RC22727 (W0302-088) February 20, 2003
Computer Science

IBM Research Report

A Simple Method of Data Perturbation and Efficient Algorithms for Density Estimation in Privacy Preserving Data Mining

Chai Wah Wu
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

A simple method for data perturbation and efficient algorithms for density estimation in privacy preserving data mining.

Chai Wah Wu
IBM Research Division, Thomas J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598, U.S.A.

Abstract

In recent years, there have been privacy concerns over the proliferation of gathering of personal information by various institutions and merchants over the internet. This has led to the development of data mining algorithms which preserve the privacy of the people whose personal data are collected and analyzed. A novel approach to such privacy preserving data mining algorithms was recently proposed where the individual data in a large data set is perturbed by adding a random value from a known distribution. This perturbation is performed by the user so that the true value of the data is not known to the data mining algorithm. In these applications, the distribution of the original data set is important and estimating it is part of the goals of the data mining algorithm. This distribution is estimated via an iterative algorithm. An algorithm based on the Expectation Maximization (EM) algorithm was subsequently shown to have desirable properties such as the ability to have low privacy loss and high fidelity estimates of the distribution of the data set. Both these algorithms are iterative in nature and each iteration requires computation which is proportional to the size of the data set and to the number of points in the estimate. This can require large computation time to estimate the distribution. In this paper we propose two methods to reduce the amount of computation. The first method constructs in one step an initial estimate of the distribution to aid the iterative algorithm in order to reduce the number of iterations. In the second method, we propose another scheme for perturbing data which also has the nice properties of allowing arbitrarily small privacy loss and arbitrarily high fidelity in the estimate (i.e. zero information loss). The main advantage of this proposed scheme is the simplicity of the estimation algorithm. In contrast to iterative algorithms such as EM, the proposed scheme admits an algorithm which estimates the unknown distribution in one step. This is significant in applications where the data set is very large or when the data mining algorithm is run in an online environment.

Keywords: privacy, data mining, density estimation, orthogonal series estimators, Expectation Maximization

1 Introduction

The proliferation of data mining algorithms to extract statistics and trends from large sets of user supplied data results in the need for data mining algorithms which preserve privacy. In [1, 2, 3] this problem is addressed from a cryptographic standpoint where data mining computations among several parties are performed on the combined data sets of the parties without revealing each party's data to the other parties and the problem is solved using multi-party secure computation. This approach is not suitable for scenarios where a single party collects data from many users who do not talk to each other and where this single party performs data mining operation on this data. An example of this is a survey that is conducted over the internet. In [4], a novel approach to address this problem is proposed where the data is perturbed by a random value from a known distribution. The specific random value is generated at the user's site and is unknown to the data mining algorithm. This random value is tied to the data, so that repeated queries return the same perturbed value. The data mining algorithm can reconstruct or estimate the distribution of the original data set, yet does not know the exact value of the individual items. The estimation algorithm is an iterative algorithm derived from Bayesian analysis. In [5], this problem is solved with a Expectation Maximization (EM) estimation algorithm which has better convergence properties and can be shown to converge to the maximum likelihood estimate (MLE). Furthermore, [5] introduced two new metrics, namely privacy loss and information loss to capture the amount of data in an individual record leaked to the data mining algorithm and the fidelity of the estimate respectively. One of the design goals of such privacy preserving data mining algorithms is to derive algorithms which can have a small privacy loss and a small information loss. It is clear that privacy loss is small when the perturbation is large. In [5], using a convergence result of EM, it was shown that the information loss is small when the data set is large. Both the estimation algorithms in [4] and in [5] are iterative in nature. The algorithm is run until a stopping criterion is met at which point an estimate for the distribution of the data set is obtained. The entire data set is used in each iteration. Thus the number of computations at each iteration is proportional to the size of the data set and the number of bins used in the estimate. The purpose of this paper is to present two ways to reduce the amount of computation in the estimation algorithm. In the first way, we

generate in one step a good estimate of the distribution in order to reduce the number of iterations needed or eliminate the iterative step completely (Section 3).

