Dakshi Agrawal and Charu C. Aggarwal
T J Watson Research Center
IBM Corporation
Hawthorne, NY 10532
USA

On the Design and Quantification of Privacy Preserving Data Mining Algorithms

http://www.research.ibm.com/people/a/agrawal

Outline

- Problem Statement
- Quantification of Privacy
- Quantification of Information Loss
- EM Algorithm for Distribution Reconstruction
- Empirical Results
- Discussion

Motivation

Problem Statement:

Consider a set of n original data values $x_1, x_2, \ldots, x_n$, drawn independently according to the density function $f_X(x)$, and a set of n perturbation values $y_1, y_2, \ldots, y_n$, drawn independently according to the density function $f_Y(y)$. Given the perturbed values $z_1 = x_1 + y_1, z_2 = x_2 + y_2, \ldots, z_n = x_n + y_n$, and the density function $f_Y(y)$, estimate the density function $f_X(x)$.

Notes:

- Agrawal and Srikant provide an algorithm (AS algorithm) to estimate $f_X(x)$, and quantify its performance (privacy versus information-loss trade-off) by using heuristic metrics.

Quantification of Privacy

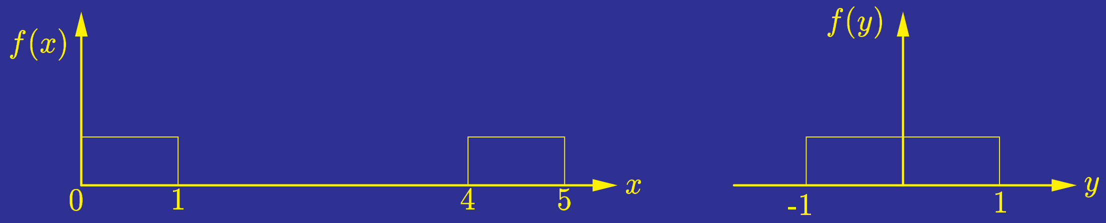
AS Metric of Privacy:

If it [an attribute] can be estimated with $c\%$ confidence that a value x lies in the interval $[x_1, x_2]i$, then the interval width $(x_2 - x_1)$ defines the amount of privacy at $c\%$ confidence level.

Distribution	50% Confidence	95% Confidence	99.9 % Confidence
Uniform (2α)	$0.5 \times (2\alpha)$	$0.95 \times (2\alpha)$	$0.999 \times (2\alpha)$

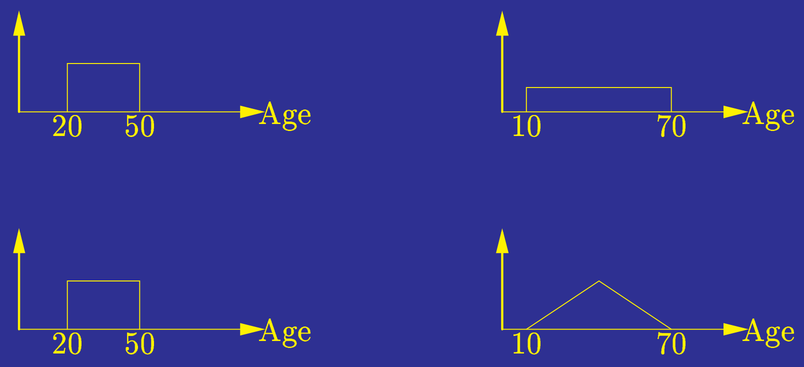
- **AS Metric is intuitive but not precise**

Quantification of Privacy
(Example)



- Privacy (AS Metric) = 2 at confidence level 100%
- However, it is easy to see that in the above example, the privacy is really 1 at confidence level 100%.
- AS metric of privacy does not depend on the data distribution $f(x)$.
- We need a more precise quantification of privacy.

Quantification of Privacy



$$h(A) = -\int_{\Omega_A} f_A(a) \log_2 f_A(a)\ da$$

- Differential entropy is not an intuitive measure. In particular, it could be negative.

