Simulation-based Approaches for Privacy Preservation

Chih-Ming Hsu
Electrical Engineering Department
National Taiwan University
Taipei, Taiwan
ming@arbor.ee.ntu.edu.tw

Ming-Syan Chen
Electrical Engineering Department
National Taiwan University
Taipei, Taiwan
mschen@cc.ee.ntu.edu.tw

ABSTRACT

Privacy preservation is a fundamental issue in data mining and knowledge discovery. The primary objective of privacy preservation is to protect an individual's confidential information in released data sets. In recent years, several simulation-based approaches for privacy preservation have been proposed. The idea is to generate a synthetic data set with the constraint that the probability distribution is as close as possible to that of the original set. In this paper, we propose two frameworks for simulation-based privacy preservation of multivariate numerical data. The first framework, called PRIMP (PRivacy preserving by Independent coMPonents), is based on independent component analysis (ICA). It is shown empirically that PRIMP outperforms other simulation-based approaches in terms of Spearman's rank correlation and Kendall's tau correlation. In addition, we prove that the synthetic data generated by PRIMP is sufficiently different from the original data; thus, we are able to protect confidential information in the original data. Also, PRIMP is easy to implement and very effective because, by using FastICA, its run time is linear to the size of the input. The second approach proposed is a hybrid method that combines PRIMP and Cholesky's decomposition technique. It is shown empirically that the hybrid method preserves the covariance matrix of the original data exactly. The method also resolves the problem of generating good seeds for the Cholesky-based approach. Although, the empirical results show that the hybrid approach is not always better than the PRIMP in terms of Spearman's rank correlation and Kendall's tau correlation, in theory, the risk of information leakage under the hybrid approach is less than that under PRIMP.

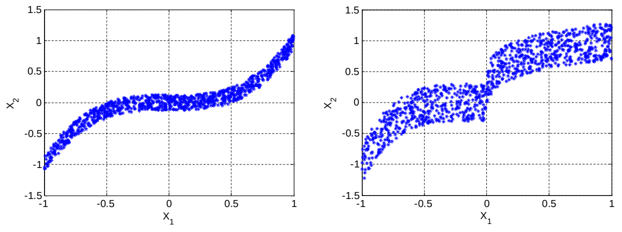
1. INTRODUCTION

The primary objective of privacy preservation, which is an important issue in the data mining and knowledge discovery field, is to protect an individual's confidential information in released data sets. When a data set is to be released for public use, individual confidentiality must be ensured. Thus, the purpose of privacy preserving techniques is twofold: they must prevent the disclosure of an individual's identity; and the published data should preserve as many statistical properties of the original data set as possible[21]. Explicitly, privacy preserving techniques have to strike a balance between two opposing factors: privacy loss and information loss.

The most popular approach for privacy preservation, called data distortion, destroys the structure of the data. For example, random data perturbation involves the addition of random noise to the values of sensitive attributes [1] [2]. However, such perturbation may lead to high information loss [1] [23], and the original data may be recovered from the perturbed data by random matrix-based spectral filtering [19] or principle component analysis [15]. Therefore, the generation of synthetic data, which preserves some statistical characteristics of the original data, has been studied extensively as an alternative means of protecting an individual's confidential information. This approach is called *simulation-based privacy preservation*. The idea is to randomly generate data with the constraint that the statistical characteristics and internal relationships of the attributes are as close as possible to those of the original data set. We know that all the statistical characteristics of any data set can be represented by the data's probability distribution. Therefore, from the viewpoint of data utility, synthetic data sets whose probability distributions are closest to the probability distribution of the original data are better. Even so, we require that the synthetic data is sufficiently different from the original data to protect the privacy of individuals.

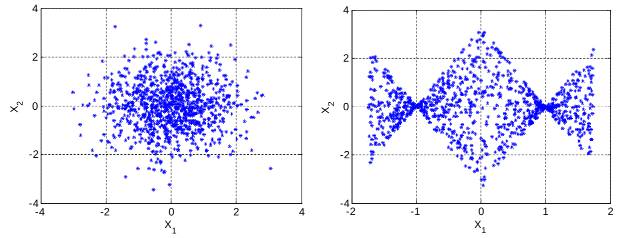
The naive method of simulation-based privacy preservation derives the probability density of the sensitive data set, and then generates the synthetic data set from the estimated density. However, multi-dimensional density estimation is time-intensive and infeasible in high dimensional space [27]. McKay et al. developed a technique, called Latin Hypercube Sampling (LHS), to generate a synthetic data set for a group of *uncorrelated* variables in which the univariate characteristics of the original data are reproduced [22]. Later, Iman and Conver refined the LHS algorithm to produce a synthetic data set that imitates Spearman's rank correlation structure of the original data [17]. In [8], Dandekar et al. used the LHS technique to create a synthetic set of the sensitive data set for privacy preservation. Although Spearman's rank correlation is one of well-known summary of the relationships of two variables, it can not completely capture the full dependency structure of the variables. For example, consider the two bivariate data sets shown in Figure

VLDB '08, August 24-30, 2008, Auckland, New Zealand



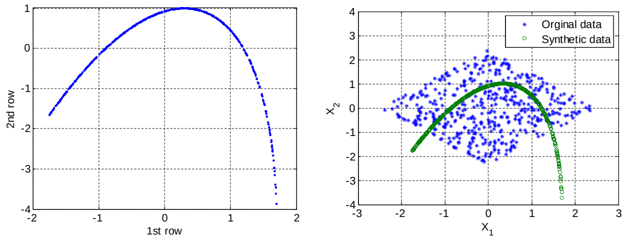
(a) Data set 1 (b) Data set 2

Figure 1: Different bivariate distributions with the same Spearman's rank correlation matrix.



(a) Data set 1 (b) Data set 2

Figure 2: Different bivariate distributions with the same Pearson's covariance matrix.



(a) The scatter plot of a (2×500) generated seeds with identity covariance for the Cholesky-based approach

(b) The scatter plot of the original confidential data set and the synthetic data generated by the Cholesky-based approach

Figure 3: An example of the Cholesky-based approach.

1[1]. From the scatter plots of both data sets, we can conclude that their distributions are significantly different, even though both have the same Spearman rank correlation matrices. The above example shows that merely preserving the Spearman's rank correlation structure of attributes does not satisfy the information loss (or data utility) requirements of multivariate data sets.

Recently, Mateo-Sanz proposed a privacy preserving technique that uses Cholesky's decomposition [14] of the covariance matrix of a sensitive data set to generate disguised synthetic data [21]. We can verify that the covariance matrix and correlation matrix[2] of sensitive data are preserved exactly in a synthetic data set generated by the Cholesky-based approach. (An outline and the properties of the approach are presented in Section 4.) Like Spearman's rank correlation, Pearson's covariance is another well-known summary of the linear relationship of two variables, but the covariance structure is not sufficient to give us the full dependency structure of the variables [5]. In the example shown in Figure 2, both bivariate data sets have the same Pearson correlation matrix. However, from the scatter plots of both data sets, we observe that their distributions are significantly different. The attributes of data set 1 in Figure 2(a) are independent, whereas the attributes of data set 2 in Figure 2(b) are dependent on one another. This example shows that Cholesky-based privacy preservation may also be prone to high information loss, even if the covariance structure can be preserved exactly. Moreover, as shown in Figure 3, the major problem with the Cholesky-based approach is that the quality of the synthetic data set depends on that of the generated seeds[3]. Figure 3(a) is the scatter plot of some 2×500 generated seeds G, while Figure 3(b) shows the original data set and the synthetic data generated by the Cholesky-based approach from G. In this example, the synthetic data set clearly misrepresents the probability distribution of the original data set X. Observably, the joint distribution of the synthetic data set is sensitive to the generated seeds G. Also, the spread shape of the synthetic data shall imitate the spread shape of the generated seeds. Since the distribution of X is invalid, generating good seeds could be very difficult if the Cholesky-based approach is used for privacy preserving. How to develop good seeds for the Cholesky-based approach still is an open problem.

Generally, existing simulation-based privacy preserving techniques only preserve some specific bivariate relationships of the attributes of the original data, such as the covariance structure in the Cholesky-based approach, and the rank correlation structure in the LHS-based approach. However, as discussed above, the probability distributions of a synthetic data and an original data derived by these approaches differ significantly in many cases. Consequently, we can not guarantee that synthetic data provides analytical validity, i.e., it represents the extent to which analyzing synthetic data

[1] These plots are called *scatter plots*. A scatter plot helps to visualize the relationship between two variables. If the probability distributions of two multi-dimensional random vectors are very close, then both scatter plots of each corresponding pair of attributes must be very similar, and vice versa [28]. Therefore, scatter plot is an effective visualization technique for comparing the similarity between two distributions.

[2] More precisely, the correlation matrix of a d-dimensional random vector $X^* = (X_1^*, \cdots, X_d^*)^T$ is defined as the expectation of $(\tilde{X}_1^*, \cdots, \tilde{X}_d^*)^T(\tilde{X}_1^*, \cdots, \tilde{X}_d^*)$, where $\tilde{X}_i^* = (X_i^* - E[X_i^*])/var(X_i^*)$, $E[X_i^*]$ and $var(X_i^*)$ are, respectively, the expected value and variance of random variable X_i^*, for $i = 1, \cdots, d$. Since we do not know the actual distribution of the sensitive data set, we can not compute its correlation matrix. For convenience, we use the same term to refer to the sample estimation, as shown in Equation (6) in Section 5, of the correlation matrix from a finite sample. Also, in this paper, the covariance matrix of a sensitive data set is referred to its sample covariance matrix.

