Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

An Efficient Distance Calculation Method for Uncertain Objects

Lurong Xiao
Department of Computing
Hong Kong Polytechnic University
Hung Hom, Hong Kong
Email: cslxiao@comp.polyu.edu.hk

Edward Hung
Department of Computing
Hong Kong Polytechnic University
Hung Hom, Hong Kong
Email: csehung@comp.polyu.edu.hk

***Abstract*— Recently the academic communities have paid more attention to the queries and mining on uncertain data. In the tasks such as clustering or nearest-neighbor queries, expected distance is often used as a distance measurement among uncertain data objects. Traditional database systems store uncertain objects using their expected (average) location in the data space. Distances can be calculated easily from the expected locations, but it poorly approximates the real expected distance values. Recent research work calculates the expected distance by calculating the weighted average of the pair-wise distances among samples of two uncertain objects. However the pair-wise distance calculations take much longer time than the the former method. In this paper, we propose an efficient method *Approximation by Single Gaussian (ASG)* to calculate the expected distance by a function of the means and variances of samples of uncertain objects. Theoretical and experimental studies show that ASG has both advantages of the latter method's high accuracy and the former method's fast execution time. We suggest that ASG plays an important role in reducing computational costs significantly in query processing and various data mining tasks such as clustering and outlier detection.**

I. Introduction

Recently, the proliferation of areas such as sensor networks and image processing gains the attention of researchers to work on how to support various kinds of interesting queries and data mining on these uncertain data. While there has been a large amount of research work done on mining and queries on relational databases, these works were done on databases that store data in exact values. However, in many real-life applications, the raw data are usually uncertain when they are collected or produced. Sources of uncertain data include readings from sensors, information extracted using probabilistic parsing of input sources, classification results of image processing using statistical classifiers, results from predictive programs used for stock market, and weather predictions in meteorology, etc. These uncertain data may be in the form of an exact value with margins of error, sometimes with or without a probability distribution (or density) function. The result may also be represented as an interval or a set of values, one of which may be the real value. However, since traditional databases only store exact values, uncertain data are usually transformed into exact data by, for example, taking the weighted average or mean value (for quantitative attributes) or by taking the value with the highest frequency or possibility. This makes the storage, query and mining much simpler by using existing commercial database systems and mining techniques, but the shortcomings are obvious:

- By approximating the uncertain source data values, the intermediate and final results from the mining tasks and queries will also be approximate and may be wrong. For example, the locations of centroids of clusters become deviated from the real ones, or some data may be assigned to the wrong clusters.

Distance between two data objects is a very important measurement used in various queries and data mining tasks such as nearest-neighbor queries and clustering (e.g., K-means clustering [1]). While it is very simple to calculate the distance between two exact data objects by applying a distance formula, it is not trivial when the two data objects' locations are uncertain. An uncertain object has more than one possible location. If each object o_i has n_i possible locations, then we have $n_1 n_2$ possible distances between objects o_1 and o_2 for every possible pair-wise combinations of their locations. Given a probability distribution P_i of the possible locations of object o_i, we can calculate a probability distribution of the possible distances. This result is very informative, but it is very expensive to compute and unnecessary in most queries and mining tasks. Instead, as in [2], an *expected distance* is used by calculating the average of all the possible distances weighted by their probabilities. The expected distance can be directly used, for example, in nearest-neighbor queries for finding the nearest neighbors, in clustering for finding the cluster centroid closest to an object, and in outlier detection for finding outliers which do not have enough neighbors within a specified threshold distance [3].

Recent research work related to data mining on uncertain data such as [2] obtains the expected distance by assuming that the information of the precise probabilities of all possible locations are known in advance. The information is either represented as (i) a probability distribution function where probabilities are given on the finite set of possible locations, or (ii) a probability density function where the probability density is defined on a region. In the first case, they calculate the weighted average of the distances between all pair-wise combinations of locations of two objects. In the second case,

either (1) a grid consisting of a finite number of cells is formed on the region and the probability of each cell is estimated by sampling, or (2) sampling is done on the region so that more samples appear in areas of higher density. Then, again, the weighted average of the distances is calculated. As discussed above, this method to calculate the expected distance is expensive between it involves a large number of distance calculations which increases quadratically to the number of possible locations or grid cells or samples.

Therefore, we have investigated into the problem of efficient computation of expected distances in various aspects. First, we derived analytic solutions for some cases such as the expected distance between point/line/circle/sphere objects in uniform/Gaussian probability density functions. We then proposed a general method *Approximation by Single Gaussian (ASG)* for arbitrary probabilistic objects, which significantly reduces the computational cost. Experimental results show that ASG can obtain accuracy very close to the calculation methods used by recent research work while the execution time can be significantly reduced.

Note that ASG can be applied to any general arbitrary uncertain objects, even a certain object, which is a special case of an uncertain object with its uncertainty domain (range of possible locations) equal to its exact location. ASG can represent a certain object by a single Gaussian distribution with its mean equal to the exact location and its variance equal to zero. As a result, ASG can calculate the distance between two uncertain objects, between an uncertain object and a certain object, as well as between two certain objects.

