# On the Effects of Dimensionality Reduction on High Dimensional Similarity Search

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
charu@us.ibm.com

## ABSTRACT

The dimensionality curse has profound effects on the effectiveness of high-dimensional similarity indexing from the performance perspective. One of the well known techniques for improving the indexing performance is the method of dimensionality reduction. In this technique, the data is transformed to a lower dimensional space by finding a new axis-system in which most of the data variance is preserved in a few dimensions. This reduction may also have a positive effect on the quality of similarity for certain data domains such as text. For other domains, it may lead to loss of information and degradation of search quality. Recent research indicates that the improvement for the text domain is caused by the re-enforcement of the semantic concepts in the data. In this paper, we provide an intuitive model of the effects of dimensionality reduction on arbitrary high dimensional problems. We provide an effective diagnosis of the causality behind the qualitative effects of dimensionality reduction on a given data set. The analysis suggests that these effects are very data dependent. Our analysis also indicates that currently accepted techniques of picking the reduction which results in the least loss of information are useful for maximizing precision and recall, but are not necessarily optimum from a qualitative perspective. We demonstrate that by making simple changes to the implementation details of dimensionality reduction techniques, we can considerably improve the quality of similarity search.

## Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications

## General Terms

Algorithms, Experimentation, Theory

## 1. INTRODUCTION

In recent years, content based retrieval of high dimensional problems has become an interesting and challenging problem. A number of applications such as multimedia, text, collaborative filtering and market basket applications require the use of high dimensional methods. Often, in content-based retrieval applications, an important query is to provide capabilities for finding data objects which are similar by content.

A number of techniques such as KDB-Trees, kd-Trees, and Grid-Files are discussed in the classical database literature [19] for indexing multidimensional data. Many of these techniques were initially proposed in the context of low-dimensional spatial applications. Starting with the seminal work of Guttman on R-Trees [10], considerable work has been done on finding multi-dimensional index structures [4, 10, 14, 20] for improving query performance. These methods generally work well for very low dimensional problems, though they degrade rapidly with increasing dimensionality, so that each query requires the access of almost all the data. A number of recent research results [5, 21] have illustrated the negative effects of increasing dimensionality on index structures both theoretically and empirically.

One solution for achieving better query scalability is by reducing the data dimensionality. The idea is to condense the data into a few dimensions [9, 12, 17] by applying data transformation methods such as Principal Component Analysis (PCA). This helps in several ways: first, it reduces the storage requirement; something which translates directly into improved performance scalability. Lower dimensionality also improves the retrieval efficiency of indexing structures [9, 17].

The focus of dimensionality reduction methods in previous research [17] has been to reduce the dimensionality only to a point where the query precision and recall continue to be high. Therefore, the discarded dimensions are the ones along which the variance in the data distribution is the least. An interesting side effect of careful implementations of dimensionality reduction techniques (such as Latent Semantic Indexing) has been observed for some data domains such as text. It has been observed that the dimensionality reduction process actually leads to a *qualitative improvement* in the relevance of the nearest neighbors found. Domains such as text have large amounts of redundancies and ambiguities (synonymy/polysemy) among the attributes which result in considerable noise effects for similarity queries. The dimensionality reduction technique is able to effectively improve the data representation by understanding the data in terms of concepts rather than words. Some analytical models suggest [8, 13, 16] that this is because picking the directions with greatest variance results in the use of *semantic* aspects of the data.

The analysis presented in [16] is based on a data model for text; the causality of the improvements (if any) is still largely an open problem for arbitrary high dimensional data mining problems. Our results indicate that this analysis exposes the desirable effects which *ought* to be achieved, but are not fully achieved by current techniques on all kinds of data sets. In this paper, we discuss a model for studying the effects of dimensionality reduction on similarity search; our model qualifies some widely held beliefs. For example, our model indicates that picking the dimensions which result in the least loss of information is not necessarily the best solution in every case; our empirical results validate the results from our model. We also show that the effects of dimensionality reduction are largely data dependent and that simple changes in the implementation details can often lead to considerable improvements.

An important observation is that distance functions for high dimensional applications are often heuristically designed, and there is no firm consensus on how similarity may be measured for an arbitrary data set. Often, there are considerable dependencies among attributes in high dimensional data; using distance functions such as the Euclidean metric on the original representation ignore these dependencies and result in poor measurements of similarity [3]. An interesting side effect of the results of this paper is that it shows how to perform dimensionality reduction effectively for high dimensional applications in such a way that it results in an automatic distance function correction: the resulting distance function on the reduced data set is much more effective because it measures distances in terms of the independent *concepts* (created by aggregate behavior and correlations) in the data set. We note that for a data set of fixed dimensionality, the *implicit dimensionality* increases when the dimensions are relatively uncorrelated to one another, because there are a larger number of independent concepts. We will provide a theoretical analysis of the effects of increasing implicit dimensionality on the dimensionality reduction problem.

This paper is organized as follows. In the remaining part of this section, we discuss the effects of high dimensionality on the behavior of similarity search applications. In section 2, we discuss the concept of coherence probability which is used to quantify the effects of the dimensionality reduction process. In section 3, we discuss the effects of increasing implicit data dimensionality on the value of the coherence probability. In section 4 we discuss the empirical results, whereas section 5 contains the conclusions and summary.

### 1.1 On High Dimensional Sparsity

Recent results [5] have shown that the maximum and minimum distances to a given query point in high dimensional space are almost the same for a wide variety of distance functions and data distributions. This makes a proximity query meaningless and unstable because there is poor discrimination between the nearest and furthest neighbor [1, 5, 11]. Thus, a small relative perturbation of the target in a direction away from the nearest neighbor could easily change the nearest neighbor into the furthest neighbor and vice-versa. A salient observation is that a dimensionality reduction process which tries to preserve the maximum amount of information from the original data does not help to alleviate this instability problem.

The results in [5] are valuable not just from the perspective of meaningfulness but also from the performance perspective of indexing. Most indexing methods work by using some kind of partitioning (hierarchical or flat) of the data set. This partitioning is then used in order to perform effective pruning of the data set. The idea is that if it is already known that some neighbor is close enough to the target, then one can prune away an entire partition by showing that the optimistic bound to that partition is no better than the nearest neighbor [18]. The results in [5] show that in high dimensionality the nearest and farthest neighbor have very similar relative distances to the target. Consequently, the optimistic bounds used by most index structures are usually not sharp enough for any kind of effective pruning in partition based methods. Therefore, the dimensionality reduction process provides a real chance to make high dimensional index structures practical.

Often real data shows considerable inter-attribute correlations [15]. Such correlations are of considerable utility in measuring similarities among records. One way of performing the dimensionality reduction process is to only reduce the number of dimensions sufficiently so that the precision and recall is not affected [17]; a more relevant goal would be to be aggressive in reducing the number of dimensions so that the noise effects are removed and the resulting distance measurements are more coherent and accurate. Such a dimensionality reduction process is also more valuable from the perspective of index structures, since greater aggression in dimensionality reduction translates to better performance.

