On Multidimensional Sharpening of Uncertain Data

Charu C. Aggarwal*

Abstract

In this paper, we will propose a technique for multidimensional enhancement of uncertain data. In many applications, the uncertainty in the different dimensions is caused by independent factors, especially if the different dimensions have been collected from independent sources. In such cases, it is possible to enhance the quality of the data and reduce the underlying uncertainty by using multidimensional uncertainty analysis. In this paper, we will discuss techniques for uncertainty reduction of multidimensional uncertain data. We will examine the effectiveness of the approach over a variety of real and synthetic data sets.

1 Introduction

In recent years, many new techniques have been developed for extracting uncertain data from a wide variety of applications. This has resulted a need for developing a variety of techniques for managing and mining uncertain data [4, 5, 13]. For example, in sensor networks, the data collected is often uncertain because of errors in the underlying readings. Often, the uncertainty is captured with the use of sensor modeling techniques. In techniques such as privacy-preserving data mining, the errors may be intentionally added in order to increase the uncertainty in the data. Some recent techniques [2] explicitly model the data in uncertain format. In many applications such as forecasting, the data may be synthetically constructed. In such cases, the uncertainty in the data may be modeled by using the known statistical characteristics of the underlying methodology of synthetic data construction.

The field of uncertain data has seen a revival in recent years, because of new ways of collecting such information. Newly developed techniques include those of clustering [3, 12, 15], classification [1], indexing [7, 8, 14] and query processing [6, 9]. Clearly, the quality of the final results are dependent upon the level of uncertainty present in the data. The presence of less uncertainty will create better quality results and vice-versa. Therefore, uncertainty reduction is useful in practical applications.

In many data mining and management models such as clustering, classification, and indexing [3, 12, 7, 8], it is often assumed that the uncertainty in the different attributes are independent of one another. This is often a direct result of how the data is collected, since different attributes in the data may be obtained from different sources. In practice, most real data contains considerable correlations, as a result of which the different data values are not independent of one another. This structure in the data is useful information, when the different attributes in the data have been collected independently, and it is possible to use the correlation structure in the data in order to reduce the uncertainty. A related work [11] shows how to attack privacy of perturbed data values, when the *entire distribution* of the perturbation is known, and all record are perturbed using the same distribution.

In this paper, we will examine the use of the correlation structure in order to reduce the uncertainty in the underlying data representation. This enhances the quality of the data for mining and management purposes. We will show the effectiveness of the approach on a number of real and synthetic data sets.

This paper is organized as follows. In the next section, we will propose techniques for reducing the uncertainty in the underlying data. In section 3, we will discuss the experimental results. Section 4 contains the conclusions and summary.

2 Sharpening the Uncertain Representation

We will first introduce some notations and definitions. It is assumed that the database $\mathcal{D}$ contains N uncertain records for which the average values are denoted by $\overline{X_1} \ldots \overline{X_N}$. The dimensionality of each record is denoted by d. The corresponding probability distributions are denoted by $\overline{f_1(\cdot)} \ldots \overline{f_N(\cdot)}$. Thus, $\overline{X_i}$ is the mean value of the probability distribution $f_i(\cdot)$. It is assumed that the jth component of record $\overline{X_i}$is denoted by x_{ij}. Similarly, the probability distribution for the jth component of the ith record is denoted by $f_{ij}(\cdot)$.

A key observation is that the uncertainty in the data may often result from independent sources. For example, different attributes may be collected from different measuring instruments, sensors, or other agents for which the underlying uncertainty will be clearly in-

*IBM T. J. Watson Research Center, charu@us.ibm.com

dependent, since they are specific to the approach being used for data collection. Yet, the true values in the data will continue to retain the underlying correlations, since this is a natural property of most multidimensional data sets. Furthermore, while the observed values clearly depend upon the magnitude of the underlying noise, we can assume that the noise is independent of the true (unknown) values of the underlying data, since the noise is assumed to be specific to the data collection methodology rather than the value of the data collected.

In order to reduce the uncertainty in the underlying data, we will use an approach which is based on singular value decomposition of the underlying data. The point of using singular value decomposition is to determine the hidden structure in the data, and exploit it in order to sharpen the underlying uncertain representation. The first step is to determine the covariance structure of the underlying data records. This is a challenge, since we do not have the true values of the records available, but only the uncertain representations. Let us assume that the original (unknown) value of the jth dimension for record $\overline{X_i}$ is denoted by z_{ij}. Let $\mathcal{D}^z$ be the (unknown) database of original values. Then, the value of z_{ij} is obtained by adding r_{ij} to x_{ij}. Thus, r_{ij} represents the noise in modeling the mean value of the distribution $f_{ij}(\cdot)$. Therefore, we have:

$$z_{ij} = x_{ij} + r_{ij}$$
$$x_{ij} = z_{ij} - r_{ij}$$

Let us denote that *random variable* corresponding to the jth dimension of the database $\overline{X_1} \ldots \overline{X_n}$ by $\hat{\mathcal{X}}_j$. Thus, there are N possible instantiations of this record present in the data, which are denoted by $x_{1j} \ldots x_{Nj}$ respectively. Note that while $\overline{X_i}$ represents the ith **row** (or instantiation vector across all dimensions), the notation $\hat{\mathcal{X}}_j$ represents the jth **column** (or dimension) of a *random variable*. It is important to not confuse between these two notations, since one represents an instantiation, whereas another represents a random variable. Furthermore, one corresponds to a row vector of $[x_{ij}]$, whereas the (instantiation of the) other correspond to a column vector of $[x_{ij}]$.

