数理解析研究所講究録
第1620巻 2009年 162-173



# Privacy Preserving Independent Component Analysis

Natsuki Sano and Yasusi Sinohara
System Engineering Research Laboratory
Central Research Institute of Electric Power Industry

**Abstract**

Privacy Preserving Data Mining is the methodology used to discover knowledge of distributed databases with providing guarantees concerning the non-disclosure of the data. We will propose the Privacy Preserving Independent Component Analysis (PPICA) in order to conduct ICA in a privacy preserving manner. The proposed method can compute a separating matrix in order to reconstruct sources from all databases without disclosing the data of each database to others. Although the issue in PPICA is the significant communication traffic, the proposed method overcome this and has property equivalent to low communication traffic. We assume that the database is horizontally partitioned, whereby each database has the same attributes in common for different samples.

**Keywords:** Privacy Preserving Data Mining, Independent Component Analysis, Principal Component Analysis, Cryptographic Communication.

## 1 Introduction

Data Mining is a technique used to discover useful knowledge from significant amounts of data. Generally, the larger amount of data we can use, the more informative results we can obtain. Consider the situation whereby certain databases are distributed at local sites. If several databases are available for data analysis, the results will be more informative than analysis using a single database.

However this is usually difficult, especially where the databases are owned by others. Data in a database cannot be provided in certain cases where there are contractual restrictions or in others, where the owner prefers to keep the data content secret to others from operational perspectives.

If there are no restrictions in terms of confidentiality, all the databases are merged and allowing analysts at local sites to analyze the merged database and obtain the results. Privacy Preserving Data Mining (PPDM) aims to obtain the same results discovered using the merged database without leaking the data of each database to analysts at other sites. The key technique used to achieve these seemingly contradictory requirements simultaneously is "to embed the cryptographic techniques into the target data analysis algorithms". Therefore privacy preserving data-mining algorithms are algorithm-specific.

In this paper, we will present an efficient Privacy Preserving Independent Component Analysis (PPICA) on a horizontally partitioned database setting, whereby each database has the same attributes in common for different samples. Efficient privacy preserving algorithms are proposed for major data analysis algorithms such as linear regressions [2], k-means clustering [7] and Support Vector Machines [8], however,

the privacy preserving algorithm for Independent Component Analysis (ICA) is never shown. Independent Component Analysis is a method used to recover "sources" from observation data, based on the assumption that observation is a mixture of "independent sources", while Principal Component Analysis (PCA) merely assumes that observation is a mixture of "uncorrelated sources". ICA can be viewed as a special type of factor analysis, which has various application fields such as Image Processing, Signal Processing and Psychology.

The difficulty in designing an efficient privacy preserving ICA is its high communication traffic. The majority of ICA algorithms repeatedly involve reconstructing tentative "sources" from the original data. In the privacy preserving setting, the original data are saved at each site and they cannot be transferred elsewhere. Therefore the privacy preserving version of these algorithms must communicate repeatedly with the local site. This has two major downsides, namely, the high communication costs and the potentially high risk of leakage of original data. We will solve this issue through a Two Line ICA, which executes ICA via twice Eigen Value Decomposition. The proposed PPICA algorithm is one with low communication traffic.

In Section 2, we will explain ICA and PCA which is executed as a preprocessing algorithm of ICA. In Section 3, the basic cryptographic protocol "Secure Sum" is introduced and the Privacy Preserving PCA will be explained. In Section 4, the issue of PPICA is raised and the Two Line ICA and the proposed PPICA algorithm will be explained. In Section 5, the conclusion and issues for future research are described.

# 2 PCA and ICA

## 2.1 Principal Component Analysis

Principal Component Analysis (PCA) [4] is one of the oldest and best known techniques used for multivariate analysis. To compare with Independent Component Analysis (ICA), we will view PCA as a method to recover sources from observations based on the assumption that observations are mixtures of uncorrelated sources with significant variances.

PCA and ICA assume the basic linear statistical model

$$\mathbf{x} = \mathbf{B}\mathbf{s}, \tag{1}$$

in which $\mathbf{x} \in \mathbb{R}^d$ represents the observations, and $\mathbf{s} \in \mathbb{R}^d$ the sources. For the sake of convenience, we assume the dimension of observations and sources to be equal and denoted by $d$. $\mathbf{B} \in \mathbb{R}^{d \times d}$, meanwhile, represents 'mixing matrix', and its entries $b_{ij}$ indicate the extent to which the $j$th source component contributes to the $i$th observation channel ($1 \leq i,\ j \leq d$), i.e. they determine how the sources are 'mixed' in the observations.

