exists at all.

Theorem 1 (Fisher and Tippett, 1928). *If* $(\max^{(n)}\{Z\} - a_n)/b_n \xrightarrow{d} W$*, then the cdf* $G(x)$ *of* W *must be of one of the following types, where* α *is a constant greater than zero [1, Theorem 8.3.1] [8]:*

$$G_1(x;\alpha) = \exp(-x^{-\alpha}) \text{ for } x > 0 \text{ and } 0 \text{ otherwise;} \tag{1}$$

$$G_2(x;\alpha) = \exp(-(-x)^{\alpha}) \text{ for } x < 0 \text{ and } 1 \text{ otherwise; or} \tag{2}$$

$$G_3(x) = \exp(-e^{-x}) \ . \tag{3}$$

2 Intrinsic Dimensionality of Random Vectors

Even though intrinsic dimensionality is most important for non-vector spaces, like strings with edit distance, we wish to know the behaviour of the intrinsic dimensionality statistic on familiar vector spaces so we can do meaningful comparisons. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be random n-component vectors as described above, using the L_p metric. We will compute the asymptotic behaviour of the intrinsic dimensionality $\rho_p(n)$ as n goes to infinity, based on the distribution of $|X - Y|$. Let μ'_k represent the k-th raw moment of $|X - Y|$; that is, the expected value of $|X - Y|^k$.

2.1 The L_p Metric for Finite p

We would like the intrinsic dimensionality statistic to be proportional to the length of the vectors when applied to random vectors with iid components and distance measured by L_p metrics. For finite p as the number of components goes to infinity, it does indeed behave that way.

Theorem 2. *With the L_p metric for fixed finite p, when the $|X_i - Y_i|$ are iid with raw moments μ'_k, then $\rho_p(n) \to [p^2(\mu'_p)^2/(2(\mu'_{2p} - (\mu'_p)^2))]n$.*

Proof. The L_p metric for finite p is computed by taking the sum of random variables $|X_i - Y_i|^p$; call the result S. Then the metric is $S^{1/p}$. We have $V[|X_i - Y_i|^p] = \mu'_{2p} - (\mu'_p)^2$, $E[S] = n\mu'_p$, and $V[S] = n(\mu'_{2p} - (\mu'_p)^2)$.

Since the mean and variance both increase linearly with n, the standard deviation will eventually become small in relation to the mean. For large n we can approximate the function $x^{1/p}$ with a tangent line:

$$E[S^{1/p}] \to E[S]^{1/p} = n^{1/p}(\mu'_p)^{1/p} \tag{4}$$

$$V[S^{1/p}] \to V[S]\left(\frac{d}{dS}S^{1/p}\right)^2\Bigg|_{S=E[S]} = \frac{\mu'_{2p} - (\mu'_p)^2}{np^2(\mu'_p)^2}n^{2/p}(\mu'_p)^{2/p} \tag{5}$$

$$\rho_p(n) = \frac{E[S^{1/p}]^2}{2V[S^{1/p}]} \to n\frac{p^2(\mu'_p)^2}{2(\mu'_{2p} - (\mu'_p)^2)} \tag{6}$$

□

If we are using the L_1 metric, the analysis is even better:

Corollary 1. *When $p = 1$, the approximation given by Theorem 2 becomes exact: $\rho_1(n) = [(\mu'_1)^2/(2(\mu'_2 - (\mu'_1)^2))]n$.*

Proof. When $p = 1$, then $E[S^{1/p}] = E[S] = E[S]^{1/p}$ and $V[S^{1/p}] = V[S] = V[S]^{1/p}$, regardless of n, and the limits for large n in the proof of Theorem 2 become equalities. □

2.2 Binary Strings with Hamming Distance

Binary strings under Hamming distance are an easy case for the theory, and are of interest in applications like the Nilsimsa spam filter [7]. We can find the intrinsic dimensionality of the space of n-bit binary strings under Hamming distance by treating the strings as vectors with each component a Bernoulli random variable, equal to one with probability q and zero otherwise. Then the Hamming distance (number of bits with differing values) is the same as the L_1 distance (sum of absolute component-wise differences), and by Corollary 1, $\rho_1(n) = nq(1-q)/(1-2q+2q^2)$.