For the second way, we modify the protocol of data perturbation and propose a novel privacy preserving scheme for data mining applications which also can have arbitrarily small privacy loss and information loss when the data set is large (Section 4). The main advantage of this proposed scheme is that the estimation algorithm is an extremely simple one-step process. This has significant performance advantages, especially when the data set is very large or when the data mining is done in an online dynamic environment. The simplicity of the estimation algorithm also makes it amenable to a simple analysis and correctness proof.

Finally, we propose other metrics for quantifying privacy loss and information loss and show that the proposed scheme can achieve arbitrarily small privacy loss and information loss. To quantify the loss of privacy, we study the probability in which an estimate of a single sample is correct. We also consider a slightly different information loss metric than the one proposed in [5] which we believe is more accurate in categorizing the loss of information due to the perturbation.

2 Problem definition

The basic problem can be abstracted into the following mathematical problem. Consider a set of n original data values $x_1, \ldots, x_n$, each considered as samples taken independently from the same random variable X. To create the perturbation, n values $y_1, \ldots, y_n$ are taken independently from the same random variable Y and the perturbed values $z_i = g(x_i, y_i)$ are created. We also assume that X and Y are independent. Given these perturbed values and the density function $f_Y(y)$ of Y, the goal is to estimate the density function $f_X(x)$ of X. In the example of the internet survey, x_i correspond to the participants' answers, y_i correspond to the perturbations generated and z_i correspond to the perturbed answers which are sent to the server for collection. In [4, 5] x_i and y_i are real numbers and the composition function g is simply addition, i.e. $z_i = x_i + y_i$.

Here we take the more general view that x_i, y_i and z_i are not necessarily real numbers, but live in some spaces S_X, S_Y and S_Z respectively. In Section 4 we present a protocol using a more complicated composition function g where y_i and z_i are both vectors. It may seem unintuitive, but by using a more complicated function g, the corresponding estimation algorithm is very simple and efficient.

3 Calculating the Fourier coefficients of f_X

Consider the case as in [4, 5] where x_i and y_i are real numbers and $z_i = x_i + y_i$. In this section we show how the Fourier coefficients of f_X can be computed in one step to generate an estimate of f_X. Estimating the Fourier coefficients of f_X in order to estimate f_X belong the the class of orthogonal series estimators [6, 7]. This estimate can be further refined using the iterative methods of [4, 5]. When the one-step estimate is close to f_X, the number of iterations needed in the refinement step is smaller than in [4, 5] (see example below).

Assume that the data is properly scaled so that the support of X is a subset of $[0, 1]$. We want to express f_X as a Fourier series defined on the interval $[0, 1]$: $f_X(x) = a_0 + \sum_{i=1}^{\infty} a_i \sin(2\pi i x) + \sum_{i=1}^{\infty} b_i \cos(2\pi i x)$. Since f_X is a probability density function, $a_0 = 1$. As $\sqrt{2}\sin(2\pi i x)$ and $\sqrt{2}\cos(2\pi i x)$ are orthonormal in $f([0, 1])$, it follows that $a_i = 2E(\sin(2\pi i X))$ and $b_i = 2E(\cos(2\pi i X)$. Because of the independence of X and Y, we have

$$\begin{aligned} E(\sin(2\pi i Z)) &= E(\sin(2\pi i(Y + X))) \\ &= E(\sin(2\pi i Y))E(\cos(2\pi i X)) + E(\cos(2\pi i Y))E(\sin(2\pi i X)) \\ E(\cos(2\pi i Z)) &= E(\cos(2\pi i(Y + X))) \\ &= E(\cos(2\pi i Y))E(\cos(2\pi i X)) - E(\sin(2\pi i Y))E(\sin(2\pi i X)) \end{aligned}$$

Therefore

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} = 2 \begin{pmatrix} E(\sin(2\pi i X)) \\ E(\cos(2\pi i X)) \end{pmatrix} = 2A_i^{-1} \begin{pmatrix} E(\sin(2\pi i Z)) \\ E(\cos(2\pi i Z)) \end{pmatrix}$$

where

$$A_i = \begin{pmatrix} E(\cos(2\pi i Y)) & E(\sin(2\pi i Y)) \\ -E(\sin(2\pi i Y)) & E(\cos(2\pi i Y)) \end{pmatrix}$$

