How to Approximate the Inner-product: Fast Dynamic Algorithms for Euclidean Similarity

Ömer Eğecioğlu
Department of Computer Science
University of California at Santa Barbara
omer@cs.ucsb.edu

Abstract

We develop dynamic dimensionality reduction based on the approximation of the standard inner-product. This results in a family of fast algorithms for checking similarity of objects whose feature representations are large dimensional real vectors, a common situation in various multimedia databases.

The method uses the power symmetric functions of the components of the vectors, which are powers of the p-norms of the vectors for $p = 1, 2, \ldots, m$. The number m of such norms used is a parameter of the algorithm whose simplest instance gives a first-order approximation implied by the Cauchy-Schwarz inequality. We show how to compute fixed coefficients that work as universal weights based on the moments of the probability density function assumed for the distribution of the components of the input vectors in the data set. If the distribution of the components of the vectors is not known we show how the method can be adapted to work dynamically by incremental adjustment of the parameters.

1 Introduction

Modern databases use various kinds of data such as images, audio, video, maps, etc. Example of these applications include Geographical Information Systems (GIS) [8], Multimedia Information Systems [10, 18], CAD/CAM [6], medical imaging [9]. A general search approach is to represent the data objects as multidimensional points and to define the similarity between objects by the distance between the corresponding multidimensional points. It is assumed that the closer the points, the more similar the data objects. Both the dimensionality and the amount of data that need to be processed increases very rapidly and consequently it becomes important to support high dimensional similarity searching in large-scale systems. This support depends on the development of efficient techniques to support approximate search techniques for high dimensional data sets. There have been several index structures for retrieval of multidimensional data. Examples of these include kdb-trees [17], hB-tree [14], R-tree [12], R*-tree [1], SS-tree [19], TV-tree [13], X-tree [5], and the Pyramid Technique [4]. Various algorithms for similarity searching have been developed associated to these indexing mechanisms. However, as dimensionality increases, query performance in these techniques degrades very significantly [5, 3]. This problem is referred as the *dimensionality curse* [11] and has attracted the attention of several researchers.

Requiring exact results is unrealistic in several applications. One obvious reason is the semantics expected from new applications. Most of the new applications need the notion of similarity to define reasonable queries. For instance, image databases may organize their data based on the content of the images. The user may want to search the database by describing some general properties, or by providing a sample image and search for similar images to the described properties or similar to the image. For example, the QBIC project at IBM provides the ability to run queries based on colors, shapes, and sketches [16, 10]. Similarly, Alexandria Project at UC Santa Barbara provides similarity queries for texture data [15]. Another reason results from the time consumed to run queries on databases having huge amount of data. Exact queries in many data-intensive applications require a long period of time to execute. Moreover, as mentioned in Asilomar Report [7], imprecise information will not only appear as the output of queries, it already appears in data sources as well. For several applications, it is reasonable to define the queries with approximations. Consider a user submitting a query such as "Are there any really good Italian restaurants close to where I live?" [7]. There is no exact answer to this query since it is impossible to give a perfect definition of goodness and even closeness.

In this report we present a family of fast approximation techniques which can be used for similarity checking of objects whose representations are large dimensional real vectors. We treat the case which the components of the vectors are drawn independently and identically from a probability distribution on the real line with density. This symmetry assumption is a special case of a more general formulation which can be analyzed in a least-squares sense in a similar manner.

2 Problem Definition

For a given pair of integers $n, i > 0$ let

$$\psi_i(z) = z_1^i + z_2^i + \cdots + z_n^i. \tag{1}$$

This is the i-th power symmetric function in the variables $z = (z_1, z_2, \ldots, z_n)$. Let I^n denote the n-dimensional unit cube defined by $0 \leq z_j \leq 1$ for $j = 1, 2, \ldots, n$. Suppose $x, y \in I^n$ are two n-dimensional real vectors with inner product $< x, y >= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$. We look for an approximation for $< x, y >^m$ of the form

$$< x, y >^m \;\approx\; b_1 \psi_1(x) \psi_1(y) + b_2 \psi_2(x) \psi_2(y) + \cdots + b_m \psi_m(x) \psi_m(y) \tag{2}$$

for large n, where the b_i are constants independent of x and y. The motivation for this problem is as follows. Suppose we have table of large number of n-dimensional vectors $x = (x_1, , x_2, \ldots, x_n)$ whose components are drawn from a distribution with (possibly unknown) density $f(t)$. In other words each x_i is drawn independently of other coordinates from the corresponding distribution $F(t)$. Given an arbitrary input vector $y = (y_1, y_2, \ldots, y_n)$, the main problem is to find the vectors x in the table minimizing (with high probability) the inner product $< x, y >$ without actually calculating all inner products. In the general case, the components do not need to satisfy $0 \leq z_j \leq 1$, nor do they have to be distributed identically.

A secondary problem is dynamic in nature. When the contents of the table changes, for instance by adding new vectors, how can the parameters used for the approximation problem to the inner product calculation be recomputed/adjusted efficiently?

We consider approximations of the form (2) by means of finding *the best set of constants* $b_1, b_2, \dots, b_m$ for the approximation. We calculate and keep m real numbers for each vector x, alongside the fixed b_j's that are used for each approximation. For the input vector y, we calculate the corresponding quantities $\psi_1(y), \psi_2(y), \dots, \psi_m(y)$ and use (2). If m can be taken much smaller than the dimension n with reasonable approximation to the inner product, then we have an overall gain on the computation time for similarity checking.

Note that in this formulation the quantities $\psi_i(z) = z_1^i + z_2^i + \cdots + z_n^i$ used in (2) are symmetric functions of the coordinates. A more general class of algorithms are obtained by taking instead $\psi_i(pz)$ in (2) where $pz = (p_1 z_1, p_2 z_2, \dots, p_n z_n)$ with $p_j \geq 0$ and $p_1 + p_2 + \cdots + p_n = 1$. This has the effect of giving a degree of importance (weight) to individual features (components) of the vectors x and y. For computational simplicity we first look at the symmetric case $\psi_i(z)$ as given in (1). By taking each $p_j = 1/n$, we can write $\psi_i(z) = n^i \psi_i(pz)$, so this calculation is a special case.