The rest of this paper is organized as follows. In Section II we discuss some related work on data mining metrics and uncertain data applications. In Section III we formally define the distance between uncertain objects, derive the results for some special cases and propose ASG for all general cases. In that section we also describe four other distance calculation methods for comparison. We state and prove a theorem showing that the results of two of the four methods above are equivalent to the result of ASG (given the same sample set). This shows that high accuracy can be maintained with a much lower cost. In Section IV, we demonstrate the effectiveness and efficiency of ASG from the results of four experiments where uniform, Gaussian mixture, and totally arbitrary distributions are used. In Section V we conclude the paper.

II. Related Works

Researches on probabilistic databases began in 1980s. An earlier attempt was done to incorporate probabilities on disjoint events (tuples) [4] or attributes [5] into the relational data model. [5]'s algebra and independence assumption among attributes were extended respectively by [6] with new operations and by [7] with different probabilistic strategies and interval probabilities. Aggregate operations were then considered in [8]. The research on uncertain data management was further extended to other kinds of databases such as temporal databases [9] and object-oriented databases [10]. The semi-structured (XML) databases were also extended with independence assumption [11], arbitrary probabilistic distributions among children with a formal theory and algebra [12] as well as interval probabilities [13]. While there has been a great deal of work on supporting uncertainty in databases, there is little work on updating or proposing new measurement definitions for uncertain objects. Traditional data mining processes often use distance as a metric to measure how different two objects are. Different distance measures, like city-block distance or Minkowski-distance, have been used in measuring the similarity between interval data [14], but the pdfs of the intervals are not taken into account in most of the metrics. [15] proposes to use probabilistic distance functions to measure the similarity between uncertain objects. Each uncertain data item is modeled as a set of sampling points over the uncertainty region, and the Monte-Carlo method is used to retrieve the data. While the paper says that the probabilistic similarity join can be used to develop clustering methods for uncertain data, it is not clear how this can be done. [2] probably first exploits the *expected distance* to improve the efficiency of their clustering algorithm. However, their expected distance calculations are very expensive since they have to compute pair-wise distances between all pairs of possible locations or grid cells or samples. In this paper, we will propose a much more efficient method to calculate the expected distance.

III. Distance Measure Between Uncertain Objects

In this section we provide a formal definition of expected distance used in this paper and other recent research papers [2]. We will then present theoretical results of analytic solutions for special cases like uniform and Gaussian distributions for spheres in multidimensional spaces. A method to calculate the expected distance in general cases is then proposed with some other methods. These methods will be compared in the experimental section.

A. Problem Definition

If we view an attribute as a dimension, then the union of the domains of all attributes produces a multidimensional space where a certain object is represented as a point. Due to the uncertain nature or actual system limitation in the data collection phase, the imperfect data quality leads to uncertain attribute values of an object. Therefore an uncertain object can be represented as a set of points, each of which is a possible location of the object. A PDF (probability distribution function) is used to represent the distribution of the probabilities of the possible locations. Alternatively, an uncertain object can also be represented as a (finite or infinite) region, which covers the possible locations of the object (especially the number of possible locations is not finite). We call this region the uncertainty domain of object o_i, denoted as $UD(o_i)$. A pdf (probability density function), p_i, is used to indicate the probability density of each possible location within the region, i.e., $\int_{UD(o_i)} p_i(\mathbf{x})\, d\mathbf{x} = 1$.

Consider two objects o_i, o_j, whose pdfs are $p_i(\mathbf{x}_i)$,$p_j(\mathbf{x}_j)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are locations of o_i and o_j. $UD(o_i)$ and $UD(o_j)$ are the uncertainty domains of o_i and o_j. $D(\mathbf{x}_i, \mathbf{x}_j)$

is the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. The following gives the expected distance and the pdf of the distance between o_i and o_j.

$$E(D(o_i, o_j)) = \int_{UD(o_i)} \int_{UD(o_j)} D(\mathbf{x}_i, \mathbf{x}_j) p_i(\mathbf{x}_i) p_j(\mathbf{x}_j)\, d\mathbf{x}_i\, d\mathbf{x}_j \quad (1)$$

We can define a pdf $D_{i,j}$ which returns the probability of a distance value as follows:

$$D_{i,j}(s) = \int_{UD(o_i)} \int_{UD(o_j)} F(D(\mathbf{x}_i, \mathbf{x}_j), s) p_i(\mathbf{x}_i) p_j(\mathbf{x}_j)\, d\mathbf{x}_i\, d\mathbf{x}_j \quad (2)$$

where s is a non-negative real number; $F(x, y) = 1$ if $x = y$; $F(x, y) = 0$ otherwise. In the other words, $D_{i,j}(s)$ returns the probability that the distance between objects o_i, o_j is actually s.

We can represent the expected distance between o_i and o_j in terms of $D_{i,j}(s)$:

$$E(D(o_i, o_j)) = \int_0^\infty D_{i,j}(s)\, s\, ds \quad (3)$$

When PDFs (e.g. P_i) are used instead of pdfs (e.g.,p_i), $\int_{UD(o_i)} \int_{UD(o_j)} p_i(\mathbf{x}_i) p_j(\mathbf{x}_j)\, d\mathbf{x}_i\, d\mathbf{x}_j$ is changed to $\sum_{UD(o_i)} \sum_{UD(o_j)} P_i(\mathbf{x}_i) P_j(\mathbf{x}_j)$.

We choose *squared Euclidean distance* as the distance function D in this paper because of its easier integration compared with Euclidean distance or Manhattan distance. The following section presents the analytic solutions of expected distance of some special cases.

Note that for certain objects, the uncertain domains in the above formulae become their exact locations.

