Reconstructing Data Perturbed by Random Projections When the Mixing Matrix Is Known[*]

Yingpeng Sang[1], Hong Shen[1], and Hui Tian[2]

[1] School of Computer Science, The University of Adelaide, SA 5005, Australia
[2] School of Mathematical Science, The University of Adelaide, SA 5005, Australia
{yingpeng.sang,hong.shen,hui.tian}@adelaide.edu.au

Abstract. Random Projection ($\mathcal{RP}$) has drawn great interest from the research of privacy-preserving data mining due to its high efficiency and security. It was proposed in [27] where the original data set composed of m attributes, is multiplied with a mixing matrix of dimensions $k \times m$ $(m > k)$ which is random and orthogonal on expectation, and then the k series of perturbed data are released for mining purposes. To our knowledge little work has been done from the view of the attacker, to reconstruct the original data to get some sensitive information, given the data perturbed by $\mathcal{RP}$ and some priori knowledge, e.g. the mixing matrix, the means and variances of the original data. In the case that the attributes of the original data are mutually independent and sparse, the reconstruction can be treated as a problem of Underdetermined Independent Component Analysis (UICA), but UICA has some permutation and scaling ambiguities. In this paper we propose a reconstruction framework based on UICA and also some techniques to reduce the ambiguities. The cases that the attributes of the original data are correlated and not sparse are also common in data mining. We also propose a reconstruction method for the typical case of Multivariate Gaussian Distribution, based on the method of Maximum A Posterior (MAP). Our experiments show that our reconstructions can achieve high recovery rates, and outperform the reconstructions based on Principle Component Analysis (PCA).

Keywords: Privacy-preserving Data Mining, Data Perturbation, Data Reconstruction, Underdetermined Independent Component Analysis, Maximum A Posteriori, Principle Component Analysis.

1 Introduction

Privacy-preserving Data Mining (PPDM) concerns the problems of completing data mining tasks without any direct access to the original data sets, because the providers claim privacy on their data. The general mining tasks include classification, clustering and association rule mining. PPDM can be treated as a subset within the problems of Secure Multi-party Computation (SMC) ([19], [20], [26], [40], etc). The cryptographic techniques from SMC provide solutions which always demand high computation cost,

[*] This work is partially supported by Australian Research Council Discovery Project grant #DP0985063.

W. Buntine et al. (Eds.): ECML PKDD 2009, Part II, LNAI 5782, pp. 334–349, 2009.

especially on processing volumes of data in data mining applications. Alternative approaches are based on data perturbation techniques which aim to be much more efficient than techniques of SMC.

Additive data perturbation, i.e. adding random data to the original data, was used in [3] to build decision tree classifiers, but in [17] and [22] random additive noise was questioned and pointed out that it can be easily filtered out, and thus lead to compromising of privacy. Multiplicative perturbation was used in [32] where the original data of each data provider is multiplied with the same matrix which is random and orthogonal before released, while in [28] this kind of perturbation is easily reconstructed by methods such as Principle Component Analysis (PCA), i.e. recovering the original data by analyzing the covariance matrix of the perturbed data.

In [27], an improved multiplicative data perturbation was proposed, in which the original data set X with m attributes is multiplied with a $k \times m$ $(m > 2k - 1)$ matrix R, each entry of which is an independently and identically distributed (i.i.d.) random number with the zero means. We name this method *Random Projection* ($\mathcal{RP}$) following [27] to avoid confusions with the method in [32]. The security claim of RX in $\mathcal{RP}$ is based on the structure of R and the fact that there does not exist a matrix T such that the product TR is a partition matrix and TRX is a separation of some attributes of X. However in the research field of Independent Component Analysis (ICA), the separation of m series of data $\hat{X}$ from k $(m > k)$ series of linearly mixed data RX is treated as the problem of Underdetermined ICA (UICA). Plenty of methods have proposed for UICA ([31]), and most of them are not seeking for the partition matrix T, but the possible values of X with maximum probability given only RX, and they have been successful in the case that the m original sources in X are mutually independent and sparse, except some permutation and scaling ambiguities.

In [8] and [16] reconstructions based on ICA were employed to attack the perturbation method of [32]. To our knowledge the only work on attacking the $\mathcal{RP}$ of [27] was proposed in [29] or [2] (Chapter 15), which was based on Maximum A Posterior (MAP). This attack only assumed the original data are uniformly distributed, but in practice many data properties can be assumed as normally (or approximately normally) distributed (e.g. personal heights, weights, financial variables), or are sparse enough to be modeled by the Laplace distribution (e.g. the voice or image data, financial data). It is not difficult for an attacker to obtain these priori knowledge on the original data, such as whether they are sparse, or whether they are normally distributed, given enough samples extracted from the same pool where the original data are extracted. It is also possible that the attacker may know the mixing matrix R by colluding with one data provider. Little work has been done to address these considerations.

In this paper, we will propose some attacking techniques on the $\mathcal{RP}$ method of [27] under some practical scenarios. We name the recovery of the original data by the attacker who is given the perturbed data as "*reconstruction*" following [3] and [17]. We also assume the attacker has a collusion with one of the data providers from which he can know the mixing matrix in $\mathcal{RP}$, the attacker has also obtained enough samples with identical distribution with the original data set, and thus some necessary priori knowledge on the original data, including whether the data attributes are mutually independent, whether they are sparse, their means and covariance matrix. Based on these assumptions we propose the following reconstruction methods from the view of the attacker:

1) If the attributes are mutually independent and sparse, we propose *Underdetermined Independent Component Analysis (UICA) based reconstruction* for the case that the attacker knows the mixing matrix, which outperforms the reconstruction based on PCA.
2) If the attributes are not mutually independent, where the ICA-based reconstruction will not be effective, we propose *Maximum A Posterior (MAP) based Reconstruction* for the case that the attacker knows the mixing matrix, and the original data following the Multivariate Gaussian Distribution. Our reconstruction outperforms the reconstruction based on PCA.