## 2. MODELING THE EFFECTS OF DIMENSIONALITY REDUCTION

The database literature has focussed largely on the use of dimensionality reduction as a vehicle to reduce the size of the data set and improve indexing performance without losing query precision. A broader view of dimensionality reduction is one in which it is used in order to perform effective *conceptual* similarity search. Clearly, the objective of maximizing query precision with respect to the original data set may not necessarily provide the most effective conceptual representations. In this section, we will try to model and quantify the meaningfulness of each of the dimensions created as a result of our transformation techniques.

Dimensionality reduction techniques such as principal component analysis determine the $d$ (orthonormal) eigenvectors of the $d * d$ covariance matrix of a $d$-dimensional data set. The covariance matrix $C$ of a $d$-dimensional data set is a $d * d$ matrix in which the $(i, j)$th entry consists of the covariance between the dimensions $i$ and $j$. (Since the covariance matrix is positive semi-definite, the eigenvectors are orthonormal.) The eigenvectors and eigenvalues are obtained by diagonalizing it as $C = P \Delta P^T$, where $\Delta$ is diagonal matrix containing the eigenvalues, and $P$ is a $d * d$ matrix whose columns consists of the orthonormal eigenvectors of

$C$. Let $\{\overline{e_1} \ldots \overline{e_d}\}$ be the set of eigenvectors corresponding to the columns in $P$. The eigenvalue corresponding to each eigenvector is equal to the variance of the data, when the entire data set is projected onto the 1-dimensional space corresponding to that eigenvector. Thus, the sum of the eigenvalues is equal to the mean square deviation from the centroid using the euclidean distance metric. This is equal to the trace of both the matrices $\Delta$ and $C$ and is invariant on rotation of the axis system.

The first step for performing the dimensionality reduction process is to transform the data onto the new axis system $\{\overline{e_1} \ldots \overline{e_d}\}$. Thus, for a given data point $\overline{X} = (x_1 \ldots x_d)$, the coordinates in this new axis system are given by $(\overline{X} \cdot \overline{e_1} \ldots \overline{X} \cdot \overline{e_d})$. In order to actually perform the dimensionality reduction we retain the $k$ eigenvectors corresponding to the $k$-largest eigenvalues. These $k$ eigenvectors correspond to a $k$-dimensional axis system $(\overline{e_{i_1}} \ldots \overline{e_{i_k}})$. When the data is projected onto this subspace, the corresponding coordinates are $(\overline{X} \cdot \overline{e_{i_1}} \ldots \overline{X} \cdot \overline{e_{i_k}})$. On performing the projection, the amount of variance (or energy) lost from the data is equal to the sum of the smallest $(d-k)$-eigenvalues. This results in a data set which is a good approximation of the original data in the reduced space; it has high precision with respect to the original set of nearest neighbors. In some cases, the change in the nearest neighbor is actually an improvement; it has been observed for the text domain [7] that by retaining a very small fraction of the vectors with largest variance, it is possible to get considerable improvements in the quality of the nearest neighbors because of the re-enforcement of the semantic aspects of the data. Some analytical evidence of the causality has been provided in [16] using a semantic generation model for the textual domain. In this paper, we would like to test whether the process of dimensionality reduction has similar effects on arbitrary high dimensional problems without assuming anything about the nature of the underlying distribution. The resulting analysis provides some interesting facts about the effects of dimensionality reduction on similarity search for arbitrary high dimensional problems. For example, our results show that the widely used techniques of performing the dimensionality reduction in order to preserve the greatest amount of information from the original data set is not always the most effective approach.

In order to do so, we will construct a model which evaluates the significance of each eigenvector with respect to how well it picks up coherent correlations from the different dimensions. Let $\overline{X} = (x_1, \ldots x_d)$ be a given point. Let us consider the eigenvector $\overline{e_i}$ which is determined by the principal component analysis technique. Then the coordinate component of $\overline{X}$ along the transformed dimension $\overline{e_i}$ is given by $\overline{X} \cdot \overline{e_i}$. In order to obtain a closer look at how the different dimensions contribute to the similarity we decompose the value $\overline{X} \cdot \overline{e_i}$ into the contributions of the different attributes.

Specifically, let us decompose the data point $\overline{X} = (x_1 \ldots x_d)$ into $\overline{X_1} = (x_1, 0, \ldots, 0)$, $\overline{X_2} = (0, x_2, 0, \ldots 0)$, ..., $\overline{X_d} = (0, \ldots 0, x_d)$. Then, the value $\overline{X} \cdot e_i$ may be decomposed into the various components as follows:

$$\overline{X} \cdot \overline{e_i} = \overline{X_1} \cdot \overline{e_i} + \ldots \overline{X_d} \cdot \overline{e_i} \tag{1}$$

How coherent are these different contributions? If these contributions are highly correlated with one another then the coordinate value $\overline{X} \cdot \overline{e_i}$ is the result of several meaningful "agreements" among the correlated characteristics in the $d$-dimensional data set. On the other hand, if the vector $\overline{e_i}$ is not meaningful, and the various contributions are randomly related to one another, then any deviations from average behavior along the dimension $\overline{e_i}$ may be justified by the variance among the different components. If this is the case for most data points, then we may say that the vector $\overline{e_i}$ is incoherent and meaningless for measuring proximity.

In order to understand this point a little bit better, let us assume (without loss of generality) that the mean of the data set is centered at the origin $(0, \ldots, 0)$. Let $\overline{Y^1} \ldots \overline{Y^N}$ be a database containing $N$ points. Then, this means that for each $i \in \{1, \ldots, d\}$, we have $\sum_{j=1}^{N} \overline{Y^j} \cdot \overline{e_i} = 0$. The individual values of $\overline{Y^j} \cdot \overline{e_i}$ would be (typically) unequal to zero. How much of this variation may be attributed to true correlation effects, and how much to random noise? In order to do so, we need to look at the components which are contributed by the individual dimensions. Therefore, for each individual data point $\overline{X}$ and eigenvector $\overline{e_i}$, we formulate the following null hypothesis:

HYPOTHESIS 2.1. *The contributions $\overline{X_1} \cdot \overline{e_i}, \ldots, \overline{X_d} \cdot \overline{e_i}$ to the value $\overline{X} \cdot \overline{e_i}$ can be modeled as statistically independent instances drawn from a distribution centered at 0. Thus, the deviation of the absolute value of $\overline{X} \cdot e_i$ from 0 may be justified by the random variation of the sum of these different statistically independent instances.*

Note that if the above hypothesis is correct and the deviation of the absolute coordinate $\overline{X} \cdot \overline{e_i}$ from the mean can be justified by random variations only, then we can characterize eigenvector $e_i$ as noisy for the data point $\overline{X}$. On the other hand, if there are indeed enough correlations in the contributions of the different dimensions, then this variation cannot be justified by noise. Let us denote the component $\overline{X}_j \cdot \overline{e_i}$ contributed by the $j$th dimension in the original database representation of $\overline{X}$ to a given vector $\overline{e_i}$ by $c_j^{(i,\overline{X})}$. Thus, the null hypothesis assumes that the mean of the contributions $c_j^{(i,\overline{X})}$ (for varying values of $j$ and fixed $i$ and $\overline{X}$) is zero, and a deviation from 0 can be attribute to the noise effects of the variations. In order to compute this, let us assume that $\sigma(\overline{e_i}, \overline{X})$ be the mean and standard deviation of the distributions of $c_j^{(i,\overline{X})}$ (for varying $j$ and fixed $i$ and $\overline{X}$) about the mean value of $\mu = 0$. Then, the actual mean of the different values of $c_j^{(i,\overline{X})}$ is given by $\frac{\sum_{j=1}^{d} c_j^{(i,\overline{X})}}{d} = \frac{\overline{X} \cdot \overline{e_i}}{d}$. Note that under the null hypothesis assumption any deviation of this value from 0 must be because of noise effects; in order to test this assumption, we will estimate the value of $\sigma(\overline{e_i}, \overline{X}) = \sqrt{\frac{\sum_{j=1}^{d} (c_j^{(i,\overline{X})})^2}{d}}$, which is the root-mean-square deviation of the different values of $c_j^{(i,\overline{X})}$ about the null-hypothesis mean value of zero.