Similarly, we can define the random variable corresponding to the true value $[z_{ij}]$ of the jth dimension of the original data by $\hat{\mathcal{Z}}_j$. The jth dimension of the random variable corresponding to $[r_{ij}]$ is denoted by $\hat{\mathcal{R}}_j$. Then, we have:

$$\hat{\mathcal{X}}_j = \hat{\mathcal{Z}}_j - \hat{\mathcal{R}}_j \tag{2.1}$$

We note that the random variable $\hat{\mathcal{R}}_j$ is independent of the random variable $\hat{\mathcal{Z}}_j$ corresponding to the true record value. This is a key assumption which is required in order to prove the result. We would like to determine the covariance between the dimensions j and k on the original data in terms of the covariances of the observed data and the noise in the underlying data. The covariance between the dimensions j and k in the original data is represented by $Cov(\hat{\mathcal{Z}}_j, \hat{\mathcal{Z}}_k)$. We would like to express $Cov(\hat{\mathcal{Z}}_j, \hat{\mathcal{Z}}_k)$ in terms of $Cov(\hat{\mathcal{X}}_j, \hat{\mathcal{X}}_k)$ and $Cov(\hat{\mathcal{R}}_j, \hat{\mathcal{R}}_k)$. This will be helpful in determining the true covariance of the underlying data, and will be exploited for sharpening purposes. We claim the following result:

LEMMA 2.1. *The covariance matrices for the variables $[x_{ij}]$, $[z_{ij}]$ and $[r_{ij}]$ are related as follows:*

$$Cov(\hat{\mathcal{X}}_j, \hat{\mathcal{X}}_k) = Cov(\hat{\mathcal{Z}}_j, \hat{\mathcal{Z}}_k) + Cov(\hat{\mathcal{R}}_j, \hat{\mathcal{R}}_k) \tag{2.2}$$

Proof. By expanding $\hat{\mathcal{X}}_j = \hat{\mathcal{Z}}_j - \hat{\mathcal{R}}_j$, we get:

$$Cov(\hat{\mathcal{X}}_j, \hat{\mathcal{X}}_k) = Cov((\hat{\mathcal{Z}}_j - \hat{\mathcal{R}}_j), (\hat{\mathcal{Z}}_k - \hat{\mathcal{R}}_j)) \tag{2.3}$$

By expanding the expression on the right, we get:

$$Cov(\hat{\mathcal{X}}_j, \hat{\mathcal{X}}_k) = Cov(\hat{\mathcal{Z}}_j, \hat{\mathcal{Z}}_k) - Cov(\hat{\mathcal{Z}}_j, \hat{\mathcal{R}}_k) - \\ -Cov(\hat{\mathcal{Z}}_k, \hat{\mathcal{R}}_j) + Cov(\hat{\mathcal{R}}_j, \hat{\mathcal{R}}_k)$$

Since the true (unknown) values $[z_{ij}]$ and the noise values $[r_{ij}]$ are assumed to be independent of one another, we know that $Cov(\hat{\mathcal{Z}}_j, \hat{\mathcal{R}}_k) = 0$ and $Cov(\hat{\mathcal{Z}}_k, \hat{\mathcal{R}}_j) = 0$. Therefore, we can simplify the expression above as follows:

$$Cov(\hat{\mathcal{X}}_j, \hat{\mathcal{X}}_k) = Cov(\hat{\mathcal{Z}}_j, \hat{\mathcal{Z}}_k) + Cov(\hat{\mathcal{R}}_j, \hat{\mathcal{R}}_k) \tag{2.4}$$

This proves the result.

We would like to use the above expression in order to estimate the covariance $Cov(\hat{\mathcal{Z}}_j, \hat{\mathcal{Z}}_k)$. While the value of $Cov(\hat{\mathcal{X}}_j, \hat{\mathcal{X}}_k)$ can be estimated from the underlying data, the estimation of $Cov(\hat{\mathcal{R}}_j, \hat{\mathcal{R}}_k)$ requires some further explanation. When $j \neq k$, we have $Cov(\hat{\mathcal{R}}_j, \hat{\mathcal{R}}_k) = 0$, because the noise from different data sources are assumed to be independent. On the other hand, when $j = k$, the value of $Cov(\hat{\mathcal{R}}_j, \hat{\mathcal{R}}_k)$ is simply the variance of $\hat{\mathcal{R}}_j$. We denote this by $var(\hat{\mathcal{R}}_j)$.

Let us assume that the standard deviation of the function $f_{ij}(\cdot)$ is denoted by σ_{ij}. Then, the value of $var(\hat{\mathcal{R}}_j)$ of the jth dimension of $[r_{ij}]$ is given by the average of the corresponding variances of the corresponding probability density functions. Therefore, we estimate $var(\hat{\mathcal{R}}_j)$ as follows:

$$var(\hat{\mathcal{R}}_j) = \sum_{i=1}^{N} \sigma_{ii}^2 / N \tag{2.5}$$

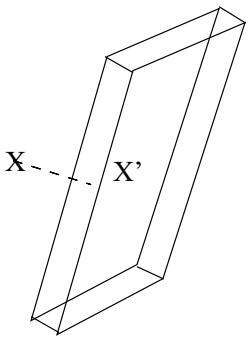

Figure 1: Uncertain Data Projection onto Principal Components

The result of Equation 2.2 can then be used in order in order to estimate the value of $Cov(\hat{\mathcal{Z}}_j, \hat{\mathcal{Z}}_k)$, which can be use to construct the covariance matrix of $[z_{ij}]$. The only difference between the covariance matrices of $[x_{ij}]$ and $[z_{ij}]$ are the lower values on the diagonal matrix of the latter. All other covariance values are the same. Thus, we first determine the covariance matrix of the $[x_{ij}]$, and then subtract the corresponding variances $var(\hat{\mathcal{R}}_j)$ from the diagonal entries (j, j).

Since z_{ij} is obtained by adding r_{ij} to x_{ij}, it would seem counter-intuitive that the variances of the covariance matrix of $[z_{ij}]$ are lower than those of $[x_{ij}]$. However, we note that the noise r_{ij} is assumed to be independent of the true (unknown) value of the uncertain data, and not the estimated average of the probability density function. The estimated average of the probability density function contains added noise because of modeling assumptions. This independence assumption is critical in resulting in lower variances of the diagonal entries in the covariance matrix of $[z_{ij}]$. Intuitively, the variances for $[z_{ij}]$ are lower, since these are the true values without the added noise, whereas the values $[x_{ij}]$ include any noise from the uncertain measurements. We further note that the covariance matrix for $[z_{ij}]$ is only estimated from the available data, and can be somewhat inaccurate for small data sets. Since the variance of the diagonal entries of $[z_{ij}]$ are obtained by subtracting the variances from the corresponding diagonal entries of $[x_{ij}]$, it is possible that estimation inaccuracies may lead to some diagonal values being slightly less than 0. Since all diagonal values are variances, they cannot be less than zero. In order to deal with such cases, we simply set any negative diagonal entry to 0. Let C^z denote the corresponding covariance matrix. This covariance matrix can be diagonalized using the following expression:

$$C^z = P \cdot D \cdot P^T \tag{2.6}$$

Since C^z is a covariance-matrix, it is positive-semidefinite, and it can be diagonalized with non-negative eigenvalues. In this case, P is an orthonormal matrix for which the columns are the orthonormal eigenvectors of C^z. The matrix D is a diagonal matrix which contains the corresponding eigenvalues. The eigenvectors are the principal components in the data, along which the second-order correlations are zero. The eigenvalues are the variances along these different principal components in the data. It can also be shown that for any given number k of dimensions, it is possible to maximize the energy in the projected data by picking only the eigenvectors which have zero second-order correlations.