Since f_Y is known, A_i^{-1} can be computed in advance and $E(\sin(2\pi i Z))$ and $E(\cos(2\pi i Z))$ are estimated as $\frac{1}{n}\sum_{j=1}^{n} \sin(2\pi i z_j)$ and $\frac{1}{n}\sum_{j=1}^{n} \cos(2\pi i z_j)$ respectively. One of the drawback of this method is that it works as long as A_i is not close to being singular for the coefficients that we are interested in. In other words, this method works well if f_Y has higher frequency components than f_X. On the other hand, the estimated density function is independent of the number of bins used in the iterative algorithms of [4, 5]. As is common in orthogonal series estimators, the Fourier coefficients need to be smoothed e.g., via finite truncation of the Fourier series or by weighting the coefficients [8].

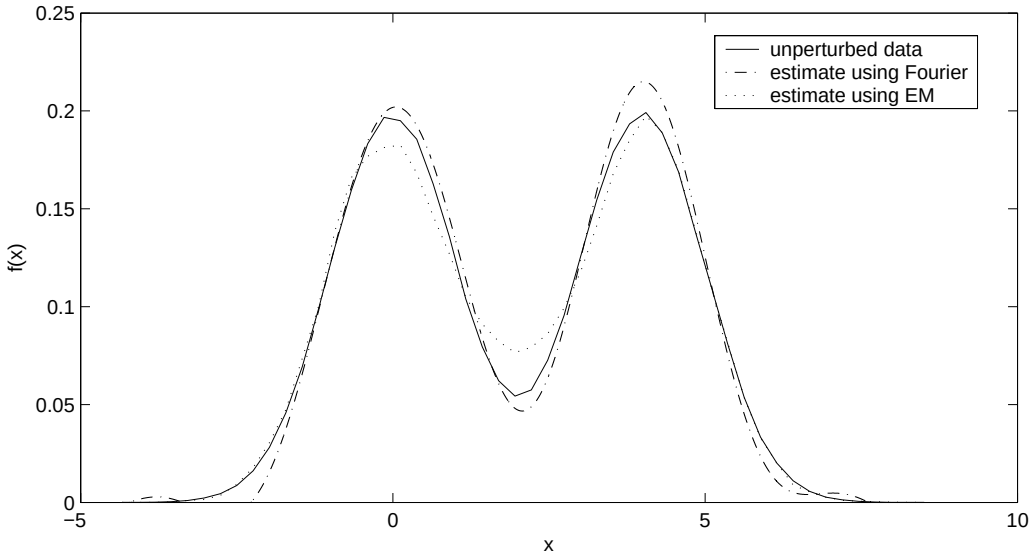

Figure 1: Plot of distribution of unperturbed data and the estimated distributions using 1) Fourier coefficients based method and 2) EM method. The number of samples is 500000. The perturbation Y is a uniform distribution in the range $[0, 20]$. The EM algorithm is run for 100 iterations.

To illustrate the effectiveness of this estimation method, consider the following example: X is a bimodal distribution, and Y is a uniformly distributed random variable in the range [0, 20]. Fig. 1 shows the estimated probability distribution and the unperturbed distribution using 500000 samples. The unperturbed distribution is computed using $k = 50$ bins. The density function is estimated using Fourier coefficients a_i, b_i ,$i = 1, 2, ..., 12$ and forced to be nonnegative. Also shown in Fig. 1 is the estimated distribution using the EM algorithm in [5] after 100 iterations. We see that the Fourier-based estimate, which is computed in a much shorter time, is comparable in quality to the EM-based estimate.