3 Determination of best parameters

The best approximation in the least square sense minimizes

$$\int \left[< x, y >^m - \sum_{j=1}^{m} b_j \psi_j(x) \psi_j(y) \right]^2 dx dy \tag{3}$$

where $dx = dx_1 dx_2 \cdots dx_n$, $dy = dy_1 dy_2 \cdots dy_n$, and the integral is over the $2n$-dimensional unit cube I^{2n}. The normal equations that $b_1, b_2, \dots, b_m$ must satisfy are found by differentiating (3) with respect to each b_i, and setting the resulting expressions to zero. Calculating,

$$\begin{aligned} \frac{\partial}{\partial b_i} \int \left[< x, y >^m - \sum_{j=1}^{m} b_j \psi_j(x) \psi_j(y) \right]^2 dx dy &= \int \frac{\partial}{\partial b_i} \left[< x, y >^m - \sum_{j=1}^{m} b_j \psi_j(x) \psi_j(y) \right]^2 dx dy \\ &= \int 2 \left[< x, y >^m - \sum_{j=1}^{m} b_j \psi_j(x) \psi_j(y) \right] \psi_i(x) \psi_i(y) \; dx dy \end{aligned} \tag{4}$$

Setting these expressions to 0 for $i = 1, 2, \dots, m$ we obtain the linear system $b_1, \dots, b_m$ must satisfy as

$$\sum_{j=1}^{m} \left[\int \psi_j(x) \psi_j(y) \psi_i(x) \psi_i(y) \; dx dy \right] b_j = \int < x, y >^m \psi_i(x) \psi_i(y) \; dx dy$$

where $1 \leq i \leq m$. This linear system satisfied by $b_1, \dots, b_m$ can be written in the form $Ab = c$ where $A = \|a_{i,j}\|$ is an $m \times m$ matrix of real numbers $a_{i,j}$

$$a_{i,j} = \int \psi_j(x) \psi_j(y) \psi_i(x) \psi_i(y) \; dx dy \tag{5}$$

for $1 \leq i,j \leq m$, $c = [c_1, \ldots, c_m]^t$ with

$$c_i = \int < x, y >^m \psi_i(x)\psi_i(y)\ dxdy \tag{6}$$

for $1 \leq i \leq m$, and $b = [b_1, \ldots, b_m]^t$ is the column vector of unknowns.

4 The $m = 2$ case

Before we present the general case, we consider the 2×2 system that arises for $m = 2$ and work out the derivation of the asymptotic expansion coefficients b_1, b_2 in (2). The linear system satisfied by b_1, b_2 is

$$\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

with

$$\begin{aligned} a_{1,1} &= \int_{I^{2n}} \psi_1(x)\psi_1(y)\psi_1(x)\psi_1(y)dxdy\ , \qquad a_{2,2} = \int_{I^{2n}} \psi_2(x)\psi_2(y)\psi_2(x)\psi_2(y)dxdy \\ a_{1,2} = a_{2,1} &= \int_{I^{2n}} \psi_1(x)\psi_1(y)\psi_2(x)\psi_2(y)dxdy \\ c_1 &= \int_{I^{2n}} < x, y >^2 \psi_1(x)\psi_1(y)dxdy\ , \qquad c_2 = \int_{I^{2n}} < x, y >^2 \psi_2(x)\psi_2(y)dxdy. \end{aligned}$$

These quantities can be computed exactly as functions of n. First of all

$$\begin{aligned} \int_{I^n} \psi_1(x)\psi_1(x)dx &= \sum_{k=1}^{n} \int_{I^n} x_k\psi_1(x)dx \\ &= n(\frac{n-1}{4} + \frac{1}{3}). \end{aligned}$$

Similarly,

$$\begin{aligned} \int_{I^n} \psi_1(x)\psi_2(x)dx &= n(\frac{n-1}{6} + \frac{1}{4}), \\ \int_{I^n} \psi_2(x)\psi_2(x)dx &= n(\frac{n-1}{9} + \frac{1}{5}). \end{aligned}$$

Therefore

$$\begin{aligned} a_{1,1} &= \int_{I^n} \psi_1(x)\psi_1(x)dx \int_{I^n} \psi_1(x)\psi_1(y)dy \\ &= \left(\int_{I^n} \psi_1(x)\psi_1(x)dx\right)^2 = n^2(\frac{3n+1}{12})^2. \end{aligned}$$

By a similar computation for $a_{2,2}$ and $a_{1,2}$, we find that the matrix of coefficients is

$$\begin{bmatrix} n^2(\frac{3n+1}{12})^2 & n^2(\frac{2n+1}{12})^2 \\ \\ n^2(\frac{2n+1}{12})^2 & n^2(\frac{5n+4}{45})^2 \end{bmatrix}$$

Next we calculate c_1 and c_2.

$$c_1 = \int_{I^{2n}} (\sum_{k=1}^{n} x_k y_k)^2 \psi_1(x)\psi_1(y)dxdy$$

There are two kinds of terms arising from the expansion of $(\sum x_k y_k)^2$. Diagonal terms of the form $x_r^2 y_r^2$, and off-diagonal terms of the form $x_r y_r x_s y_s$ for $r \neq s$. The contribution of the first kind of terms to c_1 is

$$\begin{aligned} n\int x_1^2 y_1^2 \psi_i(x)\psi_i(y)dxdy &= n\left(\int x_1^2\psi_i(x)dx\right)^2 \\ &= n\left(\frac{n-1}{6}+\frac{1}{4}\right)^2 = n\left(\frac{2n+1}{12}\right)^2. \end{aligned}$$