[3] More details of Cholesky-based shall be preserented in Section 4.

provides similar results to those obtained by analyzing the original data [24]. To resolve these problems, we propose two simulation-based frameworks for preserving the privacy of multivariate numerical data. The first framework, called PRivacy preserving by Independent coMPonents (PRIMP), is based on independent component analysis (ICA). The second is a hybrid method that combines PRIMP and the Cholesky-based approach.

Because the internal relationships of attributes are always complex and polymorphous, it is difficult to generate a synthetic data set that preserves the statistical characteristics of some given data set without information about the data's distribution [11]. However, the inter-relationships and statistical characteristics of the multivariate attributes shall be simulated easily if the attributes are mutually independent. In this case, we do not need to care the inter-relationships of attributes. Then, the statistical characteristics of the each attribute should be well approximated by using some resampling technique, such as bootstrapping [10], or LHS, on individual attributes to generate the synthetic data. In light of the above observations, we use ICA to support privacy preservation and information preservation in our simulation-based approaches. The objective of ICA is to design a filter, with the original data as its input, such that the output data is as *independent* as possible. The outputs of ICA are called source signals or independent components. The mixing relationship between the inputs and the independent components can be captured by a mixing function estimated by ICA. Because of the "independence" property, the independent relationships and statistical characteristics of the source signals do not be distorted by randomly shuffling the realizations of each independent component. PRIMP uses the mixing function to mix the shuffled source signals to get a synthetic data set. This technique preserves the multivariate relationships among the attributes as well as the distribution of the original data. Using the good information-preserving properties of PRIMP, we develop a hybrid method that combines PRIMP and the Cholesky-based approach for privacy preservation. The basic idea of the hybrid method is that we use PRIMP to generate seeds for the Cholesky-based approach because it effectively approximates the distribution of the original data. Then, we use the Cholesky decomposition to adjust the covariance structure of the synthetic data because it can exactly preserve the covariance structure of the original data.

Specifically, our work makes the following contributions: We introduce a new application of ICA. Explicitly, we propose two simulation-based privacy preserving frameworks for multivariate numerical data sets. The first framework is called PRIMP. The synthetic data set generated by PRIMP can effectively approximate the probability distributions of the original data set, hence it provides analytical validity. In addition, the risk of PRIMP leaking confidential can be analyzed. We prove that the synthetic data generated by PRIMP is sufficiently different from the original data; thus, the privacy of the original data is protected. In addition, PRIMP is easy to implement and efficient. Its run time is linear to the size of the input data set by using the FastICA algorithm. It is shown empirically that PRIMP significantly outperforms other simulation-based approaches in terms of Spearman's correlation and Kendall's tau correlation. The second framework is a hybrid method that combines PRIMP and the Cholesky-based approach for privacy preserving. The method resolves the problem of generating good seeds for the Cholesky-based approach. Although the empirical results show that the hybrid approach is not always better than PRIMP in terms of Spearman's rank correlation and Kendall's tau correlation it preserves the covariance structure of the original data exactly like the Cholesky-based approach. Also, in theory, the risk of the hybrid approach leaking confidential information is less than that of PRIMP.

The remainder of the paper is organized as follows. Section 2 describes related works. The PRIMP framework is discussed in Section 3. We present the hybrid PRIMA and Cholesky-based approach in Section 4. The experiment results are detailed and analyzed in Section 5. Finally, in Section 6, we present our conclusions.

2. RELATED WORKS

Protecting an individual's confidential information in multivariate numerical data has attracted a great deal of attention in recent years. The most relevant works are those based on perturbation techniques, which can be classified into two categories. The first consist of random data perturbation [1] [2] and random rotation perturbation [4], which disguise sensitive data by distorting the structure of the data. The second category generates a new synthetic data set that preserves certain statistical requirements.

Random data perturbation is the most popular technique for privacy preservation because it is easy to implement [1] [2]. Let X be a d-dimensional data set with sensitive attributes. The naive method of random data perturbation involves the addition of random noise to each attribute. Formally, the perturbed data can be represented as $\hat{X} = X + \epsilon$, where ϵ is a d-dimensional random vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I_d$ for some positive constant σ. The perturbation method, also known as the independent noise method, gives biased responses to some queries, and thus changes the statistical relationships among the attributes [23]. In addition, the original data may be recovered from the perturbed data by random matrix-based spectral filtering [19] or principle component analysis [15]. The correlated noise method, a variation of the independent noise method [15][23] specifies the noisy term so that ϵ has a mean $\mathbf{0}$ and covariance matrix $\sigma^2 cov(X)$ for some positive constant σ. Although the correlation structure of the perturbed attributes remains the same as that of the original attributes, the variance of an individual perturbed attribute should be $(1 + \sigma^2)$ times that of the corresponding original attribute. To eliminate this bias, a modification of the correlated noise method, called bias corrected correlated noise, was proposed by Tendick and Matloff [29]; however, it can not solve the bias problem in non-Gaussian cases. Recently, Chen and Liu [4] proposed a random rotation perturbation approach for data classification that is suitable for some rotation invariant classifiers. However, the perturbed data can not provide analytical validity, so it is difficult to explain and apply in many data mining algorithms.

Simulated-based approaches release a synthetic data set for public use. Synthetic data is sufficiently different to the original data, but it retains the statistical characteristics of the original data as much as possible. Liew et al. proposed a simulation-based scheme, called data distortion by probability distribution, which generates a synthetic data set by randomly drawing from the underlying distribution of the original data set [20]. Because the distribution of the original data set is invalid, in the first step, the Komlogorov-Smirnov test [26] is used to identify the underlying density function of the original data set from some pre-determined set and estimate the parameters of the function. In this framework, the candidates for the density function of the original data set should be empirically selected by the user in advance. Because of the tremendous number of possible data distributions, in most cases, we can not find a suitable density function in the pre-determined set to fit the data. Furthermore, the synthetic data generated by the Liew et al. proposed technique does not preserve the internal relationships among the attributes of a multivariate data set, because the process of identifying the underlying density function deals with each attribute separately. Consequently, synthetic data generated based on the probability distribution leads to high information loss in many cases. In contrast, Dandekar et al. [8] used the Latin Hypercube Sampling (LHS) technique to generate synthetic data for privacy preserving. It maintains Spearman's rank correlation structure of the original data. Mateo-Sanz proposed another simulated-based approach based on Cholesky's decomposition of the covariance matrix of sensitive data. The major advantage of this approach is that Pearson's covariance matrix is preserved exactly in the synthetic data set.

Existing simulation-based approaches only preserve some specific statistics of the original data set. Explicitly, synthetic data generated by these simulated-based approaches can not provide analytical validity. In [7], the author asked: Why not directly publish the statistics one wants to preserve directly, rather than release a synthetic data set? To remedy the data utility problem in these approaches, we propose two simulation-based frameworks using independent component analysis (ICA).

3. PRIVACY PRESERVATION WITH SYNTHETIC DATA

In this section, we introduce the simulation-based privacy preserving framework PRIMP for multivariate numerical data sets. Preliminaries are given in Section 3.1. Section 3.2 contains some necessary background information about independent component analysis. In Section 3.3, we describe PRIMP in detail, and then discuss some theoretical properties of PRIMP in Section 3.4.

3.1 Simulation-Based Privacy Preservation

We assume that users must be given access to released data without restrictions [24]. The goal of privacy preservation is to provide released data that has analytical validity, but does not disclose confidential information about individuals. The *analytical validity* represents the extent to which analyzing the synthetic data provides similar results to those obtained by analyzing the original data [24]. Simulation-based privacy preservation generates a synthetic data set that preserves the statistical characteristics of the original data set for release to the public. If the synthetic data is sufficiently different from the original data, then the privacy of the original data is said to be preserved.

Let X be an original sensitive data set, with n records and d attributes. For ease of exposition, X can be viewed as a $d \times n$ matrix whose columns and rows represent, respectively, a record and the collection of samples for an attribute of X. For $j = 1, 2, \cdots, d$, we assume that the jth row of X (,i.e. X_j) is sampled from some unknown random variable $\mathbf{X}_j$. In fact, we have no information about the probability distribution of $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_d})^T$ except for the data set X. We know that all the statistical characteristics of any data set are determined by its probability distribution. Therefore, the simulation-based approach to database privacy preservation involves generating a synthetic data set $\hat{X}$ with the same joint distributions as the original sensitive data set X. Explicitly, our purpose is to design an effective and efficient simulating schema based only on the observed data. A list of symbols used throughout this paper is given in Table 1.

Table 1: List of symbols.