In most real data sets, a large fraction of these eigenvalues are close to zero, since most of the information in the data can be represented along a small number k of principal components. However, since the uncertainty in the *unknown true values of the* data is independent of the values in the original data, the *average values in the collected data* may not lie along these principal components. This provides information about how the true values in the uncertain data are distributed. In order to illustrate this point, we have shown an example in Figure 1, in which the data points are mostly distributed on a 2-dimensional plane in 3-dimensional space. However, the average of the probability distribution of the data point $\overline{X}$ does not lie on this plane. Clearly, the probability distribution of $\overline{X}$ can be sharpened to bring it closer to the plane. For example, if X is projected onto X' this brings it closer to the *global data distribution* with the use of the principal components. A variation of this broad principle is to try to determine a fit for the average value based on *both the relative magnitude of the eigenvalues*, and the local distribution of the provided density function. We will use this variation in order to further improve the accuracy of representation.

In practice, we may not choose to use a hard threshold in order to project the data point X onto X', but we may choose to use a more refined approach with the help of the underlying eigenvalues. Our goal here is to combine the local data distribution of each point with the global data distribution of the entire data set in order to find an *optimally sharpened local uncertain distribution*. We will show that such an approach has considerable advantages in improving the quality of the underlying data set. Next, we will discuss how the sharpened uncertain distribution may be determined. The process of determination of the covariance matrix of the underlying data values is illustrated in Figure 2.

2.1 Determining an Optimally Fitting Distribution The process of determining an optimally fitting

Algorithm *DataCovariance*($\overline{X_1} \ldots \overline{X_N}$, $\overline{f_1(\cdot)} \ldots \overline{f_N(\cdot)}$)
begin
Let C^x be covariance matrix of
$\overline{X_1} \ldots \overline{X_N}$;
Let σ_{ij}^2 be the variance of the
density function $f_{ij}(\cdot)$;
Construct the $d * d$ matrix C^r
such that the ith diagonal entry
is $\sum_{j=1}^{d} \sigma_{ij}^2$ and
all other entries are 0;
$C^z = C^x - C^r$;
Diagonalize $C^z = P \cdot D \cdot P^T$;
return(P, D);
end

Figure 2: Determining Covariance Matrix

Algorithm *DetermineSharp*($\overline{X_i}$, $f_i(\cdot)$,
$\lambda_1 \ldots \lambda_d$, $\overline{e_1} \ldots \overline{e_d}$);
begin
Let $\overline{X'}$ be the projection of $\overline{X}$
onto $\overline{e_1} \ldots \overline{e_r}$;
$\overline{Y} = \overline{X'}$;
while not(termination) **do**
begin
Determine c_t such that
$\mathcal{F}(\overline{Y} + c_t \cdot \nabla \mathcal{F}(\overline{Y}))$ is maximized;
$\overline{Y} = \overline{Y} + c_t \cdot \nabla \mathcal{F}(\overline{Y})$;
end
Let $g_{ij}^q(\cdot)$ be the density function which
is the same as function $f_{ij}(\cdot)$, except that
it is centered at y_j;
Pick q using binary search, so that the
fit of x_{ij} with $g_{ij}^q(\cdot)$
is maximized;
end

Figure 3: Determining the Sharpened Distribution

distribution consists of two steps: **(1)** We first determine the optimal central point of the sharpened probability density function. This is done by finding a point at which the combination of global and local fits is maximized. **(2)** For simplicity, the overall shape of the new distribution is assumed to be the same as that of the original distribution. However, the variances are optimized in order to fit the old observation (central point) with the newly computed central point of the distribution.

In order to determine the global fit, we use our earlier methodology for determining the principal components in the data. We first construct the covariance matrix and diagonalize it using principal component analysis. Let $\overline{e_1} \ldots \overline{e_d}$ be the eigenvectors along the d different directions in the data, and $\lambda_1 \ldots \lambda_d$ be the d different eigenvalues. Without loss of generality, we can assume that that the eigenvectors are arranged in order of decreasing eigenvalues. Then, we define the *global data distribution* as a gaussian distribution along d different directions with variances $\lambda_1 \ldots \lambda_d$. It is assumed that the variances along these N different directions are independent of one another. We will refer to this distribution as $G(\cdot)$. In typical real data sets, the inter-attribute correlations ensure that most of the variance is preserved in a small number of eigenvectors $\overline{e_1} \ldots \overline{e_r}$. These are also the eigenvectors with the largest eigenvalues $\lambda_1 \ldots \lambda_r$. The fact that only a small number of the eigenvectors contain most of the variance can be used in order to sharpen the accuracy of representation.

Now, let us consider the data point $\overline{X_i}$ with probability density function which is denoted by $f_i(\cdot)$. Let $\overline{Y}$ be the newly estimated central point of the data distribution. This newly estimated central point may be determined by calculating the combined fits of the data point $\overline{Y}$ to the distributions $f_i(\cdot)$ and $G(\cdot)$. We define the combined fit $\mathcal{F}(\overline{Y})$ of the data point $\overline{Y}$ as follows:

DEFINITION 1. *The combined fit $\mathcal{F}(\overline{Y})$ for the new center $\overline{Y}$ for the uncertain data point $(\overline{X_i}, f_i(\cdot))$ in a data set with global distribution $G(\cdot)$ is equal to $log(G(\overline{Y})) + log(f_i(\overline{Y}))$.*

We note that the above is simply an indirect representation of the product of the corresponding probability density functions at data point $\overline{Y}$. By using the logarithm of $G(\overline{Y}) \cdot f_i(\overline{Y})$, it is possible to avoid numerical errors and achieve better accuracy of representation and computation. In order to maximize the sharpening of the uncertain representation, we would like to find the optimal point $\overline{Y}$, at which $\mathcal{F}(Y)$ is maximized. We state this problem as follows:

PROBLEM 2.1. *Find the point $\overline{Y}$ at which $\mathcal{F}(\overline{Y})$ is maximized.*

Note that this is a difficult problem, since we are trying to optimize a nonlinear objection function. A natural choice is to use a gradient descent (ascent)[1] approach in order to find the optimal value of the data point $\overline{Y}$. The gradient of the fit $\mathcal{F}(\overline{Y})$ is defined by $\nabla \mathcal{F}(\overline{Y})$, and defines a direction along which incremental changes maximize the increase in the objection function value of $\mathcal{F}(\overline{Y})$. The length c_t of the tth step creates a step which is denoted by $\mathcal{F}(\overline{Y}) + c_t \cdot \nabla \mathcal{F}(\overline{Y})$. The value of c_t is picked using binary search in order to maximize the improvement in a single step. The termination criterion for the algorithm is defined by its convergence behavior. The algorithm converges when its improvement in a

[1] Gradient descent is used for minimization, whereas gradient ascent is used for maximization.

given step is less than its improvement in the first step by a factor of less than 1%. The final data point thus determined is denoted by $\overline{Y} = (y_1 \ldots y_d)$.