Off-diagonal terms contribute

$$\begin{aligned} n(n-1)\int x_1 y_1 x_2 y_2 \psi_i(x)\psi_i(y)dxdy &= n(n-1)\left(\int x_1 x_2 \psi_i(x)dx\right)^2 \\ &= n(n-1)\left(\frac{n-2}{8}+\frac{1}{6}+\frac{1}{6}\right)^2 \\ &= n(n-1)\left(\frac{3n+2}{24}\right)^2. \end{aligned}$$

Therefore

$$c_1 = n(\frac{2n+1}{12})^2 + n(n-1)(\frac{3n+2}{24})^2. \tag{7}$$

By a similar calculation, we find

$$c_2 = n(\frac{5n+4}{45})^2 + n(n-1)(\frac{n+1}{12})^2. \tag{8}$$

Therefore we have the following system for b_1, b_2:

$$\begin{aligned} n^2(\frac{3n+1}{12})^2 b_1 \; + \; n^2(\frac{2n+1}{12})^2 b_2 &= n(\frac{2n+1}{12})^2 + n(n-1)(\frac{3n+2}{24})^2 \\ n^2(\frac{2n+1}{12})^2 b_1 \; + \; n^2(\frac{5n+4}{45})^2 b_2 &= n(\frac{5n+4}{45})^2 + n(n-1)(\frac{n+1}{12})^2 \end{aligned} \tag{9}$$

Since we are looking for an asymptotic formula, it is tempting to let $n \to \infty$ in (9) and then solve the resulting numerical system for b_1, b_2 directly. Attempting to do this and simplifying the resulting equations gives the system

$$\begin{aligned} \frac{1}{4^2}b_1 + \frac{1}{6^2}b_2 &= \frac{1}{8^2} \\ \frac{1}{6^2}b_1 + \frac{1}{9^2}b_2 &= \frac{1}{12^2} \end{aligned}$$

which has determinant $6^2 12^2 - 8^2 9^2 = 0$ and therefore singular. To circumvent this problem, we need to include not only the highest order term in n, but the second highest as well. This results in the (asymptotic) system

$$\begin{array}{rcrcl}(\frac{n}{16}+\frac{1}{24})b_1 &+& (\frac{n}{36}+\frac{1}{36})b_2 &=& \frac{n}{64}+\frac{19}{576} \\ (\frac{n}{36}+\frac{1}{36})b_1 &+& (\frac{n}{81}+\frac{8}{405})b_2 &=& \frac{n}{144}+\frac{25}{1296}\end{array} \tag{10}$$

which is nonsingular for every n. Solving (10) symbolically for b_1 and b_2, we find

$$b_1 = \frac{9-n}{4(4n+1)}\ , \quad b_2 = \frac{5(9n-7)}{16(4n+1)}\ .$$

Therefore the limiting values are

$$b_1 = -\frac{1}{16}\ , \quad b_2 = \frac{45}{64}. \tag{11}$$

This means that for $m = 2$, we approximate $< x, y >$ by the expression

$$\sqrt{\left|-\frac{1}{16}\psi_1(x)\psi_1(y) + \frac{45}{64}\psi_2(x)\psi_2(y)\right|} \tag{12}$$

The graph of the average relative error made appears in figure 1. The dimension n ranged from 2^4 to 2^{11}. For each dimension n, 100 pairs of vectors $x, y \in I^n$ were independently generated by drawing each coordinate from the uniform distribution on the unit interval I. The error calculated for n is the average relative error of these 100 experiments where the relative error of a single experiment is

$$\left| \ < x,y > - \sqrt{\left|-\frac{1}{16}\psi_1(x)\psi_1(y) + \frac{45}{64}\psi_2(x)\psi_2(y)\right|}\right| / < x,y > \ .$$

These are then accumulated and divided by the number of experiments.

5 The general m

Now we turn to the formulation of the general case.

Lemma 1

$$a_{i,j} = \int_{I^{2n}} \psi_i(x)\psi_i(y)\psi_j(x)\psi_j(y)dxdy \ = \ \frac{n^2(ij+n(i+j+1))^2}{(i+1)^2(j+1)^2(i+j+1)^2} \tag{13}$$

Proof Since

$$\int_{I^{2n}} \psi_i(x)\psi_i(y)\psi_j(x)\psi_j(y)dxdy \ = \ \left[\int_{I^n} \psi_i(x)\psi_j(x)dx\right]\left[\int_{I^n} \psi_i(y)\psi_j(y)dy\right] \ = \ \left[\int_{I^n} \psi_i(x)\psi_j(x)dx\right]^2,$$

it suffices to prove

$$\int_{I^n} \psi_i(x)\psi_j(x)dx \ = \ \frac{n(ij+n(i+j+1))}{(i+1)(j+1)(i+j+1)} \tag{14}$$

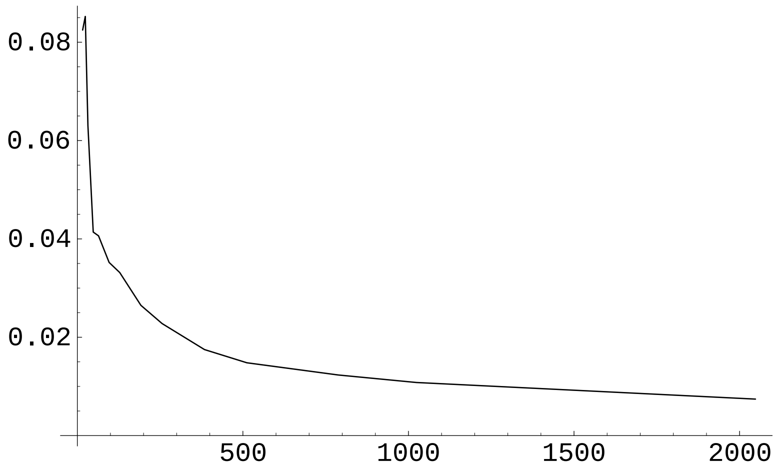

Figure 1: Average relative error versus dimension n, $16 \leq n \leq 2048$ for vectors from the uniform distribution with $m = 2$.