Notation	Definition
X	The d-dimensional sensitive set
$\hat{X}$	The disguised synthetic data set of X
S	The m-dimensional source signal data set estimated by ICA, $m \leq d$
n	The number of records of X
$\mathbf{g}$	The mixing function of the general ICA model: $X = \mathbf{g}(S)$
$\mathbf{g^{-1}}$	The unmixing function of the general ICA model: $X = \mathbf{g}(S)$ satisfies $S = \mathbf{g^{-1}}(X)$
A	The mixing matrix of the linear ICA model: $X = AS$
W	The unmixing matrix of the linear ICA model: $X = AS$ satisfies $WA = I_m$ and $AW = I_d$
M^T	The transpose of the data matrix M
M_i	The ith row of the data matrix M, i.e., the samples of ith attribute of M
$M(j)$	The jth column of the data matrix M, i.e., the jth record of M
$E[M_i]$	The sample expectation of the ith row of the data matrix M
$var(M_i)$	The sample variance of the ith row of the data matrix M
$cov(M)$	The (Pearson's) covariance matrix of the data matrix M
$\mathcal{P}\{e\}$	The probability of an event e
I_m	An $m \times m$ identity matrix

3.2 Independent Component Analysis

Here, we introduce some basic concepts of independent component analysis (ICA). A detailed study of ICA can be found in [16]. ICA defines a generative model for observed multivariate data. The model's data variables are assumed to be linear or nonlinear mixtures of some unknown *independent* latent variables, and the mixing system is also unknown [16]. The latent variables are called *independent components* or *source signals.* Let x be a d-dimensional observation. Explicitly, ICA is a problem of recovering a latent random vector s from observations of unknown m-variate ($m \leq d$) functions of that vector. The latent vector s is assumed to be sampled from some unknown random vector, i.e., source signals, whose elements are *mutually independent.* Formally, we can write the relationship of the observed data x and the realization of the source signals s in a general form

$$x = \mathbf{g}(s), \tag{1}$$

where $\mathbf{g} : R^m \to R^d$ is an unknown real-value d-component

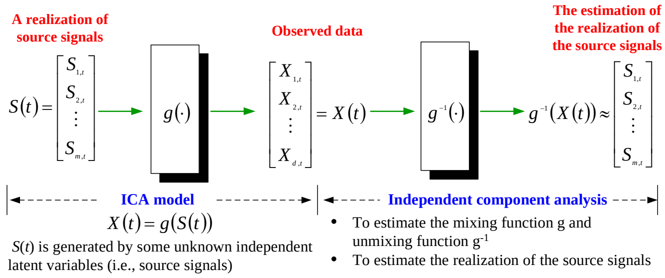

A realization of source signals
Observed data
The estimation of the realization of the source signals
$S(t)=\begin{bmatrix}S_{1,t}\\S_{2,t}\\\vdots\\S_{m,t}\end{bmatrix}$
$g(\cdot)$
$\begin{bmatrix}X_{1,t}\\X_{2,t}\\\vdots\\X_{d,t}\end{bmatrix}=X(t)$
$g^{-1}(\cdot)$
$g^{-1}(X(t))\approx\begin{bmatrix}S_{1,t}\\S_{2,t}\\\vdots\\S_{m,t}\end{bmatrix}$
ICA model
$X(t)=g(S(t))$
S(t) is generated by some unknown independent latent variables (i.e., source signals)
Independent component analysis
To estimate the mixing function g and unmixing function g-1
To estimate the realization of the source signals

Figure 4: Diagram of independent component analysis.

vector function. Given n independent, identically distributed observations $\{X(1), X(2), \cdots, X(n)\}$[4], ICA tries to find an inverse mapping $\mathbf{g^{-1}}: R^d \rightarrow R^m$, which estimates of n realizations of the source signals by

$$S(t) = \mathbf{g^{-1}}(X(t)) \tag{2}$$

for any particular $X(t)$, $t = 1, 2, \cdots, n$. The functions $\mathbf{g}$ and $\mathbf{g^{-1}}$ are called a *mixing function* and an *unmixing function*, respectively. Figure 4 shows the diagram of ICA.

In some cases, we may consider linear functions because interpreting the representation and its computation are simpler [16]. As such, the linear ICA model of Equation (1) can be expressed as a matrix formula:

$$X = \mathbf{A}S, \tag{3}$$

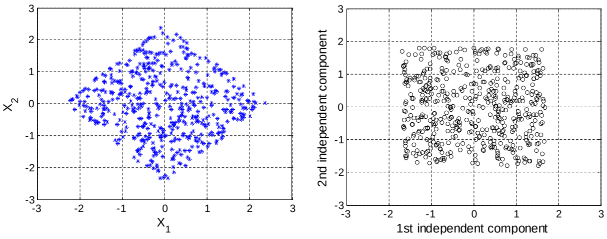
where $\mathbf{A}$ is a $d \times m$ matrix of parameters, $X = (X(1), X(2), \cdots, X(n))$, and $S = (S(1), S(2), \cdots, S(n))$. Independent source signals are reconstructed through a linear projection of:

$$S = \mathbf{W}X, \tag{4}$$

where $\mathbf{W}$ is an $m \times d$ matrix of parameters with $\mathbf{WA} = \mathbf{I}_m$ and $\mathbf{AW} = \mathbf{I}_d$. The matrices $\mathbf{A}$ and $\mathbf{W}$ are called a *mixing matrix* and an *unmixing matrix*, respectively. Note that, in some cases, such as the data set in Figure 1(b), the linear ICA model in Equation (3) is not suitable for extracting the data structure. The selection strategies of using linear or nonlinear models are beyond the scope of this paper. We refer the readers to Hyvärinen et al.'s book[16] for further details.

Example 1: Consider the scenario of a 2-dimensional data set X with attributes X_1 and X_2, as shown in Figure 5(a). Clearly, the attributes X_1 and X_2 of X are not independent; thus, it is difficult to predict the value of one attribute from the value of the other. It is also very difficult to generate a synthetic data set to preserve the joint distribution of X based on the sampled data set only [11].

We use FastICA[5] to estimate the source signals and the mixing matrix. The joint distribution of the *independent* source signals is shown in Figure 5(b). Clearly, the joint distribution of the source signals is uniform on a square, so the joint probability density of the signals is simply the product of their marginal probability densities. Hence, we can use a re-sampling technique or LHS to generate a synthetic data set with the joint distribution as the source signals. Furthermore, we can use the mixing matrix and the synthetic data generated from the source signals to obtain a synthetic data set of X, which preserves the distribution of X. □



(a) Observed data (b) Source signals

Figure 5: (a) The scatter plot of the observed data X. (b) The scatter plot of the independent source signals estimated by using the FastICA algorithm.

The above results show that the "independence" property of the source signals estimated by ICA simplifies the problem of generating a synthetic data set based on sampled (or observed) data only. In addition, in terms of the re-identification disclosure risk[6], it is better to distort or transform the data structure in the independent component space, rather than in the original data set. Based on these properties of ICA, we develop two new simulation-based approaches for preserving privacy. We discuss these approaches in Sections 3.3 and 4.

3.3 ICA-based Approach for Privacy Preservation

In this section, we describe how ICA can be extended to a new application for privacy preservation. We assume that the normalized sensitive data set X satisfies the ICA model, as in Equation (1).

The outline of PRIMP is given in Table 2. Whitening is useful as a preprocessing step in ICA [16]. A zero-mean random vector is said to be *white* if its elements are uncorrelated and have unit variance. Because whitening is essentially decorrelation followed by scaling, the technique of principle component analysis is often used [16], but it is not advisable to use a covariance matrix to find the principle components when the attributes' variances are very different [18]. Even so, normalization can reduce the complexity and improve the quality of ICA substantially [16]. Therefore, the first step of PRIMP normalizes the original data set.

As mentioned earlier, how to generate a multivariate synthetic data set with the same distribution of input is still an open question. Because of the tremendous number of inter-

[4] Note that we do not assume the attributes of data are independent. Essentially, the using of ICA is pointless if the attributes of the data are mutually independent.

[5] FastICA is one of the widely used algorithms for performing linear ICA. The fundamental mathematical background and implementation techniques of it can be found in [16]. The free FastICA package can be downloaded from http://www.cis.hut.fi/projects/ica/fastica.

[6] The re-identification disclosure risk is the risk of disclosing a one-to-one relationship between the synthetic data set and the original data set [9].

Table 2: The PRIMP algorithm.

```
Algorithm: PRIMP($X$, $m$)
Inputs: $X$ (d-dimensional data set),
        $m$ (the dimensions of source signals, the default value is $d$, $m \leq d$ )
Output: $\hat{X}$ (d-dimensional synthetic data set)
 1. Normalization. For each $j = 1, 2, \cdots, d$,
        we normalize the projected data $X_j$ of the $j$th attribute of $X$
        to give $\tilde{X_j}$ with $E[\tilde{X_j}] = 0$ and $var(\tilde{X_j}) = 1$.
 2. Decoupling. Use ICA to estimate
        the source signals $S = (S_1, S_2, \cdots, S_m)^T$
        and the mixing function $\mathbf{g}$ of the normalized data set $\tilde{X} = (\tilde{X_1}, \tilde{X_2}, \cdots, \tilde{X_d})^T$.
 3. Permutation. For each $j = 1, 2, \cdots, m$,
        we randomly permutate the independent component $S_j$ to obtain $\tilde{S_j}$.
 4. Coupling. Compute $\hat{X} = \Big(\hat{X}(1), \hat{X}(2), \cdots, \hat{X}(n)\Big)$,
        where $\hat{X}(j) = \mathbf{g}(\tilde{S}(j))$ $(j = 1, 2, \cdots, n)$.
 5. Rescaling. For each $j = 1, 2, \cdots, d$, $\hat{X_j} = \hat{X_j} \times \sqrt{var(X_j)} + E[X_j]$.
```

nal relationships among attributes, generating a synthetic data set that preserves all the statistical characteristics of a multivariate data set without distribution information is almost impossible [11]. However, if the attributes of the input are mutually independent, generating a synthetic data set based solely on the sampled data set becomes straightforward.