Using the same distribution for X and Y, in Fig. 2 we show the unperturbed distribution using 100000 samples. Also shown is the estimated distribution from Fourier coefficients up to the 10th harmonic and refined using the EM algorithm for 50 iterations. To compare with the method in [5] where EM is used starting from a uniform distribution, we also plot the estimated distribution starting from a uniform distribution after 92 iterations of EM. The same stopping criteria for EM are used in both cases. We see that using the Fourier-based estimate as an initial guess to EM results in a superior estimate, i.e. a smaller information loss (Section 7), while requiring less iterations.

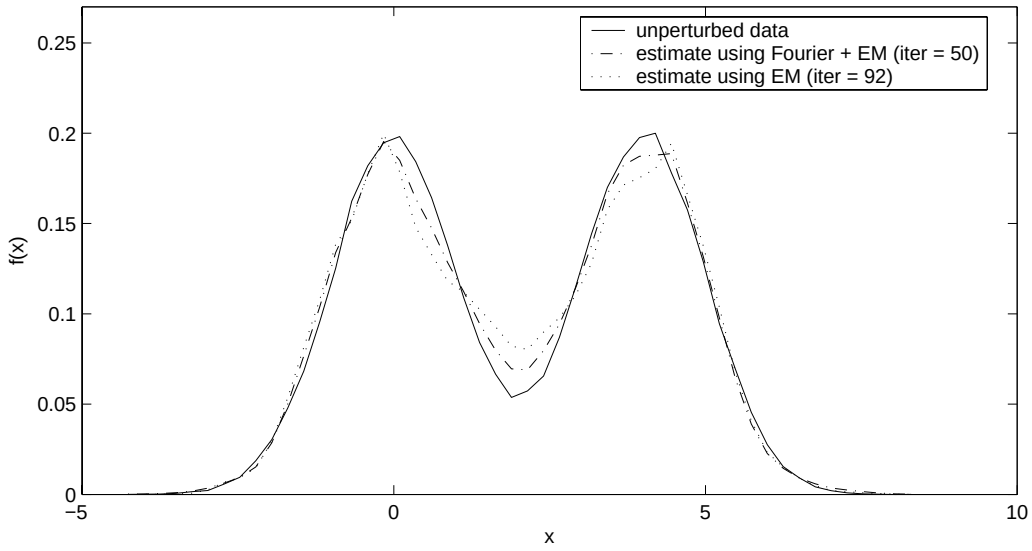

Figure 2: Plot of distribution of unperturbed data and the estimated distributions using 1) Fourier based method as an initial guess to EM and 2) using the uniform distribution as a initial guess to EM. The number of samples is 100000 samples. The perturbation Y is a uniform distribution in the range $[0, 20]$.

4 A data perturbation scheme with a one-step estimation algorithm

Even though the Fourier series based estimation method in Section 3 gives an estimate of the density function of X in a single step, its main drawback is that it does not estimate the coefficients well when the corresponding matrix A_i is close to singular. In this section, we modify the protocol of data perturbation, so that a simple one-step estimation algorithm exists which is guaranteed to approximate the unperturbed distribution given enough data samples. First note that one way to estimate $f_X(x)$ is to decompose S_X into k regions or bins, and count the number of times x_i lies in each region. Normalized by n and the measure of the regions, these numbers give estimates of $f_X(x)$ at the centers of these regions. We propose a data perturbation protocol which estimate f_X in this way, except that the count is perturbed. Let us denote the regions as $R_1, \ldots R_k$, and the j-th characteristic function of x as $\chi_j(x) = 1$ if and only if $x \in R_j$ and 0 otherwise. Then the estimate of $f_X(m_j)$ where m_j is the center of R_j is simply $\frac{1}{n}\sum_{i=1}^{n} \chi_j(x_i)/m(R_j)$ where $m(R_j)$ is the measure of R_j. Without loss of generality, let us assume that $m(R_j) = 1$.We call $\chi(x) = (\chi_1(x), \ldots, \chi_k(x))$ the *characteristic vector* of x. We assume that the regions are non-overlapping and cover S_X. This

implies that $\chi(x)$ is a unit coordinate vector, a vector with zeros and a single 1. One can view $\chi(x)$ as a quantization of x, with the quantization becoming finer as the number of bins increases. We construct y_i as an k-vector of iid samples from a distribution W with mean μ. The vector z_i is defined as the k-vector $z_i = \chi(x_i) + y_i$. It's intuitive that by making the variance of W large, the loss of privacy, i.e. the ability to reconstruct x_i from z_i can be made small.