By direct calculation we find

$$\begin{aligned} \int_{I^n} \psi_i(x)\psi_j(x)dx &= \sum_{k=1}^{n} \int_{I^n} x_k^i \psi_j(x)dx \\ &= n\left(\frac{n-1}{(i+1)(j+1)} + \frac{1}{i+j+1}\right) \end{aligned}$$

which simplifies to (14). □

Lemma 2

$$\int_{I^{2n}} < x, y >^m \psi_i(x)\psi_i(y)dxdy = \alpha^2 n^{m+2} + \beta^2 n^{m+1} + O(n^m)$$

where

$$\alpha = \frac{1}{2^m(i+1)} \tag{15}$$

$$\beta = \frac{m}{2^{m-1}(i+2)} - \frac{m}{2^m(i+1)} + \frac{m(m-1)}{3\cdot 2^{m-1}(i+1)} - \frac{m(m-1)}{2^{m+1}(i+1)} \tag{16}$$

Proof Proof goes in here. □

Remarks

In the case of general m, we see from the expression (13) for (i, j)-th entry $a_{i,j}$ of the coefficient matrix that

$$a_{i,j} \sim \frac{n^4}{(i+1)^2(j+1)^2}.$$

But the matrix

$$\|\frac{1}{(i+1)^2(j+1)^2}\|$$

has rank 1, and therefore the system obtained by ignoring all but the highest degree of n that appears in the system we are required to solve is singular for $m > 1$.

6 Other distributions

Suppose now that the coordinates of the vectors x and y are drawn from not the uniform distribution on the unit interval I, but some other distribution F on the real line. We assume that F has density f. Thus

$$F(t) = \int_{-\infty}^{t} f(x)dx$$

with

$$\int_{-\infty}^{\infty} f(x)dx \;=\; 1.$$

and

$$Pr\{a < x < b\} = \int_a^b f(x)dx.$$

The i-th moment μ_i of f (about the origin) is defined by

$$\mu_i = \int_{-\infty}^{\infty} x^i f(x)dx$$

In minimizing the least squared error

$$\int \left[< x, y >^m - \sum_{j=1}^{m} b_j \psi_j(x) \psi_j(y) \right]^2 dF(x)dF(y) \tag{17}$$

where the integral is over $\mathbb{R}^{2n}$ and

$$dF(x) = dF(x_1)dF(x_2)\cdots dF(x_n) = f(x_1)f(x_2)\cdots f(x_n)dx_1dx_2\cdots dx_n.$$

$dF(y)$ is defined similarly. The coefficients $b_1, \ldots, b_m$ to be determined satisfy the linear system $Ab = c$ where

$$a_{i,j} \;=\; \int \psi_j(x)\psi_j(y)\psi_i(x)\psi_i(y)dF(x)dF(y) \tag{18}$$

$$c_i \;=\; \int < x, y >^m \psi_i(x)\psi_i(y)dF(x)dF(y). \tag{19}$$

Lemma 3 *Suppose $a_{i,j}$ is as given in (18). Then*

$$a_{i,j} = n^2(\mu_{i+j} + (n-1)\mu_i\mu_j)^2.$$

Proof As before,

$$a_{i,j} = \left(\int_{\mathbb{R}^n} \psi_i(x)\psi_j(x)dF(x) \right)^2$$

and

$$\int \psi_i(x)\psi_j(x)dF(x) \;=\; n \int x_1^i \psi_j(x)dF(x) \tag{20}$$

$$= n \int x_1^{i+j} f(x_1)dx_1 + n(n-1) \int x_1^i x_2^j f(x_1)f(x_2)dx_1dx_2 \tag{21}$$

$$= n \int t^{i+j} f(t)dt + n(n-1) \left(\int t^i f(t)dt \right) \left(\int t^j f(t)dt \right) \tag{22}$$

$$= n\mu_{i+j} + n(n-1)\mu_i\mu_j. \tag{23}$$

$\square$

Next, we calculate c_1 for the $m = 2$ case to illustrate the calculation for general m. Again, we separate the terms in $< x, y >^2$ into two types: diagonal (terms of the form $x_r^2 y_r^2$), and off-diagonal (terms of the form $x_r y_r x_s y_s$ with $r \neq s$).

$$\begin{aligned} n \int x_1^2 y_1^2 \psi_i(x)\psi_i(y) dF(x)dF(y) &= n \left(\int x_1^2 \psi_i(x) dF(x) \right)^2 \\ &= n \left((n-1) \int x_1^2 x_2 f(x_1) f(x_2) dx_1 dx_2 + \int x_1^3 f(x_1) dx_1 \right)^2 \\ &= n \left[(n-1)\mu_1\mu_2 + \mu_3 \right]^2 . \end{aligned}$$

Off-diagonal terms contribute

$$\begin{aligned} n(n-1) \int x_1 y_1 x_2 y_2 \psi_i(x)\psi_i(y) dF(x)dF(y) &= n(n-1) \left(\int x_1 x_2 \psi_i(x) dF(x) \right)^2 \\ &= n(n-1) \left[(n-2)\mu_1^3 + 2\mu_1\mu_2 \right]^2 \end{aligned}$$

Therefore

$$c_1 = n \left[(n-1)\mu_1\mu_2 + \mu_3 \right]^2 + n(n-1) \left[(n-2)\mu_1^3 + 2\mu_1\mu_2 \right]^2 . \tag{24}$$

Similarly we find

$$c_2 = n \left[(n-1)\mu_2^2 + \mu_4 \right]^2 + n(n-1) \left[(n-2)\mu_1^2\mu_2 + 2\mu_1\mu_3 \right]^2 . \tag{25}$$

This calculation is for c_1, c_2 is a special case of the general result given below in lemma 4. However to state it we need some definitions. A sequence of integers $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r)$ is a *partition* of m if

1. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$,
2. $\lambda_1 + \lambda_2 + \cdots + \lambda_r = m$.