PCA estimates the mixing matrix $\mathbf{B}$ and/or the corresponding realizations of the sources $\mathbf{s}$, given only realizations of the observations $\mathbf{x}$, and under the following assumptions:

- The components of $\mathbf{s}$ are mutually uncorrelated, i.e. $\mathrm{cov}(\mathbf{s}) = \mathbf{D}$ is diagonal.
- Sources are zero-mean, i.e. $E(\mathbf{s}) = \mathbf{0}$.

Source signals are reconstructed from the observations as follows:

$$\mathbf{s} = \mathbf{U}^T(\mathbf{x} - \bar{\mathbf{x}}), \tag{2}$$

where $\bar{\mathbf{x}}$ is the mean of the observations and $\mathbf{U} \in \mathbb{R}^{d \times d}$ is the rotation matrix such that $\mathbf{U}\mathbf{U}^T = \mathbf{I}_d$ where $I_d$ is a $d \times d$ identity matrix.

The rotation matrix $\mathbf{U}$ is computed by the Eigen Value Decomposition (EVD) of the covariance matrix of observations $\mathbf{C}_\mathbf{x} \in \mathbb{R}^{d \times d}$ as follows:

$$\mathbf{C}_\mathbf{x} = \mathbf{U}\mathbf{D}\mathbf{U}^T. \tag{3}$$

The mean vector and covariance matrix are computed from the realized observations matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ as follows:

$$\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i / n, \quad \mathbf{C}_\mathbf{x} = (\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T / n, \tag{4}$$

where $\bar{\mathbf{X}} = [\bar{\mathbf{x}}, \bar{\mathbf{x}}, \ldots, \bar{\mathbf{x}}] \in \mathbb{R}^{d \times n}$ and $n$ represent the mean matrix and the number of observations respectively.

## 2.2 Independent Component Analysis

### 2.2.1 Problem Setting

In the same way in PCA, ICA estimates the mixing matrix $\mathbf{B}$ and/or the corresponding realizations of the sources $\mathbf{s}$, given only realizations of the observations $\mathbf{x}$, under the linear model by Equation (1).

However, ICA assumes the following assumptions (cf. Assumption of PCA):

- The components of $\mathbf{s}$ are mutually statistically independent.
- Sources are zero-mean and unit variances, i.e. $E(\mathbf{s}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{s}) = \mathbf{I}_d$.

The mixing matrix, $\mathbf{B}$, is estimated and the source signals will be constructed from the observations as follows:

$$\mathbf{s} = \mathbf{B}^{-1}(\mathbf{x} - \bar{\mathbf{x}}), \tag{5}$$

where $\mathbf{B}^{-1}$ is called a separating matrix.

In contrast to PCA, ICA assumes the observations are mixtures of "independent" sources. It is noted that if data are independent, then they are "uncorrelated" but not vice versa. Therefore ICA imposes a stronger assumption than PCA with respect to independence.

### 2.2.2 Outline of Popular ICA Algorithms

In this subsection, we will explain ICA in general terms. The ICA problem is most often solved by a two stage algorithm consisting of a whitening (PCA) stage and a finding proper rotation stage. An outline of the procedure is presented in Figure 1.

#### Whitening Stage

The whitening stage amounts to a principal component analysis (PCA) of the observations. Briefly, the goal is to transform the observations $\mathbf{x}$ into another vector $\mathbf{y}$ having a unit covariance matrix. This involves the multiplication of $\mathbf{y}$ with the inverse of the square root of its covariance matrix $\mathbf{C_y} \in \mathbb{R}^{d \times d}$.

Firstly, we observe that the covariance matrix $\mathbf{C_x}$ takes the form

$$\mathbf{C_x} = \mathbf{B}\mathbf{C_s}\mathbf{B}^T \tag{6}$$

in which the covariance of $\mathbf{s}$, $\mathbf{C_s} \in \mathbb{R}^{d \times d}$, is diagonal, since, the sources are uncorrelated. Based on the assumption that the sources have unit variance, we have

$$\mathbf{C_x} = \mathbf{B}\mathbf{B}^T. \tag{7}$$

Substitution of the SVD of the mixing matrix $\mathbf{B} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^T$ shows that the EVD of the observed covariance (PCA) allows us to estimate the components $\mathbf{U}$ and $\mathbf{D}$ whilst the rotation matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$ remains unknown;

$$\mathbf{C_x} = \mathbf{U}\mathbf{D}\mathbf{U}^T = (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{U}\mathbf{D}^{1/2})^T. \tag{8}$$

Hence $\mathbf{U}$ and $\mathbf{D}$ can be estimated from the second-order statistics of the observations, but the actual mixing matrix remains unknown up to an orthogonal factor.

