Privacy preserving data mining: a signal processing perspective and a simple data perturbation protocol

Chai Wah Wu
IBM Research Division, Thomas J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598, U.S.A.
E-mail: chaiwah@watson.ibm.com

Abstract

Privacy concerns over the proliferation of gathering of personal information by various institutions over the internet led to the development of data mining algorithms that preserve the privacy of those whose personal data are collected and analyzed. A novel approach to such privacy preserving data mining algorithms was proposed where the individual datum in a data set is perturbed by adding a random value from a known distribution. In these applications, the distribution of the original data set is important and estimating it is one of the goals of the data mining algorithm. This distribution is estimated via an iterative algorithm such as the Expectation Maximization (EM) algorithm which was shown to have desirable properties such as low privacy loss and high fidelity estimates of the distribution. Each iteration of EM requires computation that is proportional to the size of the data set and can require large computation time to estimate the distribution. In this paper we propose two ways to reduce the amount of computation. First, we show that the problem can be recast as a deconvolution problem and signal processing algorithms can be applied to solve this problem. In particular we consider both a direct method and iterative methods which are more robust against noise and ill-conditioning. We show that the Richardson-Lucy deblurring algorithm is equivalent to EM after quantization. The signal processing approach also shows how the choice of perturbation affects information loss and privacy loss and allows us to clarify some points made in the literature.

In the second part of this paper, we propose a scheme for perturbing data which also has the nice properties of arbitrarily small privacy loss and arbitrarily high fidelity in the estimate. The main advantage of the proposed scheme is the simplicity of the estimation algorithm. In contrast to iterative algorithms such as EM, the proposed scheme 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.

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 [3, 4, 6] 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 operations on this data. An example of this is a survey that is conducted over the internet. In [2], 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 [1], 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, [1] 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 is to derive efficient 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 [1], using a convergence result of EM, it was claimed that the information loss is arbitrarily small when the data set is large. We will show that information loss cannot always be made small.

The EM algorithm is iterative in nature and the entire data set is used in each iteration. 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 approaches to reduce the amount of computation in the estimation algorithm. In the first approach, the problem is studied from a signal processing viewpoint and algorithms are proposed to reduce the computation in the original protocol of perturbation. In particular, the problem is recast as a deconvolution or deblurring problem. We present a Fourier series based method to compute in one step a good initial estimate of the distribution in order to reduce the number of iteration or eliminate the iterative step completely (Section 3.1). This can be considered a form of direct convolution and we also propose iterative deconvolution algorithms to compute estimates (Section 3.2). These methods also provide a signal processing perspective to the choice of the perturbation in order to minimize information loss and privacy loss which allows us to explain some of the experimental findings in [1]. We also show that the algorithm in [2] which corresponds to quantization of the perturbed data before applying the EM algorithm is equivalent to the Richardson-Lucy deblurring algorithm (Section 3.4). By casting the problem as a deblurring problem, regularization procedures useful in image deblurring can be used here to avoid numerical instability when the problem is ill-posed.

For the second approach, we modify the protocol of data perturbation to move beyond simple perturbation via additive noise and propose a novel privacy preserving scheme for data mining applications which also has arbitrarily small privacy loss and information loss when the data set is large (Section 4). The main advantage of the 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 [1] which we believe is more accurate in quantifying 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 z_i 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 computer server for collection. In $[1,2]$ 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 The case of additive noise ($g(x, y) = x + y$)

Consider the case as in $[1,2]$ where x_i and y_i are real numbers and $z_i = x_i + y_i$. First let us restate the EM algorithm proposed in [1] for solving this case. The estimated distribution is a piecewise-constant function of the form: $\sum_{i=1}^{K} \theta_i I_{R_i}$ where R_i are fixed non-overlapping intervals partitioning the support of the distribution of X and I_{R_i} is the indicator function of R_i. The measure of R_i is m_i. The determination of the parameters θ_i is computed via the following iterative algorithm:

1. Initialize $\theta_i^0 = \frac{1}{K}$.
2. $\theta_i^{l+1} = \frac{\theta_i^l}{m_i n} \sum_{j=1}^{n} \frac{Pr(Y \in z_j - R_i)}{\sum_{i=1}^{K} \theta_i^l Pr(Y \in z_j - R_i)}$
3. $l = l + 1$
4. If termination criterion is not met, return to Step 2.