We denote this by $\lambda \vdash m$. Each λ_i is called a *part* of λ. The number of parts of λ is denoted by $r = r_\lambda$. $\binom{n}{r}$ is the binomial coefficient $n!/r!(n-r)!$, and $\binom{m}{\lambda_1, \ldots, \lambda_r}$ is the multinomial coefficient $m!/\lambda_1! \cdots \lambda_r!$.

Lemma 4 *Suppose c_i is as defined in (19). Then*

$$c_i = \sum_{\lambda \vdash m} \binom{n}{r} \binom{m}{\lambda_1, \ldots, \lambda_r} [\mu_{\lambda_1+i}\mu_{\lambda_2} \cdots \mu_{\lambda_r} + \mu_{\lambda_1}\mu_{\lambda_2+i} \cdots \mu_{\lambda_r} + \cdots + \mu_{\lambda_1}\mu_{\lambda_2} \cdots \mu_{\lambda_r+i} + (n-r)\mu_{\lambda_1}\mu_{\lambda_2} \cdots \mu_{\lambda_r}\mu_i]^2$$

Proof The proof is straightforward but tedious, and is omitted. $\square$

Note that for $m = 2$, the summation is over two partitions: (2) with $r = 1$, and $(1,1)$ with $r = 2$. Consequently

$$c_i = n[\mu_{2+i} + (n-1)\mu_2\mu_i]^2 + n(n-1)[2\mu_1\mu_{1+i} + (n-2)\mu_1^2\mu_i]^2$$

which reduces to (24) for $i = 1$, and to (25) for $i = 2$.

Using the expressions in (24) and (25) for c_1, c_2 and the 2×2 matrix of coefficients from lemma 3 for $m = 2$, we have

$$\begin{bmatrix} n^2[\mu_2 + (n-1)\mu_1^2]^2 & n^2[\mu_3 + (n-1)\mu_1\mu_2]^2 \\ \\ n^2[\mu_3 + (n-1)\mu_1\mu_2]^2 & n^2[\mu_4 + (n-1)\mu_2^2]^2 \end{bmatrix} \begin{bmatrix} b_1 \\ \\ b_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ \\ c_2 \end{bmatrix}$$

Inverting this system and letting $n \to \infty$ we obtain the following result:

Theorem 1 *The constants b_1, b_2 are functions of the first four moments of the density $f(x)$. They are given by the formulas*

$$b_1 = \mu_1^2 \cdot \frac{2\mu_2^3 + \mu_1^2\mu_4 - 3\mu_1\mu_2\mu_3}{\mu_2^3 + \mu_1^2\mu_4 - 2\mu_1\mu_2\mu_3}\ , \qquad b_2 = \frac{\mu_1^4}{\mu_2} \cdot \frac{\mu_1\mu_3 - \mu_2^2}{\mu_2^3 + \mu_1^2\mu_4 - 2\mu_1\mu_2\mu_3}\ . \tag{26}$$

7 Approximation coefficients b_1, b_2 for various distributions

Suppose the coordinates of $x, y \in \mathbb{R}^n$ are drawn identically and independently from a probability distribution with density function $f(x)$. In view of theorem 1 and the explicit formulas for the approximation coefficients b_1, b_2 in the expansion

$$< x, y >^2 \approx b_1\psi_1(x)\psi_1(y) + b_2\psi_2(x)\psi_2(y)$$

can be found using (26) as soon as the first four moments of the density are known. For most common distributions, these moments are well known as explicit functions of the parameters of the distribution (see, for example [2]). Below, we give a number of examples.

7.1 Power distribution

For a shape parameter c, the distribution function on $0 \le x \le 1$ is given by $F(x) = x^c$, with density function $f(x) = cx^{c-1}$. The i-th moment of $f(x)$ (around the origin) is given by

$$\mu_i = \frac{c}{c+i}. \tag{27}$$

From theorem 1 we get

$$b_1 = -\frac{2c^3}{(c+1)^2(c^2+3c+4)}\ , \quad b_2 = \frac{c^2(c+2)^2(c+4)}{(c+1)^2(c^2+3c+4)}\ . \tag{28}$$

For $c = 1$, $f(x) = 1$ on $0 \le x \le 1$ and the distribution is uniform. In this case the formulas in (28) specialize to

$$b_1 = -\frac{1}{16}\ , \quad b_2 = \frac{45}{64}$$

which are the previously computed values for the uniform distribution given in (11).

7.2 Exponential distribution

For a scale parameter b, the distribution function on $0 \le x \le \infty$ is given by $F(x) = 1 - \exp(-x/b)$, with density function $f(x) = (1/b)\exp(-x/b)$. The i-th moment of $f(x)$ (around the origin) is

$$\mu_i = i!\, b^i. \tag{29}$$

From theorem 1 we get

$$b_1 = \frac{b^2}{2}\,, \qquad b_2 = \frac{1}{8}\,. \tag{30}$$

7.3 Binomial distribution

Let $0 \le x \le N$ be the number of successes in N independent Bernoulli trials where the probability of success at each trial is p and the probability of failure is $q = 1 - p$. The distribution function is $\sum_{i=0}^{x} \binom{N}{i} p^i q^{N-x}$ and the probability density function is $f(x) = \binom{N}{x} p^x q^{N-x}$. The first four moments of of $f(x)$ (around the origin) are given by

$$\begin{aligned}
\mu_1 &= Np \\
\mu_2 &= Np((N-1)p+1) \\
\mu_3 &= Np((N-1)(N-2)p^2+3(N-1)p+1) \\
\mu_4 &= Np((N-1)(N-2)(N-3)p^3+6(N-1)(N-2)p^2+7(N-1)p+1).
\end{aligned} \tag{31}$$

From theorem 1 we get

$$b_1 = \frac{N^2p^2(1-2p)}{np-3p+2}\,, \qquad b_2 = \frac{N^2p^2}{(np-p+1)(np-3p+2)}. \tag{32}$$