The proposed approach views the sensitive data set generated by an unknown function of some latent independent random process. Explicitly, we assume that the normalized data set $\tilde{X}$ belonging to the sensitive data set X satisfies the equation $\tilde{X} = \mathbf{g}(S)$, where S is the latent independent random process. Under this assumption, we gain two important insights for designing a simulation-based privacy preserving technique. First, for privacy preservation, we can transform the source signals obtained by ICA into protected data by some distortion technique. Hence, we hide any clue about the distortion technique in the source signals to prevent a hacker accessing the sensitive data. (Note that most existing distortion techniques disguise the sensitive attributes in original data. Consequently, clues about adapted distortion techniques can be obtained by comparing the difference between the released data and the original data [2][15][19].) Second, the "independence" property of source signals simplifies the problem of generating a synthetic data for source signals. If the attributes of the original data set are mutually independent, we can apply a re-sampling technique, such as bootstrap [10], or LHS [22] to individual attributes to generate the synthetic data. In theory, the inter-relationships and statistical characteristics among attributes should be well maintained by such re-sampling techniques, because we do not need to consider the inter-relationships of attributes. In other words, using the ICA technique to generate a synthetic data set for the original data set requires little effort. In PRIMP, we randomly shuffle the realizations[7] of the estimated source signals to preserve privacy. Because of the "independence" property, the shuffled source signals preserve the overall statistical characteristics of the original source signals.

We mix the the shuffled source signals by mixing function, as shown in Equation (1), to obtain the (normalized) synthetic data set $\hat{X}$. Therefore, the statistical characteristics of the (normalized) synthetic data set can approximate the (normalized) original data set X. The quality of the synthetic data in terms of its data utility depends on the approximation of independence determined by ICA. Based on the foregoing, the PRIMP process is shown in Figure 6. Note that, the first step of PRIMP normalizes the original data set. Consequently, adjusting the synthetic data to preserve the mean and variance of each attribute of X is the final step of PRIMP.

The following example compares the qualities of PRIMP in terms of data utility with some popular simulation-based privacy preserving methods.

Example 2: Recall the data set X in Example 1. Figure 7(a) shows the scatter plot of the synthetic data generated by PRIMP. Because the assumption of multivariate normal distribution for the original data set is made by some privacy preserving techniques [15] [23], we generate a multivariate normal data set with the same mean and covariance structure as X, as shown in Figure 7(b). Figures 7(c) and 7(d) present the scatter plots of the synthetic data sets generated by the LHS-based and Cholesky-based approaches, respectively. As the figures show, all the sets misrepresent the joint distribution of the original data X, except for the set generated by PRIMP. □

3.4 Properties of PRIMP

As mentioned earlier, privacy preserving techniques have to strike a balance between two opposing forces: privacy loss and information loss. In fact, none of the existing privacy preserving techniques, including PRIMP, can achieve the dual goals of one hundred percent data utility and zero privacy loss. The distribution of records leaked by using PRIMP will be derived in Theorem 1. The theorem states that the synthetic data sets generated by PRIMP are sufficiently different from the original data set; thus, they are able to protect the privacy of the original data. In addition, we discuss the time complexity of PRIMP, which is presented in Section 3.4.2. Not that, for interest of space, proofs of those properties of PRIMP are given in Appendix.

Here, we introduce a few terms to save space later. Let

[7] The term "realizations" of a random variable means some experiment outcomes described by a random variable.

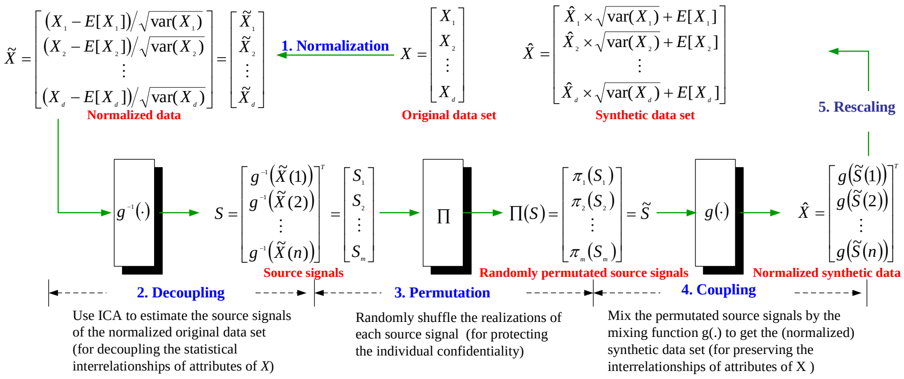
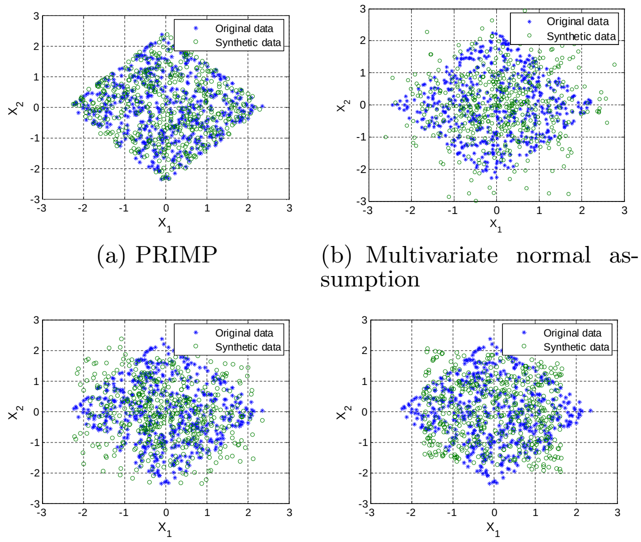

1. Normalization
Normalized data
Original data set
Synthetic data set
5. Rescaling
$g^{-1}(\cdot)$
Source signals
Π
Randomly permutated source signals
$g(\cdot)$
Normalized synthetic data
2. Decoupling
3. Permutation
4. Coupling
Use ICA to estimate the source signals of the normalized original data set (for decoupling the statistical interrelationships of attributes of X)
Randomly shuffle the realizations of each source signal (for protecting the individual confidentiality)
Mix the permutated source signals by the mixing function g(.) to get the (normalized) synthetic data set (for preserving the interrelationships of attributes of X)

Figure 6: Diagram of PRIMP.



(a) PRIMP (b) Multivariate normal assumption

(c) LHS-based (d) Cholesky-based

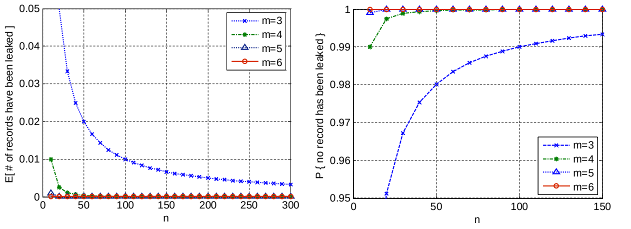
Figure 7: The synthetic data sets for the original data set X in Example 1.

$\Pi = \{\pi_1, \pi_2, \cdots, \pi_m\}$ be a set of random permutations of $\{1, 2, \cdots, n\}$, $S_i = (S_{i,1}, S_{i,2}, \cdots, S_{i,n})$ be the ith row of the source signals matrix S, and $S(j) = (S_{1,j}, S_{2,j}, \cdots, S_{m,j})^T$ be the jth column of the source signals matrix S. For $i = 1, 2, \cdots, m$, we denote $\pi_i(S_i) = (S_{i,\pi_i(1)}, S_{i,\pi_i(2)}, \cdots, S_{i,\pi_i(n)})$ as the permutated vector of S_i whose elements' positions are permutated randomly by π_i.
$\Pi(S(i)) = (S_{1,\pi_1(i)}, S_{2,\pi_2(i)}, \cdots, S_{m,\pi_m(i)})^T$ denotes the ith sample vector of the permutated source signals.

3.4.1 Privacy Loss of PRIMP

We now discuss the privacy loss of PRIMP.

Definition 1. *The privacy of record $X(i)$ should be leaked under permutation Π if it satisfies $\pi_1(i) = \pi_2(i) = \pi_3(i) = \cdots = \pi_m(i) = j$ for some j, $1 \le j \le n$. Thus $\hat{X}(i) = \mathbf{g}(\Pi(S(i)))$, the original record $X(i)$ of X should appear in the synthetic data set by using PRIMP under permutation Π. We denote the probability that at least one of the records has been leaked by PRIMP as the leakage risk.*



(a) (b)

Figure 8: (a) The theoretical expected number of records leaked by PRIMP. (b) The probability that no records have been leaked by PRIMP. (Note that $\mathcal{P}$\{no records has been leaked\} $= 1 -$ leakage risk.)

Now, we derive the leakage risk of PRIMP.

Theorem 1. *Let $\mathcal{N}$ denote the number of records leaked by PRIMP. The probability mass function, the mean, and the variance of $\mathcal{N}$ are as follows:*

$$\mathcal{P}\{\mathcal{N} = k\} = \frac{1}{k!}\left[1 - \sum_{l=1}^{n-k}(-1)^{l+1}R_m(l)\right], \text{ for } k = 0, 1, 2, \cdots, n,$$

where $R_m(l) = \frac{1}{l!}\left(\frac{(n-l)!}{n!}\right)^{m-2}$.

$$E[\mathcal{N}] = \left(\frac{1}{n}\right)^{m-2}.$$

$$var(\mathcal{N}) = \left(\frac{1}{n}\right)^{m-2}\left[1 - \frac{1}{n^{m-2}} + \frac{1}{(n-1)^{m-2}}\right].$$

Theorem 2. *The leakage risk of PRIMP is* $\sum_{l=1}^{n}(-1)^{l+1}R_m(l)$.