Note that $q = 1/2$ produces the maximum value of $\rho_1(n)$, namely $n/2$. Substituting into the lower bound of Chávez and Navarro, we find that a pivot-based algorithm using random pivots on a database of m strings each n bits long, with the Hamming metric, must use at least $\frac{1}{2}(\sqrt{n} - 1/\sqrt{f})^2 \ln m$ distance evaluations on average per query, to satisfy random queries returning at most a fraction f of the database [6].

2.3 The L_∞ Metric

The distance $D_{\infty,n}$ is the maximum of n variables drawn from $|X - Y|$. We can eliminate the absolute value function with the following lemma.

Lemma 1. *If Z is a real variate with distribution symmetric about zero, and W, a_n, and b_n exist with $(\max^{(n)}\{Z\} - a_n)/b_n \xrightarrow{d} W$, then $\max^{(n)}\{|Z|\} \overset{d}{\leftrightarrow} \max^{(2n)}\{Z\}$.*

Proof. Instead of taking the maximum absolute value of a set of n variables from Z, we could find the maximum and the negative of the minimum and then take the maximum of those two. But as described by Arnold, Balakrishnan, and Nagaraja, the maximum and minimum of a collection of random variables are asymptotically independent [1, Theorem 8.4.3]. Thus $\max^{(n)}\{|Z|\} \overset{d}{\leftrightarrow} \max\{\max^{(n)}\{Z\}, -\min^{(n)}\{Z\}\}$; and by symmetry of Z,

$$\max\{\max^{(n)}\{Z\}, -\min^{(n)}\{Z\}\} \overset{d}{=} \max^{(2n)}\{Z\} \tag{7}$$

□

Given the distribution of $X - Y$ or $|X - Y|$, we can obtain the limiting distribution for $D_{\infty,n} = \max^{(n)}\{|X - Y|\}$; and if it exists, it will be in one of the three forms stated in Theorem 1. We can then integrate to find the expectation and variance, and standard results give acceptable choices for the norming constants a_n and b_n, giving the following theorem.

Theorem 3. *For random vectors with the L_∞ metric, when Theorem 1 applies to $\max^{(n)}\{|X - Y|\}$, we have:*

$$\rho_\infty(n) \to \frac{(a_n + b_n\Gamma(1 - 1/\alpha))^2}{2b_n^2(\Gamma(1 - 2/\alpha) - \Gamma^2(1 - 1/\alpha))} \text{ for } G_1(x;\alpha), \alpha > 2; \tag{8}$$

$$\rho_\infty(n) \to \frac{(a_n + b_n\Gamma(1 + 1/\alpha))^2}{2b_n^2(\Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha))} \text{ for } G_2(x;\alpha); \text{ and} \tag{9}$$

$$\rho_\infty(n) \to \frac{3(a_n + b_n\gamma)^2}{b_n^2\pi^2} \text{ for } G_3(x); \tag{10}$$

where $\gamma = 0.57721\,56649\,015\ldots$, the Euler-Mascheroni constant. □

Unlike in the finite-p case, $\rho_\infty(n)$ does not necessarily approach a line.

2.4 Uniform Vectors

Let X and Y be uniform real random variates with the range $[0, 1)$, as used by Chávez and Navarro in their experiment [6]. The pdf of $|X - Y|$ is $2 - 2x$ for $0 \le x < 1$. Simple integration gives the raw moments $\mu'_p = 2/(p + 1)(p + 2)$ and $\mu'_{2p} = 1/(2p + 1)(p + 1)$, and then by Theorem 2, $\rho_p(n) \to [(4p + 2)/(p + 5)]n$.