If the null hypothesis is true, then we can assume that the values of $c_j^{(i,\overline{X})}$ are statistically independent instances drawn from some hypothetical data distribution $\mathcal{Q}(i, \overline{X})$ with mean 0 and standard deviation $\sigma(\overline{e_i}, \overline{X})$. In such a situation, for high enough dimensionality $d$, the mean of the distribution created by averaging the different values of $c_j^{(i,\overline{X})}$ (varying $j$, fixed $i$ and $\overline{X}$) converges to a normal distribution $\mathcal{N}(i,\overline{X})$ with mean 0 and standard deviation $\sigma(\overline{e_i},\overline{X})/\sqrt{d}$ because of the central limit theorem. Therefore, we calculate the final *coherence-factor* for a given data point $\overline{X}$ and vector $\overline{e_i}$ as follows:

$$CoherenceFactor(\overline{X},\overline{e_i}) = \frac{|\overline{X}\cdot\overline{e_i}|/d}{\sigma(\overline{e_i},\overline{X})/\sqrt{d}} \qquad (2)$$

Intuitively, the coherence factor is equal to the number of standard deviations by which $\overline{X}\cdot\overline{e_i}/d$ is greater than the mean 0 of the distribution $\mathcal{N}(i,\overline{X})$. When the different dimensions behave in a coherent and correlated way along an eigenvector, then this value increases. The higher the value of the coherence factor, the more likely that the null hypothesis is indeed incorrect and the eigenvector $\overline{e_i}$ exposes interesting correlations among the different dimensions. The coherence factor can directly characterize the coherence probability by the use of normal distribution tables on the distribution $\mathcal{N}(i,\overline{X})$. Let the (cumulative) normal distribution function be denoted by $\Phi(\cdot)$. (Specifically, the value of $\Phi(z)$ denotes the percentage of the data which is less than $z$ standard deviations more than the mean of the distribution.) Then, the *coherence probability* for a given data point $\overline{X}$ and vector $\overline{e_i}$ is given by twice the percentage of a normal distribution which lies between 0 and $z = CoherenceFactor(\overline{X},\overline{e_i})$ standard deviations from the mean.

$$CoherenceProbability(\overline{X},\overline{e_i}) =$$
$$= 2\cdot(\Phi(CoherenceFactor(\overline{X},\overline{e_i})) - \Phi(0))$$
$$= 2\cdot\Phi(CoherenceFactor(\overline{X},\overline{e_i})) - 1$$

Thus, the coherence probability is a measure of how well the vector $\overline{e_i}$ picks up the directions of the correlations among the different dimensions in $\overline{X}$. Similarily, we may characterize the expected coherence probability $\mathcal{P}(\mathcal{D},\overline{e_i})$ of the vector $\overline{e_i}$ with respect to the *entire* data set $\mathcal{D} = \overline{Y^1}\ldots\overline{Y^N}$ as follows:

$$\mathcal{P}(\mathcal{D},\overline{e_i}) = \sum_{i=1}^{N} CoherenceProbability(\overline{Y^i},\overline{e_i})/N \qquad (3)$$

Thus, the coherence probability $\mathcal{P}(\mathcal{D},\overline{e_i})$ characterizes how meaningful the vector $\overline{e_i}$ may be for the data set $\mathcal{D}$. Thus leads to the following natural selection rule: *Pick the vectors with the largest coherence probability as the set of dimensions along which the data set should be represented.*

Note that the vectors with the largest coherence probability may not always be ones which have the largest eigenvalues. In the next section, we will provide a more detailed look at this possibility.

## 2.1 Relating Coherence Probability and Eigenvalues

Why is it that algorithms which pick the directions with the largest eigenvalues show qualitative improvements in the measurement of similarity? Note that when the components of the individual dimensions on a given eigenvector are highly correlated, then the absolute values of the eigenvalues are likely to be high. However, this is not necessarily true. Consider, for example, the case illustrated in Figure 1. This figure illustrates the contributions of the different dimensions to the distance between a data point $\overline{X}$ and the mean of the data set in the one-dimensional projection for two eigenvectors $A$ and $B$. We assume in this figure that the mean of the data set is 0. In this particular case, we assume that both curves have the bell shape of a normal distribution. Let $\delta_A(\overline{X}) > 0$ be the absolute coordinate of the data point $X$ along eigenvector $A$, and let $\delta_B(\overline{X}) > 0$ be the absolute coordinate of the data point $\overline{X}$ along eigenvector $B$. Note that the mean square values of $\delta_A(\overline{X})$ and $\delta_B(\overline{X})$ over the different data points (different values of $\overline{X}$) are the eigenvalues for the eigenvectors $A$ and $B$. In Figure 1, we have illustrated the distribution of the components contributed by the different dimensions to $\delta_A(\overline{X})$ and $\delta_B(\overline{X})$ respectively. The sum of the different components illustrated by curve $A$ is equal to $\delta_A(\overline{X})$ and the sum of the different components illustrated by curve $B$ is equal to $\delta_B(\overline{X})$. (Thus, the means (peaks) of the distributions illustrated in Figure 1 are $\delta_A(\overline{X})/d$ and $\delta_B(\overline{X})/d$ respectively.) The absolute deviation from the mean is lower for the case of eigenvector $B$ than eigenvector $A$ ($\delta_B(\overline{X}) < \delta_A(\overline{X})$), however the coherence probability for the data point $X$ is larger for eigenvector $B$. This is because the contributions of the different dimensions show smaller variance. If this is the case with a majority of the data points, then it may happen that even though the eigenvalue for $B$ is smaller than $A$, the coherence probability of $B$ is larger. In such cases, it would seem that it is better to pick eigenvector $B$ rather than eigenvector $A$. As we shall see in the empirical section, such situations do not arise too often in real data-usually eigenvectors with high magnitudes also have high coherence probabilities; however in those cases when they do arise there are clear advantages in picking the eigenvectors with the largest coherence probabilities.

