Privacy-Preserving Discovery of Multivariate Linear Relationship

Ningning Wu
Information Science Department, University of Arkansas at Little Rock
Little Rock, AR 72204, USA

Jing Zhang
System Engineering Department, University of Arkansas at Little Rock
Little Rock, AR 72204, USA

and

Ning Li
Applied Science Department, University of Arkansas at Little Rock
Little Rock, AR 72204, USA

ABSTRACT

The fast development of network and database techniques makes the data collecting and storing much easy and convenient. With more data being collected and available, there come the increasing requirements and huge opportunities for cooperative computation, where data are distributed across sites, and each site holds a portion of the data and wishes to collaborate to detect globally valid multivariate linear relationship.

This paper considers the privacy-preserving cooperative linear system of equations (PPC-LSE) problem in a large, heterogeneous, distributed database scenario, in which two parties would like to conduct cooperative computation from their private database while keeping their own data secret. The paper proposes a privacy-preserving algorithm to discover multivariate linear relationship based on factor analysis. Compared with other PPC_LSE algorithms, the proposed algorithm not only significantly reduces the communication cost, but also avoids the random matrix generation of either party to hide private information.

Keywords: privacy preserving data mining, multivariate linear relationship.

1 INTRODUCTION

The fast advances in network and database technologies have dramatically increased our ability to collect, store, and share the data. With more and more data being available, there are increasing needs for sharing and "making the sense" out of the data by data mining. Data mining, as an efficient way of exploiting large databases, has been widely used for extracting useful knowledge from the data that was not known before. Traditional data mining research and development focus on efficient and scalable techniques that can handle huge datasets. As data tends to be collected and scattered across different places, in many occasions, multiple data sources owned by different parties are needed in order to extract hidden knowledge, thus data privacy becomes a major concern. Without proper control data mining can easily leak the secret information from the data.

Privacy preserving data mining is a very active research field of data mining. Its goal is to discover new and useful knowledge buried in the sheer amount of data while protecting private information from being disclosed at the same time.

The paper considers the privacy-preserving cooperative-data-mining problem, in which multiple parties want to conduct data mining on their databases to find mutually beneficial information. These parties may be trusted, partially trusted, mutually uncommitted, or even competitive. When the parties trust each other, the cooperative data mining is straightforward: it only requires knowing inputs from all partners. However, the situation could become much complicated if no trust could be assumed. This gives rise to the need for privacy-preserving cooperative data mining. The following two examples explain such need.

- Two clinics conjecture that two diseases may be related. Each clinic has the patient data of one disease. Both would like to conduct a joint investigation on their patient data to verify their conjecture. Since each clinic is required to protect their patients' data according to the privacy regulation, they need to find a way to analyze their patient data without disclosing the patients' information.
- After a costly market research, company A decided that expanding its market share in some region will be very beneficial. However A is aware that another competing company B is also planning to expand its market share in some region. Strategically, A and B do not want to compete against each other in the same region, so they want to know whether their regions overlap without giving away location information (not only would disclosure of this information cost both companies a lot of money, it can also cause significant damage to the company if it is disclosed to other parties, e.g. another bigger competitor could then immediately occupy the market there before A or B even starts; or some real estate company could actually raise their price during the negotiation if they know A or B is very interested in that location). Therefore, they need a way to solve the problem while maintaining the privacy of their locations [11].

These applications require conducting cooperative computation based on each party's private inputs, but neither party is willing to disclose its own information. The problem of how to provide the gains of shared data without "giving away the store" [5] triggers the privacy-preserving data mining research.

In this paper, we consider the privacy-preserving cooperative linear system of equations (PPC-LSE) problem in a large,

heterogeneous, distributed database scenario. The definition of PPC-LSE problem [10] can be summarized as follows.

A matrix M and a vector b represent a set of linear constraints. There exist two basic models in the cooperative computation.