7.4 Normal distribution

Normal distribution with mean μ and standard deviation σ has the probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

for $-\infty \le x \le \infty$. The first four moments of of $f(x)$ (around the origin) are given by

$$\begin{aligned}
\mu_1 &= \mu \\
\mu_2 &= \mu^2+\sigma^2 \\
\mu_3 &= \mu^3+3\mu\sigma^2 \\
\mu_4 &= \mu^4+6\mu^2\sigma^2+3\sigma^4.
\end{aligned} \tag{33}$$

From theorem 1 we get

$$b_1 = \frac{2\mu^2\sigma^4}{\mu^4+\sigma^4}\,, \qquad b_2 = \frac{\mu^4(\mu^2-\sigma^2)}{(\mu^2+\sigma^2)(\mu^4+\sigma^4)}. \tag{34}$$

7.5 Poisson distribution

Poisson distribution with parameter $\lambda > 0$ (the mean) has density function $f(x) = \lambda^x \exp(-\lambda)/x!$ for integer x in the range $0 \leq x \leq \infty$. The first four moments of $f(x)$ (around the origin) are given by

$$\begin{aligned}
\mu_1 &= \lambda \\
\mu_2 &= \lambda(\lambda+1) \\
\mu_3 &= \lambda(\lambda^2+3\lambda+1) \\
\mu_4 &= \lambda(\lambda^3+6\lambda^2+7\lambda+1).
\end{aligned} \tag{35}$$

From theorem 1 we get

$$b_1 = \frac{\lambda^2}{\lambda+2}\,, \qquad b_2 = \frac{\lambda^2}{(\lambda+2)(\lambda+1)}. \tag{36}$$

7.6 Beta distribution

Beta distribution on $0 \leq x \leq 1$ has two shape parameters $v, w > 0$, and density function

$$f(x) = \frac{(v+w-1)!x^{v-1}(1-x)^{w-1}}{(v-1)!(w-1)!}$$

(for integer v, w). The i-th moment about the origin is given by

$$\mu_i = \prod_{r=0}^{i-1} (v+r)/(v+w+r).$$

Using the first four of these moments in theorem 1 we get

$$b_1 = \frac{2v^2(w-v-1)}{(v+w)^2((v+1)^2+(v+3)w)}\,, \qquad b_2 = \frac{v^2(w+v+1)^2(w+v+3)}{(v+w)^2((v+1)^3+(v+1)(v+3)w)}. \tag{37}$$

A summary of these calculations for $m = 2$ appears in Figure 2.

8 Non-parametric case: estimating the moments

If the coordinates of each vector x are drawn from a known parametric distribution family, then the parameters can be estimated by various methods and the moments computed. Here we describe a method to estimate (and also incrementally update) the moments μ_i when the components are drawn from a distribution with an unknown density $f(t)$. As before, we assume that each coordinate of x is drawn independently from the corresponding distribution. By a transformation of the real line, we may also assume that f is identically zero outside the interval $0 \leq t \leq 1$.

The problem we address is the following. Suppose we know the empirical moments $\bar{\mu_i} = \bar{\mu_i}(N)$ of density $f(t)$, $0 \leq t \leq 1$, based on samples $t_1, t_2, \ldots, t_N$. Given t_{N+1}, how do we update $\bar{\mu_i}(N)$ to obtain the estimate $\bar{\mu_i}(N+1)$?

The idea is based on the following approximation

Distribution	Density $f(x)$	Range	b_1	b_2
Uniform	1	$0 \leq x \leq 1$	$-\frac{1}{16}$	$\frac{45}{64}$
Power	cx^{c-1}	$0 \leq x \leq 1$	$-\frac{2c^3}{(c+1)^2(c^2+3c+4)}$	$\frac{c^2(c+2)^2(c+4)}{(c+1)^2(c^2+3c+4)}$
Exponential	$(1/b)\exp(-x/b)$	$0 \leq x \leq \infty$	$\frac{b^2}{2}$	$\frac{1}{8}$
Binomial	$\binom{N}{x}p^x q^{N-x}$	$0 \leq x \leq N$	$\frac{N^2p^2(1-2p)}{np-3p+2}$	$\frac{N^2p^2}{(np-p+1)(np-3p+2}$
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp(\frac{-(x-\mu)^2}{2\sigma^2})$	$-\infty \leq x \leq \infty$	$\frac{2\mu^2\sigma^4}{\mu^4+\sigma^4}$	$\frac{\mu^4(\mu^2-\sigma^2)}{(\mu^2+\sigma^2)(\mu^4+\sigma^4)}$
Poisson	$\lambda^x \exp(-\lambda)/x!$	$0 \leq x \leq \infty$	$\frac{\lambda^2}{\lambda+2}$	$\frac{\lambda^2}{(\lambda+2)(\lambda+1)}$
Beta	$\frac{(v+w-1)!x^{v-1}(1-x)^{w-1}}{(v-1)!(w-1)!}$	$0 \leq x \leq 1$	$\frac{2v^2(w-v-1)}{(v+w)^2((v+1)^2+(v+3)w)}$	$\frac{v^2(w+v+1)^2(w+v+3)}{(v+w)^2((v+1)^3+(v+1)(v+3)w)}$

Figure 2: $< x, y >^2 \approx b_1\psi_1(x)\psi_1(y) + b_2\psi_2(x)\psi_2(y)$: asymptotic expansion coefficients b_1, b_2 for various distributions.