B. Analytic Solutions for Uniform and Gaussian Distributions (AS)

In this section, we have chosen to report the analytic solutions of spherical objects in different dimensions with uniform/Gaussian distribution with the following reasons. First, in practice, uniform distribution will be assumed if we have no information of the probability distribution of an uncertain object. Second, a pdf that is a step function can also be decomposed into several overlapping regions with uniform distribution. Third, Gaussian distribution can well approximate a variety of psychological test scores and physical phenomena in the behavioural and natural sciences. Fourth, Gaussian distribution maximizes information entropy among all distributions with known mean and variance, which is naturally chosen as the underlying distribution where we can just store the mean and variance of the samples of an uncertain object [16]. Fifth, arbitrary distributions could be approximated by a mixture of Gaussian distributions.

As shown in Figure 1, for uniform distribution, the expected squared distance $ED_{AS}(o_i, o_j)$ between a point object o_i and an uncertain object o_j whose uncertainty domain is (1) a line is $c^2 + (a^2 - ab + b^2)/3$; (2) a circle is $c^2 + r^2/2$; (3) a sphere is $c^2 + 3r^2/5$. The expected squared distance between

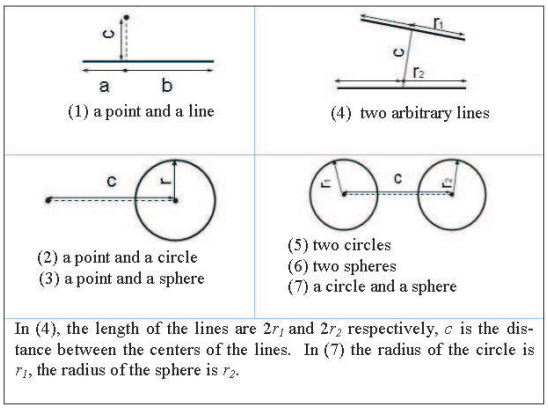

Fig. 1. Special Cases of Probabilistic Objects

(4) two arbitrary lines is $c^2 + (r_1^2 + r_2^2)/3$; (5) two circles is $c^2 + (r_1^2 + r_2^2)/2$; (6) two spheres is $c^2 + 3(r_1^2 + r_2^2)/5$; (7) a circle and a sphere is $c^2 + r_1^2/2 + +3r_2^2/5$.

Assume objects o_i follow Gaussian distribution $N(\mu_i, \Sigma_i)$ where μ_i is a $d \times 1$ mean vector, Σ_i is a $d \times d$ covariance matrix. The expected distance between objects o_i, o_j is

$$ED_{AS}(o_i, o_j) = \|\mu_i - \mu_j\|^2 + trace(\Sigma_i) + trace(\Sigma_j) \quad (4)$$

where $trace(\Sigma_i)$ is sum of all diagonal elements in Σ_i.

Due to the page limitation, the details of the derivation are not included here but they can be found at http://www.comp.polyu.edu.hk/∼csehung/paper/epdist-tech.ps

C. Approximation Methods for General Cases

Since it is usually difficult or impossible to derive analytic solutions for pdf other than uniform and Gaussian distributions, we have considered several methods to calculate the expected distance between uncertain objects in arbitrary distributions.

1) Distance between Means (DM): Just like the distance between certain data in traditional data mining applications, the expected distance between two uncertain objects o_i, o_j is calculated from the squared Euclidean distance between their means:

$$ED_{DM}(o_i, o_j) = \|\mu_i - \mu_j\|^2 \quad (5)$$

where μ_i and μ_j are the means of the object o_i and o_j. The mean of an uncertain object can be derived from the pdf or by taking average of the samples. For objects in d dimensions, the computation cost takes only $O((n_i + n_j)d)$, where n_i and n_j are number of samples of object o_i and o_j. However, this method does not consider the probability distributions of the uncertain data objects and naturally leads to low accuracy.

2) Pair-wise between Random Samples (PRS): This method takes random samples according to the pdf so that more samples appear in areas with higher probability density. All samples carry identical weights and the expected distance can be calculated by taking the average of the distances of all possible pair-wise combinations of samples of the two uncertain objects as follows:

$$ED_{PRS}(o_i, o_j) = \frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}||\mathbf{x}_{i,u} - \mathbf{x}_{j,v}||^2$$
$$= \frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}\sum_{w=1}^{d}(x_{i,u,w} - x_{j,v,w})^2 \quad (6)$$

where $x_{i,u,w}$ is the value of the wth attribute of uth sample of object o_i.

The time complexity of calculation of the pair-wise distance is $O(n_i n_j d)$ where d is the number of dimensions, n_i and n_j are sample numbers from object o_i and object o_j respectively.

3) Grid Approximation and Pair-wise between Samples (GAPS): This method divides the uncertainty domain of an uncertain object into a number of grid cells. The probability of each cell is approximated by multiplying the area of the cell and the probability density at the center of the cell. Another method of approximating the probability of a cell is by random sampling so that the probability of a cell is determined based on the number of samples in it. The first experiment we will report in the experimental section is based on the first method. (Note the first method described above is a very slight modification of that from [2].)