The organization of this paper is as following. In Section 2 we briefly review the related work. In Section 3 we give formal definitions on the problems of Data Perturbation and Data Reconstruction. In Section 4 we talk about how to obtain the necessary priori knowledge. In Section 5 and Section 6 we propose the ICA-based and MAP-based Reconstructions respectively. In Section 7 we conduct some experiments to evaluate our reconstruction methods, and compare them with reconstructions based on PCA. Section 8 concludes the paper.

2 Related Work

2.1 Data Perturbation

According to the taxonomy of [1], two families of approaches, *query restriction* and *data perturbation*, are usually used to provide statistical information (sum, count, average, etc) without compromising sensitive information about individuals. The query restriction family includes restricting the size of query result, controlling the overlap amongst successive queries, suppression of data cells, clustering entities into mutually exclusive atomic populations, etc. The data perturbation family includes the methods, loosely speaking, which perturb the original value of the data, X, into a random value Y.

The perturbation methods employed until now consist of data replacement ([1], [25], [23]), data swapping ([10], [13]), additive value distortion ([3], [22], [17]), and multiplicative value distortion ([32], [27]), random perturbation on categorical or boolean data ([12],[33], [4]), etc. k-anonymity ([35], [30], [24]) and sensitive rule hiding ([5], [34], [39]) have also been employed in PPDM. Details on these methods can be found in the given references. In this paper we only focus on the multiplicative value distortion (or multiplicative data perturbation).

Multiplicative Data Perturbation of [32]. Multiplicative data perturbation was used in [32] in which the original data set $X = (x_1, ..., x_m)'$ is multiplied by a random and orthogonal matrix R of dimensions $m \times m$, and perturbed into the set $U = RX$. Given $U_1 = RX_1$ and $U_2 = RX_2$, obviously $U_2' \cdot U_1 = X_2' \cdot X_1$, i.e. the inner products $X_2' \cdot X_1$ are the same as the inner products $U_2' \cdot U_1$. The distance-related metrics such as

Euclidean distance between X_1 and X_2 can be computed based on the inner products $U_2' \cdot U_1$, $U_1' \cdot U_1$, $U_2' \cdot U_2$, and data mining tasks can continue by these metrics without knowing the private X_1 and X_2.

Random Projection of [27]. In [27], a similar R is used in their multiplicative perturbation, but is different from [32] in that R is rectangle, which is $k \times m(m \geq 2k-1, m \geq 2)$ with i.i.d entries from $N(0, \sigma_r^2)$. Two data owners respectively compute $U = \frac{1}{\sqrt{k}\sigma_r} RX$ and $V = \frac{1}{\sqrt{k}\sigma_r} RY$, and then send them to the miner. Because $U'V = \frac{1}{k\sigma_r^2} X'R'RY$, and $E(R'R) = k\sigma_r^2 I$ (by Lemma 5.2 of [27]), then $E(U'V) = X'Y$. By this statistical result, the inner product $x \cdot y$ $(\forall x \in X, \forall y \in Y)$ can be computed without knowing X and Y.

The security of this $\mathcal{RP}$ method is based on the fact that if R is a rectangle $(k \times m)$ matrix and $m \geq 2k-1$, there does not exist a matrix T such that the product TR becomes a partition matrix which has at most one nonzero element in each column, and separates any single independent signal in TRX. Details on the partition matrix and signal separation can be referred to [7].

2.2 Reconstructions on Multiplicative Data Perturbation

Without the transformation matrix R, the recovery of X from U is infeasible given only the linear system $RX = U$. If in the $k \times m$ matrix R $k < m$, given R and $RX = U$, X can not also be determined from solving the linear system. However, the solutions can be sought from clues in some priori knowledge on X.

Reconstruction on the Perturbation of [32]. According to [28], if some input-output pairs "(x_i, u_i)" (i.e. $x_i \in X$, $u_i \in U$, and $u_i = Rx_i$), or some samples $x_i \in X$, are known apriori by an attacker, he may approximately recover X. Given some input-output pairs, the attacker can uniformly select an R to meet some criterion function, but this kind of attack can be prevented by deleting these pairs from the data owners, or de-identifying the output set U by randomly mixing the records $(u_1, ..., u_i, ...)$, before they are released for mining.

Another reconstruction of [28] is based on PCA, and does not requires the input-output pairs, but some general samples from the same pool where X is originated. By analyzing these samples, the attacker can estimate the covariance matrix of X, i.e. Σ_X. By analyzing the covariance matrix of the perturbation U, i.e. Σ_U, the attacker can obtain the information on R since $\Sigma_U = R\Sigma_X R'$. Specifically, suppose the eigenvalue decompositions of Σ_X and Σ_U are $Q_X E_X Q_X'$ and $Q_U E_U Q_U'$, then the diagonal matrices $E_X = E_U$, and the orthogonal matrices $Q_U = RQ_X$. When R is know in this way, the attacker can then have an estimate on X.

This attack based on PCA will not be effective on reconstructing the $\mathcal{RP}$ perturbation of [27], since in this method R is rectangle $(k < m)$, Σ_U is very similar as a diagonal matrix as proved by [11], and it may be difficult to make an eigenvalue decomposition on Σ_U. What's more, given Σ_U and Σ_X, it is difficult to obtain the information on the rectangle R from $\Sigma_U = R\Sigma_X R'$.

Attacks based on ICA were proposed in [2] (Chapter 15), [8] and [16], under various assumptions on the attacker's priori knowledge. They assumed an $m \times m$ mixing matrix R, thus were not suitable to attack the $\mathcal{RP}$ in [27].

Underdetermined Independent Component Analysis (UICA). Underdetermined (or overcomplete) independent component analysis (or blind source separation) has been under years of research in the field of signal processing, which addresses the problem that, X composed of m sources is linearly mixed by a $k \times m$ matrix R, and given only the mixed data RX, without knowing R, the sources are required to be separated out. A survey on the research can be referred to [31]. The methods of UICA have been mostly successful in the case that the m sources are mutually independent and sparse. They generally require two steps: 1) recovering the mixing matrix R, and 2) given R, recovering the sources. For Step 1) a lot of improvement work have been continuously done (e.g. [6], [24], [37] and [41]). For Step 2) L_1-norm minimization is the standard method that has been widely used.