3.1 Calculating the Fourier coefficients of f_X

In this section we show how the Fourier coefficients of f_X can be computed in one step to generate an initial estimate of f_X. Estimating the Fourier coefficients of f_X belongs to the class of orthogonal series estimators $[10, 13]$. This estimate can be further refined using the iterative methods of $[1,2]$. When the one-step estimate is close to f_X, the number of iterations needed in the refinement step is smaller than in [1, 2] (see example below).

Assume that the data are 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 $[0, 1]$: $f_X(x) = a_0 + \sum_{i=1}^{\infty} a_i \sin(2\pi ix) + \sum_{i=1}^{\infty} b_i \cos(2\pi ix)$. Since f_X is a probability density function, $a_0 = 1$. As $\sqrt{2}\sin(2\pi ix)$ and $\sqrt{2}\cos(2\pi ix)$ are orthonormal in $f([0, 1])$, it follows that $a_i = 2E(\sin(2\pi iX))$ and $b_i = 2E(\cos(2\pi iX))$. Because of the independence of X and Y,

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

Therefore

$$\begin{aligned}
\begin{pmatrix} a_i \\ b_i \end{pmatrix} &= 2\begin{pmatrix} E(\sin(2\pi iX)) \\ E(\cos(2\pi iX)) \end{pmatrix} \\
&= 2A_i^{-1}\begin{pmatrix} E(\sin(2\pi iZ)) \\ E(\cos(2\pi iZ)) \end{pmatrix}
\end{aligned}$$

where

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

Since f_Y is known, A_i^{-1} can be computed in advance and $E(\sin(2\pi iZ))$ and $E(\cos(2\pi iZ))$ are estimated as $\frac{1}{n}\sum_{j=1}^{n}\sin(2\pi iz_j)$ and $\frac{1}{n}\sum_{j=1}^{n}\cos(2\pi iz_j)$ respectively. One of the restriction of this method is that it works well 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, i.e. has a small information loss, if f_Y has higher frequency components than f_X. We will see this restriction again in Section 3.2.

Note that the estimated density function is independent of the number of bins used in the iterative algorithms of [1, 2]. 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 [11].

To illustrate this estimation method, consider the following example where 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 (after rescaling) using Fourier coefficients a_i, b_i, $i = 1, 2, ..., 12$ and is forced to be non-negative. Also shown in Fig. 1 is the estimated distribution using the EM algorithm in [1] 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.

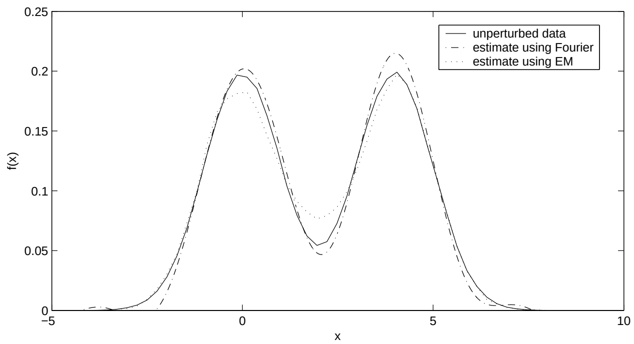

Figure 1. Plot of distribution of unperturbed data and the estimated distributions using 1) Fourier coefficients based method and 2) EM method. The number of samples is 500000. The perturbation Y is a uniform distribution in the range $[0, 20]$. The EM algorithm is run for 100 iterations.