An important issue in the effective implementation of the algorithm is the creation of a good starting point. In order to pick a good starting point, we simply pick the eigenvectors $\{\overline{e_1} \ldots \overline{e_r}\}$ which contain 99% of the variance and project the data point $\overline{X_i}$ onto the hyperplane created by this set of eigenvectors denoted by $\{\overline{e_1} \ldots \overline{e_r}\}$. If $\overline{X_i'}$ be the corresponding projected data point, then it is used as the starting point for the gradient ascent method. We note that a good starting point ensures that the gradient descent method converges quickly to an accurate solution. We note that more effective solutions are possible for picking the starting point. When an eigenvector is excluded from the selection, we lose information which is proportional to the square root of the eigenvalue (standard deviation of the original data along that eigenvector), but we also lose noise which is equal to the sum of the uncertainty standard deviation projections of the different dimensions along that eigenvector. As long as the information loss is less than the noise less, the corresponding eigenvector can be projected out. Next, we will discuss the gradient descent method for determining the optimal position of the final data point.

Let $g_{ij}^q(.)$ represent the same function as $f_{ij}(.)$ except that it is centered at y_j instead of x_{ij}, and the standard deviation is given by $q \cdot \sigma_{ij}$, where q is a proportionality constant. For simplicity, we are assuming that the overall shape of the uncertain data distribution along a given dimension remains the same. It remains to determine q. Since the function $g_{ij}^q(\cdot)$ represents the local data distribution of point x_{ij} (with the addition of global distribution information), we would like to pick q such that the density of our original model mean x_{ij} is maximized. The idea here is that the original model mean was an instantiation of the uncertain data behavior, and we would like to maximize its fit with the new model. Therefore, we need to determine the value of q such that $g_{ij}^q(\overline{X_i})$ is maximized.

PROBLEM 2.2. *Determine the value of q such that $g_{ij}^q(x_{ij})$ is maximized.*

For general distributions, the value of q may be determined by binary search, since the value of $g_{ij}^q(x_{ij})$ first increases with increasing q and then reduces. Therefore, by starting off with a value of q much smaller than the distance between x_{ij} and y_j (in terms of standard deviations), and doubling in each iteration, it is possible to determine a range in which $g_{ij}^q(\cdot)$ peaks. Once this range has been determined, we can use a bracket bisection technique in order to narrow down to the final value of q to any desired degree of accuracy. The overall process of determination of the density function is illustrated in Figure 3.

For some special distributions, it is possible to compute the value of q in closed form. Closed form solutions are always desirable for ease in computation. Two such distributions are the gaussian distribution and the uniform distribution. For the case of the gaussian distribution, the density fit $g_{ij}^q(x_{ij}$ at the point x_{ij} is as follows:

$$g_{ij}^q(\cdot) = \frac{1}{\sqrt{2 \cdot \pi} \cdot q \cdot \sigma_{ij}} e^{-\frac{(x_{ij}-y_j)^2}{2 \cdot q^2 \cdot \sigma_{ij}^2}} \tag{2.7}$$

Since this expression needs to maximized with respect to the variable q, we can differentiate with respect to q and set the resulting expression to zero. Therefore, we have $\delta g_{ij}^q(\cdot) \delta q = 0$. It follows that:

$$(-1/q^2 + (x_{ij}-y_j)^2/(q^4 \cdot \sigma_{ij}^2)) \cdot \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma_{ij}} e^{-\frac{(x_{ij}-y_j)^2}{2 \cdot q^2 \cdot \sigma_{ij}^2}} = 0$$

$$q = |(x_{ij} - y_j)|/\sigma_{ij}$$

Therefore, we summarize as follows:

LEMMA 2.2. *The optimal value of q which optimizes Problem 2.2 for the gaussian distribution $q_{ij}^q(\cdot)$ is given by:*

$$q = |(x_{ij} - y_j)|/\sigma_{ij} \tag{2.8}$$

We can also derive the range of the sharpened distribution for the case when the $f_{ij}(\cdot)$ is a uniform distribution. This case is much easer to derive, since the range of the new distribution is given by $\sigma_{ij} \cdot q \cdot \sqrt{12}$, and the corresponding probability density is given by $1/(\sigma_{ij} \cdot q \cdot \sqrt{12})$. This density is a monotonically decreasing function of q. Therefore, we wish to pick the smallest value of q such at $\overline{x_{ij}}$ is included in a distribution with range $\sigma_{ij} \cdot q \cdot \sqrt{12}$, and centered at y_j. Therefore, the minimum such range of the new distribution is given by $2 \cdot |x_{ij} - y_j|$, in order to include x_{ij} within the end points of this range. Therefore, it follows that:

$$\sigma_{ij} \cdot q \cdot \sqrt{12} = 2 \cdot |x_{ij} - y_j| \tag{2.9}$$

$$q = |(x_{ij} - y_j)|/(\sigma_{ij} \cdot \sqrt{3}) \tag{2.10}$$

Therefore, we summarize as follows:

LEMMA 2.3. *The optimal value of q which optimizes Problem 2.2 for the uniform distribution $q_{ij}^q(\cdot)$ is given by:*

$$q = |(x_{ij} - y_j)|/(\sigma_{ij} \cdot \sqrt{3}) \tag{2.11}$$

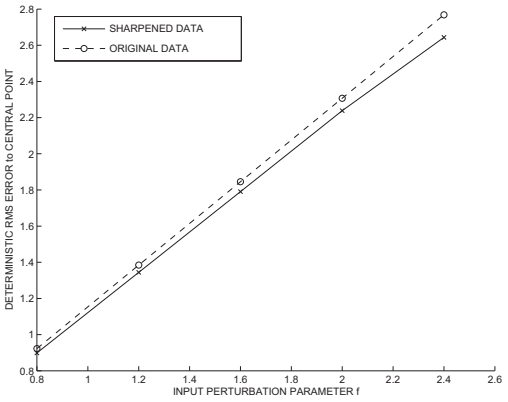

Figure 4: Average (per dimension) deterministic RMS distance of mid-point of uncertain data distribution from true data point (Corel.U(f) Data Set)