For the L_∞ metric, we note that the cdf of $|X - Y|$ is $F(x) = 2x - x^2$. Then standard results on extreme order statistics [1, Theorems 8.3.2(ii), 8.3.4(ii)] give

us that $(\max^{(n)}\{|X-Y|\}-a_n)/n_b \xrightarrow{d} W$ where $a_n = 1, b_n = 1/\sqrt{n}$, and the cdf of W is $G_2(x;\alpha)$ with $\alpha = 2$. By (9), $\rho_\infty(n) \to n/(2-(\pi/2))$. So as n increases, $\rho_\infty(n)$ approaches a line with slope $1/(2-(\pi/2)) = 2.32989\,61831\,6\ldots$; the same line approached by $\rho_{\tilde{p}}(n)$ where $\tilde{p} = (1+\pi)/(7-2\pi) = 5.77777\,31051\,9\ldots$.

We repeated the experiment described by Chávez and Navarro [6, Fig. 3], of randomly choosing one million pairs of points, finding their distances, and computing the intrinsic dimensionality. The results are shown in Fig. 1. Examination reveals an apparent linear trend for each metric, but the points seem to be on much shallower lines than the theory predicts. The points for L_{256} match those for L_∞, supporting the intuition that L_p for large p should have the same asymptotic behaviour as L_∞.

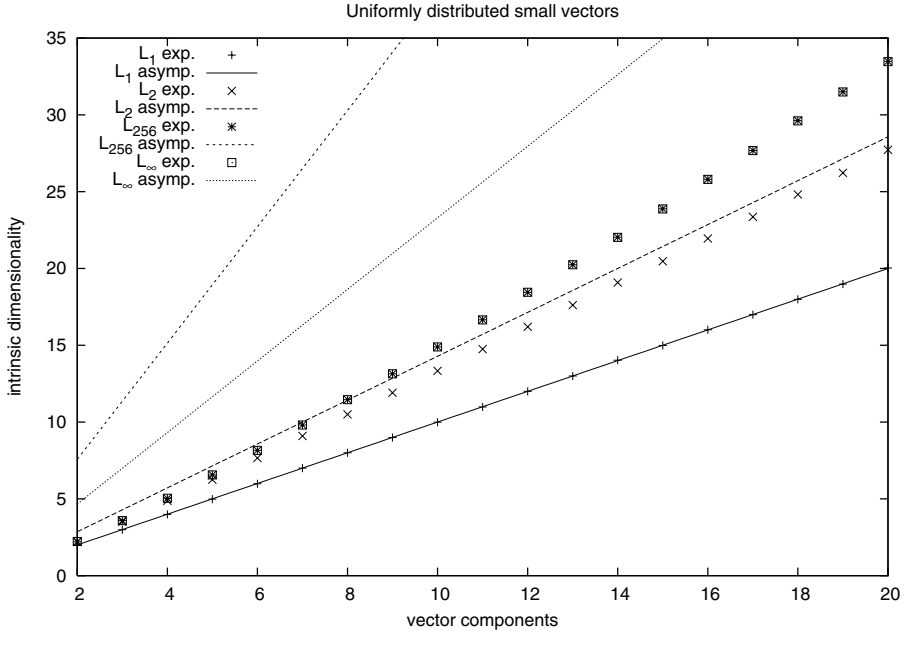

Fig. 1. Experimental results for short vectors with uniform random components

Intuition turns out to be wrong. Repeating the experiment with vectors of up to one million components (Fig. 2), we see that the line for L_p does approach a slope of four as p increases, but with the L_∞ metric, the line drops to coincide with the line for $L_{\tilde{p}}$, $\tilde{p} \approx 5.778$, just as predicted by the theory. This phenomenon is actually not quite so strange as it may seem: this is simply a situation where we are taking two limits and it matters which order we take them.