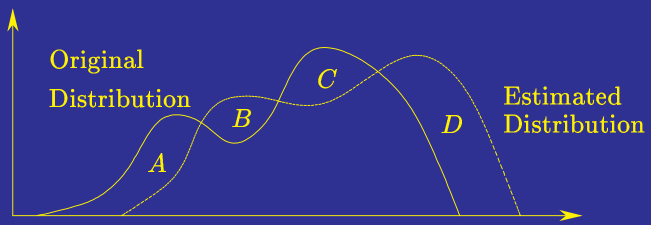
- $\Pi(A) = 2^{h(A)}$

Quantification of Privacy

$$
\begin{aligned}
h(A|B) &= -\int_{\Omega_{A,B}} f_{A,B}(a,b) \log_2 f_{A|B=b}(a)\ da\ db \\
\Pi(A|B) &= 2^{h(A|B)}
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
\mathcal{P}(A|B) &= \frac{\Pi(A) - \Pi(A|B)}{\Pi(A)} = 1 - 2^{h(A|B)}/2^{h(A)} \\
&= 1 - 2^{-I(A;B)},
\end{aligned} \tag{2}
$$

Quantification of Information Loss



$$\begin{aligned}\text{Information Loss} &= \text{half the sum of the mismatched areas}\\ &= 1 - \text{area shared by both distributions}\end{aligned}$$

$$\mathcal{I}(f_X, \widehat{f}_X) = \frac{1}{2} E\left[\int_{\Omega_X} \left| f_X(x) - \widehat{f}_X(x) \right| \, dx\right]$$

- $0 \leqslant \mathcal{I}(f_X, \widehat{f}_X) \leqslant 1$
- $\mathcal{I}(f_X, \widehat{f}_X) = 0$ implies perfect reconstruction of $f_X(x)$ and $\mathcal{I}(f_X, \widehat{f}_X) = 1$ implies that there is no overlap between $f_X(x)$ and its estimate $\widehat{f}_X(x)$

EM Algorithm

Properties

- EM algorithm converges to the Maximum-Likelihood Estimate (MLE).
- MLE is *consistent*.
- With a large number of data observations, the EM algorithm will provide zero information loss.

EM Algorithm for Distribution Reconstruction

1. E-step: Compute

$$Q(\Theta, \Theta^k) = E\Big[\ln f_{\mathbf{X};\Theta}(\mathbf{X}) \;\Big|\; \mathbf{Z} = \mathbf{z}; \Theta^k\Big]$$

2. M-step: Update

$$\Theta^{k+1} = \operatorname*{argmax}_{\Theta} Q(\Theta, \Theta^k)$$

EM Reconstruction Algorithm

1. Initialize $\theta_i^0 = \frac{1}{K}$, $i = 1, 2, \ldots, K$; $k = 0$;
2. Update Θ as follows:

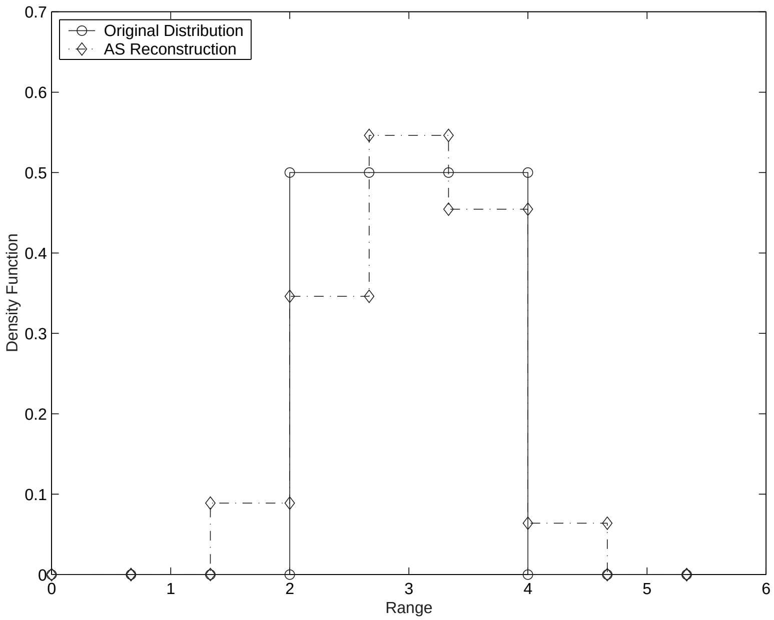
$$\theta_i^{(k+1)} = \frac{\theta_i^k \sum_{j=1}^N \frac{\Pr(Y \in z_j - \Omega_i)}{f_{Z;\Theta^k}(z_j)}}{m_i N};$$

3. $k = k + 1$;
4. If *not termination-criterion* then return to Step 2.