Based on the inverse of the square root of $\mathbf{D}$ and $\mathbf{U}$, a whitened vector $\mathbf{y}$ can be defined as:

$$\mathbf{y} = \mathbf{D}^{-1/2}\mathbf{U}^T\mathbf{x}. \tag{9}$$

#### Finding the Proper Rotation Stage

If the proper rotation matrix $\mathbf{V}$ is decided, we can find the mixing matrix $\mathbf{B} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^T$. In general, the proper rotation matrix is estimated iteratively.

First, tentative sources are computed based on a tentative rotation matrix as follows:

$$\mathbf{s} = \mathbf{V}\mathbf{y}. \tag{10}$$

A certain independence measure of the sources $L(\mathbf{s})$ is evaluated based on tentative sources and $\mathbf{V}$ is updated in the direction in order to improve the independence measure. This procedure is repeated until $L(\mathbf{s})$ converges.

In popular ICA algorithms, as an independent measure of the sources $L(\mathbf{s})$, Amari et.al [1] take Kullback-Leibler divergence, Hyvärinen et al (FastICA) [3] take 4-*th* order kurtosis and Jutten et al [5] take non-linear cross-correlations.

**Step 1** Whitening Stage (PCA)

(a) Compute the sample mean $\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{n}$ from $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$.

(b) Compute the covariance matrix $\mathbf{C_x} = (\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T / n \in \mathbb{R}^{d \times d}$, where $\bar{\mathbf{X}} = [\bar{\mathbf{x}}, \bar{\mathbf{x}}, \ldots, \bar{\mathbf{x}}] \in \mathbb{R}^{d \times n}$.

(c) Eigen Value Decomposition of $\mathbf{C_x}$ into $\mathbf{UDU}^T$ with $\mathbf{U}, \mathbf{D} \in \mathbb{R}^{d \times d}$.

(d) Whiten observed data $\mathbf{Y} = \mathbf{D}^{-1/2}\mathbf{U}^T(\mathbf{X} - \bar{\mathbf{X}}) \in \mathbb{R}^{d \times n}$.

**Step 2** Finding the Proper Rotation Stage

Initialize a rotation matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$.

**Repeat**

(a) Reconstruct the source data $\mathbf{S} = \mathbf{VY}$ where $\mathbf{S} \in \mathbb{R}^{d \times n}$.

(b) Compute the independence measure $L(\mathbf{S})$.

(c) Renew $\mathbf{V}$.

**Until** $L$ converges

Compute a mixture matrix $\mathbf{B} = \mathbf{UD}^{1/2}\mathbf{V}^T$.

Figure 1: Outline of popular ICA Algorithms

# 3 Privacy Preserving PCA

## 3.1 Secure Sum

As mentioned in Section 2, ICA must compute the EVD of the observed covariance during the whitening stage. In order to compute the covariance matrix securely, namely, where only the covariance matrix can be computed between sites and each of the observations is secret to the other sites, the use of a technique to compute sums securely is important.

Secure Sum is a technique used to compute sums securely, which was originally proposed by Schneier [6]. We will use Secure Sum to compute the covariance matrix and that which we will use is revised to be strong for collusion.

We will explain the (revised) Secure Sum algorithm. We assume there are $N$ sites and that the 1st site is the master site, which is faithful, conforms to the determined protocols and does not collude or engage in tapping. Other sites conform to the protocols but may collude with each other or tap the communication line.

**Step 1** Firstly, the master site generates one random number $R_i$ for each site $i$ and sends it to the corresponding site using SSL (secure socket layer) or other secure cryptographic communication channels.

**Step 2** Each site $i$ computes the encrypted value $E_i$ by adding its local data $Z_i$ to the

received random number $R_i$

$$E_i = Z_i + R_i \mod p,$$

where $p$ is assumed to be greater than $\sum_{i=1}^{N} Z_i$.

**Step 3** Site 1 sends its encrypted value $E_1$ to the next site 2. After Site 1, site $i$ sends the sum of the received data and its encrypted value, $\sum_{j=1}^{i-1} E_j + E_i \mod p$, to the next site $i+1$. The last site sends the sum of the received data and its encrypted value, $\sum_{j=1}^{N-1} E_j + E_N \mod p$, to the master site 1.

**Step 4** The value received at site 1 is the sum of the local values and the random numbers $\sum_{i=1}^{N} Z_i + \sum_{i=1}^{N} R_i \mod p$. The sum of random numbers is known to the master. Therefore, the sum of the local values can be obtained by subtracting the sum of the random numbers.

The figure in the case of $N = 4$ is shown in Figure 2.