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 [1] 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 (as defined in Section 5) while requiring less iterations.

3.2 Estimating f_X by iterative deconvolution

Since $Z = X + Y$, the density function of Z is the convolution of the density functions of X and Y, i.e. $f_Z = f_X * f_Y$. From a signal processing perspective, if we consider f_Y as a filter, the goal of reconstructing f_X from f_Z is a deconvolution or deblurring problem. Since the density function of Y is known, by estimating a density function of Z using the techniques in [11], we can estimate the density function of X by means of deconvolution or deblurring algorithms. The method in Section 3.1 can be considered as direct deconvolution by expanding f_X into a Fourier series. If f_Y corresponds to a low pass filter, then direct deconvolution means filtering f_Z with a high pass filter (the inverse of f_Y) to obtain f_X. As is well known in deblurring, this is problematic since noise is amplified by the high pass filter.

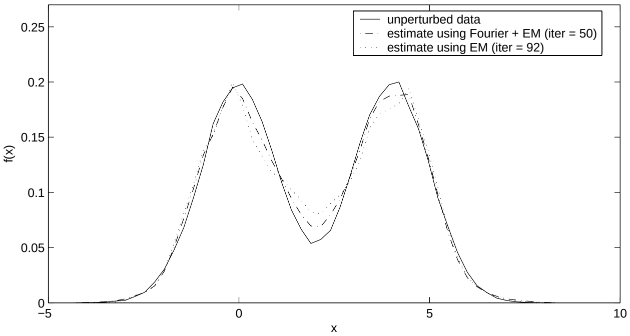

Figure 2. Plot of distribution of unperturbed data and the estimated distributions using 1) Fourier based method as an initial guess to EM and 2) using the uniform distribution as a initial guess to EM. The number of samples is 100000 samples. The perturbation Y is a uniform distribution in the range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

In our case, noise corresponds to inaccuracy in estimating f_Z due to the finite number of samples, and is larger as the number of samples n is smaller. This can be dealt with by limiting the magnitude of the frequency response of the inverse filter (which corresponds to limiting $\|A_i^{-1}\|$ in Section 3.1). Another solution is to use iterative deconvolution algorithms [5, 8].

3.3 Spectral properties of f_X and f_Y and information loss

Since the Fourier transform of f_Z (denoted as F_Z) is the multiplication of the Fourier transforms of f_X and f_Y (denoted as F_X and F_Y resp.), exact reconstruction of f_X is not possible (i.e. there is information loss) when $F_Y(\omega) = 0$ and $F_X(\omega) \neq 0$ for some frequency ω. In this case, regularization techniques need to be used to the deblurring algorithm to avoid ill-conditioning and instability in the reconstruction process [12] which occurs when $|F_X(\omega)| >> |F_Y(\omega)|$ for some frequencies ω. This usually comes with a cost of loss of resolution.

Therefore to achieve zero information loss, the support of F_X should be a subset of the support of F_Y. This is the same restriction encountered in Section 3.1. To minimize information loss, the support of F_Y should be large, whereas to minimize privacy loss, the support of f_Y should be large. In [1] it was shown that the EM algorithm constructs the MLE estimate of f_X and claims that this results in zero information loss. The above discussion shows that zero information loss is not always possible, i.e. reconstructing the MLE estimate does not always imply zero information loss. To illustrate, consider the following example where f_X is a uniform distribution on $[0, 25]$ and f_Y is a Gaussian distribution with $\mu = 0$, $\sigma = 1$, i.e. the perturbation is much smaller than the unperturbed data. Figure 3 shows the reconstructed distribution after 500 iterations of EM on 100000 samples with 100 bins which illustrates the numerical instability due to the ill-conditioning of the problem.