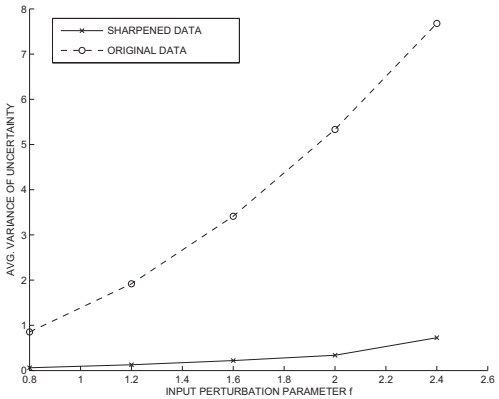

Figure 5: Variance (per dimension) of uncertain data point (Corel.U(f) Data Set)

For the case of general distributions, a closed form solution may not be derived, but we can use binary search in order to derive the value of q. This is used to reconstruct the final distribution.

3 Experimental Results

We designed a number of experimental tests in order to measure the effectiveness of the algorithm for sharpening the underlying data. One inherent difficulty with measuring the effectiveness of the sharpening process, is that in most uncertain data applications (such as those involving measurement or hardware errors), only the uncertain data distribution is known, though the true data values may not be known. In order to develop a benchmark for effectiveness, we need to create the uncertainty synthetically, so that the true data values are known. In order to achieve this goal, we will add uncertainty to the real data sets from the UCI machine learning repository, and then apply our algorithms for sharpening the resulting representation. This is a particularly effective way to test the effectiveness of sharpening, since the *true data set* is known before the uncertainty was added. Therefore, it is possible to meaningfully measure the effects of the sharpening process.

Each data set was normalized, so that the standard deviation along each dimension was one unit. This was done in order to provide better interpretability to the errors in the results in terms of the standard deviations along each dimension. Errors were added to the data set with the use of a normal distribution with zero mean, and a standard deviation whose parameter was chosen as follows. For each entry, the standard deviation *parameter* of the normal distribution was chosen from

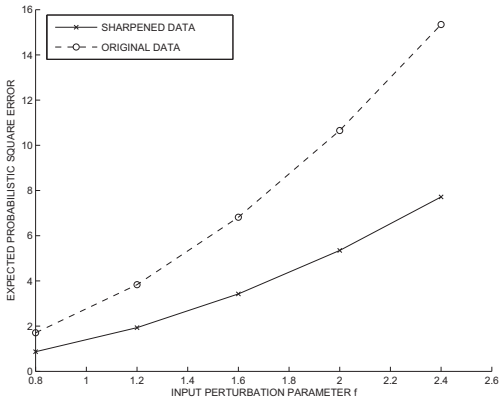

Figure 6: Expected probabilistic mean square error (per dimension) between uncertain data point and true data point (Corel Data Set)

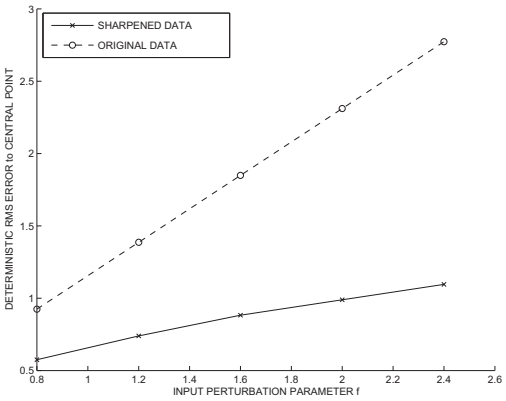

Figure 7: Average (per dimension) deterministic RMS distance of mid-point of uncertain data distribution from true data point (Musk Data Set)

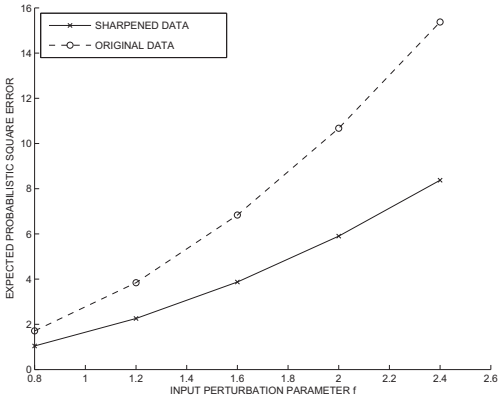

Figure 9: Expected probabilistic mean square error (per dimension) between uncertain data point and true data point (Musk Data Set)

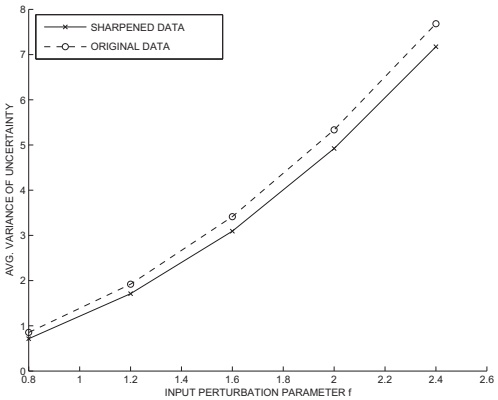

Figure 8: Variance (per dimension) of uncertain data point (Musk Data Set)

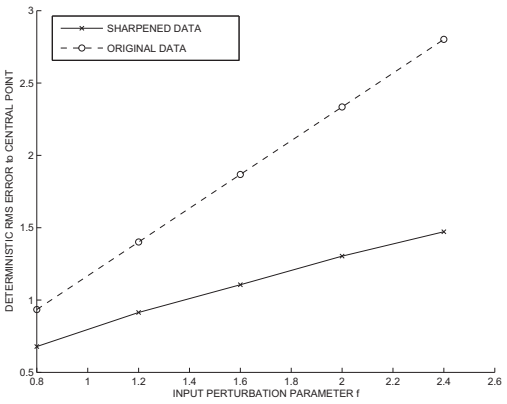

Figure 10: Average (per dimension) deterministic RMS distance of mid-point of uncertain data distribution from true data point (Wisconsin Breast Cancer Data Set)