2.5 Normal Vectors

Consider a similar case but let X and Y be standard normal random variates. Since X and Y are standard normal, their difference $X-Y$ is normal with mean

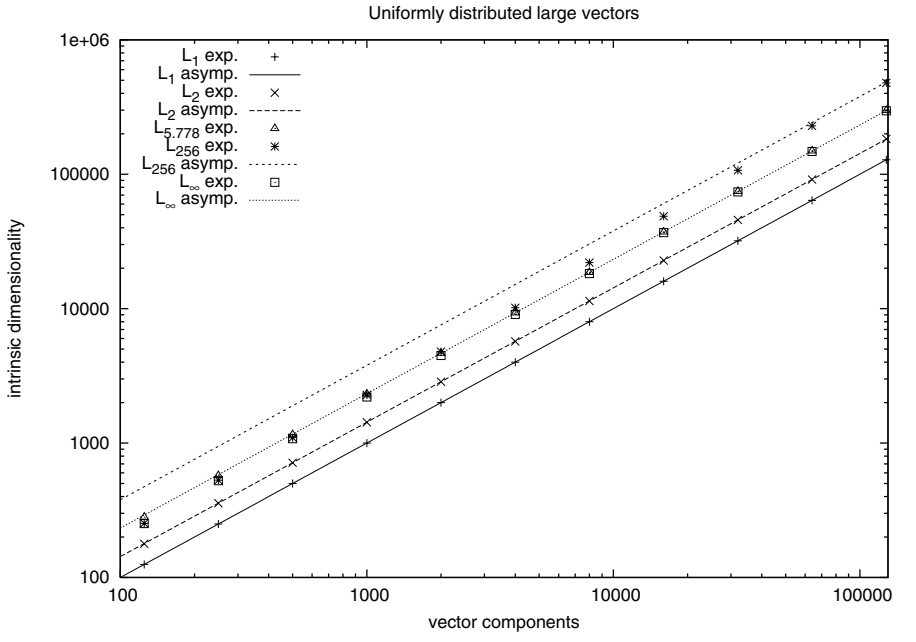

Fig. 2. Experimental results for long vectors with uniform random components

zero and variance two. Then each $|X - Y|$ has a "half-normal" distribution with pdf $(\sqrt{2}/\pi)e^{-x^2/2\pi}$. As before, we compute the raw moments and substitute into the intrinsic dimensionality formula, finding that $\mu'_p = \pi^{(p-1)/2}2^{p/2}\Gamma((p+1)/2)$, $\mu'_{2p} = \pi^{p-1/2}2^p\Gamma(p+1/2)$, and so $\rho_p(n) \to n[p^2\Gamma^2((p+1)/2)/2(\sqrt{\pi}\Gamma(p+1/2) - \Gamma^2((p+1)/2))]$. As in the uniform case, $\rho_p(n) = \Theta(n)$, but the slope is quite different. The maximum slope is one, with the L_2 metric; L_1 and L_3 give slopes of approximately 0.9; and for larger p the slope rapidly approaches zero.

Now, $D_{\infty,n} = \max^{(n)}\{|X - Y|\}$. By Lemma 1 we can instead consider $\max^{(2n)}\{X - Y\}$. Each $X - Y$ is normal with mean zero and variance two. Standard results on the maximum of normal random variables give us that $(D_{\infty,n} - a_{2n})/b_{2n} \xrightarrow{d} W$ where the cdf of W is $G_3(x) = \exp(-e^{-x})$ and the norming constants are $a_{2n} = 2\sqrt{\log 2n} - (\log(4\pi \log 2n))/2\sqrt{\log 2n}$ and $b_{2n} = 1/\sqrt{\log 2n}$ [1,9,12]. Then we can substitute into (10) to find the asymptotic intrinsic dimensionality $\rho_\infty(n) \to (3/4\pi^2) \cdot [4 \log n - \log\log 2n + \log(4/\pi) + 2\gamma]^2$, which is $\Theta(\log^2 n)$.