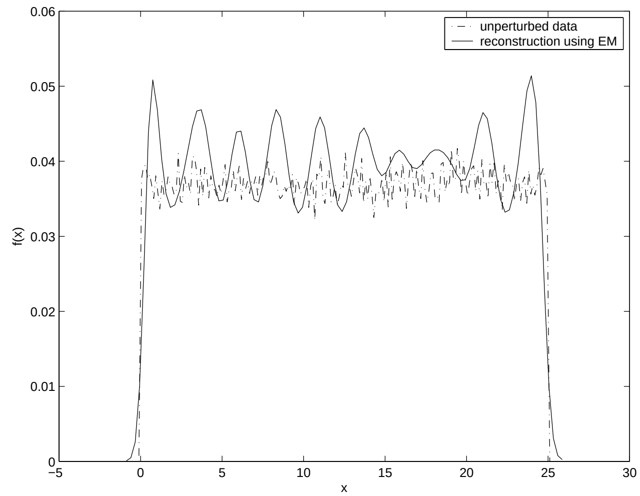

Figure 3. Plot of distribution of unperturbed data and the estimated distribution using EM. The number of samples is 100000. The distribution of the unperturbed data X is a uniform distribution in the range $[0, 25]$. The perturbation Y is a Gaussian distribution with variance 1. The EM algorithm is run for 500 iterations. We observe Gibbs-like phenomena due to ill-conditioning of the problem.

In [1] it was also observed via simulation that when f_X is Gaussian, the choice of whether f_Y is uniform or Gaussian has little effect on the information loss, whereas when f_X is uniform, the information loss for uniform f_Y is smaller than the information loss for Gaussian f_Y. This observation can be explained by noting that the the spectrum of a Gaussian distribution drops off faster than the spectrum of a uniform distribution. This means than it is easier to reconstruct f_X when f_Y is uniform and it is harder to reconstruct f_X when f_Y is Gaussian, especially when f_X is uniform.

3.4 Quantizing z_i and the Richardson-Lucy algorithm

The EM algorithm can be sped up by quantizing the perturbed values z_i. Let Δ_m, $m = 1, \ldots, M$ be non-overlapping intervals partitioning the support of Z and t_m be the center of the interval Δ_m. Let n_m be the number of samples z_i which lie in Δ_m. Quantizing z_i to t_m if and only if $z_i \in \Delta_m$ and then applying the EM algorithm results in the modified step 2:

$$\theta_i^{l+1} = \frac{\theta_i^l}{m_i n} \sum_{m=1}^{M} \frac{n_m Pr(Y \in t_m - R_i)}{\sum_{i=1}^{K} \theta_i^l Pr(Y \in t_m - R_i)}$$

By choosing M to be much smaller than and independent of n, the resulting algorithm, which was proposed in [2], executes much faster than the EM algorithm. It is easy to see that this modified EM algorithm is equivalent to the Richardson-Lucy algorithm [12] commonly used in deconvolution problems. In other words, by using the histogram of z_i as an estimator of f_Z and then applying the Richardson-Lucy deblurring algorithm, we obtain the modified EM algorithm in [2].

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

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 recall that the histogram method 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 these regions. We propose a data perturbation protocol which estimate f_X in this way, except that the count is perturbed. As before, we denote the regions as $R_1, \dots R_k$ with measures m_j and indicator functions I_{R_j}. Then the estimate of f_X is simply $\sum_j \theta_j I_{R_j}$ where $\theta_j = \frac{1}{m_j n} \sum_{i=1}^n I_{R_j}(x_i)$. Without loss of generality, let us assume that $m_j = 1$. We call $\chi(x) = (I_{R_1}(x), \dots, I_{R_k}(x))$ the *indicator 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 is 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 $\frac{1}{n}\sum_{i=1}^n z_{ij} - \mu = \frac{1}{n}\sum_{i=1}^n I_{R_j}(x_i) + \frac{1}{n}\sum_{i=1}^n y_{ij} - \mu$ which converges to the correct estimate $\theta_j = \frac{1}{n}\sum_{i=1}^n I_{R_j}(x_i)$ 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 $\max(0, \frac{1}{n}\sum_{i=1}^n z_{ij} - \mu)$. Note that because of the use of the indicator vector, this data perturbation algorithm is not equivalent to data perturbation via additive noise (Section 3).