L_1-norm minimization comes from the general approach of Maximum A Posterior (MAP). Considering $u = Rx$ and neglecting any additional noise, the probability of observing a vector x given R and a vector u is $p(x|R, u)$, and by Bayes Theorem

$$p(x|R, u) = \frac{p(u|R, x)p(x)}{p(u)}$$

By MAP x will be the vector in the Euclidean space of $\mathbb{R}^m$ that maximize the above equation. In the searching of this vector, $p(u)$ is a constant, $p(u|R, x)$ can be viewed as a constraint $Rx = u$, a sparse source x_i $(i = 1, ..., m)$ can be modeled by the Laplace distribution, i.e. $p(x_i) \propto e^{-|x_i|}$ (assuming they have zero means and identical variances). Then x will be the solution of the following constrained linear programming problem:

$$\begin{aligned} x &= \arg \max_{Rx=u} p(x) \\ &= \arg \max_{Rx=u} e^{-|x_1|-...-|x_m|} \\ &= \arg \min_{Rx=u} \sum_{i=1}^{m} |x_i| \end{aligned} \tag{1}$$

In Eq. (1) $\sum_{i=1}^{m} |x_i|$ is the L_1-norm of the vector $x = (x_1, ..., x_m)'$, therefore x will be the solution of the constrained L_1-norm minimization problem. There have been many methods to solve this problem and a survey of them can be referred to [9].

When the methods of UICA are used for reconstructions on the $\mathcal{RP}$-perturbed data, there is no additional noise N such as $U = RS + N$ and no need to reduce N, but they still have the following limitations:

1) When R is not known, these methods have permutation and scaling ambiguities. For each estimate of R, e.g. $\hat{R}$, there is infinite equivalent matrices $\tilde{R} = \hat{R}PL$ in which P is a permutation matrix of dimensions $m \times m$, L is a nonsingular diagonal matrix of dimensions $m \times m$ (scaling matrix), and $\tilde{R}$ is also an estimate of R. Thus the recovering of X will also have permutation and scaling ambiguities.

2) When R is known, there is no permutation ambiguity (the detailed reason is postponed to Section 5), but the means and variances of x_i should not be neglected in the constrained L_1-norm minimization of Eq. (1), since in practical scenarios their means may not be zero, and their variances may be not identical.
3) These methods generally require the original sources are mutually independent and sparse. In many scenarios of data mining, the attributes of the original data are correlated and not sparse. One typical model of these data is the Gaussian Mixture Model (GMM).

2.3 Disclosure Risk of the Distances

The risks of disclosing the mutual distances between data objects were investigated in [38]. They proposed two reconstruction methods based on the mutual distances of the data objects. The first one needs some known samples in their original forms and perturbed forms, so this method will not be effective when all perturbed data objects are mixed arbitrarily and de-identified before released, and an attacker can not find the corresponding perturbed data of the known samples. The second one needs no known sample, but makes a PCA on the perturbed data. However, this PCA will not be effective on analyzing the data perturbed by the method of [27], because R is rectangle, the covariance matrix of $U = RX$ is very similar as a diagonal matrix by [11], and it will be difficult to make a desired eigenvalue decomposition of the covariance matrix.

3 Problem Statement

3.1 Database Model

For convenience we consider a two-party case in which Alice and Bob share a distributed database. The reconstructions under the cases of more than two parties are similar as the two-party case, since in all the cases the parties use the same mixing matrix R for $\mathcal{RP}$. Suppose Alice and Bob have the data set X and Y respectively. Suppose the database has m attributes. If the database is horizontally distributed on the two parties, Alice has n_1 records, Bob has n_2 records, then X is an $m \times n_1$ matrix $[[x_{i,j}]_{i=1}^{m}]_{j=1}^{n_1}$, Y is an $m \times n_2$ matrix $[[y_{i,j}]_{i=1}^{m}]_{j=1}^{n_1}$.

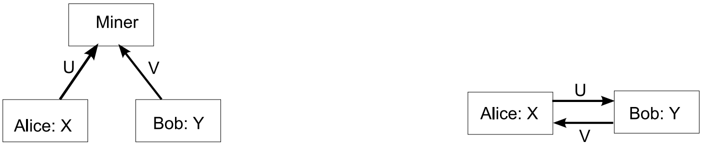
If the database is vertically distributed and the database has n records, Alice has m_1 attributes, Bob has m_2 attributes, then X is an $n \times m_1$ matrix $[[x_{j,i}]_{j=1}^{n}]_{i=1}^{m_1}$, Y is an $n \times m_2$ matrix $[[y_{j,i}]_{j=1}^{n}]_{i=1}^{m_2}$. In this paper we only focus on the horizontally distributed database. In the $\mathcal{RP}$ of [27] the vertically distributed databases are perturbed as similar as the horizontally distributed databases are perturbed, so our proposed reconstruction methods can be easily extended to the vertical cases.

3.2 Network Model

We consider two kinds of network models in this paper:

1) *Centralized model*: There is an independent miner who receives perturbed data from the data owners Alice and Bob (as in Fig. 1(a)). Scenarios are some companies under the investigation of a governmental organization. The companies are data providers, and the governmental organization wants to mine their private data.

2) *Distributed model*: There is no independent miner. All the data owners, Alice and Bob, act as miners on the perturbed data of their own and those received from the other party (as in Fig. 1(b)). Scenarios are some companies which want to share their data with each other to complete the data mining tasks. Each company is simultaneously a data provider and miner.



(a) Centralized Model

(b) Distributed Model

Fig. 1. Two Network Models for PPDM

3.3 Adversary Model

Adversary models have been theoretically defined in SMC ([14]), and also extensively used in PPDM. Depending on whether the participants merely gather information, or take active steps to disrupt the execution of the protocol, there are usually two types of adversaries:

1) *Semi-honest* participants, which are assumed to execute the solution exactly as what is prescribed, but may collude and analyze all the intermediate computations.
2) *Malicious* participants, which may arbitrarily deviate from the specified solution, e.g. generate arbitrary inputs, substitute the intermediate computations, or prematurely quit.