Lemma 5 *An estimate of the moment μ_i under the assumptions above is given by*

$$\bar{\mu_i}(N) = \frac{1}{N}\sum_{j=1}^{N} x_j^i \tag{38}$$

Proof An estimate $F_N(t)$ for the distribution given the samples $t_1, t_2, \ldots, t_N$ is the histogram

$$F_N(t) = \frac{1}{N}|\{t_j \mid t_j < t\}|$$

where the bars denote cardinality. Integrating by parts, we find that

$$\bar{\mu_i}(N) = \int_0^1 t^i dF_N(t) = 1 - i\int_0^1 t^{i-1}F_N(t)dt \; = 1 - \frac{1}{N}\sum_{j=1}^{N} j\left(t_{j+1}^i - t_j^i\right)$$

where $t_{N+1} = 1$. This sum simplifies to (38) as stated. □

Using lemma 5, we can write $\bar{\mu_i}(N+1)$ in terms of $\bar{\mu_i}(N)$ and the $(N+1)$-st sample t_{N+1} as

$$\bar{\mu_i}(N+1) = \frac{1}{N+1}\left(N\bar{\mu_i}(N) + t_{N+1}^i\right). \tag{39}$$

This update rule takes on a particularly nice form when we think of a table of vectors and run this update rule for every vector incrementally, instead of individual components. Suppose currently there are r vectors present in the table, each n-dimensional with entries in the unit interval, drawn from a distribution with unknown density. Let $\bar{\mu_i}[r]$ denote the estimate of the i-th moment of this density obtained from the $N = nr$ samples which are the components of these r vectors. If x is the $(r+1)$-st vector, then the new estimate is obtained by the update rule

$$\bar{\mu_i}[r+1] = \frac{1}{r+1}\left(r\bar{\mu_i}[r] + \frac{1}{n}\psi_i(x)\right). \tag{40}$$

Note that for the $m = 2$ approximation we need to compute $\psi_1(x), \psi_2(x)$ anyway. To estimate the moments up to $i = 4$ (which are needed for the calculation of b_1, b_2 by theorem 1), by (40) we also compute $\psi_3(x), \psi_4(x)$, and $\psi_5(x)$.

9 Approximations for various values of m

For $m = 2$ with the uniform distribution, we obtain the approximation

$$< x, y >^2 \quad \approx \quad -\frac{1}{16}\psi_1(x)\psi_2(y) + \frac{45}{64}\psi_2(x)\psi_2(y), \tag{41}$$

in which the dimension n does not appear in the asymptotic expansion coefficients. For $m = 3$

$$< x, y >^3 \quad \approx \quad -\frac{5}{16}n\psi_1(x)\psi_1(x) + \frac{3}{2}n\psi_2(x)\psi_2(y) - \frac{7}{6}n\psi_3(x)\psi_3(y), \tag{42}$$

and for $m = 4$

$$< x, y >^4 \quad \approx \quad -\frac{59}{256}n^2\psi_1(x)\psi_1(x) + \frac{1575}{1024}n^2\psi_2(x)\psi_2(y) - \frac{175}{64}n^2\psi_3(x)\psi_3(y) + \frac{1575}{1024}n^2\psi_4(x)\psi_4(y). \tag{43}$$

Values of $b_1, \ldots, b_m$ for various values of m for the uniform distribution appear in 3.

10 Experiments

In the experiments, 100 pairs of vectors $x, y \in I^n$ were independently generated from the uniform distribution on I^n (each coordinate was drawn from the uniform distribution on the unit interval I). The dimension n ranged from 16 to 2048. The error calculated is the average relative error of these 100 experiments where the relative error of a single experiment is given by

$$\left| < x, y > - \left| \sum_{j=1}^{m} b_j \psi_j(x)\psi_j(y) \right|^{1/m} \right| / < x, y >$$

These relative errors are then accumulated and divided by the number of experiments. Figures 4 and 5 are the plots of the errors versus the dimension for the approximations corresponding to $m = 2, 4, 6, 8$.

m	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
2	$-\frac{1}{16}$	$\frac{45}{64}$						
3	$-\frac{5}{16}n$	$\frac{3}{2}n$	$-\frac{7}{6}n$					
4	$-\frac{59}{256}n^2$	$\frac{1575}{1024}n^2$	$-\frac{175}{64}n^2$	$\frac{1575}{1024}n^2$				
5	$-\frac{31}{256}n^3$	$\frac{9}{8}n^3$	$-\frac{27}{8}n^3$	$\frac{135}{32}n^3$	$-\frac{297}{160}n^3$			
6	$-\frac{221}{4096}n^4$	$\frac{11025}{16384}n^4$	$-\frac{6125}{2048}n^4$	$\frac{202125}{32768}n^4$	$-\frac{24255}{4096}n^4$	$\frac{35035}{16384}n^4$		
7	$-\frac{89}{4096}n^5$	$\frac{45}{128}n^5$	$-\frac{275}{128}n^5$	$\frac{825}{128}n^5$	$-\frac{1287}{128}n^5$	$\frac{1001}{128}n^5$	$-\frac{2145}{896}n^5$	
8	$-\frac{535}{65536}n^6$	$\frac{43659}{262144}n^6$	$-\frac{43659}{32768}n^6$	$\frac{2837835}{524288}n^6$	$-\frac{3972969}{327680}n^6$	$\frac{3972969}{262144}n^6$	$-\frac{81081}{8192}n^6$	$\frac{1378377}{524288}n^6$

Figure 3: $< x, y >^m \approx b_1\psi_1(x)\psi_1(y) + \cdots + b_m\psi_m(x)\psi_m(y)$: asymptotic expansion coefficients $b_1, b_2, \ldots, b_m$ for the uniform distribution.