4.1 Choosing the distribution of the random variable 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, 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 indicator vectors of the x_i's, $p(W > m)$ 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 $\gamma\text{round}(N)$, 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 $p_W(\gamma w) = \frac{\lambda^w e^{-\lambda}}{w!}$ for w a nonnegative integer or the discrete random variable with probability distribution $p_W(\gamma w) = \frac{6}{(2\pi^2-6)(|w|+1)^2}$ for w an integer[1]. 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 one cannot deduce x_i from z_i, i.e. z_i is such that different x_i's could have generated this z_i, we choose γ to be the inverse of a positive integer,

4.2 Quantifying privacy loss

One way to quantify the loss of privacy is to calculate the mutual information between Z and X [1]. 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 defined in this sense approaches 0 as the number of bins increases. First we consider a distribution for W for which we can eas-

[1] Note that this distribution does not have a mean, but by symmetry we can use $\mu = 0$ in the estimation algorithm.

ily 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.

5 Information loss metric

By adding perturbation, we lose precision in estimating f_X, the density function of X. In [1] this is referred to as information loss and a metric is proposed to quantify this. The metric, which we will denote as I, is defined as the expected value of the statistical difference (also called variation distance [9] or Kolmogorov distance [7]) between the original distribution of X and the estimated distribution. In other words, it is equal to half the expected value of the L_1 norm between the original distribution of X and the estimated distribution. 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 [1] 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 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 two calculations are the same then it is clear from the discussion in Section 4 that the information loss as defined here approaches 0 as the number of samples n increases.

We run the algorithm in Section 4 on 10000 data points from the bimodal distribution in Section 3.1. The perturbation W is the discrete normal with $\gamma = 0.5$, $\sigma = 1$. Fig. 4 shows a plot of the estimated distribution and the distribution of the unperturbed data points. The number of bins is $k = 30$.

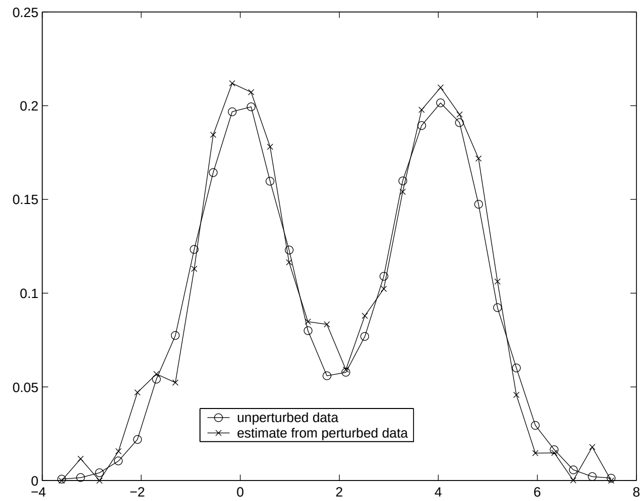

Figure 4. Plot of the estimated distribution and of the distribution of the unperturbed data points. $n = 10000$ and $k = 30$. The perturbation is the discrete normal with $\gamma = 0.5$, $\sigma = 1$.

In Fig. 5 we show the decrease in the proposed information loss metric as the number of samples n increases for various σ of the perturbation W. The perturbation W is a discrete normal with $\gamma = 1$.

6 Extensions

One way the problem in Section 2 can be generalized is by having each user generate an additional data value u_i that is used to index a parameterized family of perturbation random variables. In particular, y_i will be taken from the random variable Y_{u_i} with density function $f_{Y_{u_i}}(y) = f_Y(y, u_i)$. In this case u_i is sent along with z_i to the server. At the server, the density of the perturbation function is constructed as the average of the perturbations, i.e. $\frac{1}{n}\sum_{i=1}^{n} f_{Y_{u_i}}(y) = \frac{1}{n}\sum_{i=1}^{n} f_Y(y, u_i)$.