The estimation algorithm is very simple and its correctness easily proved. It essentially consists of calculating the average of z_i's and subtracting the mean μ. Let z_{ij} be the j-th component of the k-vector z_i. Then $\frac{1}{n}\sum_{i=1}^{n} z_{ij} - \mu = \frac{1}{n}\sum_{i=1}^{n} \chi_j(x_i) + \frac{1}{n}\sum_{i=1}^{n} y_{ij} - \mu$ which converges to the correct estimate $\frac{1}{n}\sum_{i=1}^{n} \chi_j(x_i)$ as $n \to \infty$ by the law of large numbers. In other words, the information loss is arbitrarily small for large enough n. To force the estimate of the j-th bin to be nonnegative, we set it equal to $\max(0, \frac{1}{n}\sum_{i=1}^{n} z_{ij} - \mu)$.

5 Choosing the distribution of the perturbation W

What are some of the requirements in choosing the distribution of the perturbation random variable W? A slightly simpler estimation algorithm can be obtained by requiring W to have zero mean ($\mu = 0$). This can be obtained by subtracting the mean from the samples of W. On the other hand, as we will see later, it is sometimes helpful to allow a nonzero mean in order to efficiently store and transmit the vectors z_i. To be able to mask the characteristic vector of X, $p(W > m)$ should be non-vanishing for arbitrarily large m. An obvious choice would be the normal distribution. In this case, each z_i is a k-vector of real numbers and for large k, the amount of information in the vector z_i which needs to be stored and transmitted can be too large. Therefore it is better to choose W to be a discrete random variable. This way z_i can be represented as a k-vector of integers rather than a k-vector of reals. For example, we can choose W as $W = \gamma\text{round}(N)$, where N is a normal distribution with zero mean and variance σ. We will call this the discrete normal distribution. Other possible distributions for W include the Poisson distribution $p_W(\gamma x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for x a nonnegative integer or the discrete random variable with probability distribution $p_W(\gamma x) = \frac{6}{(2\pi^2-6)(|x|+1)^2}$ for x an integer (note that this distribution does not have a mean). Here γ denotes the strength of the perturbation. The higher γ is, the stronger the perturbation is, and the smaller the privacy loss is. To ensure that z_i looks like a sample from Y, i.e. $z_i \in S_Y$, we choose γ to be the inverse of a positive integer.

6 Quantifying privacy loss

One way to quantify the loss of privacy is to calculate the mutual information between Z and X [5]. Another way to quantify the loss of privacy is by defining it as the probability in which an estimate from a single sample z_i is correct assuming we know the probability distribution of both X and W.

We show here two cases where the privacy loss approaches 0 as the number of bins increases. First we consider a distribution for W for which we can easily calculate the maximum likelihood estimate (MLE). Because $\chi(x)$ is a many-to-one map, the MLE cannot distinguish between x_i's lying in the same region R_j. Therefore we will estimate x_i by finding the region it belongs to. Let us define $S_n = \sum_{k=1}^{\infty} \frac{1}{k^n}$ which exists for $n > 1$. Consider the following mass function for $W : p_W(k) = \frac{1}{S_3 k^3}, k > 0$. The mean is $\frac{S_2}{S_3} \approx 1.368$. The MLE given a sample z_i is R_{j^*} where $j^* = \arg\min_j \{z_{ij} | z_{ij} \geq 2\}$. It can be shown that the probability that the MLE is incorrect approaches 1 as the number of bins increases, i.e. the loss of privacy approaches 0.