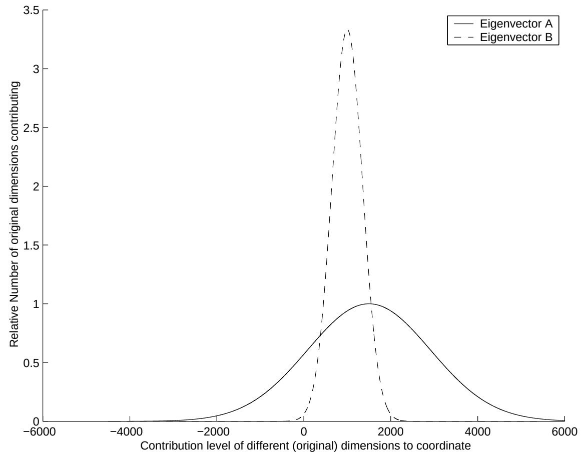

**Figure 1: Distributions of contribution of different dimensions to deviation of data point from the mean along an eigenvector (two cases)**

## 2.2 Effects of Data Representation

The dimensionality reduction process produces sets of uncorrelated[1] concepts in terms of which the data is repre-

[1] The concepts show no correlations of the second order.

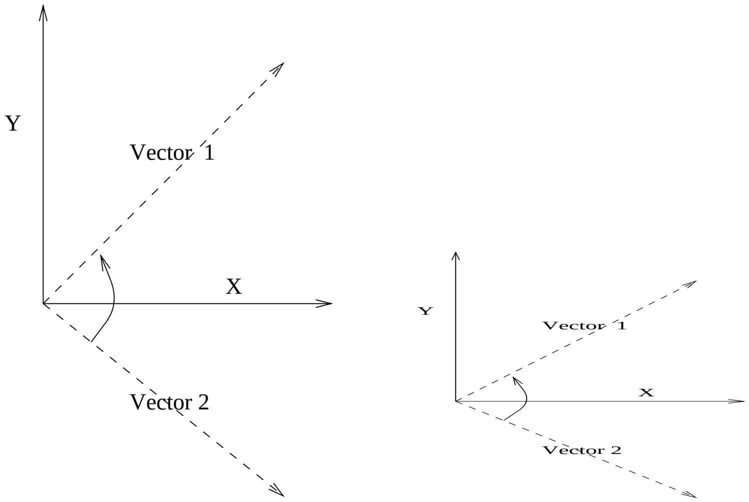

**Figure 2: Capturing the effects of data scaling**

sented. However, the exact nature of the transformation (and corresponding concepts created) is very sensitive to the initial data scaling. For example, let us consider the case when the different attributes form the demographic attributes such as age, salary, personal worth etc. In this case, the different attributes would draw from ranges which are very different. For example, the age attribute in years could range from 0 to 100, whereas the salary attribute in yearly compensation could range from 0 to 100,000 for given set of data. Different kinds of scalings would produce very different basis systems as a result of the principal component analysis. This is because after performing the scaling, an orthogonal basis system in the unscaled representation may no longer remain orthogonal. A simple example is illustrated in Figure 2, where after performing the scaling, the two vectors illustrated no longer remain orthogonal.

Scaling issues are often a problem for multi-dimensional data, when different dimensions may refer to completely different scales of reference. It makes some sense to choose a normalization of the data set in which the relative selectivity for each unit of distance along each dimension is similar. Thus, one sensible way of studentizing the data would be to compute the variance of the data set along each dimension $i$, and normalizing the data such that the variance along each dimension is one unit. (If the initial variance is zero along any dimension, then that dimension may be discarded.) This kind of scaling implicitly leads to the use of the correlation matrix as opposed to the covariance matrix of the original data set in arbitrary scales.

It is interesting to analyze what effects such a scaling would have on the actual coherence probability. Note that the coherence factor for a given eigenvector $\overline{e_i}$ increases when the standard deviation $\sigma(\overline{e_i}, \overline{X})$ reduces relative to the absolute magnitude of the projection of each data point on the eigenvector $\overline{e_i}$. It is intuitively clear that if the original set of dimensions are not scaled equally to begin with, the value of $\sigma(\overline{e_i}, \overline{X})$ for a given vector $\overline{e_i}$ and data point $\overline{X}$ are more likely to be very different for different eigenvectors. For example, if a particular dimension from the original data has very high variance compared to the other dimensions, then those eigenvectors which have larger projections (greater dot product) onto this dimension are likely to have higher values of $\sigma(.)$. This makes situations such as those illustrated in Figure 1 more likely. On the other hand, when the original dimensions are scaled equally, the eigenvalue magnitudes are likely to correlate better with the coherence probabilities.

Another interesting effect of the scaling process is on the actual magnitude of the coherence probability itself. The wildly varying scales along the different dimensions are likely to lead to widely varying values of $\overline{X}_j \cdot \overline{e_i}$. This would result in a proportionately higher second order moment $\sigma(\overline{e_i}, \overline{X})$ along a given eigenvector $\overline{e_i}$ as compared to the first-order moment $|\overline{X} \cdot \overline{e_i}|$. Thus, if the components have widely varying magnitudes, then the ratio of the absolute deviation $|\overline{X} \cdot \overline{e_i}|$ to the standard deviation is likely to be low. Since the coherence factor depends upon this ratio, this would result in a lower coherence-factor and hence a lower coherence probability. Therefore, it is intuitively clear that the process of performing the scaling is also likely to increase the absolute magnitude of the coherence probability.

## 3. EFFECTS OF IMPLICIT DIMENSIONALITY

An interesting aspect of the coherence probability is that it provides a good *absolute* measure of the level of correlation among the different dimensions. It has been discussed in earlier work that for many data mining problems, the implicit dimensionality is low; therefore the eigenvectors (which correspond to the second-order uncorrelated directions) get aligned in these directions. What happens to the coherence probability when the implicit dimensionality is high? In this section, we will discuss some of these effects by examining the case of uniformly distributed data, which is the worst case with increasing data dimensionality. We assume that the data is uniformly distributed in a unit cube centered at the origin. For the particular case of uniformly distributed data the implicit dimensionality is equal to the actual dimensionality and the original set of dimensions are one possible set of valid eigenvectors. The analysis for this case provides an insight into the behavior of the coherence probability with increasingly noisy data in which the dimensions have little correlations with one another.

Since there are no correlations among the attributes in uniformly distributed data, we can assume that the original set of vectors may be used as the transformed data. In this case, consider the point $\overline{X} = (x_1, \ldots x_d)$. Then, for the vector $\overline{e_1} = (1, 0, \ldots 0)$, the contribution of the different dimensions is given by $(x_1, 0, \ldots 0)$. Then, we have $|X \cdot \overline{e_1}|/d = |x_1/d|$. Also we have:

$$\begin{aligned}\sigma(\overline{e_1}, \overline{X}) &= \sqrt{((x_1 - 0)^2 + (0)^2 \cdot (d-1))/d} \\ &= |x_1/\sqrt{d}|\end{aligned}$$

Using these results we get the following value for the coherence probability:

$$CoherenceFactor(\overline{X}, \overline{e_1}) = 1 \qquad (4)$$

Note that the coherence factor is independent of the coordinates or dimensionality of $\overline{X}$. Similar results may be derived for the vectors $\overline{e_2} \ldots \overline{e_d}$. Thus, the coherence probability of the entire data set $\mathcal{D}(d)$ (in dimensionality $d$) and vector $\overline{e_i}$ is given by:

$$\mathcal{P}(\mathcal{D}(d), \overline{e_i}) = 2 \cdot \Phi(1) - 1 = 0.68 \qquad (5)$$

At this value of the coherence probability it cannot be said with certainty that $\overline{e_i}$ is a semantic concept in the data. At the same time the vector $\overline{e_i}$ cannot be discarded, else there is a risk of losing information. Furthermore, since the coherence probability is the same for each and every vector, all the dimensions have to be retained. This also provides an insight into the nature of noisy data sets with high implicit dimensionality; some data sets are inherently unsuited to dimensionality reduction- for such cases, the coherence probability is likely to be similar for all eigenvectors.