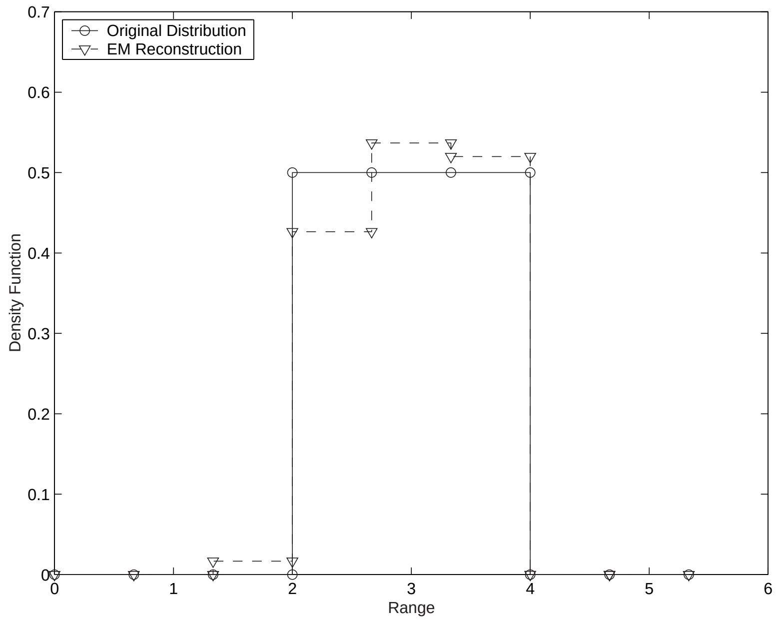
Similarities between AS and EM Algorithm

	AS Algorithm	EM Algorithm
Iteration	$\theta_i^{(k+1)} = \frac{\theta_i^k \sum_{j=1}^N \frac{\left[\mathrm{Pr}(Y \in [z_j] - \Omega_i)\right]}{\left[f_{Z;\Theta^k}([z_j])\right]}}{mN}$	$\theta_i^{(k+1)} = \frac{\theta_i^k \sum_{j=1}^N \frac{\mathrm{Pr}(Y \in z_j - \Omega_i)}{f_{Z;\Theta^k}(z_j)}}{m_i N}$
Approximations	3	1
Complexity	Lower	Higher

Empirical Results (500 data points)

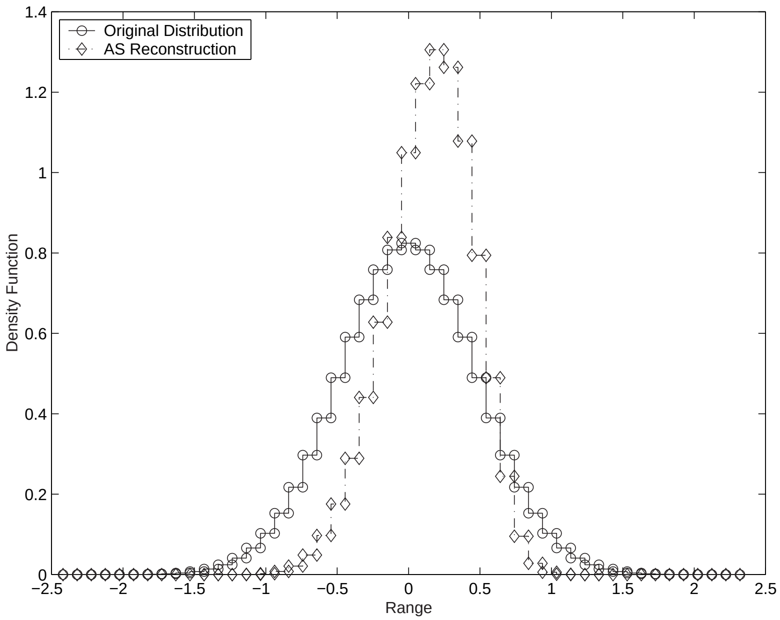


Reconstructed Uniform Distribution
(AS Algorithm 13.3% Info. Loss)