We have assumed in Section 4 that the regions R_i are disjoint. We can extend the method to the case where R_i

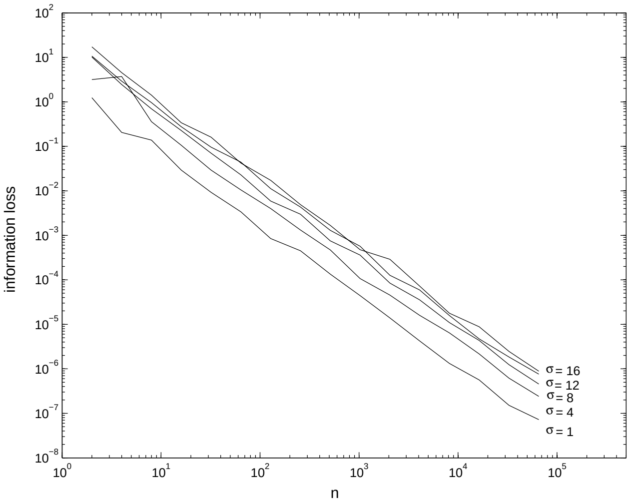

Figure 5. Decrease of information loss as n increases. The perturbation is a discrete normal with $\gamma = 1$.

are not necessarily disjoint in which case $\chi(x)$ are vectors with multiple 1's and 0's (i.e. they are not necessarily unit vectors). In this case, the data perturbation and estimation are performed as before. However, an extra step is needed to recover the histogram of x from $\frac{1}{n}\sum_i z_i - \mu$. If the regions R_i are chosen properly, this step can take the form of an simple linear matrix-vector multiplication. For instance, if R_i are the intervals $[i, i+2]$, $i = 1, \ldots m-1$ and $R_m = [m, m+1]$, then a histogram for X over the nonoverlapping regions $\tilde{R}_i = [i, i+1]$, $i = 1, \ldots, m$ can be obtained by multiplying the vector $\frac{1}{n}\sum_i z_i - \mu$ by the matrix

$$\begin{pmatrix} 1 & -1 & 1 & -1 & \ldots \\ & 1 & -1 & 1 & \ldots \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 \end{pmatrix}$$

7 Conclusions

We present two ways 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. First, we consider the problem as a deconvolution problem in order the solve it using deconvolution or deblurring algorithms. Second, 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, as opposed to 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 potential drawback of the proposed scheme in dealing with numerical data versus the scheme in [2] 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 EM algorithm where all the vectors z_i's are needed at each iteration.

Finally, we also propose metrics for quantifying privacy loss and information loss.

References

[1] 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.

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

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

[4] 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.

[5] J. S. Lim. *Two-dimensional signal and image processing*. Prentice Hall, 1990.

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

[7] M. Nielsen and I. Chuang. *Quantum computation and quantum information*. Cambridge University Press, 2000.

[8] M. Pruksch and F. Fleischmann. Positive iterative deconvolution with energy conservation. *Computers in Physics*, 12(2):182–189, 1998.

[9] A. Sahai and S. Vadhan. Manipulating statistical difference. In P. Pardalos, S. Rajasekaran, and J. Rolim, editors, *Randomization Methods in Algorithm Design (DIMACS Workshop, December 1997)*, volume 43 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 251–270. American Mathematical Society, 1999.

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

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

[12] D. L. Snyder, T. J. Schulz, and J. A. O'Sullivan. Deblurring subject to nonnegativity constraints. *IEEE Transactions on Signal Processing*, 40(5):1143–1150, 1992.

[13] N. N. Čencov. Evaluation of an unknown distribution density from observations. *Soviet Math.*, 3:1559–1562, 1962.