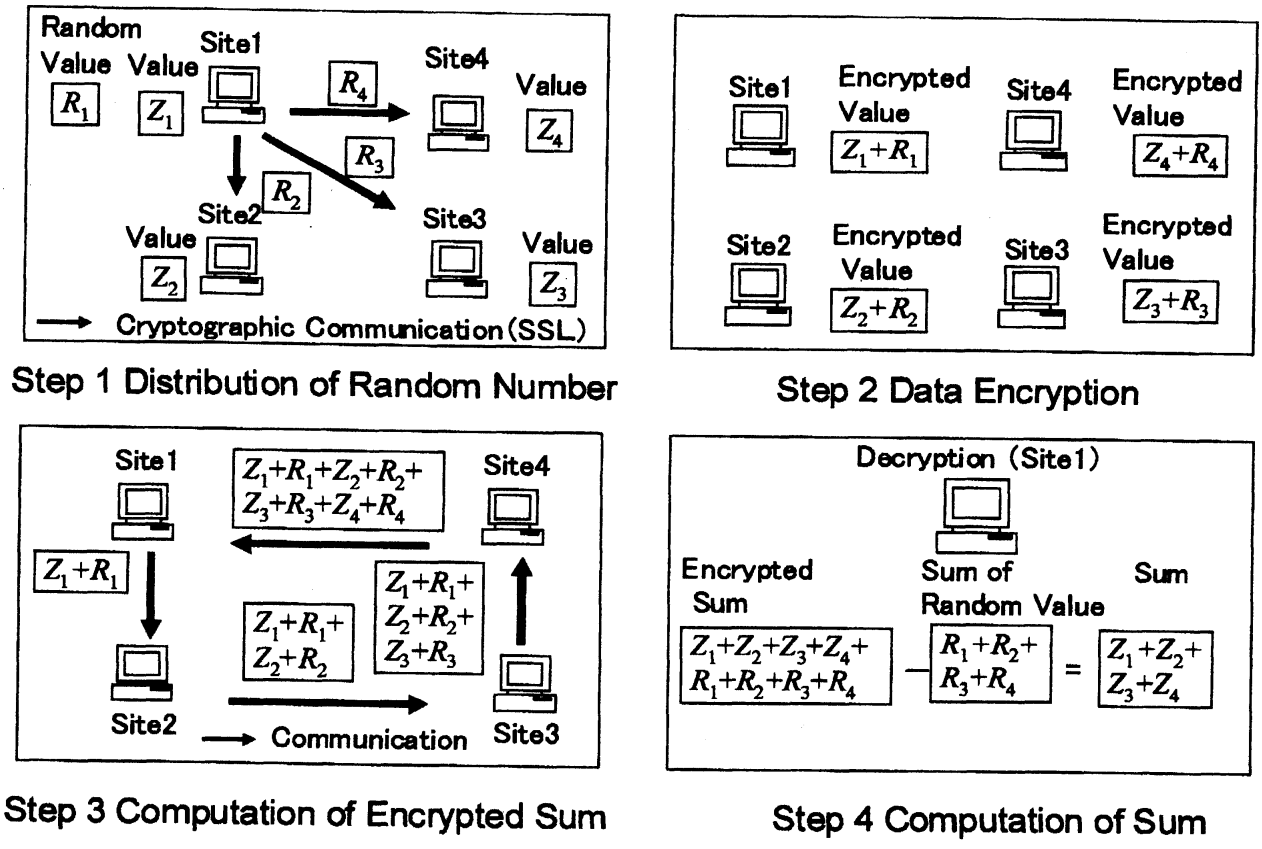


Figure 2: Secure Sum

**Safety of the Secure Sum**

As mentioned in the assumption of Secure Sum, if the master site conforms to the protocol and properly manages the encryption key $R_i$, this means local data are secure from tapping and colluding. This is because even if both of the received and sent data of the local site $i$ are known to someone, he or she will only be able to recover the encrypted value of site $i$,

$$\sum_{j=1}^{i} Z_j + R_j - \sum_{j=1}^{i-1} Z_j + R_j = Z_i + R_i,$$

since $R_i$ is the random number only known to Site $i$ and the master site.

## 3.2 Privacy Preserving PCA

In this Section, we will explain how to ensure PCA privacy is preserved, in a horizontally partitioned setting. We assume that each site $i$ has an observation matrix $\mathbf{X}_i \in \mathbb{R}^{d \times n_i}$. Using the Secure Sum mentioned in Section 3.1, each sites can whiten the local data with local data unknown to other sites. The procedure of privacy preserving PCA is as follows:

**Step 1** Each site $i$ computes the local number of observations, $n_i$, weighted local mean, $n_i\bar{\mathbf{x}}_i \in \mathbb{R}^d$ and the weighted local covariance matrix, $n_i\mathbf{C}_i \in \mathbb{R}^{d \times d}$.