We use an example of 2-dimensional object to illustrate this method. Suppose we use a grid of $\sqrt{s}\times\sqrt{s}$ cells and s samples (one sample at the center of each cell) to approximate the probability density function. The probability for each cell of the grid is the cell area multiplied by its center's probability density. The cell probabilities are then normalized so that they sum up to 1. The expected distance is calculated by the sum of the distances between all pairs of cells (samples) from the two uncertain objects, weighted by the corresponding probability densities of cells as shown in Figure 2.

$$ED_{GAPS}(o_i, o_j) = \sum_{u,v}\sum_{u',v'} P_i(\mathbf{x}_{u,v})P_j(\mathbf{x}_{u',v'})\|\mathbf{x}_{u,v} - \mathbf{x}_{u',v'}\|^2 \quad (7)$$

For simplicity, we denote $\mathbf{x}_{u,v}$ and $P_i(\mathbf{x}_{u,v})$ as the center and the center's probability of the grid cell of uth row and vth column of object o_i; similarly are defined for object o_j.

The time complexity is $O(s^2 d)$ where d is the number of dimensions and s is the number of grid cells. The time complexity does not include the preprocessing time of estimating the probabilities of grid cells.

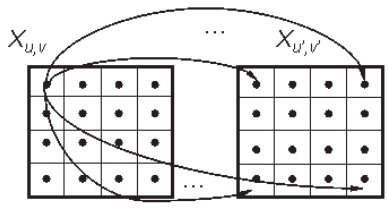

Fig. 2. Grid approximation and pair-wise between samples

4) Pair-wise between Gaussian Mixture (PGM): In this method, after random sampling in an uncertain object as described in PRS, we use K-means clustering to classify samples into a few clusters. Each cluster is approximated as a Gaussian distribution. As a result, an uncertain object is approximated by a mixture of Gaussian distributions in the form of $\sum_u^{C_i} A_{i,u}N(\mu_{i,u}, \Sigma_{i,u})$ where C_i is the number of clusters in object o_i, $\mu_{i,u}$ and $\Sigma_{i,u}$ are a $d \times 1$ mean vector and a $d \times d$ covariance matrix of the uth cluster of object o_i. $A_{i,u} = \frac{n_{k_i}}{n_i}$ is the weight of the uth cluster, where n_i is the total number of samples in o_i, n_{k_i} is the number of samples in the uth cluster of object o_i. Note that $\sum_{u=1}^{C_i} A_{i,u} = 1$.

Then, instead of pair-wisely calculating distances between samples of two objects, we pair-wisely calculate distances between clusters within two objects. The expected distance between two Gaussian distributions (clusters) from two objects can be obtained by Equation 4. The final expected distance between two uncertain objects can be obtained by taking an average of the distances above, each of which is weighted by the products of the weights of the two clusters, i.e.,

$$ED_{PGM}(o_i, o_j)$$
$$= \sum_{u=1}^{C_i}\sum_{v=1}^{C_j} A_{i,u}A_{j,v}$$
$$(\|\mu_{i,u} - \mu_{j,v}\|^2 + trace(\Sigma_{i,u}) + trace(\Sigma_{j,v})) \quad (8)$$

However, K-means clustering is very time consuming. In the coming section, we will find that the result has no relationships with K in K-means, which gives light to improve the calculation efficiency.

5) Approximation by Single Gaussian (ASG): Theoretically as we are going to show in the next section, and experimentally as in the experimental section, this last method ASG is the lowest in computational cost compared with all other methods (except method DM), which still gains the same accuracy as the previous method PGM and PRS.

Similar to PGM and PRS, we do random sampling in an uncertain object. We then approximate the object by a single Gaussian distribution $N(\mu_i, \Sigma_i)$ where μ_i and Σ_i are the mean and the covariance of the samples. The expected distance between two objects can be obtained by Equation 4, i.e.,

$$ED_{ASG}(o_i, o_j) = \|\mu_i - \mu_j\|^2 + trace(\Sigma_i) + trace(\Sigma_j) \quad (9)$$

where $trace(\Sigma_i)$ is sum of all diagonal elements in Σ_i.

The time complexities of ASG to compute the two mean vectors, the two variances and Equation 9 are $O((n_i + n_j)d)$, $O((n_i + n_j)d)$, and $O(d)$ respectively. Thus, the total complexity is $O((n_i + n_j)d)$.

D. Equivalence of PRS, PGM and ASG

The theorem below states that the results of PRS, PGM and ASG are equivalent to each other, given the same sample set.

Theorem 1: Given any uncertain objects o_i, o_j and their samples $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n_i}$, $\mathbf{x}_{j,1}, \ldots, \mathbf{x}_{j,n_j}$, $ED_{PRS}(o_i, o_j) = ED_{PGM}(o_i, o_j) = ED_{ASG}(o_i, o_j)$.

Proof of Theorem 1:

(i) The following well-known lemma in statistics will be used in our proof.

Lemma 1: Suppose we have taken n samples of a value a, and the samples are $a_1, \ldots, a_n$, then $\frac{1}{n}\sum_{u=1}^{n} a_u^2 = \bar{a}^2 + \frac{1}{n}\sum_{u=1}^{n}(a_u - \bar{a})^2$ where $\bar{a}$ is the sample mean, i.e., $\bar{a} = \frac{1}{n}\sum_{u=1}^{n} a_u$.

This lemma is a sample-based reflection of the following theorem:

$$E(a^2) = (E(a))^2 + Var(a) \tag{10}$$

where a is a random variable.