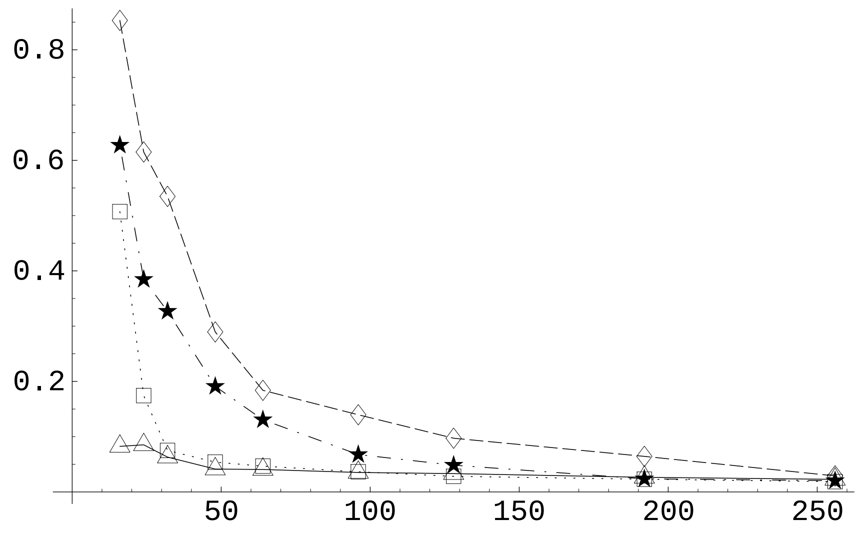

Figure 4: Average relative error versus dimension n: $16 \leq n \leq 256$. (Legend: $\triangle$ $m = 2$, $\square$ $m = 4$, $\star$ $m = 6$, $\diamond$ $m = 8$)

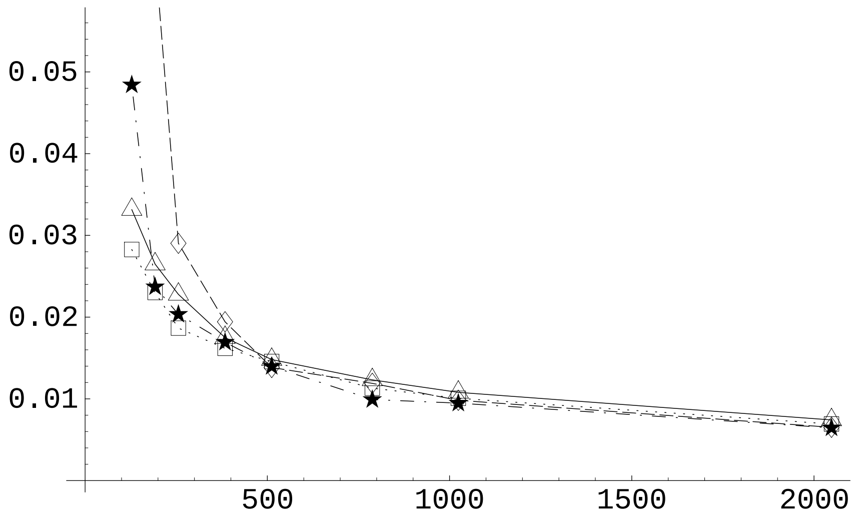

Figure 5: Average relative error versus dimension n: $128 \leq n \leq 2048$. (Legend: $\triangle$ $m = 2$, $\square$ $m = 4$, $\star$ $m = 6$, $\diamond$ $m = 8$)

References

[1] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The R* tree: An efficient and robust access method for points and rectangles. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 322–331, May 23-25 1990.

[2] N.A.J. Hastings and J.B. Peacock. *Statistical Distributions,* Halsted Press, New York, 1975.

[3] S. Berchtold, C. Bohm, D. Keim, and H. Kriegel. A cost model for nearest neighbor search in high-dimensional data space. In *Proc. ACM Symp. on Principles of Database Systems*, Tuscon, Arizona, 1997.

[4] S. Berchtold, C. Bohm, and H.-P. Kriegel. The Pyramid-Technique: Towards breaking the curse of dimensionality. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 142–153, Seattle, Washington, USA, June 1998.

[5] S. Berchtold, D. Keim, and H. Kriegel. The x-tree: An index structure for high-dimensional data. In *Proceedings of the Int. Conf. on Very Large Data Bases*, pages 28–39, Bombay, India, 1996.

[6] Kriegel H.-P. Berchtold S. S3: Similarity search in cad database systems. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 564–567, Tuscon, Arizona, 1997.

[7] P. Bernstein, M. Brodie, S. Ceri, D. DeWitt, M. Franklin, H. Garcia-Molina, J. Gray, J. Held, J. Hellerstein, H. Jagadish, M. Lesk, D. Maier, J. Naughton, H. Pirahesh, M. Stonebraker, and J. Ullman. The Asilomar report on database research, December 1998.

[8] X. Cheng, R. Dolin, M. Neary, S. Prabhakar, K. Ravikanth, D. Wu, D. Agrawal, A. El Abbadi, M. Freeston, A. Singh, T. Smith, and J. Su. Scalable access within the context of digital libraries. In *IEEE Proceedings of the International Conference on Advances in Digital Libraries, ADL*, pages 70–81, Washington, D.C., 1997.

[9] Korn F., Sidiropoulos N., Faloutsos C., Siegel E., and Protopapas Z. Fast nearest neighbor search in medical image databases. In *Proceedings of the Int. Conf. on Very Large Data Bases*, pages 215–226, Mumbai, India, 1996.

[10] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3:231–262, 1994.

[11] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 419–429, Minneapolis, May 1994.

[12] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 47–57, 1984.

[13] K. Lin, H. V. Jagadish, and C. Faloutsos. The TV-tree: An index structure for high-dimensional data. *VLDB Journal*, 3:517–542, 1995.

[14] D. B. Lomet and B. Salzberg. The hb-tree: A multi-attribute indexing method with good guaranteed performance. *ACM Transactions on Database Systems*, 15(4):625–658, December 1990.

[15] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–42, August 1996.

[16] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, and P. Yanker. The QBIC project: Querying images by content using color, texture and shape. In *Proc. of the SPIE Conf. 1908 on Storage and Retrieval for Image and Video Databases*, volume 1908, pages 173–187, February 1993.

[17] J. T. Robinson. The kdb-tree: A search structure for large multi-dimensional dynamic indexes. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 10–18, 1981.

[18] T. Seidl and Kriegel H.-P. Efficient user-adaptable similarity search in large multimedia databases. In *Proceedings of the Int. Conf. on Very Large Data Bases*, pages 506–515, Athens, Greece, 1997.

[19] D. White and R. Jain. Similarity indexing with the SS-tree. In *Proc. Int. Conf. Data Engineering*, pages 516–523, 1996.