3.4 Problem Definition

Definition 1 - Privacy-preserving Data Mining based on Data Perturbation: *A database is distributed on two parties, Alice and Bob. The two providers respectively perturb their data matrices X and Y into U and V, and publish U, V to the miner. The miner may be an independent party, or replaced by Alice and Bob. All of them may be semi-honest or malicious. Privacy-preserving data ming on the miner should satisfy the following two requirements:*

1) *Privacy Requirement: No sensitive information on X and Y should be inferred from U and V by the miner.*
2) *Accuracy Requirement: The mining tasks including classification, clustering, etc, on U and V, should have statistically the same results as directly mining X and Y.*

It is worthy to note that in Definition 1 we do not specify the method of perturbing X and Y into U and V. Different perturbation methods possess different properties on privacy and accuracy, so the definition is made inclusive so as to cover as many perturbation methods as possible. In addition, to achieve the accuracy requirement, an accurate computation on the inner product of two vectors, such as $x'y(\forall x \in X, \forall y \in Y)$, is enough. Distance-related metrics like Euclidean distance, required in both the horizontally and vertically partitioned data mining, can be computed based on those inner products, the details of which can be referred to [27].

Definition 2 - Data Reconstruction: *An attacker obtains the perturbed data U and V from the data providers Alice and Bob. He wants to recover as many as possible the entries of X and Y. The attacker and any of the providers may be semi-honest or malicious.*

In the centralized model, if the miner and any of the providers are semi-honest, they may collude to reconstruct the data of the other providers. If the miner is malicious, he may not communicate the correct data mining results with the data providers.

In the distributed model, if one of the provider is semi-honest, he may reconstruct the data of another provider using R. If he is malicious, he may arbitrarily substitute his original data and publish them to another provider. Malicious attackers are not the focus of this paper.

Figure 2 shows the two mutually inverse processes, data perturbation and reconstruction.

X
Perturbation
U
U
Reconstruction
$\mathcal{X}$

(a) Data Perturbation by the owner Alice (b) Data Reconstruction by an adversarial miner

Fig. 2. Data Perturbation and Reconstruction for PPDM

Definition 3 - Recovery Rate: *Suppose $\hat{X}$ is a reconstruction of the original data X, $\hat{X} = [[\hat{x}_{i,j}]_{i=1}^{m}]_{j=1}^{n_1}$, and $X = [[x_{i,j}]_{i=1}^{m}]_{j=1}^{n_1}$. The Recovery Rate, $r(\hat{X}, \epsilon)$ with a given threshold ϵ, is the percentage of reconstructed entries whose relative errors are within ϵ, i.e.*

$$r(\hat{X}, \epsilon) = \frac{\#\{\hat{x}_{i,j} : |\frac{x_{i,j} - \hat{x}_{i,j}}{x_{i,j}}| \leq \epsilon, i = 1, ..., m, j = 1, ..., n_1\}}{m * n_1} \quad (2)$$

We will use the recovery rate in this definition to evaluate the performance of our reconstruction methods.

4 Obtaining the Priori Knowledge

In this section we discuss how the attacker can obtain the necessary priori knowledge on the original data, including their mean values, covariance matrix, whether they are mutually independent, under the condition that he has got enough samples, i.e. m-dimension vectors like $v = (v_1, ..., v_m)$, which are identically and independently selected from the multivariate distribution of the original data X, i.e. $p(X) = p(x_1, ..., x_m)$.

Given enough samples, the attacker can compute the *sample means* $\overline{X} = (\overline{x}_1, ..., \overline{x}_m)$ in which $\overline{x}_i$ is an estimate of x_i's mean u_i, and he can also compute the *sample covariance matrix* $\Sigma_X = [[cov(x_i, x_j)]_{i=1}^m]_{j=1}^m$ in which $cov(x_i, x_j)$ is an estimate of the covariance $E[(x_i - u_i)(x_j - u_j)]$.

When the estimated covariance matrix is diagonal, $x_1, ..., x_m$ are uncorrelated with each other. Uncorrelation is only a necessary condition of independence. In order to know the independence, the attacker should do a further test of *mutual independence* of the m attributes $(x_1, ..., x_m)$. One test method is to compute the mutual information I among the m attributes, i.e. $I(x_1, ..., x_m) = \sum_{i=1}^m H(x_i) - H(x_1, ..., x_m)$, in which $H(x_i)$ is the entropy of the i-th attribute x_i, $H(x_1, ..., x_m)$ is the joint entropy of the m attributes. I is zero if and only if the m attributes are statistically independent. There are also characteristic function-based and kernel-density based methods, which can be referred to [15] and [21].

When the attributes are mutually independent, the attacker can use statistical test, e.g. Kolmogorov-Smirnov Test, to check whether the values are following the Laplace distribution (i.e. sparse enough). When the attributes are not mutually independent, the attacker can employ some multivariate statistical test, e.g. the method of [36], to check whether the attributes are following the Multivariate Gaussian Distributions. As we discuss in Section 2.2, practically the original data may be in a Gaussian Mixture Model in which there are multiple clusters, but the attacker can easily identify some clusters given the perturbed data, and target his attacks on the original data belonging to these clusters.

In the later sections we assume the attacker has obtain all the necessary priori knowledge about whether the m attributes are sparse, whether they are in the Multivariate Gaussian Distribution, the means and covariance matrix. The reconstructions made by the attacker are summarized in Table 1.

Table 1. Types of Reconstructions

Priori Knowledge	*Reconstruction*
$I(x_1, ..., x_m) = 0$ & $p(x_i) = Laplace(\mu_i, \sigma_i)$	UICA-based
$I(x_1, ..., x_m) > 0$ & $p(X) = N(\mu, \Sigma)$	MAP-based

5 UICA-Based Reconstruction

As we have discussed in Section 2.2, UICA has permutation and scaling ambiguities in recovering X. However, in $\mathcal{RP}$ of [27], the data owner can not arbitrarily permute the rows in R and the resulting $U = RX$, before U is released, otherwise suppose P_1 and P_2 are two different permutation matrices of Alice and Bob respectively, $\mathcal{U} = P_1RX$, $\mathcal{V} = P_2RY$, then $\mathcal{V}'\mathcal{U}$ will not equal $Y'X$ on expectations. Therefore, when an attacker knows R and U, he will not have permutation ambiguity in the reconstruction of X.