Figures 8(a) and 8(b) show, respectively, the theoretical expected number of records that have been leaked and the probability that no records have been leaked for different dimensions of source signals with increasing data size. Clearly, the leakage risk decreases with increases in the data size or dimensionality. As $m = 3$, if the size of data set is larger than 100, the leakage risk of the PRIMP approaches 0. Furthermore, for any data size, the leaked risk is almost 0 as $m \geq 4$. Note that, as $m = 2$, the leakage risk should not converge to 0, even if n is very large. Explicitly, for large n and $m = 2$, the leakage risk is approximately $1 - e^{-1} \approx 0.63$ and the estimated number of leaked records is 1. Fortunately, for large n, as $m = 2$, the expected leakage ratio should be close to $1/n \approx 0$. To minimize the leakage risk, we recommend that the optimal number of source signals should equal the dimensions of the sensitive data set, i.e. $m = d$, because the risk decreases as the number of source signals increases. In summary, from Theorem 1 and Figure 8, we conclude that the synthetic data generated by PRIMP is sufficiently different from the original data. Thus, PRIMP is capable of protecting the privacy of the sensitive data set.

3.4.2 *Time Complexity of PRIMP*

We now analyze the computational complexity of PRIMP.

THEOREM 3. *The time complexity of PRIMP is $O(Knd^2)$ by using algorithm FastICA, where K is the total number of iterations required for FastICA to determine the optimal number of source signals.*

In general, the number of records, n, is much larger than the number of attributes, d, i.e., $d << n$. In addition, the total number of iterations of FastICA is always much smaller than the number of records n, i.e. $K << n$. As such, the time complexity of PRIMP, implemented by FastICA, is linear to the size of the inputs.

4. A HYBRID PRIMP AND CHOLESKY-BASED APPROACH FOR PRIVACY PRESERVATION

In this section, we develop a hybrid privacy preserving method that combines PRIMP and the Cholesky-based approach [21]. The fundamental concept of the Cholesky-based privacy preserving technique is based on the following theorem.

THEOREM 4. *Let X be any d-dimensional data set and G be any $d \times n'$ matrix whose Pearson's covariance matrix is equal to the identity matrix. Then, the covariance matrix of LG should be the same as that of X, where L is a lower triangular matrix by the equation $cov(X) = LL^T$ using Cholesky's decomposition.*

PROOF. Since $cov(LG) = Lcov(G)L^T$ and $cov(G) = I_d$, then $cov(LG) = LI_dL^T = cov(X)$, where I_d is the d-dimensional identity matrix. Hence, the covariance matrices of X and LG are the same. □

From Theorem 4, we can generate a $d \times n$ random matrix G whose Pearson's covariance matrix is equal to the identity matrix, where n and d are the size of the synthetic data set and the dimensions of the sensitive data set X, respectively. Therefore, the synthetic data set should be obtained by $\hat{X} = LG$, where L is a lower triangular matrix derived by the equation $C = LL^T$ using Cholesky's decomposition and C is the Pearson covariance matrix of X. By Theorem 4, the covariance matrix and correlation matrix of X are preserved exactly in the synthetic data set $\hat{X}$. An outline of the Cholesky-based privacy preserving approach is presented in Table 3. We call the random matrix G the *seeds* of Cholesky-based privacy preservation.

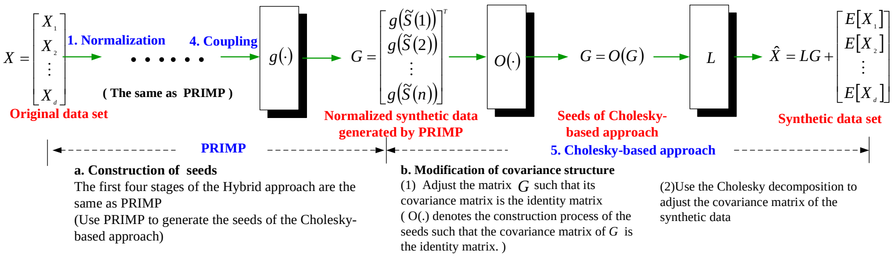
Table 3: The algorithm of Cholesky-based privacy preservation.

```
Algorithm: Cholesky-based privacy preservation
Inputs: $X$ (d-dimensional data set),
Output: $\hat{X}$ (d-dimensional synthetic data set)
  1. Generate an $n \times d$ random matrix $G$ with an
       identity matrix.
  2. Compute $C = cov(X)$.
  3. Use Cholesky's decomposition on $C$ to obtain
       $C = LL^T$, where $L$ is a lower triangular matrix.
  4. Obtain the synthetic data $\hat{X} = LG$.
  5. Adjust the mean of each attribute of the
       synthetic data by $\hat{X}_j = \hat{X}_j + E[X_j]$, for $j = 1, 2, \cdots, d$.
```

As mentioned above, the synthetic data generated by the Cholesky-based approach reproduces the covariance and correlation structures of the original data set *exactly.* However, in terms of data utility, the approach's performance depends on the generated seeds G. For example, as shown in Figures 3 and 7(d), the distributions of synthetic data sets are different significantly from those of the original data. Consequently, the synthetic data can not provide analytical valid. For example, in these cases, the data's extreme statistics and the range of each attribute differ significantly from those of the original data. Unfortunately, how to generate good seeds for the Cholesky-based approach is still an open question.

In Section 3.3, we proposed a simulation-based approach, PRIMP, for privacy preservation. In theory, the probability distribution of synthetic data generated by PRIMP can successfully approximate the probability distribution of the original data. However, in terms of preserving the covariance structure, the Cholesky-based method outperforms all existing simulation-based approaches. *Essentially, the Cholesky-based approach is a modification process that, given a $d \times d$ positive semidefinite matrix*[8]*, can translate any d-dimensional data set such that the covariance matrix of the translated data set is equal to the given matrix.* Following the above observations, we use the Cholesky-based technique to modify the covariance structure of the synthetic data generated by PRIMP for covariance matrix preservation. In the other words, because the output of PRIMP can satisfactorily preserve the distribution of the original data, we use PRIMP in the construction process to generate seeds for the Cholesky-based approach. As such, the hybrid PRIMP and Cholesky-based approach can preserve the distribution and the covariance matrix of a sensitive data set. The stages of the hybrid approach, shown in Figure 9, are the same as

[8] A $d \times d$ matrix M is *positive semidefinite* if it satisfies $a^T Ma \geq 0$, for all d-dimensional vectors a [16]. It is easy to verify that the covariance matrix of any random vector is positive semidefinite, since $var(a^T X) = a^T Ma \geq 0$ if M is the covariance matrix of some random vector X. In addition, the correlation matrix of any random vector is positive semidefinite.


$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix}$
1. Normalization
4. Coupling
(The same as PRIMP)
$g(\cdot)$
$G = \begin{bmatrix} g(\tilde{S}(1)) \\ g(\tilde{S}(2)) \\ \vdots \\ g(\tilde{S}(n)) \end{bmatrix}^T$
$O(\cdot)$
$G = O(G)$
L
$\hat{X} = LG + \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_d] \end{bmatrix}$
Original data set
Normalized synthetic data generated by PRIMP
Seeds of Cholesky-based approach
Synthetic data set
PRIMP
5. Cholesky-based approach

a. Construction of seeds
The first four stages of the Hybrid approach are the same as PRIMP
(Use PRIMP to generate the seeds of the Cholesky-based approach)

b. Modification of covariance structure
(1) Adjust the matrix G such that its covariance matrix is the identity matrix
(O(.) denotes the construction process of the seeds such that the covariance matrix of G is the identity matrix.)

(2)Use the Cholesky decomposition to adjust the covariance matrix of the synthetic data

Figure 9: Diagram of the hybrid PRIMP and the Cholesky-based approach for privacy preservation.

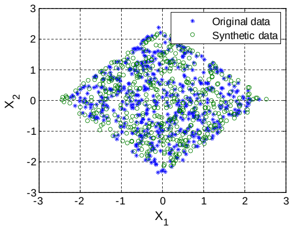


Figure 10: The synthetic data set generated by the hybrid approach for the original data set X in Example 1.

those of PRIMP, except for the rescaling procedure, which uses the Cholesky-based approach to refine the covariance structure of the synthetic data. The resulting, synthetic data preserves the covariance and correlation structures of the original data exactly like the Cholesky-based approach. The outline of the hybrid approach is summarized in Table 4.

Example 3: Again, we consider the data set X in Example 1. Figure 10 shows the scatter plots of the synthetic data generated by the hybrid method. Because PRIMP can preserve the distribution, the outputs of the hybrid method approximate the distribution of the original data sufficiently. □

In PRIMP, we hide the clues about the distortion technique in the source signals to prevent a hacker accessing the original data. Also, by Theorem 1, we prove that the synthetic data generated by PRIMP is sufficiently different from the original data. Thus, PRIMP can protect the privacy of the sensitive data set. Furthermore, since the hybrid approach uses the two distortion techniques proposed for PRIMP and the Cholesky-based approach simultaneously, it is more effective than the techniques individually. With Theorem 1, we then have that the leakage risk of the hybrid PRIMP and Cholesky-based approach should be less than $\sum_{l=1}^{n}(-1)^{l+1}R_m(l)$. In addition, as $n >> d$, by using FastICA the overall computational complexity is $O(Knd^2+d^4)$, which is the same that of PRIMP[9]. In summary, the hybrid approach combines the advantages of PRIMP and the Cholesky-based approach for privacy preservation and data utility. It also solves the problem of how to generate good seeds for the Cholesky-based approach.