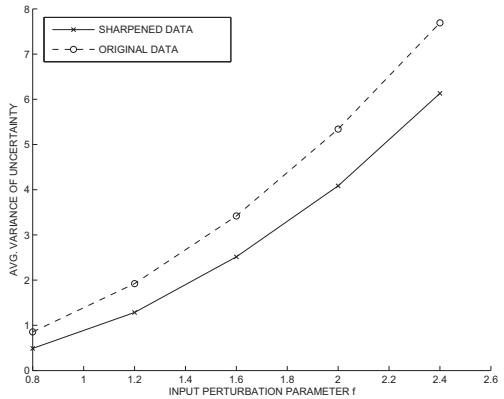

Figure 11: Variance (per dimension) of uncertain data point (Wisconsin Breast Cancer Data Set)

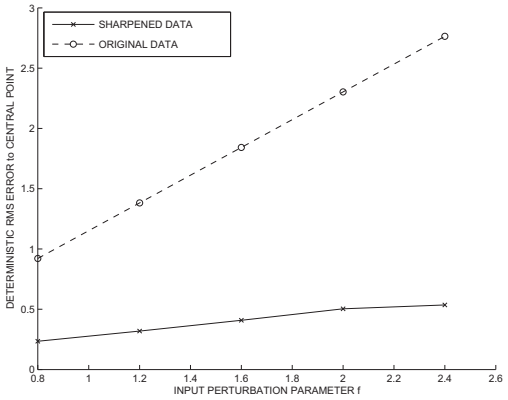

Figure 13: Average (per dimension) deterministic RMS distance of mid-point of uncertain data distribution from true data point (Synthetic Data Set D5000.d100.θ2)

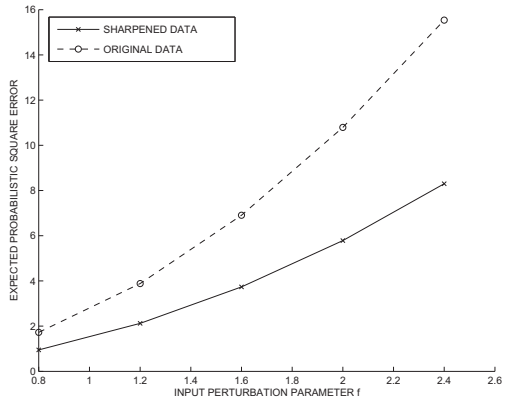

Figure 12: Expected probabilistic mean square error (per dimension) between uncertain data point and true data point (Wisconsin Breast Cancer Data Set)

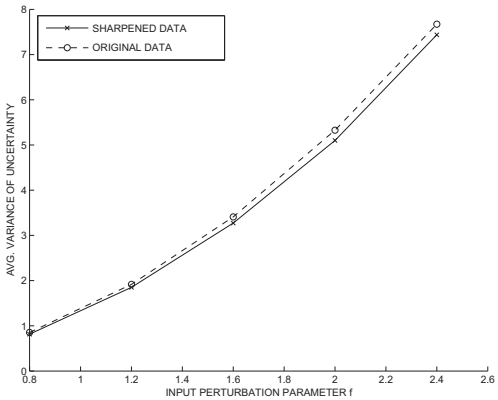

Figure 14: Variance (per dimension) of uncertain data point (Synthetic Data Set D5000.d100.θ2)

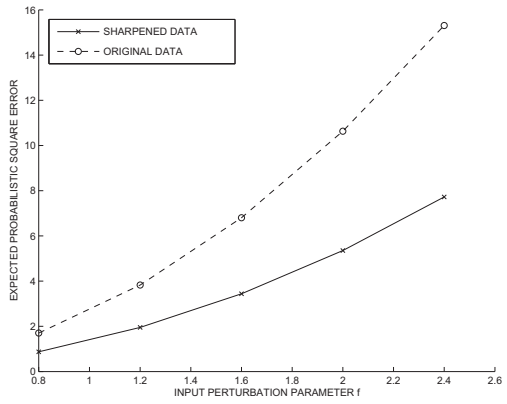

Figure 15: Expected probabilistic mean square error (per dimension) between uncertain data point and true data point (Synthetic Data Set D5000.d100.θ2)

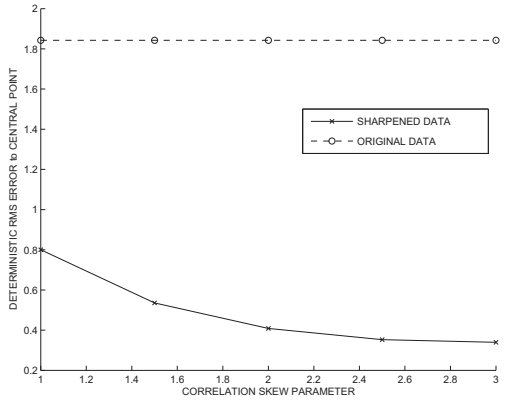

Figure 16: Average (per dimension) deterministic RMS error with increasing correlation factor of underlying data (D5000.d100.$\theta(x)$.U(1.6))

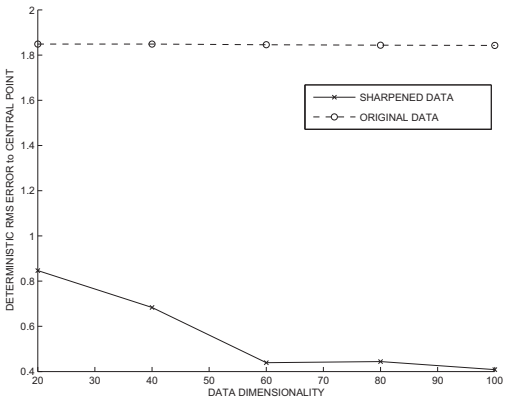

Figure 17: Average (per dimension) deterministic RMS error with increasing dimensionality of the underlying data (D5000.d(x).θ2.U(1.6))