The attacker still needs to reduce the scaling ambiguity with the priori knowledge on the mean μ_i and variance σ_i^2 of the i-th attribute x_i. In our reconstruction, we will use the same optimization function as Eq. (1), which assumes all x_i have zero means and identical variances. In order to do this, we firstly remove the means of all x_i before the use of Eq. (1), and afterwards add them again. We also change R to RL in which L is a scaling matrix whose diagonal entries are the variances, thus we can use Eq. (1) to obtain solutions with identical variances, and afterwards the solutions will be multiplied with the corresponding variances.

Specifically our UICA-based reconstruction includes the following steps:

1) *Remove the means:* Let μ_{u_i} is the sample mean of u_i, the i-th row of U $(i = 1, ..., k)$. $\mu_U = (\mu_{u_1}, ..., \mu_{u_k})'$. $\Theta = (1, ..., 1)_{n_1}$, then $\tilde{U} = U - \mu_U \Theta$.
2) *Change R to $\tilde{R}$:* $\tilde{R} = RL$, in which

$$L = \begin{pmatrix} \sigma_1 & 0 & ... & 0 \\ 0 & \sigma_2 & ... & 0 \\ ... \\ 0 & 0 & ... & \sigma_m \end{pmatrix}$$

3) *L_1-norm Minimization:* By the optimization function in Eq. (1), substitute $\tilde{R}$ into R, and each column of $\tilde{U}$ into u, search for the solution x of the function. Let $\tilde{X}$ be an $m \times n_1$ matrix, each column of it is the solution vector of the function corresponding to each column of $\tilde{U}$.
4) *Reduce the scaling ambiguity:* Let $\mu = (\mu_1, ..., \mu_m)'$ in which $\mu_i (i = 1, ..., m)$ is the sample mean of x_i, then the reconstruction is $\hat{X} = L\tilde{X} + \mu\Theta$.

6 MAP-Based Reconstruction

The UICA-based reconstruction is effective when the original data $x_1, ..., x_m$ are mutually independent and non-Gaussian. For the case that the m attributes are following the Multivariate Gaussian Distribution, we use the method of Maximum A Posterior (MAP) to estimate them. The basic idea of our MAP is similar as the constrained linear programming problem in Section 2.2, but the probability density function of the original data are different, and thus our MAP becomes a constrained *quadratic* programming problem.

6.1 Priori Knowledge

As same as UICA-based reconstruction, MAP method also requires sufficient samples from the multivariate distribution $p(x_1, ..., x_m)$, from which the means $\mu = (\mu_1, ..., \mu_m)'$, and the covariance matrix (i.e. Σ_X) of X, can be successfully estimated. By the definition of Multivariate Gaussian Distribution, Σ_X is positive definite, i.e. its eigenvalues are all positive.

6.2 Reconstruction under Collusion

Given $\mathbf{u} = (\mathbf{u}_1, ..., \mathbf{u}_k)'$ which is one column of U, the attacker can search a vector $\hat{\mathbf{x}} = (\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_m)'$ in $\mathbb{R}^m$ to maximize the posterior probability $p(\hat{\mathbf{x}}|\mathbf{u})$. Since $p(\hat{\mathbf{x}}|\mathbf{u}) = p(\hat{\mathbf{x}})p(\mathbf{u}|\hat{\mathbf{x}})/p(\mathbf{u})$, and under collusion the attacker will know R, then

$$p(\hat{\mathbf{x}}|\mathbf{u}) = \begin{cases} \frac{p(\hat{\mathbf{x}})}{p(\mathbf{u})}, & if\ R\hat{\mathbf{x}} = \mathbf{u}, \\ 0, & if\ R\hat{\mathbf{x}} \neq \mathbf{u}, \end{cases} \tag{3}$$

$p(\mathbf{u})$ can be treated as a constant in the search of $\hat{\mathbf{x}}$, then the maximization of $p(\hat{\mathbf{x}}|\mathbf{u})$ is equivalent to the following constrained optimization problem:

$$MAX_{\hat{x}}\ p(\hat{\mathbf{x}}),\ Subject\ to\ R\hat{\mathbf{x}} = \mathbf{u} \tag{4}$$

We assume $X \sim N(\mu, \Sigma_X)$, i.e. given a vector $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_m)' \in X$,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{m/2}|\Sigma_X|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma_X^{-1}(\mathbf{x}-\mu)}$$

Since the exponential function is a monotone one-to-one function, the problem in Eq. (4) is equivalent to the following Quadratic Programming (QP) problem:

$$MIN_{\hat{x}}\ f(\hat{\mathbf{x}}) = \frac{1}{2}(\hat{\mathbf{x}} - \mu)'\Sigma_X^{-1}(\hat{\mathbf{x}} - \mu),\ Subject\ to\ R\hat{\mathbf{x}} = \mathbf{u} \tag{5}$$

We have assumed the Σ_X is positive definite, so is Σ_X^{-1}, then $f(\hat{\mathbf{x}})$ is convex and has the unique global minimizer, which can be computed by the gradient of the Lagrange function:

$$L(\hat{\mathbf{x}}, \Lambda) = \frac{1}{2}(\hat{\mathbf{x}} - \mu)'\Sigma_X^{-1}(\hat{\mathbf{x}} - \mu) + \Lambda'(R\hat{\mathbf{x}} - \mathbf{u}) \tag{6}$$

in which $\Lambda = (\lambda_1, ..., \lambda_k)'$, λ_i $(i = 1, ..., k)$ are Lagrange multipliers.