## 3.1 On the Nature of Noisy Data Sets

Uniformly distributed data is an example of a "perfectly noisy" data set. The intuition derived from the above analysis is also applicable to arbitrary high dimensional noisy data sets. In noisy data sets, there is little correlation among the different dimensions; therefore it becomes difficult to have high coherence probabilities. Thus, the absolute values of the coherence probabilities derived provide insight into the nature of high dimensional data sets. It has been discussed in recent work that many high dimensional data sets often have low implicit dimensionality; this would result in a few eigenvectors having high coherence probability corresponding to the non-noisy directions; the remaining vectors with low coherence probabilities would get pruned off as noise. Furthermore, since the coherence probability provides a certain level of insight into the feasibility of performing dimensional pruning, this technique can also be used to understand which data sets are most suitable for removing dimensions. The data sets which are most suitable for dimensionality reduction are those which have a few vectors with high coherence probabilities. Thus, each vector which corresponds to a high coherence probability phenomenon is a concept in the data. The remaining vectors which have low coherence probabilities can be removed as the noise in the data sets. Many of such vectors could correspond to vectors which have large eigenvalues. As we will see in later sections, the use of coherence probability allows us to be significantly more aggressive in performing the dimensionality reduction process.

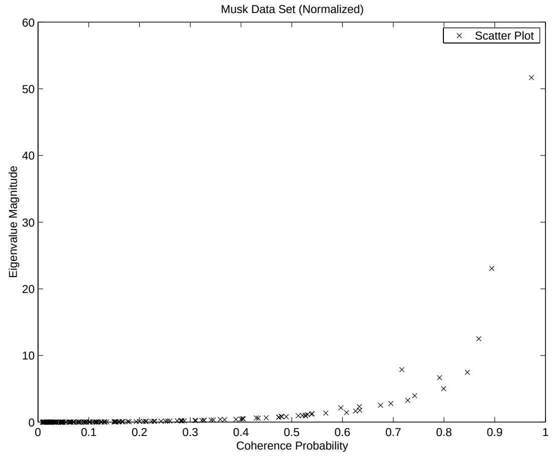

**Figure 3: Eigenvalue Magnitudes Versus Coherence Probability (Musk)**

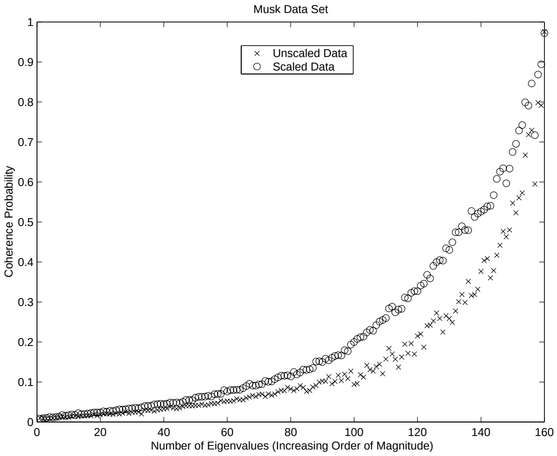

**Figure 4: Coherence Probability Distribution for Eigenvectors (Musk)**

Noisy data sets in which all the vectors have similar coherence probability are ones in which the number of independent concepts is so high that meaningful nearest neighbor search [5] may not be possible in full dimensionality using the characteristics of the entire data set. In some of these cases, alternative approaches are possible; significantly a generalized projected clustering technique [2, 6] which may be used in order to decompose the data into subsets with low implicit dimensionality and then apply the techniques discussed in this paper- a discussion of such an extension is beyond the scope of this paper.

# 4. EMPIRICAL RESULTS

In this section we will discuss the empirical results obtained from applying the dimensionality reduction algorithm to several high dimensional data sets. Since the focus of this paper is in measuring the quality and coherence of the nearest neighbor obtained after performing the dimensionality

**Table 1: Advantages of aggressive dimensionality reduction**

| **Data Set** | Full Dimensional | Optimal Quality | Optimal Quality | 1%-thresholding | 1%-thresholding |
|---|---|---|---|---|---|
| **Dimensionality** | Accuracy | (Accuracy) | (Dimensionality) | (Accuracy) | (Dimensionality) |
| **Musk (160)** | 437 | 498 | 13 | 433 | 31 |
| **Ionosphere (34)** | 891 | 934 | 10 | 892 | 32 |
| **Arrythmia (279)** | 732 | 762 | 10 | 742 | 124 |