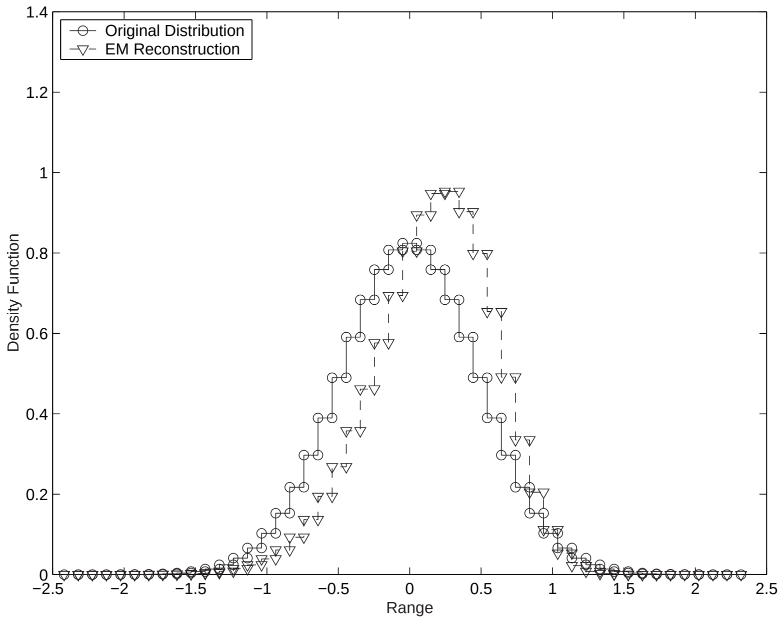


Reconstructed Uniform Distribution
(EM Algorithm 4.9% Info. Loss)

Empirical Results (500 data points)



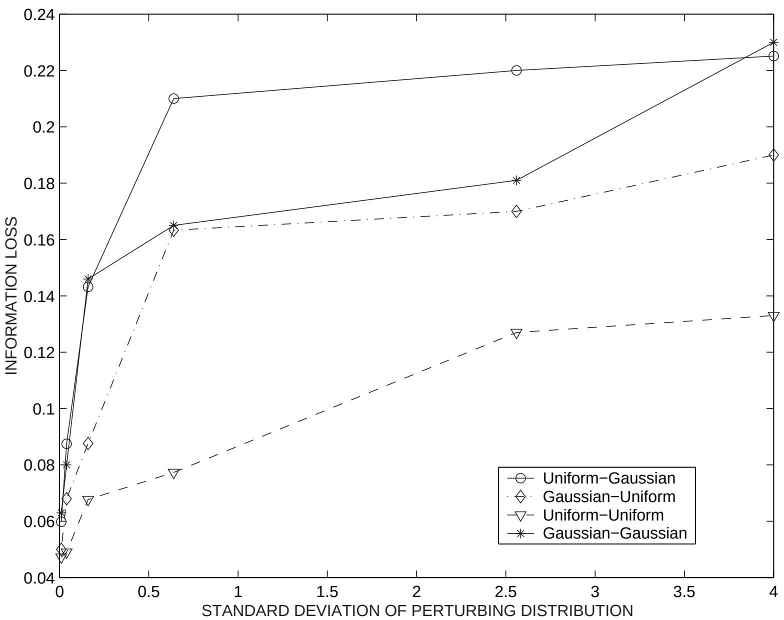
Reconstructed Gaussian Distribution (AS Algorithm 26.5% Info. Loss)



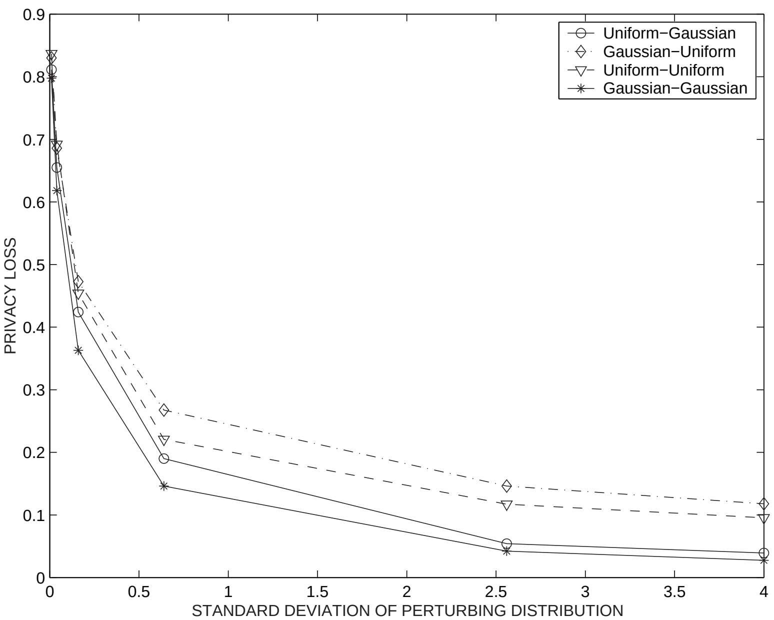
Reconstructed Gaussian Distribution (EM Algorithm 17.9% Info. Loss)