5. EXPERIMENT EVALUATIONS

To assess the performance of PRIMP and the hybrid approach in terms of data utility, we conducted a series of experiments on both artificial and real data sets. We compared the information loss in the synthetic data sets produced by PRIMP with the synthetic data sets produced by the following privacy preserving approaches: the LHS-based approach, the Cholesky-based approach, the hybrid PRIMP and Cholesky-based approach, and sampling under the multivariate normal assumption (MN).

Data sets. Data set 1 was derived from the American Housing Survey of the U. S. Census Bureau using the DataFerrett system[10]. Data sets 2 and 3 were used in [23] and [24], respectively[11]. Data sets 3 to 10 were downloaded from the UCI Machine Learning Database [12]. Data set 11, downloaded from the web site of the Computational Aspects of Statistical Confidentiality (CASC) project[12], was also used in [21], [9], and [13]. Note that, in this paper, our proposed method focuses on multivariate numerical data sets. Therefore, we removed the non-numerical attributes of the above data sets.

As noted in [16] (page 152), for linear ICA model, the independent source signals must have non-Gaussian distributions; at most, one of the independent components can have a Gaussian distribution. To evaluate the effect of the non-Gaussian assumption, we generate the following artificial data set as Data set 12, $X = (X_1, X_2, X_3)^T$. For $i = 1, 2, 3, \cdots, 2000$,

$$\begin{aligned} x_{1,i} &= s_{1,i} + 3s_{2,i} + 3s_{3,i} \\ x_{2,i} &= 4s_{1,i} + 4s_{2,i} + 6s_{3,i} \qquad (5) \\ x_{3,i} &= 12s_{1,i} + 13s_{2,i} + 17s_{3,i}, \end{aligned}$$

where $x_{i,j}$ is the j-th element of X_i, $s_{1,i}$ and $s_{2,i}$ are randomly sampled from a univariate Gaussian distribution with mean 0 and variance 1, and $s_{3,i}$ is randomly sampled uniformly from $(0, 1)$. The generated source signals

[9] The overall complexity of the Cholesky-based approach is $O(n + d^4)$ [21].

[10] The DataFerrett system can be downloaded from http://dataferrett.census.gov/.

[11] Data set 2 and Data set 3 can be downloaded from http://gatton.uky.edu/faculty/muralidhar/maskingpapers/.

[12] Data set 11 can be downloaded from http://neon.vb.cbs.nl/casc/testsets.html.

Table 4: The hybrid PRIMP and Cholesky-based approach for privacy preservation.

```
Algorithm: The hybrid approach for privacy preservation
Inputs: $X$ (d-dimensional data set),
        $m$ (the dimensions of source signals, the default value is $d$, $m \leq d$ )
Output: $\hat{X}$ (d-dimensional synthetic data set)
(Use PRIMP to generate seeds)
 1. Use PRIMP to generate a $d \times n$ synthetic data $G$.
 2. Adjust the matrix $G$ such that its covariance matrix is $I_d$.
    ($G$ is used as the seeds of the Cholesky-based approach)
(Use the Cholesky-based approach)
 2. Compute $C = cov(X)$.
 4. Use Cholesky's decomposition on $C$ to obtain $C = LL^T$,
    where $L$ is a lower triangular matrix.
 5. Obtain the synthetic data $\hat{X} = LG$.
 6. Adjust the mean of each attribute of the synthetic data by
        $\hat{X}_j = \hat{X}_j + E[X_j]$, for $j = 1, 2, \cdots, d$.
```

$S_1 = (s_{1,1}, s_{1,2}, \cdots, s_{1,2000})$, $S_2 = (s_{2,1}, s_{2,2}, \cdots, s_{2,2000})$, and $S_3 = (s_{3,1}, s_{3,2}, \cdots, s_{3,2000})$, are mutually independent. Clearly, the model (5) violates the non-Gaussian restriction of the linear ICA model. Table 5 summarizes the data sets used in our experiments.

Information loss measures. Here, we use three well-known measures of bivariate relationships: Pearson's correlation matrix, Spearman's rank correlation matrix, and Kendall's tau to evaluate the information loss of privacy preserving algorithms. Let $X_i = (x_{i1}, x_{i2}, \cdots, x_{i,n})$ and $X_j = (x_{j1}, x_{j2}, \cdots, x_{j,n})$ be the observations of the ith and jth attributes of data set X, respectively. The ijth element of (Pearson's) correlation matrix, (Spearman's) rank correlation matrix, and Kendall's Tau can be computed as follows ([5], chapter 3):

$$\text{Correlation: } \frac{\sum_{k=1}^{n}(x_{i,k} - \bar{x_i})(x_{j,k} - \bar{x_i})}{\sqrt{\sum_{k=1}^{n}(x_{i,k} - \bar{x_i})^2 \sum_{k=1}^{n}(x_{j,k} - \bar{x_j})^2}}, \quad (6)$$

$$\text{Rank correlation: } 1 - 6\frac{\sum_{k=1}^{n}(R_{i,k} - R_{j,k})^2}{n(n^2-1)}, \quad (7)$$

$$\text{Kendall's tau: } \frac{2}{n(n-1)} \sum_{k=1}^{n} \sum_{l>k} sgn\left((x_{i,k} - x_{i,l})(x_{j,k} - x_{j,l})\right), \quad (8)$$

where $\bar{x_i} = \sum_{k=1}^{n} x_{i,k}$; $R_{i,k}$ is the rank of $x_{i,k}$ in $\{x_{i,1}, x_{i,2}, \cdots, x_{i,n}\}$ in ascending order; and $sgn(\cdot)$ is the sign function, defined as $sgn(x) = 1$ if $x \geq 0$ or $sgn(x) = -1$ if $x < 0$, for $i = 1, 2, \cdots, d$, $k = 1, 2, \cdots, n$.

Let C and $\hat{C}(P)$ denote, respectively, the matrix of bivariate relationship measurements for the original data set and the synthetic data set generated by algorithm P. The *relative bias* of C for algorithm P is defined as:

$$\frac{2}{m^2+m} \sum_{i=1}^{m} \sum_{j=i}^{m} \frac{|\hat{C}_{ij}(P) - C_{ij}|}{|C_{ij}|},$$

where C_{ij} and $\hat{C}_{ij}(P)$ denote, respectively, the elements of C and $\hat{C}$ in the ith row and jth column. A smaller relative bias implies better preservation of the bivariate structure C. That is, a lower relative bias represents less information loss by the statistic C.

In our experiments, we use FastICA to evaluate the information loss of PRIMP. The elements of the seeds of the Cholesky-based approach were generated from a uniform distribution over $(0, 1)$. Then, the Gram-Schmidt process [14] was applied to transform the covariance matrix of the generated seeds into an identity matrix. We also set the number of source signals to be the same as the dimensions of the input (,i.e. $m = d$). Because the scales of the attributes vary, we normalize all of the data sets in the pre-processing stage to prevent the side effect encountered by attributes with differing variances significantly. The empirical results of 100 trials are shown in Tables 6 to 10.

First, we compare the information loss of PRIMP with that of the LHS-based approach, the Cholesky-based approach, and sampling under the multivariate normal assumption. As expected, the Cholesky-based approach preserves the correlation structure of the original data set exactly. However, the probability distribution of the synthetic data generated by the Cholesky-based approach depends on the seeds generated; thus, the output of the approach does not always preserve the distribution of the original data successfully, as shown in Figure 3(b). Compared to the relative biases of Spearman's rank correlation matrix and Kendall's Tau, the performance of PRIMP is better than those approaches in most cases. Even though data set 12 violates the non-Gaussian restriction of the linear ICA model, PRIMP outperforms the other approaches.

We also compare the hybrid method with the other approaches in terms of data utility. Although the results show that the hybrid approach is not always better than PRIMP for Spearman's rank correlation matrix and Kendall's tau matrix, like Cholesky-based approach, it can exactly preserve the covariance matrix of the original data as shown in Table 7. In the other words, the hybrid method significantly outperforms PRIMP in terms of preserving the correlation matrix. Also, as shown in Figures 10, the hybrid method solves the problem of generating seeds for the Cholesky-based approach. Furthermore, in theory, the leakage risk of the hybrid approach is less than that of PRIMP. Therefore, it is also useful for privacy preserving.

6. CONCLUSION

We have proposed two new simulation-based privacy preserving frameworks for multivariate numerical data sets. The

Table 5: The data sets used in our experiments.

Index	Name of Database	# Records	# Attributes	Source
1	American Housing Survey	20000	9	U. S. Census Bureau
2	Bank Database	50000	5	K. Muralidhar and R. Sarathy
3	Data Shuffling	50000	3	K. Muralidhar and R. Sarathy
4	Iris Plant Database	150	4	UCI machine learning repository
5	Ecoli Database	336	5	UCI machine learning repository
6	Ionosphere Database	351	32	UCI machine learning repository
7	Housing Database	351	32	UCI machine learning repository
8	Pima Indians Diabetes Database	768	8	UCI machine learning repository
9	Yeast Database	1484	6	UCI machine learning repository
10	New diagnostic database	569	30	UCI machine learning repository
11	ACAS	1080	13	U. S. Census Bureau
12	Simulated	2000	3	Simulated by model (5)

Table 6: The average relative biases of Pearson's correlation matrix, Spearman's rank correlation matrix, and Kendall's tau matrix based on PRIMP over 100 trials.