**Step 2** Using Secure Sum, Site 1 computes the total number of observations, $n_m = \sum_{i=1}^{N} n_i$, the merged mean $\bar{\mathbf{x}}_m = \sum_{i=1}^{N} n_i\bar{\mathbf{x}}_i/n_m$ and the merged covariance matrix $\mathbf{C}_m = \sum_{i=1}^{N} n_i\mathbf{C}_i/n_m \in \mathbb{R}^{d \times d}$.

**Step 3** Site 1 computes an Eigen Value Decomposition of $\mathbf{C}_m = \mathbf{UDU}^T$ and the whitening matrix $\mathbf{W} = \mathbf{D}^{-1/2}\mathbf{U}^T$ and sends $\mathbf{W} \in \mathbb{R}^{d \times d}, \bar{\mathbf{x}}_m \in \mathbb{R}^d$ to other sites.

**Step 4** Each site $i$ whitens local data $\mathbf{Y}_i = \mathbf{W}(\mathbf{X}_i - \bar{\mathbf{X}}_m) \in \mathbb{R}^{d \times n_i}$, where $\bar{\mathbf{X}}_m = [\bar{\mathbf{x}}_m, \bar{\mathbf{x}}_m, \ldots, \bar{\mathbf{x}}_m] \in \mathbb{R}^{d \times n_i}$ denotes the merged mean matrix.

The figure in the case of $N = 4$ is shown in Figure 3. By this, each site can get $\mathbf{W}$ and its local data $\mathbf{Y}_i$.

# 4 Privacy Preserving ICA

## 4.1 Issue of Privacy Preserving ICA

In this Section, we will introduce the general outline of Privacy Preserving ICA (PPICA) and describe an issue of PPICA. The PPICA procedure based on popular ICA algorithms is shown as follows:

**Step 1** Whiten local data $\mathbf{Y}_i = \mathbf{D}^{-1/2}\mathbf{U}^T(\mathbf{X}_i - \bar{\mathbf{X}}_m) \in \mathbb{R}^{d \times n_i}$ (Privacy Preserving PCA)