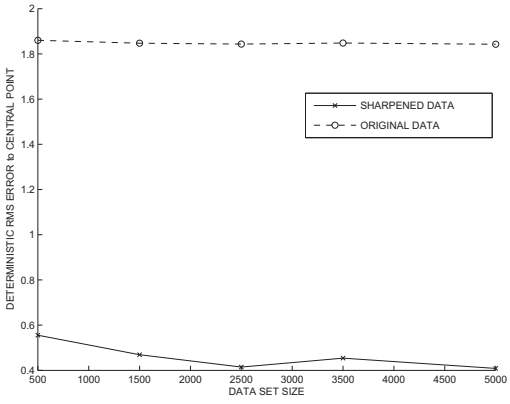

Figure 18: Average (per dimension) deterministic RMS error with increasing size of the underlying data (D(x).d100.θ2.U(1.6))

a uniform distribution in the range $[0, 2 \cdot f]$. Thus, by changing the value of f, it is possible to vary the uncertainty level in the data set. We refer to f as the *input perturbation parameter*. For each entry, we assumed that the normal distribution was centered at the point which was obtained after adding errors to the true data set from the UCI machine learning repository. We note that the errors in the central point of the uncertain data distribution may correspond to the measurement or modeling errors in a real application. The *true data point value* may never really be known in a real application. Therefore, for the purpose of the experiments we would like to distinguish between the term *original data* and *true data*. The original data corresponds to the data before sharpening, but with the uncertainty incorporated in it. This is the data available in a real application and includes both the error in modeling as well as the uncertain distribution associated with it. Therefore, it is the *original data available* in a real application. The *true data* are the true (deterministic) values from the UCI machine learning repository. In a real application, these true deterministic values may never be known. However, our uncertainty generation process provides us a benchmark to measure the effectiveness of the sharpening process. A data set $< DataSetName >$ to which the uncertainty level f was added is denoted by $< DataSetName > .U(f)$.

In order for the sharpened data to be better than the original data set, we would like the mean of the new uncertain distribution to be closer to the true values. Additionally, we would like the variance of the sharpened distribution to be competitive or lower than the original data. Finally, the expected error of the entire uncertain distribution from the true (unknown) values should be smaller than those of the original data set to which the transformation was applied. Therefore, the primary measures use to judge the quality of the final data were as follows:

- We calculated the RMS distance (averaged along each dimension and data point) between the mean of the probability distribution of both representation of the uncertain data (original data and perturbed data) with the true values (which are available from the base data) that was perturbed. We note that this is a deterministic value rather than an expected value.
- We computed the variance of the probability distribution of the original and sharpened data sets.
- This measure computed the expected probabilistic mean square error (per data point and per dimension) between the true data point and the two uncertain representations (original and sharpened) representations. Thus, for a data point with probability density function $f_{ij}(\cdot)$ and true value $\overline{Z}$, the measure $Pdist(\overline{Z}, f_{ij}(\cdot))$ for this single point and dimension j is computed as follows:

$$Pdist(\overline{Z}, f_{ij}(\cdot)) = \int_x ||\overline{Z} - x||^2 \cdot f_{ij}(x) dx \tag{3.12}$$

The corresponding mean square error MSE over all data points and dimensions is computed as follows:

$$MSE = \sum_{\text{All } \overline{Z}} \sum_{\text{All dim. } j} Pdist(\overline{Z}, f_{ij}(\cdot))^2/(N \cdot d) \tag{3.13}$$

We note that the main distance between the third measure and the first is that the latter accounts for the entire pdf of the underlying uncertainty, whereas the former only looks at the accuracy of the mid-point of the uncertain probability density function.

One of the points to be noted is that all of the above measures are characteristic of the uncertainty in the data rather than the global distribution of the data. Therefore, if the same perturbation parameter f is used in order to add uncertainty to the data sets from the UCI repository, the resulting measures will be very similar irrespective of the nature of the underlying true data values. On the other hand, we will also see that the level of sharpening is sensitive of the behavior of the underlying data. Therefore, when the above measures are plotted with respect to increasing uncertainty level, the resulting curves look very similar on the different data sets. On the other hand, the results on the sharpened data look quite different for different data sets, since the effectiveness of the sharpening is somewhat dependent on the characteristics of the base data. We will explore the sensitivity of the sharpening to the base data characteristics in the experimental section.

We used three real data sets and one synthetic data set in order to test our approach. The real data sets were the Musk, Corel histogram, and Wisconsin Breast Cancer data sets (WBC), which were obtained from the UCI machine learning repository. We further note that the intermediate estimation steps are particularly difficult for smaller data sets. Therefore two of the real data sets (Musk and WBC) were particularly small, containing 476 and 569 data points respectively.

In order to test the effects of different data set characteristics, we also generated a series of synthetic data sets. We note that the main characteristics which affect multi-dimensional sharpening are (1) Data size

(2) Dimensionality (3) Correlations between different dimensions. Therefore, we need a controlled way of varying these parameters for a given data set. In order to generate such a series of data sets, we first generated an axis system with random orientation. This axis system represents the directions of correlation. The level of correlation can be varied by changing the variances along the different axis directions. Note that in a data set with low implicit dimensionality, most of the variance is concentrated along a few of the axis-directions which are also referred to as principal components. Therefore, in order to create skew in the variance along the different principal components, we determined the standard-deviation along the ith axis direction using the Zipf distribution $1/i^{\theta}$. Therefore, the implicit dimensionality can be varied by changing the value of θ. Increasing values of θ lead to a larger level of correlation. Since the data set is synthetic, it is also easy to vary the dimensionality and number of points. When the data set is generated with N points, a dimensionality of l, and a correlation factor of c, then the corresponding data set is referred to as $D(N).d(l).\theta(c)$.

The overall deterministic errors for the mean of the distribution are illustrated in Figures 4, 7, 10 and 13 respectively. In each case, the uncertainty level f is modeled on the X-axis, whereas the deterministic error is modeled on the Y-axis. Thus, the figures can show the relative behavior of the different data sets with increasing uncertainty level. As discussed earlier, these curves represent the deterministic RMS error (per dimension) in the mean of the uncertain distribution from the true values for the sharpened and unsharpened data sets respectively. In each figure, the dotted curves represent the behavior of the unsharpened data, whereas the solid curves represent the behavior of the sharpened data. Since the measure is only dependent upon the errors in the uncertain data and not on the behavior of the base data, the dotted curves look very similar across all data sets. However, the sharpening process is dependent upon correlation characteristics of the data sets, and therefore there is variation in the behavior of the solid curve across the different data sets. However, in each case, the sharpened data has much lower error as compared to the unsharpened data. The second observation is that the difference in the quality of the sharpened and unsharpened data increases with error level. Thus, the greater the errors in the data set, the greater the utility of our technique.

In addition to the mean errors of the uncertain distribution, we also compared the variances of the unsharpened and sharpened data. The results are illustrated in Figures 5, 8, 11 and 14 respectively. In each case, the uncertainty level f is illustrated on the X-axis, and the variance of the uncertainty is illustrated on the Y-axis. In each case, the sharpened data has somewhat lower variance of the uncertainty than the unsharpened data. This is because the mean values of the new probability density functions have now been corrected, and therefore a lower variance is required in order to represent the behavior of the underlying data. Clearly, it is desirable to have lower variances on the uncertain distributions in order to obtain the most effective results for data mining and management applications.