By Eq. (6),

$$\frac{\partial L}{\partial \hat{\mathbf{x}}} = \left(\tfrac{\partial L}{\partial \hat{\mathbf{x}}_1}, ..., \tfrac{\partial L}{\partial \hat{\mathbf{x}}_m}\right)' = \Sigma_X^{-1}(\hat{\mathbf{x}} - \mu) + R'\Lambda = 0, \tag{7a}$$

$$\frac{\partial L}{\partial \Lambda} = \left(\tfrac{\partial L}{\partial \lambda_1}, ..., \tfrac{\partial L}{\partial \lambda_k}\right)' = R\hat{\mathbf{x}} - \mathbf{u} = 0 \tag{7b}$$

The $m + k$ equations in Eq. (7) can be treated as a linear system with $m + k$ variables $(\lambda_1, ..., \lambda_k, \hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_m)$:

$$R'\Lambda + \Sigma_X^{-1}\hat{\mathbf{x}} = \Sigma_X^{-1}\mu, \tag{8a}$$

$$0 \cdot \Lambda + R\hat{\mathbf{x}} = \mathbf{u} \tag{8b}$$

Let Θ_1, Θ_2 be $m \times k$ and $k \times k$ zero matrices, I be an $m \times m$ identity matrix, then by solving the above linear system,

$$\hat{\mathbf{x}} = (\Theta_1, I)\begin{pmatrix}\Lambda \\ \hat{\mathbf{x}}\end{pmatrix} = (\Theta_1, I)\,\Omega^{-1}\begin{pmatrix}\Sigma_X^{-1}\mu \\ \mathbf{u}\end{pmatrix},\ \Omega = \begin{pmatrix}R' & \Sigma_X^{-1} \\ \Theta_2 & R\end{pmatrix}. \tag{9}$$

Lemma 1. *Ω in Eq. (9) is nonsingular with high probability.*

Proof. By the Leibniz formula,

$$\det(\Omega) = \det(\Sigma_X^{-1})\det(\Theta_2 - R\Sigma_X R') = (-1)^k \det(\Sigma_X^{-1})\det(R\Sigma_X R').$$

$R\Sigma_X R' = \Sigma_U$, which is the covariance matrix of $U = RX$.

By [11], when $\mathbf{x}$ is fixed, R is a $k \times m$ matrix each entry of which is an i.i.d random number, then $\mathbf{u} = R\mathbf{x}$ is approximately Gaussian, following the distribution $N(R\mu, ||\mathbf{x}||^2 I_k)$ in which I_k is a $k \times k$ identity matrix. Therefore $\Sigma_U \approx ||\mathbf{x}||^2 I_k$, which means when $\mathbf{x}$ is not a zero vector, Σ_U will be nonsingular with high probability. Since $\det(\Sigma_X^{-1}) \neq 0$, then $\det(\Omega) \neq 0$, i.e. Ω is nonsingular. When Σ_U is singular, it is most possible that $\mathbf{x}$ is a zero vector. □

In sum, the MAP-based reconstruction includes the following steps:

1) estimate Σ_X and μ by enough samples from the same distribution as X;
2) compute $\hat{\mathbf{x}}_i$ by Eq. (9) for each column $\mathbf{u}_i$ of U, $i = 1, ..., n_1$. Let $\Theta_3 = (1, ..., 1)_{n_1}$, the reconstructed $\hat{X} = (\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_{n_1})$ can be written as:

$$\hat{X} = (\Theta_1, I)\begin{pmatrix} R' & \Sigma_X^{-1} \\ \Theta_2 & R \end{pmatrix}^{-1}\begin{pmatrix} \Gamma \\ U \end{pmatrix}, \ \Gamma = \Sigma_X^{-1}\mu\Theta_3. \tag{10}$$

7 Experiments and Comparisons

7.1 Reconstruction Based on Principle Component Analysis

As we sum in Section 2.2 the PCA-based attack of [28] is not suitable for the $\mathcal{RP}$ of [27], and to our knowledge, there is no PCA-based attack proposed for the $\mathcal{RP}$. For comparison purposes, we use the pre-whitening phase of ICA ([18]) as a PCA-based attack, which includes the following steps:

1) The attacker removes the mean of each row $u_i (i = 1, ..., k)$ of U.
2) The attacker computes the covariance matrix of $U = RX$ as $\Sigma_U = \mathbf{E}(UU')$, and makes an eigenvalue decomposition of it. Let $\Sigma_U = QDQ'$, in which Q is an orthogonal matrix, D is a diagonal matrix each entry of which is an eigenvalue of Σ_U.
3) The attacker computes $\tilde{X} = QD^{-1/2}Q'U$, in which $D^{-1/2}$ is a diagonal matrix each diagonal entry of which is the inverse of the square root of the corresponding entry of D. Let $A = QD^{-1/2}Q'$, then

$$\Sigma_{\tilde{X}} = A\Sigma_U A' = (QD^{-1/2}Q')(QDQ')(QD^{-1/2}Q') = I.$$

4) Suppose $\tilde{x}_j$ is the j-th row of $\tilde{X}$. For $i = 1, ..., m$, and $j = 1, ..., k$ the attacker computes:

$$\hat{x}_j = \sigma_i \tilde{x}_j + \mu_i, \tag{11}$$

and makes a statistical test $G(\hat{x}_j, p(x_i))$ in which $p(x_i)$ is the p.d.f of x_i, e.g. using the Two-sample K-S Test. If G outputs 1, $\hat{x}_j$ has a similar distribution to x_i, and the attacker treats $\hat{x}_j$ as an estimate of x_i.

In this method, $\Sigma_{\tilde{X}}$ is an identity matrix, so the k rows of $\tilde{X}$ will be uncorrelated. This reconstruction can be an approximate recovery when the m attributes of the original data are mutually independent, or they are not mutually independent and not having high correlations.

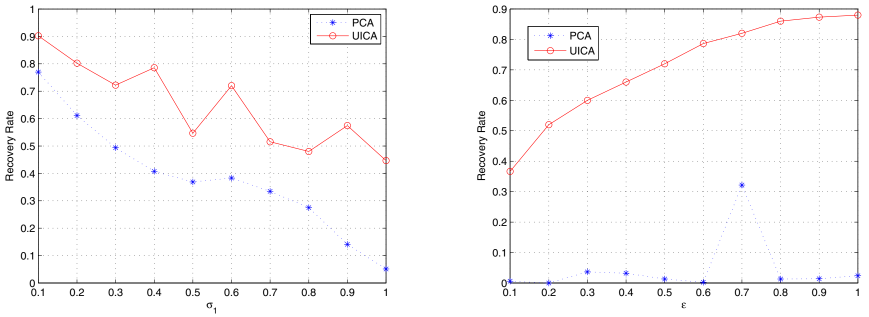
One major limitation of this method is that it can only recover k components, so it is essential for the attacker to use the priori knowledge to reduce the permutation and scaling ambiguities, as in Step 4). Another limitation is that Σ_U may be diagonal by [11], then in Step 2) Q will be an identity matrix, and in Step 3) the reconstruction result $\tilde{X}$ is simply $D^{-1/2}U$, i.e. some scaling of U.