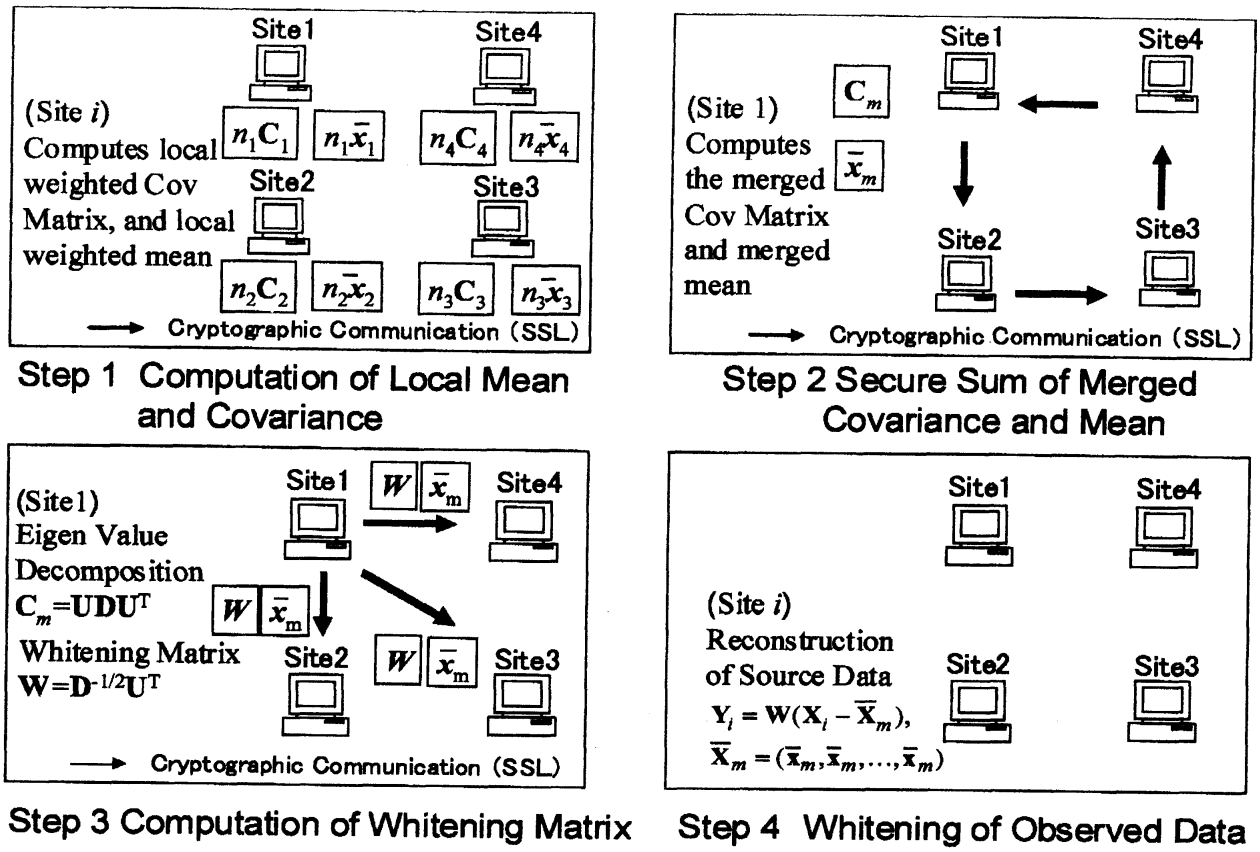


Figure 3: Privacy Preserving PCA

**Step 2** Site 1 initializes $\mathbf{V} \in \mathbb{R}^{d \times d}$, where, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_d$ and sends it to other sites.

**Step 3** Repeat

(a) Each site $i$ reconstructs the tentative sources $\mathbf{S}_i = \mathbf{V}\mathbf{Y}_i \in \mathbb{R}^{d \times n_i}$.

(b) Using Secure Sum, Site 1 computes a certain independence measure $L(\mathbf{S}_i)$ of sources in all sites.

(c) Site 1 renews $\mathbf{V}$, and sends it to other sites.

Until $L$ converges

**Step 4** Each site computes the separating matrix $\mathbf{B}^{-1} = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}^T$.

After whitening by Privacy Preserving PCA, the procedure for finding proper rotation matrix is repeatedly executed until some independence measure $L$ converges. In computing an independence measure $L(\mathbf{S}_i)$ of sources in all sites, the execution of Secure Sum is repeatedly required. This incurs a considerable communication cost and could also involve the potential risk of observations being leaked.

## 4.2 Two Line ICA

In order to avoid repetition when finding a proper rotation matrix, we use the Two Line ICA algorithm. The Two Line ICA is originally proposed by Weiss et al [9], because of its compactness. However its practical advantage has been never exploited.

Similarly to popular ICA algorithms, the Two Line ICA needs to whiten observations by PCA as preprocessing. The Two Line ICA algorithm can find a proper rotation matrix via once Eigen Value Decomposition. In total, the Two Line ICA can be executed by twice Eigen Value Decomposition without repeated evaluation of the independence measure $L$.

We extract three sources $s_{i_1}, s_{i_2}, s_{i_3}$, from $d$ independent sources $s_1, s_2, \ldots, s_d$ with replacement and assume that $E[s_i] = 0, E[s_i^2] = 1, i = 1, \ldots, d$. We also consider a moment matrix for the extracted three sources as follows:

$$M_{i_1 i_2} = \sum_{i_3} E[s_{i_1} s_{i_2} s_{i_3}^2] \quad \text{for} \quad i_1, i_2 = 1, \ldots, d \tag{11}$$

It is noted that, if $d$ sources are independent, then a moment matrix $M_{i_1 i_2}$ is a diagonal matrix. By replacing the expectation with the sample mean, a moment matrix for sources is as follows:

$$M(\mathbf{S}) = \mathbf{S}\text{diag}^{-1}\text{diag}(\mathbf{S}^T\mathbf{S})\mathbf{S}^T/n, \tag{12}$$

where $\mathbf{S} \in \mathbb{R}^{d \times n}$ represents the corresponding realizations of sources, given the observations. diag($\mathbf{A}$) returns the column vector of diagonal elements of $\mathbf{A}$ and $\text{diag}^{-1}(\mathbf{v})$ return the matrix whose diagonals are the elements of $v$.
By substituting $\mathbf{S}$ for $\mathbf{VY}$, we can find $M(\mathbf{S})$ has a relation with $M(\mathbf{Y})$ as follows:

$$M(\mathbf{S}) = \mathbf{V}M(\mathbf{Y})\mathbf{V}^T, \tag{13}$$

where $\mathbf{V} \in \mathbb{R}^{d \times d}$ is a rotation matrix and $\mathbf{Y} \in \mathbb{R}^{d \times n}$ is a whiten data matrix. We can obtain the rotation matrix $\mathbf{V}$, by Eigen Value Decomposition

$$M(\mathbf{Y}) = \mathbf{V}^T M(\mathbf{S})\mathbf{V}, \tag{14}$$

since we assume the sources to be independent and $M(\mathbf{S})$ to be diagonal. In total, we can obtain the rotation matrix $\mathbf{V}$ by twice Eigen Value Decomposition.