As with uniform vector components, the intrinsic dimensionality shows markedly different asymptotic behaviour with the L_∞ metric from its behaviour with L_p metrics for finite p; but here, instead of being linear with a surprising slope, it is not linear at all. The argument for linear behaviour from Yianilos [21, Proposition 2] only applies to finite p.

To verify these results, we generated one million pairs of randomly-chosen vectors for a number of combinations of vector length and L_p metric, and calculated the intrinsic dimensionality. The results are shown in Figs. 3 and 4 along with the theoretical asymptotes. As with uniform components, the true asymptotic behaviour for some metrics is only shown at the largest vector sizes.

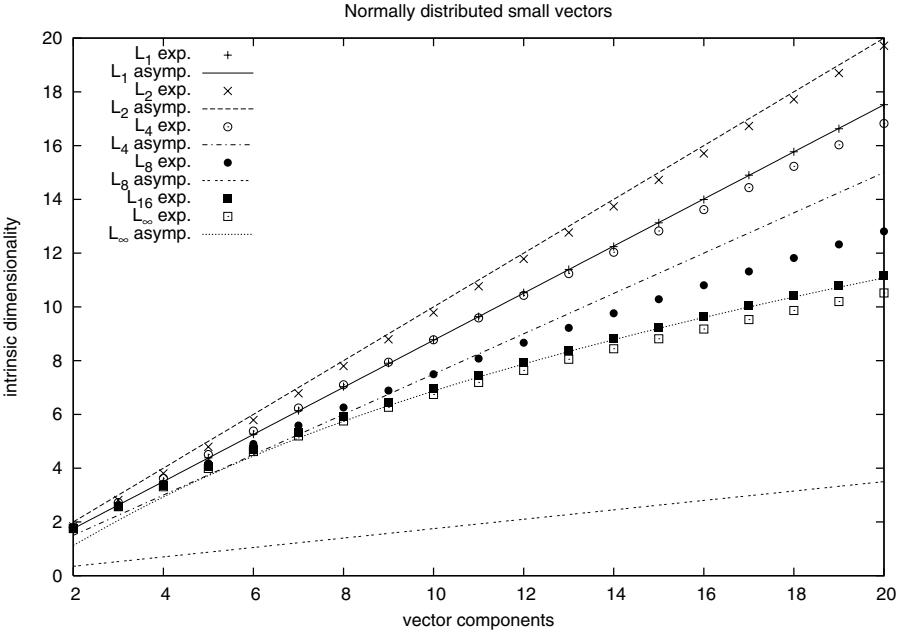

Fig. 3. Experimental results for short vectors with normal random components

Normal distributions in high-dimensional vector spaces have smaller intrinsic dimensionality than uniform distributions with vectors of the same length, when considered with L_∞ and L_p for large p. Does that mean normal distributions are easier to index, or only that intrinsic dimensionality is a poor measure of indexing difficulty? We argue that normal distributions really are easier to index.

A random vector $\boldsymbol{x}$ from a high-dimensional normal distribution will typically have many small components and one, or a few, of much greater magnitude. Comparing $\boldsymbol{x}$ to another random point $\boldsymbol{y}$, the greatest components of $\boldsymbol{x}$ will usually correspond to small components of $\boldsymbol{y}$ and vice versa, so the L_∞ distance between the two will usually be approximately equal to the one largest component of either vector. At high enough dimensions we could closely approximate the distances between points in almost all cases by only examining the index and magnitude of the single greatest component of each vector. We could achieve good indexing by just putting the points into bins according to index of greatest component, and using cheap one-dimensional data structures within bins. That is how pyramid-trees work [4], and they work well in this case.