7.2 Experiments and Comparisons for UICA-Based Reconstruction

We use the Laplace distribution to simulate 3 series of independent and sparse data, as the original data X. We generate a random R with dimensions 2×3 following the $\mathcal{RP}$ method of [27]. The means of X are $(0.3, 0.5, 0.8)$, the variances of X are changed to get different recovery rates measured by Definition 3 in Section 3.4. ϵ for the recovery rates are set to be 0.2. To search the solutions of the L_1-norm minimization problems, we use the *fmincon* function in the Optimization Toolbox of MATLAB.

In Fig. 3(a) the x-axis is the variance σ_1 of the first row x_1 of X, we make $\sigma_2 = 0.8\sigma_1$, $\sigma_3 = 0.3\sigma_1$. From Fig. 3(a) our reconstruction based on UICA achieves higher recovery rates than PCA.

3 series of financial data from the UCI Machine Learning Repository, including Attribute 1 of the Japanese Credit Screening Data Set, Attribute 1 of the Australian Credit Approval Data Set, Attribute 5 of the German Credit Data Set, are used as the original data, and perturbed by a random 2×3 matrix. They are treated as sparse data since they have high kurtosis (subtract 3), respectively 11.1, 1.17, 3.84. With different ϵ Fig. 3(b) shows the recovery rates of UICA-based and PCA-based reconstructions, and in comparison our UICA-based reconstruction performs much better than PCA.



(a) Synthetic data experiments

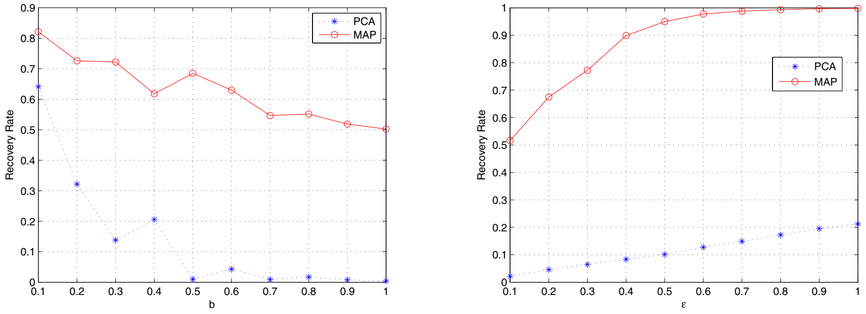
(b) Real data experiments

Fig. 3. Experiments on UICA-based and PCA-based Reconstructions

7.3 Experiments and Comparisons for MAP-Based Reconstruction

For the synthetic data experiments we assume $m = 10$, $\mu_i = 0.1 * i$ for $i = 1, ..., 10$. It is a non-trivial problem to generate Σ_X for the experiments which require variations on the structure of the covariance matrices. We use $\Sigma_X = AA'$ in which A is 10×10 matrix with i.i.d entries uniformly sampled from $[0, b]$, and b is changed (from 0.1 to 1) to get different Σ_X. We use the $mvnrnd$ function in MATLAB to generate X composed of 10 synthetic data attributes. R is a 4×10 matrix each entry of which follows $N(0, 1)$. Fig. 4(a) gives the recovery rates of our MAP-based reconstruction with different b, in comparisons with the recovery rates of PCA ($\epsilon = 0.2$). The figure shows that our method achieves higher recovery rates than PCA.

It is difficult to find real data strictly following the Multivariate Gaussian Distribution. We take the Attribute 2 ("Duration in month"), 13 ("Age in years"), 16 ("Number of existing credits at this bank") of the German Credit Data Set from the UCI Machine Learning Repository. They have $\mu = (20.9, 35.5, 1.4)$, $\Sigma_X = (145.4, -4.96, -0.08; -4.96, 129.4, 0.98; -0.08, 0.98, 0.33)$. They are perturbed by a 2×3 random R. Fig. 4(b) shows the recovery rates with different ϵ, and our MAP-based reconstruction performs much better than PCA.



(a) Synthetic data experiments

(b) Real data experiments

Fig. 4. Experiments on MAP-based and PCA-based Reconstructions

8 Conclusions

In this paper we propose two types of methods to reconstruct the original data from the data perturbed by Random Projection in [27]. Our reconstructions consider the case that the original data are mutually independent and sparse, and the case that the original data are not mutually independent and not sparse. Experiments show that our methods outperform the reconstructions based on PCA, and achieve higher recovery rates on the perturbed data. In the future work we will consider more reconstruction methods when R is not known, towards an improved perturbation method which is secure under these reconstructions.