Let $S(a) = \sum_{i=0}^{\infty} \frac{1}{a^i} = \frac{a}{a-1}$ for $a > 1$. Consider the following distribution for W: $p_W(k) = \frac{1}{S(a)a^k} = \frac{a-1}{a^{k+1}}$ for $k = 0, 1, \ldots$. For this distribution, we can easily calculate the maximum a posterior estimate (MAP). The MAP estimate given a sample z_i is R_{j^*} where $j^* = \arg\max_j \{p_X(R_j) | z_{ij} \geq 1\}$. For continuous random variables X, it is easy to see that the probability the MAP estimate is incorrect approaches 1 as the number of bins increases.

7 Information loss metric

By adding perturbation, we lose precision in estimating f_X, the density function of X. In [5] this is referred to as information loss and a metric is proposed to quantify this. The metric is defined as half the expected value of the L_1 norm between the original distribution of X and the estimated distribution, which we will denote by I. Note that this metric depends on k, the number of bins used in estimating the distribution. This metric is zero for perfect reconstruction. However, the information loss metric should measure the additional decrease in precision in estimating p_X due to the perturbation Y, and should be 0 when the perturbation Y is zero. This is not the case for the metric I in [5] since even with the unperturbed samples there is imprecision in estimating p_X, especially when the number of samples n is small. One way to remedy this is to subtract from I the value of I when the perturbation is 0. Therefore, we propose to measure the information loss by one half of the expected value of the L_1 norm between the histogram of x_i and the reconstructed

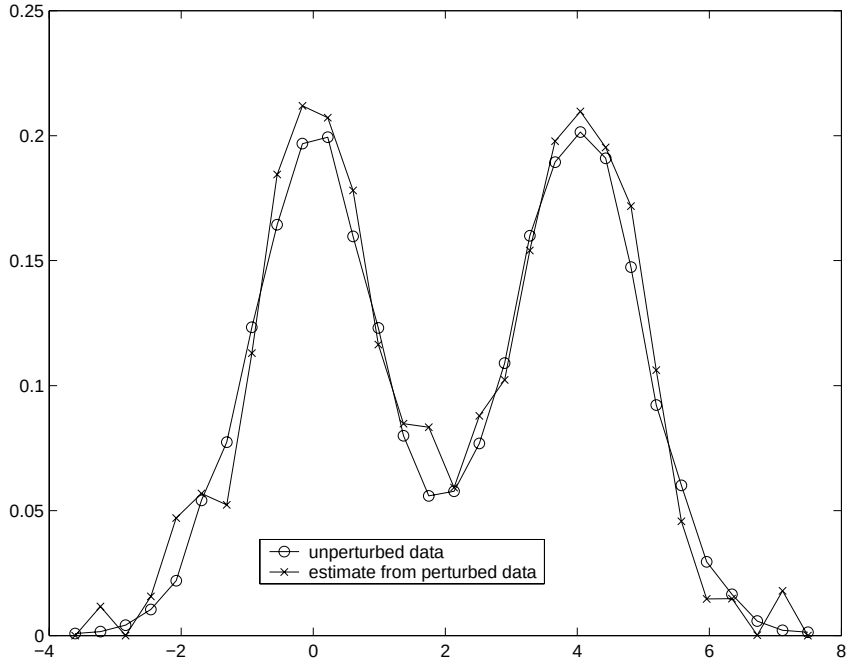

Figure 3: Plot of estimated distribution and of the distribution of the unperturbed data points. $n = 10000$ and $k = 30$. The perturbation is the discrete normal with $\gamma = 0.5$, $\sigma = 1$.

histogram. This metric also depends on k. For this metric, its value is 0 for perfectly reconstructing the histogram of x_i, i.e. its value is 0 when $Y = 0$.

In Section 4 the function $\chi(x)$ classifies points into bins and is a many-to-one function and there is some loss of information. On the other hand, calculating the histogram of the unperturbed data or estimating the distribution using EM also require classifying points into bins and thus also incur some loss of information. If the sets of bins used in these algorithms are the same then it is clear from the discussion in Section 4 that the information loss approaches 0 as the number of samples n increases.