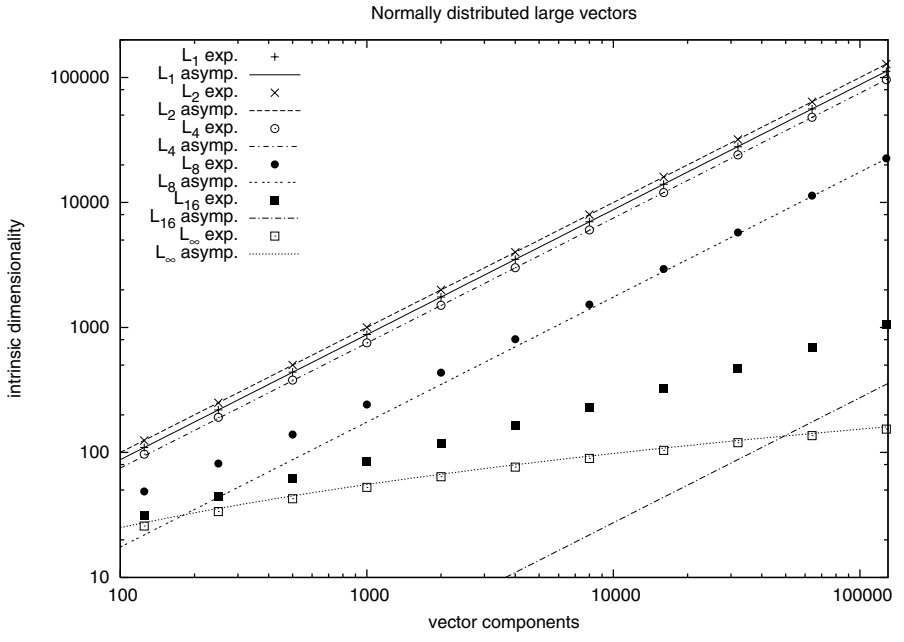

Fig. 4. Experimental results for long vectors with normal random components

However, when the vectors are selected from a uniform distribution, then componentwise differences have a triangular distribution, with heavier tails. More components of the difference vector are likely to be large and have a chance of determining the distance, so the indexing structure must represent more information per vector.

3 Other Spaces

Random vector spaces are of interest for calibrating the intrinsic dimensionality statistic, but practical spaces may be more difficult to analyse. Here we show the application of the statistic to some other spaces of interest.

3.1 Balls in Hamming Spaces

Consider a ball of radius r in the space of n-bit binary strings; that is, a fixed n-bit string $\boldsymbol{c}$ and all the strings with Hamming distance from $\boldsymbol{c}$ equal to or less than r. If we consider this set as a metric space itself, using the Hamming distance and choosing points uniformly at random from the set, what is its intrinsic dimensionality?

Theorem 4. *For a ball in the space of n-bit strings of constant radius r using the Hamming metric and choosing strings uniformly at random, $\rho \to [r/(2r+1)]n$.*

Proof. Consider how many ways we could choose i of the n bits, then j of the remaining $n-i$ bits, then k of the remaining $n-i-j$ bits. This number is given by the multichoose function $(i,j,k,n-i-j-k)! = n!/i!j!k!(n-i-j-k)!$. If we choose two strings $\boldsymbol{x}$ and $\boldsymbol{y}$ from the ball, let i be the number of bit positions where $\boldsymbol{x}$ is different from $\boldsymbol{c}$ and $\boldsymbol{y}$ is equal, let j be the number of bit positions where $\boldsymbol{y}$ is different from $\boldsymbol{c}$ and $\boldsymbol{x}$ is equal, and then let k (which must be from zero to $r - \max\{i,j\}$) be the number of bit positions where $\boldsymbol{x}$ and $\boldsymbol{y}$ are both different from $\boldsymbol{c}$ and thus equal to each other. We can count the number of ways to choose these two strings as

$$N = \sum_{i=0}^{r}\sum_{j=0}^{r}\sum_{k=0}^{r-\max\{i,j\}} (i,j,k,n-i-j-k)! \tag{11}$$

$$= \frac{1}{r!^2}n^{2r} - \frac{r-3}{r!(r-1)!}n^{2r-1} + o(n^{2r-1}) \ . \tag{12}$$