This lemma can be proved easily as follow:

$$\begin{aligned} \frac{1}{n}\sum_{u=1}^{n} a_u^2 &= \frac{1}{n}\sum_{u=1}^{n}(a_u - \bar{a} + \bar{a})^2 \\ &= \frac{1}{n}\sum_{u=1}^{n}\big((a_u - \bar{a})^2 + \bar{a}^2 - 2(a_u - \bar{a})\bar{a}\big) \end{aligned} \tag{11}$$

and

$$\sum_{i=1}^{n}(a_u - \bar{a})\bar{a} = \left(\sum_{u=1}^{n} a_u - n\bar{a}\right)\bar{a} = (n\bar{a} - n\bar{a})\bar{a} = 0 \tag{12}$$

Combine Eq.(11) and Eq.(12), we can readily obtain that

$$\frac{1}{n}\sum_{u=1}^{n} a_u^2 = \bar{a}^2 + \frac{1}{n}\sum_{u=1}^{n}(a_u - \bar{a})^2. \tag{13}$$

Thus the proof of the lemma is completed.

(ii) *Proof of* $ED_{PRS}(o_i, o_j) = ED_{ASG}(o_i, o_j)$*:*

From Equation 6, we have

$$\begin{aligned} ED_{PRS}(o_i, o_j) &= \frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}||\mathbf{x}_{i,u} - \mathbf{x}_{j,v}||^2 \\ &= \frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}\sum_{w=1}^{d}(x_{i,u,w} - x_{j,v,w})^2 \end{aligned}$$

We can easily see that the computation of ED_{PRS} can be decomposed into a superposition of the ED_{PRS} in each dimension.

$$\begin{aligned} ED_{PRS}(o_i, o_j) &= \sum_{w=1}^{d}\Big(\frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}(x_{i,u,w} - x_{j,v,w})^2\Big) \\ &= \sum_{w=1}^{d} ED_w(o_i, o_j) \end{aligned} \tag{14}$$

where

$$\begin{aligned} ED_w(o_i, o_j) &= \frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}(x_{i,u,w} - x_{j,v,w})^2 \\ &= \frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}x_{i,u,w}^2 + \frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}x_{j,v,w}^2 \\ &\quad -2\frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}x_{i,u,w}x_{j,v,w} \\ &= \frac{1}{n_i}\sum_{u=1}^{n_i}x_{i,u,w}^2 + \frac{1}{n_j}\sum_{v=1}^{n_j}x_{j,v,w}^2 \\ &\quad -2\frac{1}{n_i}\sum_{u=1}^{n_i}x_{i,u,w}\frac{1}{n_j}\sum_{v=1}^{n_j}x_{j,v,w} \end{aligned} \tag{15}$$

Based on Lemma 1, Equation 15 can be written as

$$\begin{aligned} ED_w(o_i, o_j) &= m_{i,w}^2 + s_{i,w}^2 + m_{j,w}^2 + s_{j,w}^2 - 2m_{i,w}m_{j,w} \\ &= (m_{i,w} - m_{j,w})^2 + s_{i,w}^2 + s_{j,w}^2 \end{aligned} \tag{16}$$

Here $m_{i,w}$ and $m_{j,w}$ are the sample means of the wth scalars (i.e., values of wth dimension) of o_i and o_j; and $s_{i,w}$ and $s_{j,w}$ are the sample variances:

$$m_{i,w} = \frac{1}{n_i}\sum_{u=1}^{n_i}x_{i,u,w}, \quad m_{j,w} = \frac{1}{n_j}\sum_{v=1}^{n_j}x_{j,v,w}$$

$$s_{i,w} = \frac{1}{n_i}\sum_{u=1}^{n_i}(x_{i,u,w} - m_{i,w})^2, \quad s_{j,w} = \frac{1}{n_j}\sum_{v=1}^{n_j}(x_{j,v,w} - m_{j,w})^2$$

By substituting Equation 16 to Equation 14, we have

$$\begin{aligned} ED_{PRS}(o_i, o_j) &= \sum_{w=1}^{d}\big((m_{i,w} - m_{j,w})^2 + s_{i,w}^2 + s_{j,w}^2\big) \\ &= ||\mathbf{m}_i - \mathbf{m}_j||^2 + \sum_{w=1}^{d}s_{i,w}^2 + \sum_{w=1}^{d}s_{j,w}^2 \\ &= \|\mu_i - \mu_j\|^2 + trace(\Sigma_i) + trace(\Sigma_j) \\ &= ED_{ASG}(o_i, o_j) \end{aligned} \tag{17}$$

(iii) *Proof of* $ED_{PRS}(o_i, o_j) = ED_{PGM}(o_i, o_j)$*:*

From Equation 14, we have

$$\begin{aligned} ED_w(o_i, o_j) &= \frac{1}{n_i}\frac{1}{n_j}\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}(x_{i,u,w} - x_{j,v,w})^2 \\ &= \frac{1}{n_i}\frac{1}{n_j}\sum_{k_i=1}^{C_i}\sum_{k_j=1}^{C_j}\sum_{u_{ki}=1}^{n_{k_i}}\sum_{v_{k_j}=1}^{n_{k_j}}(x_{i,u_{k_i},w} - x_{j,v_{k_j},w})^2 \\ &= \sum_{k_i=1}^{C_i}\sum_{k_j=1}^{C_j}\frac{n_{k_i}}{n_i}\frac{n_{k_j}}{n_j}\sum_{u_{k_i}=1}^{n_{k_i}}\sum_{v_{k_j}=1}^{n_{k_j}}(x_{i,u_{k_i},w} - x_{j,v_{k_j},w})^2 \end{aligned} \tag{18}$$

where C_i, C_j are the numbers of clusters (i.e., Gaussian distributions) for o_i and o_j respectively, n_{k_i} and n_{k_j} are the numbers of samples in k_ith cluster in object o_i and in k_jth cluster in object o_j. $\sum_{k_i=1}^{C_i} n_{k_i} = n_i, \sum_{k_j=1}^{C_j} n_{k_j} = n_j$.