8 Examples

We run the algorithm in Section 4 on 10000 data points from the bimodal distribution in Section 3. The perturbation W is the discrete normal with $\gamma = 0.5$, $\sigma = 1$. Fig. 3 shows a plot of the estimated distribution and the distribution of the unperturbed data points. The number of bins is $k = 30$.

In Fig. 4 we show the decrease in the proposed information loss metric as the number of samples

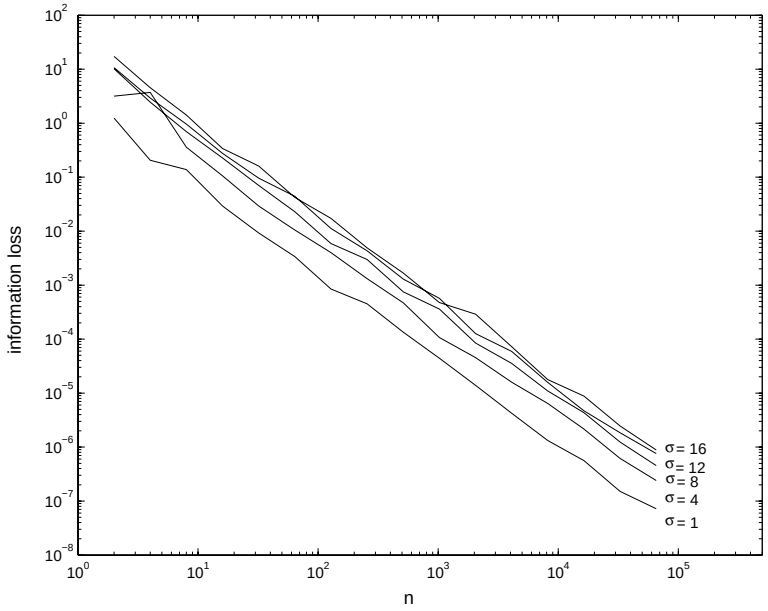

Figure 4: Decrease of information loss as n increases. The perturbation is a discrete normal with $\gamma = 1$.

n increases for various σ of the perturbation W. The perturbation W is a discrete normal with $\gamma = 1$.

9 Conclusions

We present two methods to speed up the estimation of the density function of the data in privacy preserving data mining applications where the data is perturbed by samples from a known distribution. In the first method, we estimate the Fourier coefficients of the density functions in order to obtain an estimate in one step. This estimate can be further refined by iterative methods such as EM. In the second method, we propose a novel privacy preserving data mining scheme where the density function of the original data set can be estimated using a simple one-step algorithm, compared with more complicated iterative procedures that have been proposed in the past. The presentation in Section 4 deals primarily with numerical data (i.e. x_i are real numbers), but categorical data can be handled by considering each region as a separate category. Indeed, categorical data is more natural in this setting than numerical data. One drawback of the proposed scheme in dealing with numerical data versus the scheme in [4] is that when the number of regions k is large, the number of bits needed to store/transmit the vectors z_i in the proposed scheme is also

large. In practice this is not a serious drawback, since in many cases, k is quite small (for instance, a person's salary is usually categorized into a few salary ranges or a preference is ranked on a scale of 1 to 5). Furthermore, because the estimation algorithm is essentially a summation of the vectors z_i's, only a running total of the z_i is needed in storage at any one time. This is in contrast to the iterative algorithms in [4, 5] where all the vectors z_i's are needed at each iteration.

References

[1] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Advances in Cryptology-crypto 2000*, 2000.

[2] W. Du and M. J. Atallah, "Privacy-preserving cooperative scientific computation," in *14th IEEE Computer Security Foundations Workshop*, 2001.

[3] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," in *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2002.

[4] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 439–450, ACM Press, May 2000.

[5] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Symposium on Principles of Database Systems*, 2001.

[6] N. N. Čencov, "Evaluation of an unknown distribution density from observations," *Soviet Math.*, vol. 3, pp. 1559–1562, 1962.

[7] S. C. Schwartz, "Estimation of probability density by an orthogonal series," *Annals of Mathematical Statistics*, vol. 38, pp. 1261–1265, 1967.

[8] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability, Chapman and Hall, 1986.