## 4.3 Privacy Preserving ICA

We propose PPICA based on Two Line ICA. PPICA based on popular ICA algorithms usually take more than twice the use of the Secure Sum. However the number of Secure Sums in the proposed PPICA is only twice. Therefore the proposed PPICA is the protocol with little communication traffic. The concrete procedure for privacy preserving ICA is as follows:

**Step 1** Each site $i$ computes a local number of observations, $n_i$ a weighted local mean, $n_i\bar{\mathbf{x}}_i \in \mathbb{R}^d$ and a weighted local covariance matrix, $n_i\mathbf{C}_i \in \mathbb{R}^{d \times d}$.

**Step 2** Using Secure Sum, Site 1 computes the total number of observations, $n_m = \sum_{i=1}^{N} n_i$, the merged mean $\bar{\mathbf{x}}_m = \sum_{i=1}^{N} n_i \bar{\mathbf{x}}_i / n_m \in \mathbb{R}^d$ and the merged covariance matrix $\mathbf{C}_m = \sum_{i=1}^{N} n_i \mathbf{C}_i / n_m \in \mathbb{R}^{d \times d}$.

**Step 3** Site 1 computes an Eigen Value Decomposition of $\mathbf{C}_m = \mathbf{U}\mathbf{D}\mathbf{U}^T$ and the whitening matrix $\mathbf{W} = \mathbf{D}^{-1/2}\mathbf{U}^T$ and sends $\mathbf{W} \in \mathbb{R}^{d \times d}, \bar{\mathbf{x}}_m \in \mathbb{R}^d$ to other sites.

**Step 4** Each site $i$ whitens the local data $\mathbf{Y}_i = \mathbf{W}(\mathbf{X}_i - \bar{\mathbf{X}}_m) \in \mathbb{R}^{d \times n_i}$, where $\bar{\mathbf{X}}_m = [\bar{\mathbf{x}}_m, \bar{\mathbf{x}}_m, \ldots, \bar{\mathbf{x}}_m] \in \mathbb{R}^{d \times n_i}$.

**Step 5** Each site $i$ computes a local moment matrix, $\mathbf{M}_i = \mathbf{Y}_i \mathrm{diag}^{-1} \mathrm{diag}(\mathbf{Y}_i^T \mathbf{Y}_i) \mathbf{Y}_i^T \in \mathbb{R}^{d \times d}$, where $\mathbf{Y}_i \in \mathbb{R}^{d \times n_i}$ is a local whiten data matrix.

**Step 6** Using Secure Sum, site 1 computes the merged moment matrix, $\mathbf{M}_m = \sum_{i=1}^{N} \mathbf{M}_i \in \mathbb{R}^{d \times d}$.

**Step 7** Site 1 computes the Eigen Value Decomposition of $\mathbf{M}_m = \mathbf{V}^T \mathbf{H} \mathbf{V}$ and the separating matrix $\mathbf{B}^{-1} = \mathbf{V}\mathbf{W}$ and sends $\mathbf{B}^{-1} \in \mathbb{R}^{d \times d}$ to other sites.

**Step 8** Each site $i$ computes and the local source matrix $\mathbf{S} = \mathbf{B}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}_i) \in \mathbb{R}^{d \times n_i}$, where $\bar{\mathbf{X}}_i = [\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_i, \ldots, \bar{\mathbf{x}}_i] \in \mathbb{R}^{d \times n_i}$.

The procedure from Steps 1 to Step 4 is the whitening stage, while the same procedure as PPPCA and Secure Sum is used in Step 2. The procedure from Steps 5 to Step 8 is the finding proper rotation matrix stage, while a Secure Sum is used in Step 6. We can find that the total use of the Secure Sum is only twice. The figure in the case of $N = 4$ is shown in Figure 4.

# 5 Conclusion

We have proposed a new PPICA based on Two Line ICA. Two Line ICA was originally proposed due to its compactness, however, its practical advantage has never been exploited. In this paper, its practical advantage has been shown for the first time. The proposed Secure Sum is revised to withstand collusion by assuming strong faith on one of the sites. Therefore the proposed PPICA features low communication traffic and the ability to withstand collusion.

Communication experiments of the proposed PPICA protocol are an issue for future research.