From Equations 1,16 and 17,

$$\begin{aligned} ED_{PRS}(o_i, o_j) &= \sum_{k_i=1}^{C_i}\sum_{k_j=1}^{C_j}\frac{n_{k_i}}{n_i}\frac{n_{k_j}}{n_j} \\ &\quad \sum_{w=1}^{d}\sum_{u_{k_i}=1}^{n_{k_i}}\sum_{v_{k_j}=1}^{n_{k_j}}(x_{i,u_{k_i},w} - x_{j,v_{k_j},w})^2 \\ &= \sum_{k_i=1}^{C_i}\sum_{k_j=1}^{C_j}\frac{n_{k_i}}{n_i}\frac{n_{k_j}}{n_j}(\|\mu_{i,u_{k_i}} - \mu_{j,v_{k_j}}\|^2 \\ &\quad + trace(\Sigma_{i,u_{k_i}}) + trace(\Sigma_{j,v_{k_j}})) \quad (19) \end{aligned}$$

Recall A_{i,k_i} is the weight of the k_ith cluster in o_i so that $\sum_{k_i=1}^{C_i} A_{i,k_i} = 1$, A_{j,k_j} is the weight of the k_jth cluster in o_j so that $\sum_{k_j}^{C_j} A_{j,k_j} = 1$. Therefore,

$$\begin{aligned} ED_{PRS}(o_i, o_j) &= \sum_{k_i=1}^{C_i}\sum_{k_j=1}^{C_j} A_{i,k_i}A_{j,k_j} \\ &\quad (\|\mu_{i,u_{k_i}} - \mu_{j,v_{k_j}}\|^2 \\ &\quad + trace(\Sigma_{i,u_{k_i}}) + trace(\Sigma_{j,v_{k_j}})) \\ &= ED_{PGM}(o_i, o_j) \quad (20) \end{aligned}$$

The whole proof is completed.

This theorem shows that the results of PRS, PGM and ASG are the same. Therefore, we can use the much faster method ASG, which computes the means and variances of the uncertain objects and applies Equation 9 to obtain the result which has the same accuracy as PRS and PGM.

IV. Performance Study

We have done experiments by simulations in Matlab using a PC with 1.5 GHz Intel Pentium 4 CPU, and 512MB RAM. We found that the ASG method is much faster and more accurate than other methods. Note that the sample number of GAPS refers to its number of grid cells. K in K-means clustering in PGM is randomly set in $[3, 5]$ for each object.

A. Experiment 1 (Scalability w.r.t. Number of Samples)

In the first experiment, 100 uncertain objects are generated with the mean of every uncertain object located in a 100×100 $2D$ space as shown in Figure 3. The pdf of each object is a superposition of four Gaussian distributions. The variance of any Gaussian distribution and the distance between any two Gaussian distributions' means are uniformly distributed within $[1, 10]$. The correct expected distance between two uncertain objects are calculated by the analytic solution (Equation 4) we proposed in Section III-B, using the actual means and variances of the Gaussian mixture generated. This value will be used to compare with the values calculated by approximation methods proposed in Section III-C. We will compare all the five methods for the execution time. However, only DM, GAPS and ASG will be compared with the accuracy of the expected distance calculated because we have proved that the results of PRS, PGM and ASG are equivalent.

Figure 4 shows that the accuracy of ASG is always higher than GAPS and DM. The accuracy of a method is obtained by one minus the average relative error of the expected distances of all different pairs of the uncertain objects. DM's accuracy is not high (0.856 to 0.874) because it does not count the variances of the objects. The accuracy of ASG increases from 0.977 to 0.990 with more samples while that of GAPS is from 0.946 to 0.957.

Although ASG is only about 3% to 4% more accurate than GAPS, the execution time of ASG is much shorter than that of GAPS, as shown in Figure 5. GAPS is the slowest. PRS is also time consuming while the execution time of ASG is always less than 0.02ms.

GAPS is much slower because it takes two more multiplication operations of the probabilities of a grid cell from each object. Note that the execution time in Figure 5 does not include the preprocessing time of GAPS, which estimates the probability of each grid cell. The preprocessing time of GAPS is shown in Figure 6.

B. Experiment 2 (ASG's Performance on Data Generated as in [2])

In the second experiment, data are generated in the way similar to that in [2]. All 100 uncertain objects are located in a 100×100 $2D$ space. Each object is represented by an MBR (minimum bounding rectangle), which is simply randomly positioned inside the space. Each MBR is divided into 14×14 grid cells. Each grid cell has a probability randomly generated so that the sum of the probabilities of all cells equal to one.

Let k be the sample number, i.e., a total of k points are randomly located at the centers of some grid cells according to the probabilities of the grid cells. In the other words, it is *likely* that more samples are located at cells with higher probability. GAPS provides the correct answers because the grid probabilities completely capture the pdf of the object. Therefore we would compare the results of ASG with that of GAPS to see how accurate ASG can be.