Index	Correlation	Rank correlation	Kendall's tau
1	2.202×10^{-4}	1.700×10^{-3}	1.060×10^{-2}
2	4.193×10^{-5}	3.719×10^{-4}	6.274×10^{-4}
3	1.174×10^{-4}	5.713×10^{-4}	1.400×10^{-3}
4	6.462×10^{-4}	3.936×10^{-4}	6.599×10^{-4}
5	7.088×10^{-4}	1.800×10^{-3}	1.700×10^{-3}
6	8.000×10^{-3}	1.190×10^{-2}	8.600×10^{-3}
7	7.400×10^{-3}	9.700×10^{-3}	8.500×10^{-3}
8	1.600×10^{-3}	6.700×10^{-2}	7.200×10^{-3}
9	1.000×10^{-3}	6.780×10^{-3}	7.300×10^{-3}
10	8.200×10^{-3}	2.900×10^{-3}	1.270×10^{-3}
11	2.400×10^{-3}	2.400×10^{-3}	3.100×10^{-3}
12	1.707×10^{-6}	8.544×10^{-6}	3.002×10^{-5}

Table 7: The average relative biases of Pearson's correlation matrix, Spearman's rank correlation matrix, and Kendall's tau matrix based on the hybrid PRIMP and Cholesky-based approach over 100 trials.

Index	Correlation	Rank correlation	Kendall's tau
1	6.432×10^{-17}	1.800×10^{-3}	8.800×10^{-2}
2	9.280×10^{-17}	3.979×10^{-4}	8.784×10^{-4}
3	4.693×10^{-17}	6.802×10^{-4}	1.300×10^{-3}
4	3.382×10^{-18}	4.0761×10^{-4}	5.266×10^{-4}
5	8.232×10^{-18}	1.400×10^{-3}	1.400×10^{-3}
6	2.468×10^{-17}	9.900×10^{-3}	1.840×10^{-2}
7	3.143×10^{-17}	8.700×10^{-3}	1.710×10^{-2}
8	1.501×10^{-17}	2.310×10^{-2}	8.5×10^{-2}
9	1.525×10^{-17}	9.600×10^{-3}	1.380×10^{-2}
10	3.257×10^{-16}	3.300×10^{-3}	2.300×10^{-3}
11	2.379×10^{-4}	2.900×10^{-3}	3.300×10^{-3}
12	5.530×10^{-18}	7.979×10^{-6}	1.159×10^{-5}

Table 8: The average relative biases of Pearson's correlation matrix, Spearman's rank correlation matrix, and Kendall's tau matrix based on the multivariate normal assumption over 100 trials (average/standard deviation).

Index	Correlation	Rank correlation	Kendall's tau
1	2.481×10^{-4}	3.900×10^{-3}	1.190×10^{-2}
2	8.722×10^{-5}	1.100×10^{-3}	2.500×10^{-3}
3	4.729×10^{-5}	2.900×10^{-3}	6.100×10^{-3}
4	8.761×10^{-4}	9.612×10^{-4}	1.500×10^{-3}
5	2.000×10^{-3}	3.800×10^{-3}	3.600×10^{-3}
6	8.000×10^{-3}	1.320×10^{-2}	1.790×10^{-2}
7	7.600×10^{-3}	1.070×10^{-2}	1.510×10^{-2}
8	2.700×10^{-3}	3.200×10^{-2}	1.090×10^{-2}
9	1.800×10^{-3}	1.180×10^{-2}	1.110×10^{-2}
10	7.100×10^{-3}	3.100×10^{-3}	1.500×10^{-3}
11	4.200×10^{-3}	3.300×10^{-3}	4.300×10^{-3}
12	1.179×10^{-5}	1.233×10^{-5}	5.691×10^{-5}

Table 9: The average relative biases of Pearson's correlation matrix, Spearman's rank correlation matrix, and Kendall's tau matrix based on the LHS-based approach over 100 trials (average/standard deviation).

Index	Correlation	Rank correlation	Kendall's tau
1	2.300×10^{-3}	4.100×10^{-3}	1.430×10^{-2}
2	9.809×10^{-4}	1.100×10^{-3}	2.100×10^{-3}
3	3.000×10^{-3}	2.900×10^{-3}	6.100×10^{-3}
4	3.273×10^{-4}	1.200×10^{-3}	1.700×10^{-3}
5	1.200×10^{-3}	2.100×10^{-3}	2.100×10^{-3}
6	4.400×10^{-3}	1.112×10^{-3}	1.630×10^{-2}
7	4.400×10^{-3}	9.300×10^{-3}	1.250×10^{-2}
8	1.700×10^{-3}	5.890×10^{-2}	7.700×10^{-3}
9	1.500×10^{-3}	6.800×10^{-3}	6.300×10^{-3}
10	5.500×10^{-3}	2.800×10^{-3}	1.400×10^{-3}
11	1.300×10^{-3}	2.500×10^{-3}	3.300×10^{-3}
12	4.345×10^{-6}	2.253×10^{-5}	8.044×10^{-5}

Table 10: The average relative biases of Pearson's correlation matrix, Spearman's rank correlation matrix, and Kendall's tau matrix based on the Cholesky-based approach over 100 trials (average/standard deviation).

Index	Correlation	Rank correlation	Kendall's tau
1	4.406×10^{-17}	4.000×10^{-3}	1.460×10^{-2}
2	4.937×10^{-17}	2.000×10^{-3}	2.700×10^{-3}
3	5.124×10^{-17}	2.900×10^{-3}	6.100×10^{-3}
4	3.250×10^{-18}	1.000×10^{-3}	1.900×10^{-3}
5	3.704×10^{-18}	1.900×10^{-3}	1.900×10^{-3}
6	9.552×10^{-18}	9.400×10^{-3}	9.200×10^{-3}
7	1.079×10^{-17}	9.100×10^{-3}	9.100×10^{-3}
8	1.007×10^{-17}	5.000×10^{-2}	8.000×10^{-3}
9	7.393×10^{-18}	7.700×10^{-3}	7.000×10^{-3}
10	1.072×10^{-17}	2.500×10^{-3}	1.300×10^{-3}
11	1.035×10^{-17}	2.800×10^{-3}	3.800×10^{-3}
12	8.279×10^{-18}	7.578×10^{-5}	1.146×10^{-4}

first is called PRIMP (PRivacy preserving by Independent coMPonents). It is shown empirically that the synthetic data set generated by PRIMP can preserve more statistical characteristics of the original data set than previous approaches. Furthermore, we have proved that the synthetic data generated by PRIMP is sufficiently different from the original data; thus, PRIMP can protect the privacy of the original data. The PRIMP algorithm is easy to implement and very effective because, by using FastICA, its run time is linear to the size of the input. We have also proposed a hybrid method that combines PRIMP and the Cholesky-based approach for privacy preservation. The proposed method resolves the problem of generating good seeds for the Cholesky-based approach. Although the empirical results show that the hybrid approach is not always better than PRIMP in terms of Spearman's rank correlation matrix and Kendall's tau matrix for preserving privacy, like the Cholesky-based approach, it can preserve the covariance matrix of the original data exactly. Finally, in theory, the leakage risk of the hybrid approach is less than that of PRIMP. Overall, it is also useful for privacy preservation.

7. REFERENCES

[1] N. R. Adam and J. C. Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, 21(4):515–556, 1989.

[2] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 439–450, New York, NY, USA, 2000. ACM Press.

[3] R. Boscolo, H. Pan, and V. P. Roychowdhury. Independent component analysis based on nonparametric density estimation. *IEEE Transactions on Neural Networks*, 15(1):55–65, 2004.

[4] K. Chen and L. Liu. Privacy preserving data classification with rotation perturbation. In *ICDM*, pages 589–592, 2005.

[5] U. Cherubini, E. Luciano, and W. Vecchiato. *Copula Methods in Fiance.* Wiley, 2004.

[6] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms.* The MIT Press, 2 edition, 2001.

[7] G. Crises. Synthetic microdata generation for database privacy protection. Technical Report CRIREP-04-2004, The CRISES research group, 2004.

[8] R. A. Dandekar, M. Cohen, and N. Kirkendall. Sensitive micro data protection using latin hypercube sampling technique. In *Inference Control in Statistical Databases, From Theory to Practice*, pages 117–125, London, UK, 2002. Springer-Verlag.

[9] R. A. Dandekar, J. Domingo-Ferrer, and F. Sebé. Lhs-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In *Inference Control in Statistical Databases, From Theory to Practice.*

[10] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application.* Cambridge University Press, 1997.

[11] L. Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, 1986.

[12] C. B. D.J. Newman, S. Hettich and C. Merz. UCI repository of machine learning databases, 1998.

[13] J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure control methods for microdata. In *confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 111–134. North-Holland, 2001.

[14] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[15] Z. Huang, W. Du, and B. Chen. Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 37–48, New York, NY, USA, 2005. ACM Press.

[16] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[17] R. L. Iman and W. J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics*, 11(3):311–334, 1982.

[18] I. T. Jollife. *Principle Component Analysis.* Springer-Verlag, New York, 2 edition, 2002.