Finally, we measured the expected probabilistic square error with respect to the true values. The results are illustrated in Figures 6, 9, 12 and 15 respectively. We note that this measure accounts for both the mean errors and the variances in the uncertainty. In each case, the uncertainty level f is illustrated on the X-axis, and the expected square error (per dimension) is illustrated on the Y-axis. As in the case of the deterministic measures, the sharpened data has much lower expected error. Furthermore, this difference in quality increased with increasing uncertainty level in each case.

Finally, we also tested the effectiveness of the sharpening method with the underlying characteristics of the data sets. Clearly, many of the underlying characteristics of the data sets such as the correlation factor, the dimensionality and the data set size affect the sharpening process. In Figure 16, we have illustrated the effectiveness of the sharpening process with an increasing correlation factor θ. In each case, the uncertainty level f was fixed at 1.6. For this purpose, we used the series of data sets denoted by $D5000.d100.\theta(x).U(1.6)$, where x denotes the varying value. The value of θ is illustrated on the X-axis, whereas the error of the mean of the distribution is illustrated on the Y-axis. In each case, it is clear that the quality difference between the sharpened and unsharpened data increases with increasing value of the skew parameter θ. The error levels on the original data do not change very much with increasing correlation, since the error metric is essentially independent of the underlying data. However, the quality of the sharpened data increases significantly with increasing correlation level. This is to be expected, since the sharpening process uses the relationship between the different dimensions in order to improve the quality of the representation. Therefore, a higher level of correlation leads to greater predictability of each value, since the inter-attribute correlations can be used in order increase the accuracy and reduce the uncertainty of each data value.

We also examined the behavior of the sharpening method with increasing data dimensionality. In Figure 18, we have illustrated the error level of the data set

with increasing dimensionality. In this case, the data set series was denoted by $D(100).d(x).\theta(2).U(1.6)$, where x denotes the changing variable. The dimensionality is illustrated on the X-axis, whereas the error is illustrated on the Y-axis. The errors (per dimension) on the unsharpened data do not change much with increasing data dimensionality. However, the errors reduce significantly with increasing data dimensionality, because a greater number of dimensions are now available in order to reduce the errors and uncertainty in the correlation-based sharpening process. Since high dimensional data sets are quite common in real applications, this is quite promising for the utility of the sharpening method.

In Figure 18, we have illustrated the error level of the data set with increasing number of data points. In this case, the data set series was denoted by $D(x).d100.\theta(2).U(1.6)$, where x denotes the changing variable. The number of data points are illustrated on the X-axis, whereas the error is illustrated on the Y-axis. As in the previous case, the errors on the unsharpened data do not increase with data set size. However, the errors on the unsharpened data reduce considerably with data size. This is because of the intermediate steps is approximate covariance estimation which is most accurate with increasing data size. While the accuracy increased considerably with increasing data size, it is interesting to see that the sharpening process continues to be effective even for very small data sets containing only 500 points. This illustrates the robustness of the sharpening method.

4 Conclusions and Summary

In this paper, we presented a method for multidimensional sharpening of uncertain data sets. The process of sharpening improves the quality of the underlying representation by using the correlation information which is usually available in the underlying information. Since most real data sets have underlying inter-attribute correlations, this means that the technique is usually quite helpful in improving the accuracy of representation. Our results show that the technique not only improves the mean of the representation, but also reduces the variance of the underlying uncertainty. Furthermore, the technique continues to be effective on relatively small data sets, whose statistical parameters are often difficult to estimate accurately because of the underlying uncertainty. Our results show that the technique is extremely effective on a wide variety of data sets, and improves with increasing data set size and dimensionality.

Acknowledgement

This research is continuing through participation in the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under contract number W911NF-09-2-0053.

References

[1] C. C. Aggarwal, *On Density Based Transforms for Uncertain Data Mining*, ICDE Conference, (2007), pp. 866–875.

[2] C. C. Aggarwal, *On Unifying Privacy and Uncertain Data Models*, ICDE Conference, (2008), pp. 386–395.

[3] C. C. Aggarwal, and P. S. Yu, *A Framework for Clustering Uncertain Data Streams*, ICDE Conference, (2008), pp. 150–159.

[4] C. C. Aggarwal, and P. S. Yu, *A Survey of Uncertain Data Algorithms and Applications*, IEEE Transactions on Knowledge and Data Engineering, 21(5), May 2009, pp. 609–623.

[5] C. C. Aggarwal, *Managing and Mining Uncertain Data*, (2009), Springer.

[6] D. Burdick, P. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, *OLAP Over Uncertain and Imprecise Data*, VLDB Conference, (2005), pp. 970–981.

[7] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. Vitter, Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data, VLDB Conference, (2004), pp. 876–887.

[8] R. Cheng, D. Kalashnikov, and S. Prabhakar, *Evaluating Probabilistic Queries over Imprecise Data*, SIGMOD Conference, (2003), pp. 551–562.

[9] N. Dalvi, and D. Suciu, *Efficient Query Evaluation on Probabilistic Databases*, VLDB Conference, (2004), pp. 864–875.

[10] A. Das Sarma, O. Benjelloun, A. Halevy, and J. Widom, *Working Models for Uncertain Data*, ICDE Conference, (2006), pp. 7.

[11] Z. Huang, W. Du, and B. Chen, *Deriving Private Information from Randomized Data*, ACM SIGMOD Conference, (2005), pp. 37–48.

[12] H.-P. Kriegel, and M. Pfeifle, *Density-based clustering of uncertain data*, KDD Conference, (2005) pp. 672–677.

[13] L. V. S Lakshmanan, N. Leone, R. Ross, and V. S. Subrahmanian, *ProbView: A Flexible Database System*, ACM Transactions on Database Systems, 22(3), 1997, pp. 419–469.

[14] S. Singh, C. Mayfield, S. Prabhakar, R. Shah, and S. Hambrusch, *Indexing Uncertain Categorical Data*, ICDE Conference, (2007), pp. 616–625.

[15] Y. Tao, R. Cheng, X. Xiao, W. Ngai, B. Kao, and S. Prabhakar, *Indexing Multi-dimensional Uncertain Data with Arbitrary Probability Density Functions*, VLDB Conference, (2005), pp. 922-933.