Figure 7 shows that the accuracy of ASG increases from 0.988 to 0.997 when the sample number varies between 64 and 324.

Figure 8 shows that ASG performs much faster than GAPS.

The results here show that even the pdfs of uncertain objects are so arbitrary, ASG can still perform very well. A distance calculation method very similar to GAPS was used in [2]. The high accuracy of ASG suggests that ASG can replace the method used in [2] and significantly improves the execution time of their results. Note that the execution time of distance calculation reported in [2] is much shorter than our GAPS because they computed the distance between an uncertain object and a cluster centroid which is a certain point. However, GAPS here computes distance between two uncertain objects, which involves much more distance calculations.

C. Experiment 3 (ASG's Performance on Objects with Uniform pdf)

In this experiment, the uncertain objects are generated as circles with uniform distribution. We would like to see whether

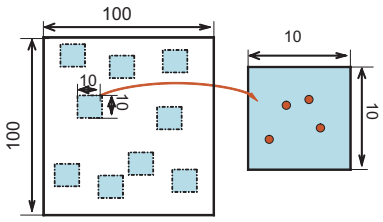

Fig. 3. Data setting in Experiment 1

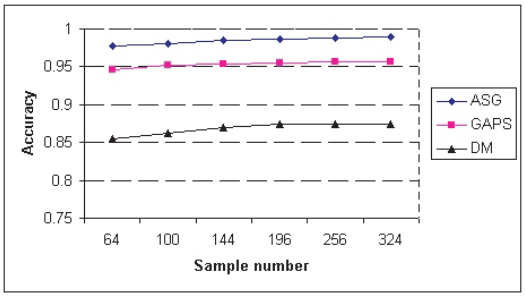

Fig. 4. Comparison of accuracy of ASG, GAPS and DM in Experiment 1

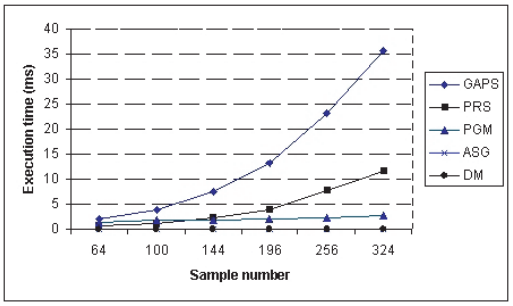

Fig. 5. Execution time of DM,PRS,GAPS,PGM and ASG in Experiment 1

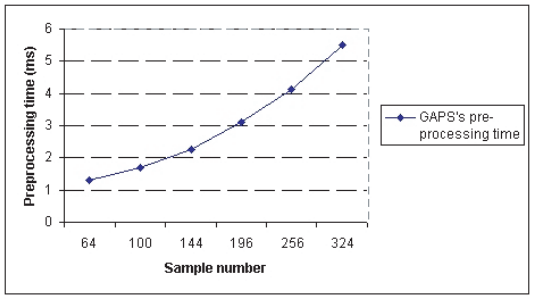

Fig. 6. Processing time of GAPS in Experiment 1

ASG, which approximates a pdf using a single Gaussian distribution, can also approximate uniform distribution well or not.

10 uncertain objects, each with radius randomly in [1,5], are randomly located in 100×100 2D space. As shown in Figure 1 (case (5)) and Section III-B, the correct expected distance can be calculated and used to compare with the answers given by ASG. ASG takes 100 samples on each object. We repeated ASG for 6 times and found that in the worst case the accuracy is over 0.98 and the average accuracy is over 0.99.

D. Experiment 4 (Scalability w.r.t. Number of Dimensions)

In this experiment, we would like to see how scalable DM, PRS, GAPS and ASG are with respect to the dimensionality. We generated 10 uncertain objects similarly as in Experiment 1. The number of samples or grid cells for two, three and four dimensional spaces are 256, 216, 256 respectively. As shown in Figure 9, we find that the accuracy of ASG is very high: from 0.97 to 0.99. GAPS's accuracy is also not low: from 0.93 to 0.96. Their accuracy decreases slightly when the number of dimensions increases because the space becomes much larger but the sample number does not change much.

Figure 10 shows that the execution time of GAPS and PRS increases almost linearly to the number of dimensions. This is probably because the number of terms in the distance calculation is directly proportional to the number of dimensions. The execution time of ASG, PGM and DM does not change much.

Among all methods, GAPS is the most time consuming and ASG is very close to the fastest method DM.

V. Conclusion

We have described the importance of expected distance calculation in queries and data mining applications on uncertain data. We provided the analytic solutions of special cases (uniform and Gaussian distributions) and proposed five approximation methods for general cases (arbitrary distributions). We have shown theoretically and experimentally that ASG, in a very short execution time, can obtain results of very high accuracy (compared with other existing methods that use sampling). This strongly suggests that ASG can replace the calculation method (similar to GAPS) used in recent research work for answering queries as well as data mining applications on uncertain data.

References

[1] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press*, 1967, vol. 1, pp. 281–297.

[2] Jacky Ngai, Ben Kao, Reynold Cheng, Michael Chow, Kevin Yip, and Chun-Kit Chui, "Efficient clustering of uncertain data," The 2006 IEEE International Conference on Data Mining (ICDM 2006), Hong Kong, December, 2006.