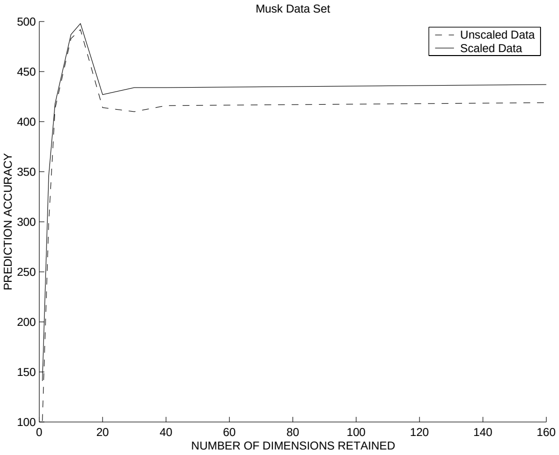

**Figure 5: Musk Data Set (Quality of Similarity Search)**

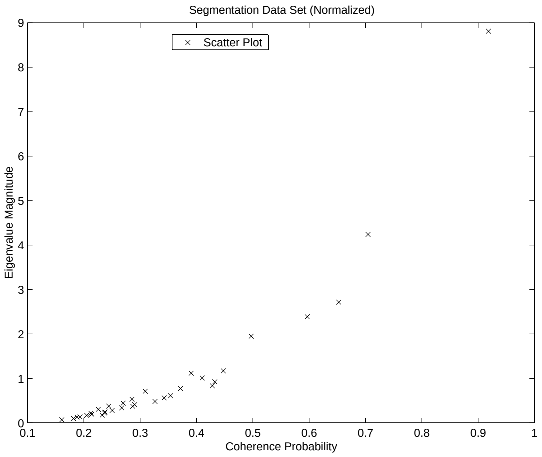

**Figure 6: Eigenvalue Magnitudes Versus Coherence Probability (Ionosphere)**

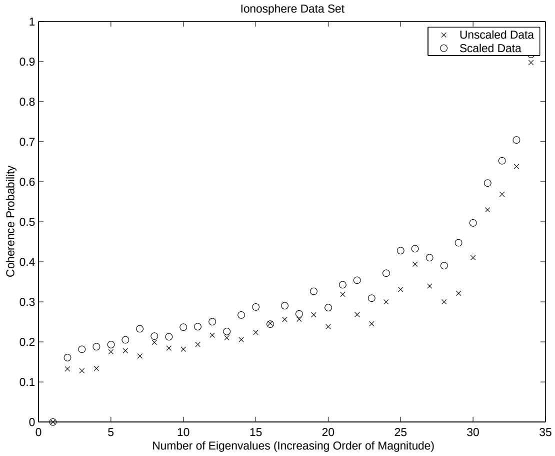

**Figure 7: Coherence Probability Distribution of Eigenvectors (Ionosphere)**

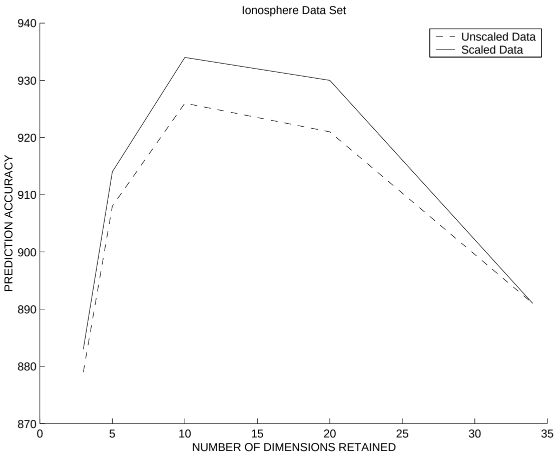

**Figure 8: Quality of Similarity Search (Ionosphere Data Set)**

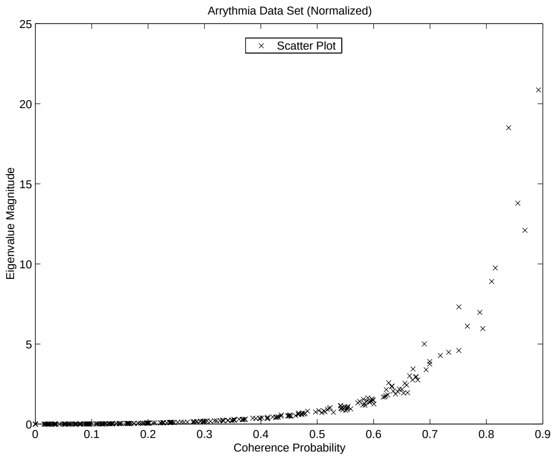

**Figure 9: Eigenvalue Magnitudes Versus Coherence Probability (Arrythmia Data Set)**

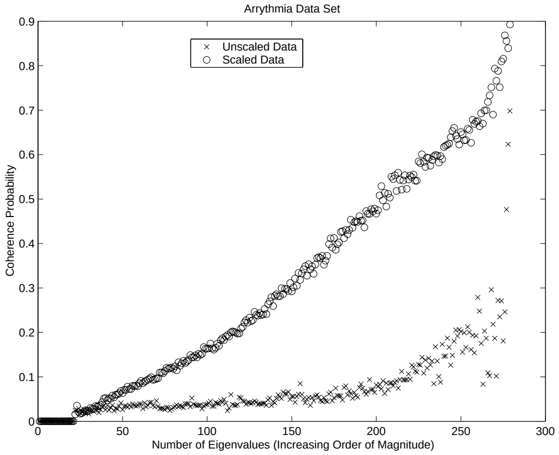

**Figure 10: Coherence Probability Distribution of Eigenvectors (Arrythmia Data Set)**

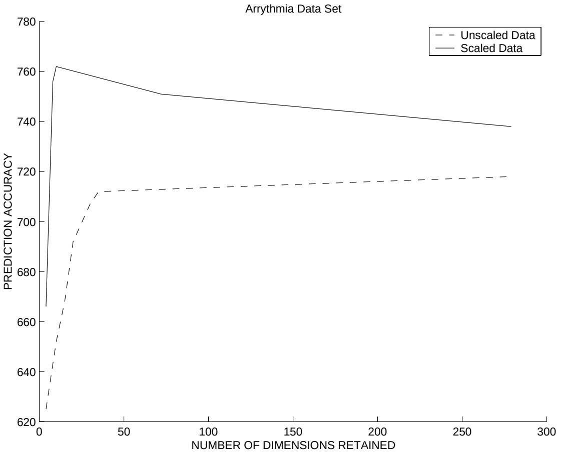

**Figure 11: Quality of Similarity Search (Arrythmia Data Set)**

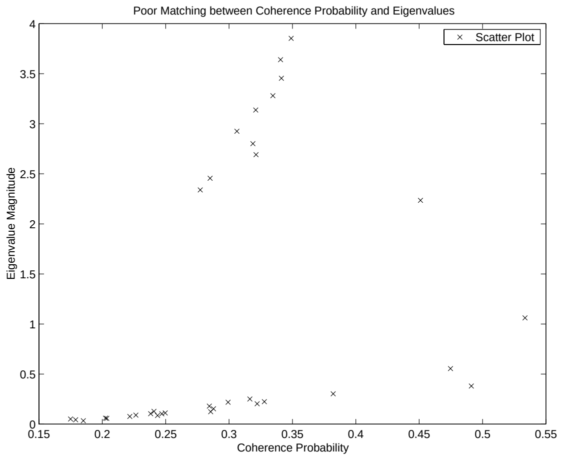

**Figure 12: Poor Matching between Coherence Probability and Eigenvalues (Noisy Data Set A)**

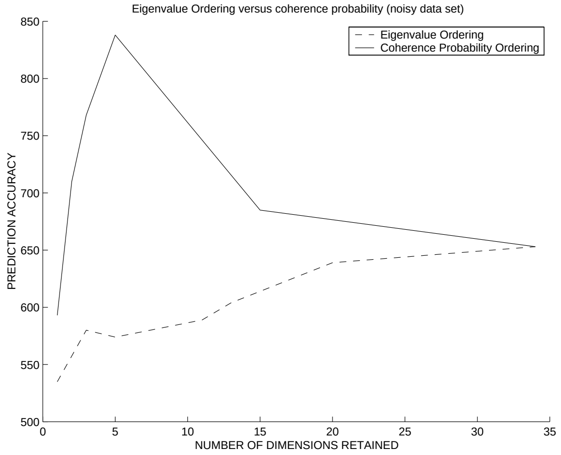

**Figure 13: Comparing Eigenvalue and Coherence Probability Ordering (Noisy Data Set A)**

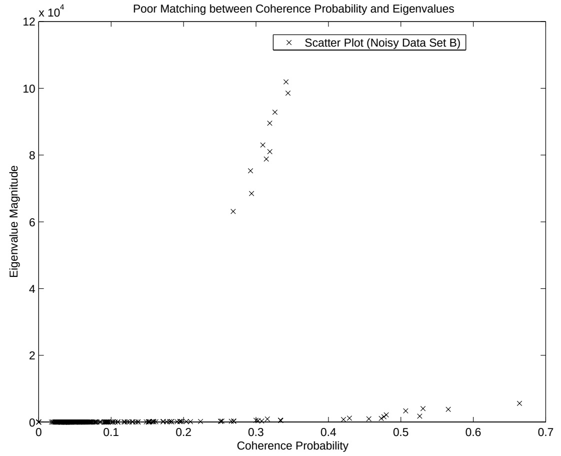

**Figure 14: Poor Matching between Coherence Probability and Eigenvalues (Noisy Data Set B)**

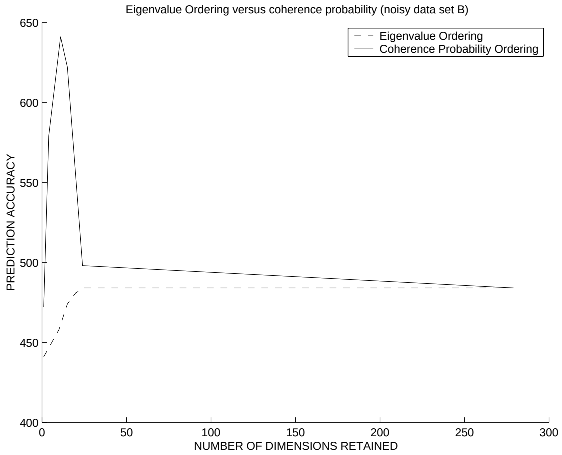

**Figure 15: Comparing Eigenvalue and Coherence Probability Ordering (Noisy Data Set B)**

reduction process, we need some hard criterion for making such qualitative measurements. Standard measures such as precision and recall are of no use for such a purpose. Ideally, such results are presented in the basis of human perception and understanding. Such judgements are easier to make in some domains such as text, where one can make semantic interpretations out of individual documents and their nearest neighbors. There is however no natural way to apply such a technique to arbitrary high dimensional data sets. Furthermore, using human relevance and judgement in estimating the quality of a nearest neighbor is fraught with a certain level of subjectivity that we would like to avoid. In order to create a technique which is somewhat analogous, we used a *feature stripping technique*. In this technique, we strip one of the features from the data set which has good semantic relationship with the rest of the attributes. Such a feature could typically be a categorical attribute or a class variable (for a classification problem.) Then, we find the nearest neighbor (for all methods) without using the information about this semantic variable. We check whether the semantic variable for the nearest neighbor matches with that of the target. When the nearest neighbor is affected by the noise in the data, the matching between the stripped attribute of the target and the nearest neighbor is likely to be poor. This is because the noise effects would mask out any relationship that the stripped attribute would have with the rest of the data set. On the other hand, when the nearest neighbor is very coherent and meaningful then the semantic variable for the nearest neighbor is also likely to match better with the semantic variable for the target. We concede that these results are somewhat evidential in nature since the exact relationship between the feature variables and the semantic attribute is unknown. However, a consistent effect of the dimensionality reduction process on the prediction accuracy does tend to be a powerful way of demonstrating the qualitative effects of dimensionality reduction on similarity search.

We used several data sets from the UCI machine learning repository [2] for our testing purposes. Each of the data sets contain a set of feature variables and a class variable. This class variable was used for our semantic interpretation. We tested three data sets from the machine learning repository in order to use various cases in terms of the overall data dimensionality. For each case, we have illustrated the scatter plot in order to show the relationship between the eigenvalue magnitude and the coherence probability. We have also illustrated the quality of similarity in terms of the class variable prediction accuracy of $\gamma = 3$ nearest neighbors drawn from the original data. The prediction accuracy is the total number of the nearest neighbors (over all queries) for which the semantic variables match between the target and nearest neighbor.