Similarly, by finding the leading terms of the sums and applying long division, we can find expressions for the first two raw moments of the distance for two strings chosen uniformly at random from the ball:

$$\mu_1' = \frac{1}{N}\sum_{i=0}^{r}\sum_{j=0}^{r}\sum_{k=0}^{r-\max\{i,j\}} (i+j)(i,j,k,n-i-j-k)! \tag{13}$$

$$= 2r - 2r(r+1)n^{-1} + o(n^{-1}) \tag{14}$$

$$\mu_2' = \frac{1}{N}\sum_{i=0}^{r}\sum_{j=0}^{r}\sum_{k=0}^{r-\max\{i,j\}} (i+j)^2(i,j,k,n-i-j-k)! \tag{15}$$

$$= 4r^2 - 2r(4r^2+2r+1)n^{-1} + o(n^{-1}) \ . \tag{16}$$

Then by substitution into the intrinsic dimensionality formula, we obtain $\rho \to [r/(2r+1)]n$. □

3.2 An Image Database

We constructed an image database by selecting frames at random from a selection of commercial DVD motion pictures, choosing each frame with 1/200 probability to create a database of 3239 images, which were converted and scaled to give 259200-element vectors representing the RGB colour values for 360×240 pixels. Sampling 10^5 pairs of these vectors using each of the L_2 and L_∞ metrics produced ρ values of 2.759 for L_2 and 38.159 for L_∞. These results suggest that the L_2 metric reveals much stronger clumping structure on this database than the L_∞ metric does; and with L_2, this database is approximately as hard to index as a three-dimensional normal distribution in L_2 space ($\rho = 2.813$, from the experiment shown in Fig. 3). If we have a choice about which metric to use, the L_2 metric will produce a much more efficient index than the L_∞ metric.

3.3 A Text Database

We obtained a sample of 28999 spam email messages from SpamArchive.org [18], and added 2885, or approximately 10 percent, non-spam messages from locally collected outgoing email, to simulate the database a practical spam-filtering application might process. We sampled 10^5 pairs of messages, computed their distances using the Perl Digest::Nilsimsa [15] 256-bit robust hash, and Hamming distance, and computed the intrinsic dimensionality $\rho = 10.338$. For the spam messages alone, and for the non-spam messages alone, we obtained ρ values of 10.292 and 11.262 respectively, again with sampling of 10^5 pairs for each database. An index of the email database based on Hamming distance of the Nilsimsa hashes would perform better than a similar index on uniform random 256-bit strings, but answering queries would still be quite difficult, a little more difficult than for random data normally distributed in 10-dimensional L_2 space.

4 Conclusions and Future Work

Intrinsic dimensionality answers questions about spaces: which spaces have comparable indexing difficulty, which metrics will allow good indexing, and lower bounds on query complexity. We have characterised the asymptotic behaviour of the intrinsic dimensionality statistic for randomly chosen vectors with the components having uniform or normal distributions, and the L_p metrics for both finite and infinite p. As our theoretical results show, uniform and normal components produce vastly different results, especially for L_p with large p and L_∞. In those metrics, high-dimensional normal distributions are easier to index than uniform distributions of the same dimension. We have also given results for more complicated spaces: balls in Hamming space, and practical databases of images and email messages, demonstrating the flexibility of the technique.

The ultimate question for indexing difficulty measurement is how much making a distance measurement reduces our uncertainty about the query point's distance to points in the index. Intrinsic dimensionality attempts to answer the question based on the mean and variance of the distribution of a single distance; but we might obtain a more useful statistic by considering the joint distribution of distances among more than two randomly chosen points. Such a statistic could allow the proof of highly general bounds on indexing performance.