[3] Edwin M. Knorr and Raymond T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, Ashish Gupta, Oded Shmueli, and Jennifer Widom, Eds. 1998, pp. 392–403, Morgan Kaufmann.

[4] Roger Cavallo and Michael Pittarelli, "The theory of probabilistic databases.," in *VLDB*, 1987, pp. 71–81.

[5] Daniel Barbará, Hector Garcia-Molina, and Daryl Porter, "The management of probabilistic data.," *IEEE Trans. Knowl. Data Eng.*, vol. 4, no. 5, pp. 487–502, 1992.

[6] Debabrata Dey and Sumit Sarkar, "A probabilistic relational model and algebra," *ACM Trans. Database Syst.*, vol. 21, no. 3, pp. 339–369, 1996.

[7] Laks V. S. Lakshmanan, Nicola Leone, Robert Ross, and V. S. Subrahmanian, "Probview: A flexible probabilistic database system.," *ACM Trans. Database Syst.*, vol. 22, no. 3, pp. 419–469, 1997.

[8] Robert Ross, V. S. Subrahmanian, and John Grant, "Aggregate operators in probabilistic databases.," *J. ACM*, vol. 52, no. 1, pp. 54–101, 2005.

[9] Alex Dekhtyar, Robert Ross, and V. S. Subrahmanian, "Probabilistic temporal databases, i: algebra.," *ACM Trans. Database Syst.*, vol. 26, no. 1, pp. 41–95, 2001.

[10] Thomas Eiter, James J. Lu, Thomas Lukasiewicz, and V. S. Subrahmanian, "Probabilistic object bases.," *ACM Trans. Database Syst.*, vol. 26, no. 3, pp. 264–312, 2001.

[11] Andrew Nierman and H. V. Jagadish, "Protdb: Probabilistic data in XML.," in *VLDB*, 2002, pp. 646–657.

[12] Edward Hung, Lise Getoor, and V. S. Subrahmanian, "PXML: A probabilistic semistructured data model and algebra.," in *ICDE*, 2003, pp. 467–478.

[13] Edward Hung, Lise Getoor, and V. S. Subrahmanian, "Probabilistic interval XML.," in *ACM Transactions on Computational Logic (TOCL)*, 2006.

[14] Renata M. C. R. de Souza and Francisco de A. T. de Carvalho, "Clustering of interval data based on city-block distances.," *Pattern Recognition Letters*, vol. 25, no. 3, pp. 353–365, 2004.

[15] Hans-Peter Kriegel, Peter Kunath, Martin Pfeifle, and Matthias Renz, "Probabilistic similarity join on uncertain data.," in *DASFAA*, 2006, pp. 295–309.

[16] Wikipedia, "http://en.wikipedia.org/wiki/normal_distribution," 2006.

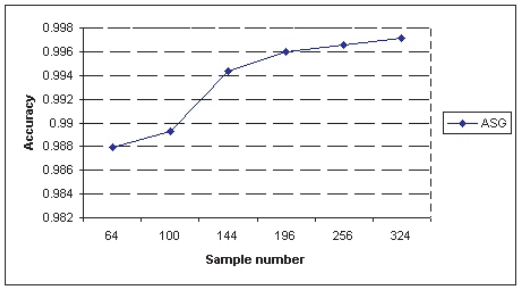

Fig. 7. Accuracy of ASG in Experiment 2

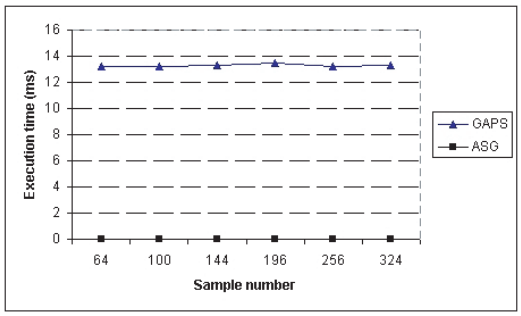

Fig. 8. Execution time of GAPS and ASG in Experiment 2

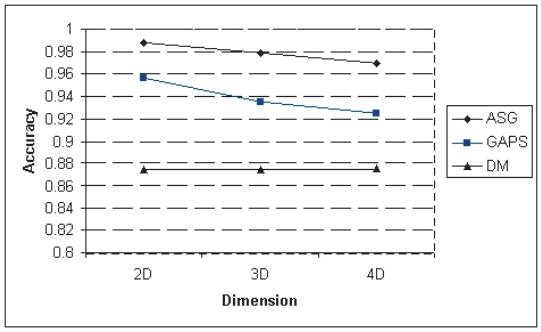

Fig. 9. Accuracy of DM, PRS, GAPS and ASG with varying dimensionality in Experiment 4

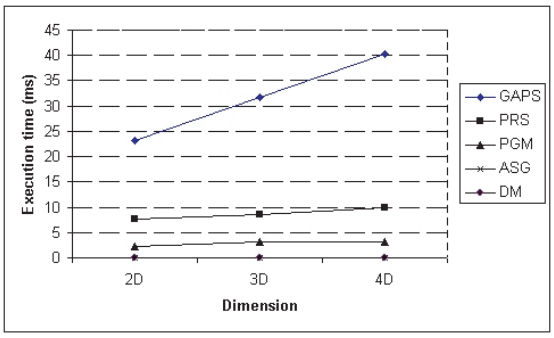

Fig. 10. Execution time of DM, PRS, GAPS and ASG with varying dimensionality in Experiment 4