References

1. Adam, N., Worthmann, J.: Security-control methods for statistical databases: a comparative study. ACM Computing Surveys 21(4), 515–556 (1989)
2. Aggarwal, C., Yu, P.S. (eds.): Privacy-Preserving Data Mining: Models and Algorithms. Springer, Heidelberg (2008)
3. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proc. of the 2000 ACM SIGMOD Conference on Management of Data, pp. 439–450. ACM, New York (2000)
4. Agrawal, S., Haritsa, J.R.: A Framework for High-Accuracy Privacy-Preserving Mining. In: Proc. 21st Int'l Conf. Data Eng. (ICDE 2005), pp. 193–204 (2005)
5. Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., Verykios, V.: Disclosure Limitation of Sensitive Rules. In: Proc. of IEEE Knowledge and Data Engineering Workshop, pp. C45–C52 (1999)
6. Bofill, P., Zibulevsky, M.: Underdetermined blind source separation using sparse representations. Signal Processing 81(11), 2353–2362 (2001)
7. Cao, X., Liu, R.: General Approach to Blind Source Separation. IEEE Transactions on Signal Processing 44(3), 562–571 (1996)
8. Chen, K., Sun, G., Liu, L.: Towards Attack-resilient Geometric Data Perturbation. In: Proceedings of the 2007 SIAM International Conference on Data Mining (SDM 2007), Minneapolis, MN (April 2007)
9. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic Decomposition by Basis Pursuit. SIAM Review 43(1), 129–159 (2001)
10. Dalenius, T., Reiss, S.P.: Data-swapping: A Technique for Disclosure Control. Journal of Statistical Planning and Inference 6, 73–85 (1982)
11. Dasgupta, S., Hsu, D., Verma, N.: A Concentration Theorem for Projections. In: Proc. the 22nd Conference in Uncertainty in Artificial Intelligence, pp. 1–17. AUAI Press (2006)
12. Evfimievski, A., Gehrke, J., Srikant, R.: Limiting privacy breaches in privacy preserving data mining. In: Proc. 22nd ACM Symposium on Principles of Database Systems (PODS 2003), pp. 211–222 (2003)
13. Fienberg, S.E., McIntyre, J.: Data Swapping: Variations on a Theme by Dalenius and Reiss. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 14–29. Springer, Heidelberg (2004)
14. Goldreich, O.: Foundations of Cryptography: Basic Applications, vol. 2. Cambridge University Press, Cambridge (2004)
15. Gretton, A., Fukumizu, K., Teo, C., Song, L., Scholkopf, B., Smola, A.: A Kernel Statistical Test of Independence. In: Proc. Advances in Neural Information Processing Systems (NIPS 2007), pp. 585–592. MIT Press, Cambridge (2007)
16. Guo, S., Wu, X.: Deriving private information from arbitrarily projected data. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 84–95. Springer, Heidelberg (2007)
17. Huang, Z., Du, W., Chen, B.: Deriving Private Information from Randomized Data. In: SIGMOD 2005, pp. 37–48. ACM, New York (2005)
18. Hyvärinen, A., Oja, E.: Independent Component Analysis: Algorithms and Applications. Neural Networks 13, 411–430 (2000)
19. Jha, S., Kruger, L., McDaniel, P.: Privacy Preserving Clustering. In: de di Vimercati, S.C., Syverson, P.F., Gollmann, D. (eds.) ESORICS 2005. LNCS, vol. 3679, pp. 397–417. Springer, Heidelberg (2005)
20. Kantarcioglu, M., Clifton, C.: Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. IEEE Transactions on Knowledge and Data Engineering 16(9), 1026–1037 (2004)

21. Kankainen, A., Ushakov, N.: A consistent modification of a test for independence based on the empirical characteristic function. Journal of Mathematical Sciences, 1–10 (1998)
22. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation techniques. In: Proc. 3rd IEEE International Conference on Data Mining (ICDM 2003), p. 99 (2003)
23. Lefons, E., Silvestri, A., Tangorra, F.: An analytic approach to statistical databases. In: Proceedings of the 9th VLDB Conference (1983)
24. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Proc. ICDE 2007, pp. 106–115 (2007)
25. Liew, C.K., Choi, U.J., Liew, C.J.: A data distortion by probability distribution. ACM Transactions on Database Systems 10(3), 395–411 (1985)
26. Lindell, Y., Pinkas, B.: Privacy Preserving Data Mining. In: Bellare, M. (ed.) CRYPTO 2000. LNCS, vol. 1880, pp. 36–54. Springer, Heidelberg (2000)
27. Liu, K., Kargupta, H., Ryan, J.: Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering 18(1), 92–106 (2006)
28. Liu, K., Giannella, C., Kargupta, H.: An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 297–308. Springer, Heidelberg (2006)
29. Liu, K.: Multiplicative Data Perturbation for Privacy Preserving Data Mining., PhD thesis, University of Maryland, Baltimore County, Baltimore, MD (January 2007)
30. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-Diversity: Privacy Beyond k-Anonymity. In: Proc. of ICDE 2006, p. 24 (2006)
31. O'Grady, P.D., Pearlmutter, B.A., Rickard, S.T.: Survey of Sparse and Non-Sparse Methods in Source Separation. International Journal of Imaging Systems and Technology 15(1), 18–33 (2005)
32. Oliveira, S.R.M., Zaïane, O.R.: A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. Computers & Security 26(1), 81–93 (2007)
33. Rizvi, S., Haritsa, J.: Maintaining Data Privacy in Association Rule Mining. In: Proc. of 28th Intl. Conf. on Very Large Databases (VLDB) (August 2002)
34. Saygin, Y., Verykios, V.S., Clifton, C.: Using unknowns to prevent discovery of association rules. ACM SIGMOD Record 30(4), 45–54 (2001)
35. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
36. Szekely, G.J., Rizzo, M.L.: Testing for Equal Distributions in High Dimension, InterStat, November (5)
37. Theis, F.J., Lang, E.W., Puntonet, C.G.: A Geometric Algorithm for Overcomplete Linear ICA. Neurocomputing 56, 381–398 (2004)
38. Turgay, E.O., Pedersen, T.B., Saygin, Y., Savas, E., Levi, A.: Disclosure Risks of Distance Preserving Data Transformations. In: Ludäscher, B., Mamoulis, N. (eds.) SSDBM 2008. LNCS, vol. 5069, pp. 79–94. Springer, Heidelberg (2008)
39. Verykios, V., Elmagarmid, A., Elisa, B., Elena, D., Saygin, Y., Dasseni, E.: Association Rule Hiding. IEEE Transactions on Knowledge and Data Engineering 16(4), 434–447 (2004)
40. Yang, Z., Zhong, S., Wright, R.N.: Privacy-Preserving Classification of Customer Data without Loss of Accuracy. In: Proc. of the 2005 SIAM International Conference on Data Mining, SDM (2005)
41. Zibulevsky, M., Pearlmutter, B.A.: Blind Source Separation by Sparse Decomposition in a Signal Dictionary. Neural Computation 13(4), 863–882 (2001)