References

1. Arnold, B.C., Balakrishnan, N., Nagaraja, H.N.: A First Course in Order Statistics. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., New York (1992)
2. Baeza-Yates, R.A., Cunto, W., Manber, U., Wu, S.: Proximity matching using fixed-queries trees. In: CPM (Combinatorial Pattern Matching). Volume 807 of Lecture Notes in Computer Science., Springer (1994) 198–212
3. Beckmann, N., Kriegel, H.P., Schneider, R., Seeger, B.: The R*-tree: An efficient and robust access method for points and rectangles. In: SIGMOD (International Conference on Management of Data). (1990) 322–331

4. Berchtold, S., Böhm, C., Kriegel, H.P.: The pyramid-tree: Breaking the curse of dimensionality. In: SIGMOD (International Conference on Management of Data). (1998) 142–153
5. Bozkaya, T., Ozsoyoglu, M.: Indexing large metric spaces for similarity search queries. ACM Transactions on Database Systems **24** (1999) 361–404
6. Chávez, E., Navarro, G.: Measuring the dimensionality of general metric spaces. Technical Report TR/DCC-00-1, Department of Computer Science, University of Chile (2000) Submitted. Online `ftp://ftp.dcc.uchile.cl/pub/users/gnavarro/metricmodel.ps.gz`.
7. Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., Samarai, P.: An open digest-based technique for spam detection. In: 2004 International Workshop on Security in Parallel and Distributed Systems, San Francisco, CA, USA (2004)
8. Fisher, R.A., Tippett, L.H.C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society **24** (1928) 180–190
9. Galambos, J.: The Asymptotic Theory of Extreme Order Statistics. Second edn. Robert E. Krieger Publishing Company, Malabar, Florida, U.S.A. (1987)
10. Grumbach, S., Tahi, F.: A new challenge for compression algorithms: Genetic sequences. Journal of Information Processing and Management **30** (1994) 875–886
11. Guttman, A.: *R*-trees: a dynamic index structure for spatial searching. SIGMOD Record (ACM Special Interest Group on Management of Data) **14** (1984) 47–57
12. Hall, P.: On the rate of convergence of normal extremes. Journal of Applied Probability **16** (1979) 433–439
13. Katayama, N., Satoh, S.: The SR-tree: an index structure for high-dimensional nearest neighbor queries. In: SIGMOD (International Conference on Management of Data). (1997) 369–380
14. Li, M., Badger, J.H., Xin, C., Kwong, S., Kearney, P., Zhang, H.: An information based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics **17** (2001) 149–154
15. Norwood, C., cmeclax: Digest::Nilsimsa 0.06. Computer software (2002) Online `http://search.cpan.org/~vipul/Digest-Nilsimsa-0.06/`.
16. Sahinalp, S.C., Tasan, M., Macker, J., Ozsoyoglu, Z.M.: Distance based indexing for string proximity search. In: ICDE (International Conference on Data Engineering), IEEE Computer Society (2003)
17. Sellis, T.K., Roussopoulos, N., Faloutsos, C.: The R+-tree: A dynamic index for multi-dimensional objects. In: VLDB'87 (International Conference on Very Large Data Bases), Morgan Kaufmann (1987) 507–518
18. SpamArchive.org: Donate your spam to science. Web site (2005) Online `http://www.spamarchive.org/`.
19. Uhlmann, J.K.: Satisfying general proximity/similarity queries with metric trees. Information Processing Letters **40** (1991) 175–179
20. Yianilos, P.N.: Data structures and algorithms for nearest neighbor search in general metric spaces. In: SODA (Symposium on Discrete Algorithms), SIAM (1993) 311–321
21. Yianilos, P.N.: Excluded middle vantage point forests for nearest neighbour search. In: ALENEX (Algorithm Engineering and Experimentation: International Workshop). (1999)