The results for the musk data set are illustrated in Figures 3, 4 and 5. This is a data set containing 160 dimensions. In Figure 3 we have illustrated the relationship between the eigenvalue magnitude and the coherence probability for the normalized data set. As we see, there is very high correlation between the eigenvalue magnitude and the coherence probability; we have also illustrated the effects of scaling on the coherence probability of the different eigenvalues in Figure 4. As expected, the process of performing the scaling significantly increases the coherence probability and also tends to increase the correlation between the absolute eigenvalue magnitude and the coherence probability. This trend is carried over to the qualitative effects illustrated in Figure 5 in which we see that the scaled data is consistently better in terms of qualitative similarity. One of the interesting aspects that we note in this case is that optimal qualitative performance is reached by picking only 13 eigenvectors out of a 160 dimensional data set. Note, that in this case, only about 11 eigenvectors are somewhat separated from the rest of the values in the scatter plot of Figure 3. On using these 11 eigenvectors only, the prediction accuracy was about 98% of the optimal prediction accuracy. The resulting accuracy is also significantly better than that provided by the full dimensional representation. Therefore, by examining the nature of the distribution of the eigenvalues and coherence probabilities, it is possible to provide a good intuitive judgement for the cut-off point.

The results for the ionosphere data set are illustrated in Figures 6, 7 and 8. In this particular case, the scatter graph of Figure 6 shows that the largest 5 eigenvalues are somewhat isolated from the rest of the data both in terms of magnitude and coherence probability. Picking these eigenvalues resulted in a data set which had higher prediction accuracy than by using the entire set of dimensions. When the next cluster of 5 eigenvalues was also included, this results in the optimal prediction accuracy. Another interesting aspect of this data set is the effect of scaling on the coherence probability. As evident from Figure 8, the effect of scaling is not present in full dimensionality; however in the reduced dimensional space, the qualitative performance of the scaled data is significantly better. This is because of the fact that the process of scaling results in an axis system with significantly higher coherence probability.

The data set with the largest dimensionality which we tested was the arrythmia data set. This data set had 279 dimensions corresponding to different attributes. The scatter plot of Figure 9 shows that the 10 eigenvectors tend to be sep-

[2] http://www.cs.uci.edu/~mlearn

arated from the rest of the data in terms of magnitudes and coherence probability. Indeed, in this case, the optimum prediction accuracy is obtained by picking the top 10 eigenvectors from the data as illustrated in Figure 11. Another interesting aspect of this data set is that the coherence probability of each vector in the transformed data representation increases significantly after performing the scaling. Correspondingly the quality of similarity search (prediction accuracy) also increases significantly. Thus the coherence probability can also be used in order to guide the process of picking a data representation which is most suitable for qualitatively effective similarity search.

Much of the recent work [17] in the database literature concentrates on reducing the dimensionality of the data set only to the extent that precision and recall are maximized. The rationale behind these methods is that any change in the nearest neighbor from the full dimensionality leads to loss of information; the rationale behind our approach is to be aggressive in removing the dimensions which have low coherence as noise; thus, on an overall basis the aggressiveness of a dimensionality reduction process which uses the coherence probability of the dimensions may lead to very low precision with respect to the original data but much higher effectiveness and coherence. In order to illustrate our point, we have indicated (in Table 1) the prediction accuracy using a 1%-thresholding technique in which only those eigenvalues which are less than 1% of the largest eigenvalue are discarded. This prediction accuracy is typically very close to the full dimensional accuracy and is significantly lower than the optimal accuracy for all 3 data sets (as illustrated in the accuracy charts of Figures 5, 8, 11). At the optimal dimensionality a very large fraction of the variance in the data set is discarded; for example, for the case of the arrythmia data set about 60% of the variance in the data was discarded. Furthermore, the precision and recall for such aggressive dimensionality reduction was often in the range of 10% or so (musk and arrythmia data sets). Thus, such a drastic reduction in dimensionality does not attempt to mirror the original nearest neighbors in the data; but rather improves their quality by removing the noise effects in high dimensionality. It is also clear from Table 1 that the optimal accuracy dimensionality is significantly lower than the 1%-thresholding method. In fact, the dimensionality for the 1%-thresholding method is quite close to the full dimensionality. In all cases, we see that the optimal accuracy dimensionality is low enough that it is possible to use traditional indexing methods in order to retrieve the data.