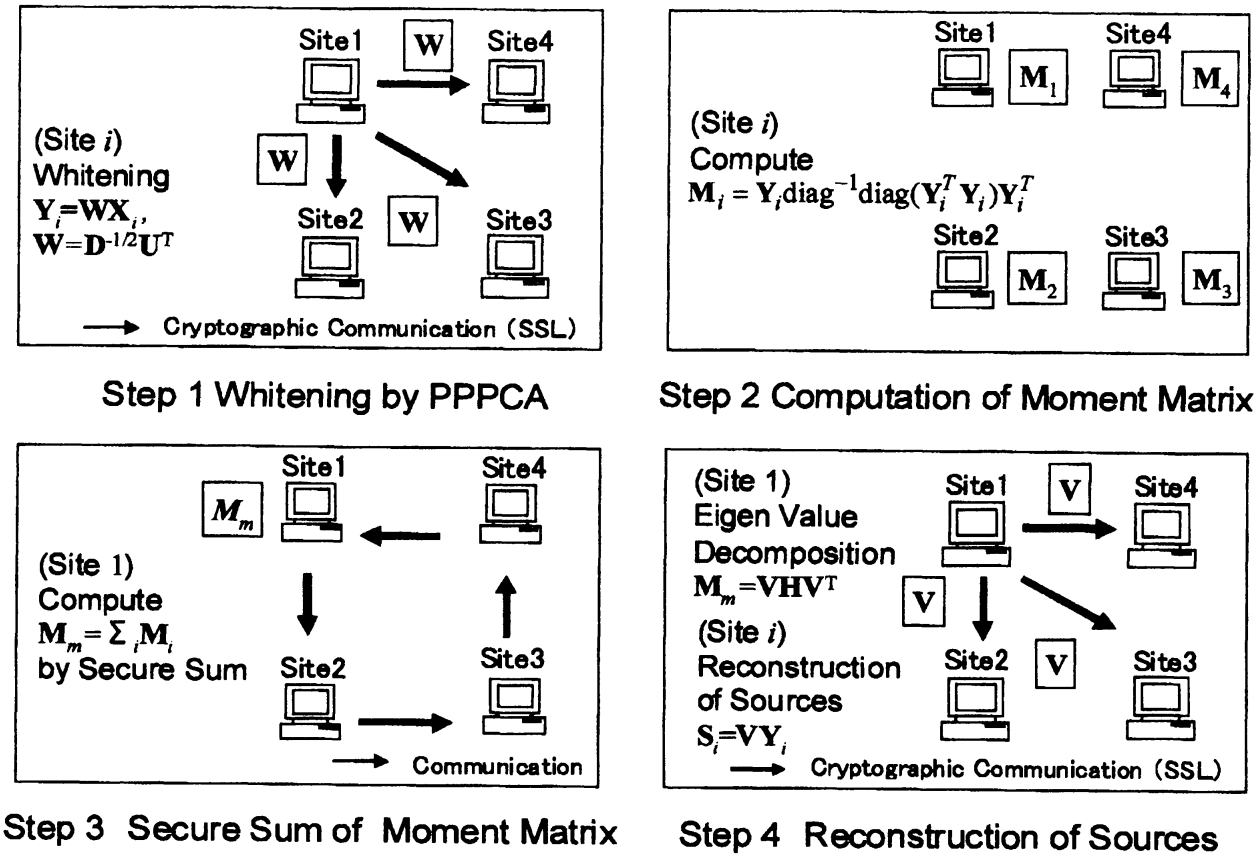


Figure 4: Privacy Preserving ICA

# References

[1] S. Amari, A. Cichocki, and H.H. Yang: A new learning algorithm for blind source separation. *In Advances in Neural Information Processing*, **8**, (MIT Press, Cambridge, MA, 1996), 757–763.

[2] W. Du, Y. S. Han, and S. Chen: Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classification. *In Proceedings of the 4th SIAM International Conference on Data Mining*, (2004), 222-233.

[3] A. Hyvärinen and E. Oja: A fast fixed-point algorithm for independent component analysis. *Neural Computation*, **9-7**(1997), 1483–1492.

[4] I. T. Jolliffe: Principal Component Analysis (Springer-Verlag, New York, 2002).

[5] C. Jutten and J. Herault: Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, **24** (1991), 1–10.

[6] B. Schneier: Applied Cryptography (John Wiley & Sons, New York, 1995).

[7] J. Vaidya and C. Clifton: Privacy-Preserving K-Means Clustering over Vertically Partitioned Data. *In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2003), 206-215.

[8] H. Yu, J. Vaidya, and X. Jiang: Privacy-Preserving SVM classification on vertically partitioned data. *In Lecture Notes in Computer Science*, **3918**, (Springer-Verlag, Berlin, Heidelberg, 2006), 647-656.

[9] http://www.cs.toronto.edu/~roweis/kica.html.