Empirical Results

trade-off between information and privacy loss

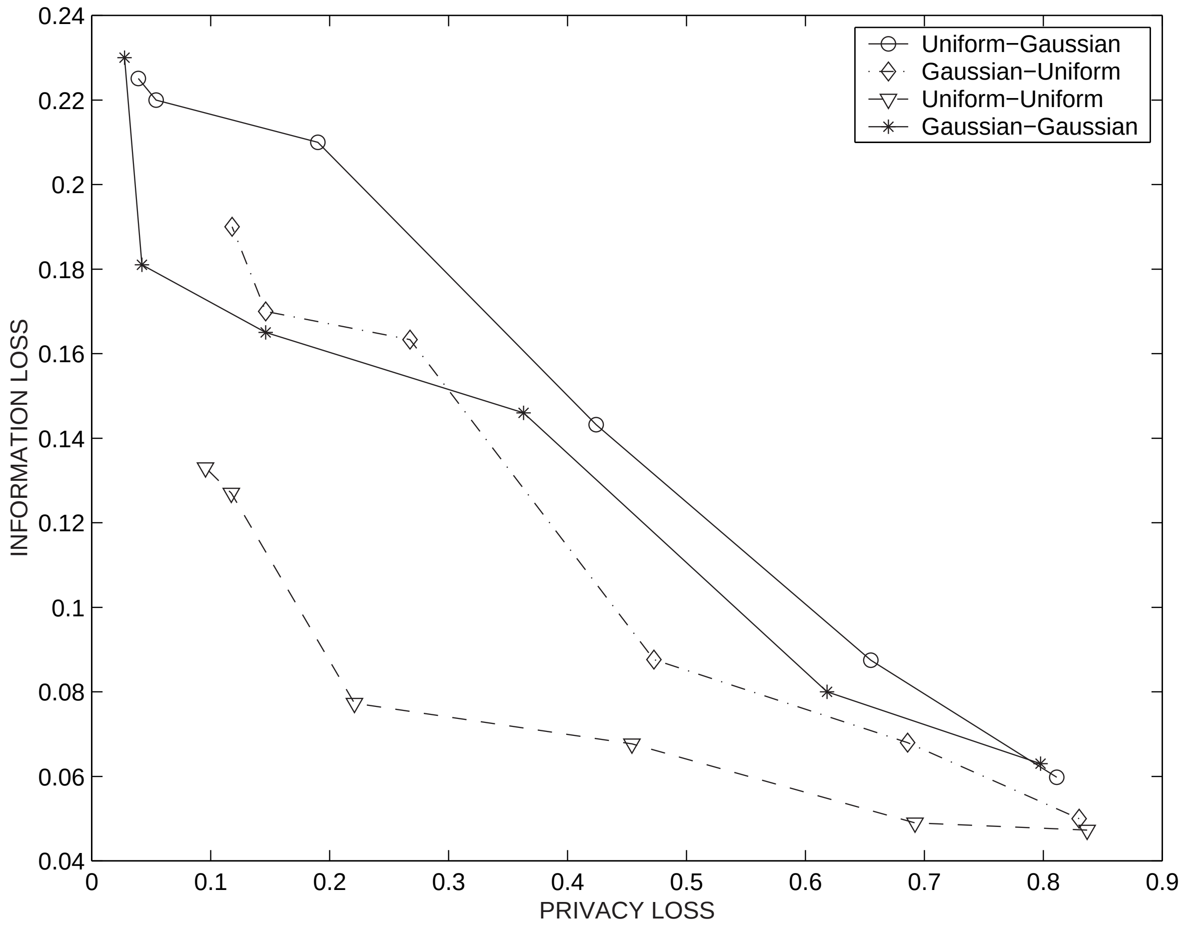


Information Loss with Increasing Perturbation



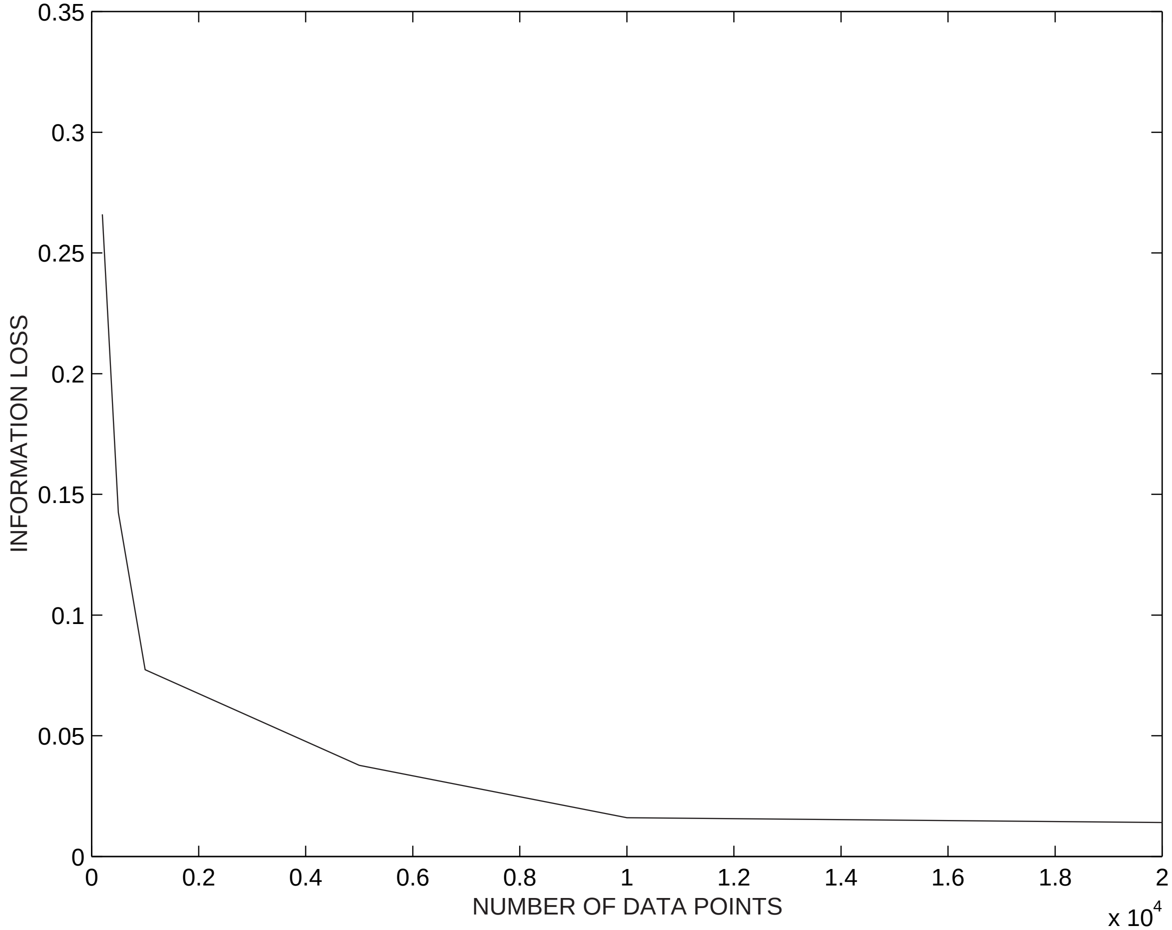
Privacy Loss with Increasing Perturbation

Empirical Results



The Tradeoff between
Information Loss and Privacy

Empirical Results



Information Loss with Number of Data Points (Constant Perturbation)

Conclusions

- Derived rigorous metrics for the quantification of privacy and information loss.
- These metrics are *universal* and they provide a sound foundation to compare privacy-preserving data mining algorithms.
- Qualified effectiveness of different perturbing distributions by using these metrics.
- The EM algorithm derived in this paper provably converges to the MLE.
- The estimate generated by EM algorithm results in very little loss for large data sets.