## 4.1 Interesting Cases with Noisy Data Sets

In all the data sets that we have illustrated in this section, the coherence probability is very closely correlated with the absolute eigenvalues; consequently there is little difference in the quality of similarity, when one or the other is used. However, it is especially interesting to test cases in which the coherence probability is poorly related to the eigenvalues. Such a situation can happen in very noisy data sets, where some of the directions with the largest variance correspond to the noise in the data. In order to illustrate this case, we generated a synthetically corrupted version of the ionosphere data set. (We shall henceforth refer to this data set as noisy data A.) In order to corrupt the data, we picked 10 of the original set of 34 dimensions in the data and replaced them with data generated from a uniform distribution with amplitude $a = 6$. In Figure 12, we have illustrated the scatter plot of the eigenvalues and the coherence probability. We note that in this case, the plot does not show as good correlation between the coherence probability and the eigenvalues as in any of the other cases. In fact, the largest few eigenvalues correspond to very low coherence probability and vice-versa. In this case, we also tested the quality of similarity based on two techniques: (a) retaining the set of dimensions with the highest eigenvalues (b) retaining the eigenvectors with the highest coherence probability. As we see from Figure 13, there were considerable qualitative advantages in picking the vectors with the highest coherence probability. In fact, the qualitative curve for the coherence probability ordering completely dominates the corresponding curve for the eigenvalue ordering. Furthermore, the curve based on the eigenvalue ordering does not peak at any point; rather in order to obtain the optimal qualitative performance all the dimensions need to be retained. On the other hand, in the case of the coherence probability ordering, the curve peaks for the use of only 5 out of the 34 dimensions. These eigenvalues correspond to those 5 cases on the extreme right of the scatter plot of Figure 12 which have high coherence probability but low eigenvalues. Another interesting observation about this data set was that at the optimal quality point, the total variance of the reduced data set was only 12.1% of the variance in the original data. This is the largest reduction among all data sets. The reason for this large reduction is that the original data set is very noisy and the dimensionality reduction process successfully extracts the relevant coherent information from the data set.

We generated a second corrupted data set (noisy data set B) from the arrythmia data set which had 279 dimensions. In order to generate the corrupted data set, we picked 10 of the original set of dimensions and replaced them with uniformly distributed data as in the previous case. In Figure 14, we have illustrated the scatter plot of the eigenvalue magnitudes versus the coherence probability. It is evident that there is poor matching between the coherence probabilities and the eigenvalues. The corresponding curve for the prediction accuracy on the quality of similarity search is indicated in Figure 15. As in the previous case, the eigenvalue ordered curve always loses information on performing the dimensionality reduction. Thus, this is another example of a data set in which straightforward techniques of applying dimensionality reduction are always detrimental for a similarity search algorithm because of the fact that the eigenvectors with the greatest magnitude are quite noisy and have little information in them. On the other hand, the curve created by using the coherence probability ordering provides much better quality of similarity search. The curve peaks just before including the outlier cluster of eigenvectors (the 11 eigenvectors in the scatter plot of Figure 14 which have very high eigenvalues) in the representation. This is the optimum cut-off point before which the noise in the data should be excluded. At this point only 11 of the original set of dimensions need to be included in the representation. The ability of the coherence probability ordering technique to always identify the (possibly) small number of directions which have the greatest amount of semantic information in them is a key advantage from the perspective of a dimensionality reduction method.

## 5. CONCLUSIONS AND SUMMARY

In this paper, we provided an analysis and understanding of the effects of dimensionality reduction on high dimensional similarity search. We showed that the effect of dimensionality reduction on similarity search may vary with the nature of the data. In order to provide a formal understanding of these effects on a given data set, we introduce a model which measures the noise effects in the behavior of the transformed dimensions. Our results indicates that those data sets which have a few eigenvectors with large coherence probability are most likely to be amenable for effective dimensionality reduction in terms of noise reduction. Currently used methods for dimensionality reduction are effective for those data sets in which there is greatest correspondence between coherence probability and eigenvalue magnitudes. For those cases in which this correspondence is not present, the quality of similarity can be significantly improved by retaining the dimensions with the greatest coherence probability. For data sets in which the implicit dimensionality is high, the coherence probability model indicates that all dimensions may need to be retained; in such cases, the curse of dimensionality applies and effective dimensionality reduction is not possible.

## 6. REFERENCES

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

[2] C. C. Aggarwal, P. S. Yu. Finding Generalized Projected Clusters in High Dimensional Spaces. *ACM SIGMOD Conference Proceedings*, 2000.

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

[4] S. Berchtold, D. A. Keim, H.-P. Kriegel. The X-Tree: An Index Structure for High Dimensional Data. *VLDB Conference Proceedings*, pages 28–39, September 1996.

[5] K. Beyer et al. When is Nearest Neighbors Meaningful? *ICDT Conference*, 1999.

[6] K. Chakrabarti, S. Mehrotra. Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces. *VLDB Conference Proceedings*, 2000.

[7] S. Deerwester et al. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6): pages 391-407, 1990.

[8] C. Ding. A Similarity Based Probability Model for Latent Semantic Indexing. *ACM SIGIR Conference Proceedings*, 1999.

[9] C. Faloutsos, K.-I. Lin. FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. *ACM SIGMOD Conference Proceedings*, 1995.

[10] A. Guttman. R-Trees: A Dynamic Index Structure for Spatial Searching. *ACM SIGMOD Conference Proceedings* 1984.

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

[12] I. T. Jolliffe. *Principal Component Analysis*, Springer-Verlag, New York, 1986.

[13] J. Kleinberg, A. Tomkins. Applications of Linear Algebra in Information Retrieval and Hypertext Analysis. *ACM PODS Conference Proceedings*, 1999.

[14] K.-I. Lin, H. V. Jagadish, C. Faloutsos. The TV-tree: An Index Structure for High Dimensional Data. *VLDB Journal*, pages 517–542, 1992.

[15] B.-U. Pagel, F. Korn, C. Faloutsos. Deflating the Dimensionality Curse Using Multiple Fractal Dimensions. *ICDE Conference Proceedings*, 2000.

[16] C.-H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala. Latent Semantic Indexing: A Probabilistic Analysis. *ACM PODS Conference Proceedings*, pages 159-168, 1998.

[17] K. V. Ravi Kanth, D. Agrawal, A. Singh. Dimensionality Reduction for Similarity Search in Dynamic Databases. *ACM SIGMOD Conference Proceedings*, 1998.

[18] N. Roussopoulos, S. Kelley, F. Vincent. Nearest Neighbor Queries. *ACM SIGMOD Conference Proceedings*, 1995.

[19] H. Samet. Design and Analysis of Spatial Data Structures. *Addison Wesley*, 1989.

[20] T. Sellis, N. Roussopoulos, C. Faloutsos. The R+ Tree: A Dynamic Index for Multidimensional Objects. *VLDB Conference Proceedings*, pages 507–518, 1987.

[21] R. Weber, H.-J. Scheck, S. Blott. A Quantitative Analysis and Performance Study for Similarity Search Methods in High Dimensional Spaces. *VLDB Conference Proceedings*, 1998.