[19] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *Proceedings of the Third IEEE International Conference on Data Mining*, pages 19–22, Washington, DC, USA, 2003. IEEE Computer Society.

[20] C. K. Liew, U. J. Choi, and C. J. Liew. A data distortion by probability distribution. *ACM Transactions Database Systems*, 10(3):395–411, 1985.

[21] J. M. Mateo-Sanz, A. Martínez-Ballesté, and J. Domingo-Ferrer. Fast generation of accurate synthetic microdata. In *Privacy in Statistical Databases*, pages 298–306, 2004.

[22] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[23] K. Muralidhar and R. Sarathy. Security of random data perturbation methods. *ACM Transactions Database Systems*, 24(4):487–493, 1999.

[24] K. Muralidhar and R. Sarathy. Data shuffling - a new masking approach for numerical data. *Management Science*, 52(5):658–670, 2006.

[25] S. Ross. *A First Course in Probability.* Prentice Hall, 2002.

[26] G. G. Roussas. *A Course in Mathematical Statistics.* Academic Press, 1997.

[27] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, New York, 1992.

[28] A. C. Tamhane and D. D. Dunlop. *Statistics and Data Analysis: from Elementary to Intermediate.* Prentice Hall, 2000.

[29] P. Tendick and N. Matloff. A modified random perturbation method for database security. *ACM Transactions Database Systems*, 19(1):47–63, 1994.

Appendix: Proofs of Leakage Risk and Time Complexity of PRIMP

In Section 3.4, we derived some theoretical properties of PRIMP. Explicitly, in Theorem 1, the distribution of records leaked by using PRIMP were derived. The theorem states that the synthetic data sets generated by PRIMP are sufficiently different from the original data set; thus, they are able to protect the privacy of the original data. In addition, we discuss the time complexity of PRIMP in Theorem 3. However, in Section 3.4, the proofs of those theorems were omitted for interest of space. Here, we present the full proofs of those theorems.

Proof of Theorem 1. As mentioned in Section 3.3, the permutation stage of PRIMP randomly shuffles the individual source signals to protect an individual's privacy. The stage is equivalent to the following process in that we fix the first row of the source signals matrix S and randomly permutate the elements of each row of S except S_1. Thus, the privacy leakage problem can be viewed as an extension of the matching problem ([25], page 42).

(1) Let us denote E_i, $i = 1, 2, \cdots, n$, as an event where the ith record of X has been leaked by PRIMP. Furthermore, let $E_{i_1}E_{i_2}\cdots E_{i_l}$ be events where the respective l records $i_1, i_2, \cdots, i_l$ have been leaked by PRIMP. Hence, based on random permutation, we have

$$\mathcal{P}\{E_{i_1}E_{i_2}\cdots E_{i_l}\} = \frac{((n-l)!)^{m-1}}{(n!)^{m-1}}.$$

Also, as there are $\binom{n}{l}$ terms in $\sum_{i_1,i_2,\cdots,i_l}\mathcal{P}\{E_{i_1}E_{i_2}\cdots E_{i_l}\}$, we observe that

$$\sum_{i_1,i_2,\cdots,i_l}\mathcal{P}\{E_{i_1}E_{i_2}\cdots E_{i_l}\} = \frac{1}{l!}\left(\frac{(n-l)!}{n!}\right)^{m-2}.$$

Thus, by the inclusion-exclusion property of the n union of n events ([25], page 34), the probability that at least one of the records has been leaked by PRIMP is given by

$$\begin{aligned}\mathcal{P}\{\bigcup_{i=1}^{n} E_i\} =& \sum_{i=1}^{n}\mathcal{P}\{E_i\} - \sum_{i_1<i_2}\mathcal{P}\{E_{i_1}E_{i_2}\} + \cdots + \\ (-1)^{l+1}&\sum_{i_1,i_2,\cdots,i_l}\mathcal{P}\{E_{i_1}E_{i_2}\cdots E_{i_l}\} + \cdots + (-1)^{n+1}\mathcal{P}\{E_1E_2\cdots E_n\} \\ =& \sum_{l=1}^{n}(-1)^{l+1}R_m(l),\end{aligned}$$

where $R_m(l) = \frac{1}{l!}\left(\frac{(n-l)!}{n!}\right)^{m-2}$. Hence, the probability that no records of X have been leaked is

$$1 - \sum_{l=1}^{n}(-1)^{l+1}R_m(l).$$

To obtain the probability that exactly k of the n records of X have been leaked, we first focus on a particular set of k records. The number of ways in which these and only these k records could have been leaked by PRIMP is equal to the number of ways in which none of the other $(n-k)$ records could have been leaked. Since

$$1 - \sum_{l=1}^{n-k}(-1)^{l+1}R_m(l)$$

is the probability that no $(n-k)$ records have been leaked, for the selections of k leaked records under consideration, there are

$$\binom{n}{k}(n-k)!\left[1 - \sum_{l=1}^{n-k}(-1)^{l+1}R_m(l)\right]$$

ways in which exactly k of the records could have been leaked. Thus,

$$\begin{aligned}\mathcal{P}\{\mathcal{N} = k\} &= \frac{\binom{n}{k}(n-k)!}{n!}\left[1 - \sum_{l=1}^{n-k}(-1)^{l+1}R_m(l)\right] \\ &= \frac{1}{k!}\left[1 - \sum_{l=1}^{n-k}(-1)^{l+1}R_m(l)\right].\end{aligned}$$

(2) For $i = 1, 2, \cdots, n$, let $\mathbf{1}_i$ be the binomial random variable defined as

$$\mathbf{1}_i = \begin{cases} 1 & \text{if the event } E_i \text{ occurs,} \\ 0 & \text{otherwise.}\end{cases}$$

Now, since $\mathcal{P}\{\mathbf{1_i} = 1\} = (1/n)^{m-1}$, then $E[\mathbf{1}_i] = (1/n)^{m-1}$ and $var(\mathbf{1}_i) = (1/n)^{m-1}[1 - (1/n)^{m-1}]$. Also, $E[\mathbf{1}_i\mathbf{1}_j] = \mathcal{P}\{\mathbf{1}_i = 1, \mathbf{1}_j = 1\} = \mathcal{P}\{\mathbf{1}_i = 1\}\mathcal{P}\{\mathbf{1}_j = 1|\mathbf{1}_i = 1\} = (\frac{1}{n})^{m-1}(\frac{1}{n-1})^{m-1}$.

Hence,

$$\begin{aligned}cov(\mathbf{1}_i, \mathbf{1}_j) &= E[\mathbf{1}_i\mathbf{1}_j] - E[\mathbf{1}_i]E[\mathbf{1}_j] \\ &= \left(\frac{1}{n}\right)^{m-1}\left[\left(\frac{1}{n-1}\right)^{m-1} - \left(\frac{1}{n}\right)^{m-1}\right].\end{aligned}$$

Therefore,

$$E[\mathcal{N}] = \sum_{i=1}^{n}E[\mathbf{1}_i] = \left(\frac{1}{n}\right)^{m-2},$$

and

$$\begin{aligned}var(\mathcal{N}) &= var(\sum_{i=1}^{n}\mathbf{1}_i) \\ &= \sum_{i=1}^{n}var(\mathbf{1}_i) + 2\sum_{i<j}\sum cov(\mathbf{1}_i, \mathbf{1}_j) \\ &= \left(\frac{1}{n}\right)^{m-2}\left(\frac{n^{m-1}-1}{n^{m-1}}\right) + \\ 2&\binom{n}{2}\left(\frac{1}{n}\right)^{m-1}\left[\left(\frac{1}{n-1}\right)^{m-1} - \left(\frac{1}{n}\right)^{m-1}\right] \\ &= \left(\frac{1}{n}\right)^{m-2}\left[1 - \frac{1}{n^{m-2}} + \frac{1}{(n-1)^{m-2}}\right].\end{aligned}$$

□

Proof of Theorem 2. According to Theorem 1, we know that the probability that no records have been leaked by PRIMP is $\mathcal{P}\{\mathcal{N} = 0\} = 1 - \sum_{l=1}^{n}(-1)^{l+1}R_m(l)$. Therefore, the leakage risk of PRIMP is $\sum_{l=1}^{n}(-1)^{l+1}R_m(l)$. □

Proof of Theorem 3. Clearly, both the normalization cost and the rescaling cost are $O(nd)$. Since the matrix product cost in the coupling stage is $O(nm^2)$, the cost of the coupling stage is $O(nm^2)$. Furthermore, the random permutation can be implemented in $O(n)$ time ([6], page102), so the cost of the permutation stage is $O(nm)$. The remainder of evaluating the cost is to analyze the computational complexity of the decoupling stage.

The first step of the FastICA algorithm decorrelates the inputs [16], which costs $O(nd^2)$ time. Essentially, the remaining processes of FastICA relate to the optimization problem. Formally, FastICA tries to find the best solution in the feasible region that has the minimum (or maximum) value of the objective function[13]. The implementation of FastICA requires a number of floating-point operations proportional to $O(nm)$ time to evaluate the object function and proportional to $O(nm^2)$ time to compute its derivatives [3]. Therefore, the overall time complexity of FastICA is $O(Knd^2)$, where K is the total number of iterations required to evaluate the cost function. □

[13] The selection strategies of object function are beyond the scope of this paper. Hyvärinen et al.'s book [16] discuss the differences on the results when using different objective functions in the FastICA algorithm. We refer the readers to Hyvärinen et al.'s book for further details.