Model 1 Homogeneous Model

Party A has a $m_1 \times n$ matrix M_1 and a vector b_1 of length m_1; party B has a $m_2 \times n$ matrix M_2 and a vector b_2 of length m_2. Without releasing private matrix, they want to solve

$$\begin{pmatrix} M_1 \\ M_2 \end{pmatrix} x = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

Model 2 Heterogeneous Model

Party A has a $m \times n_1$ matrix M_1; party B has a $m \times n_2$ matrix M_2. Both parties know a vector b of length m. Without releasing private matrix, they want to solve

$$(M_1 \quad M_2)x = b$$

The two models can be combined into a hybrid model

$$(M_1 + M_2)x = b_1 + b_2$$

PPC-LSE protocol [10] was proposed to solve this problem. It uses 1-out-of-N Oblivious Transfer protocol [12] as secure protocol. Based on 1-out-of-N Oblivious Transfer protocol, two linear equations, $(M_1 + M_2)x = b_1 + b_2$ and $P(M_1 + M_2)QQ^{-1}x = P(b_1 + b_2)$, have the same solution x, where **P** and **Q**, are $n \times n$ random matrices and **Q** is invertible. The protocol proceeds in three steps. Firstly, Party A hides $\mathbf{M}_1$ in j random matrices $\boldsymbol{D}_i$, i=1,2,…,j, and sends all to party B. Secondly, party B generates **P** and **Q**, and sends the results $\mathbf{P}(\mathbf{D}_i + \mathbf{M}_2)\mathbf{Q}$, $i = 1,2,\cdots, j$ and $\mathbf{P}(b_1 + b_2)$ to party A. In terms of 1-out-of-N Oblivious Transfer protocol, party A can know $\mathbf{P}(\mathbf{M}_1 + \mathbf{M}_2)\mathbf{Q}$, while party B doesn't know party A's choice. Thirdly, party A solves the linear equation $\mathbf{P}(\mathbf{M}_1 + \mathbf{M}_2)\mathbf{Q}\hat{\mathbf{x}} = \mathbf{P}(b_1 + b_2)$, and sends $\hat{\mathbf{x}}$ to party B. Finally, party B calculates the final results $x = \mathbf{Q}\hat{\mathbf{x}}$.

This paper focuses on the Heterogeneous Model of PPC-LSE problem. This model assumes a large, heterogeneous, distributed database scenario in which numerical data is vertically partitioned in two sites. Each site contains some elements of a transaction and shares a join key of the two databases. Without privacy concern, the problem is to mine linear regression model involving attributes other than the join key, and it can be solved using the traditional statistical method because all inputs are known. However, the assumption that all inputs are known is not true in the privacy-preserving cooperative computation situation. The paper introduces a privacy-preserving algorithm based on factor analysis to mine multivariate linear relationship from vertically partitioned data.

The remainder of this paper is organized as follows. Section 2 gives the related work. Section 3 presents the formal definition of the problem. Section 4 presents the proposed method. Section 5 analyzes the accuracy of the algorithm. Section 6 discusses security, communication and computation cost of the algorithm. Section 7 concludes the paper and gives future work.

2 RELATED WORK

Secure multiparty computation (SMC) problem was first introduced by Yao [13], and extended by Goldreich [14]. From SMC perspective, [11] lists a number of privacy-preserving problems in the area of data mining, scientific computing, statistical analysis and so on. Privacy is becoming an important issue in data mining applications. There exist three classes of privacy-preserving solutions in data mining: data obfuscation, data summarization, and data separation [6].

Data obfuscation is based on reconstructing distribution of the original data value. The techniques include swapping values between records [15, 20], adding noise to the values in the database [2, 4, 3, 16]. Data summarization provides statistical information via query restriction, for example, controlling the overlap [17], suppressing data cells of small size [18] and so on. Data separation is a kind of secure multiparty computation. Research has also been conducted to construct the decision tree [19] and exploit association rules from horizontally partitioned data [8] by using cryptographic oblivious functions. A privacy-preserving scalar product protocol was proposed in [9, 7] to mine association rules from vertically partitioned data.

From the eigensystem and statistics perspectives, this paper proposes a privacy-preserving algorithm to discover multivariate linear relationship and analyzes the algorithm's confidence.

3 PROBLEM DEFINITION

Let $I = \{i_1, i_2, \cdots, i_m\}$ be a set of m items, $T = \{t_1, t_2, \cdots, t_n\}$ be a collection of transactions, where each transaction t_i is a set of items such that $t_i \subseteq I$. Let **X** be an $n \times p$ matrix that represents n transactions defined on p correlated attributes X_1, X_2,…, X_p; **Y** be an $n \times q$ matrix that represents n transactions defined on q correlated attributes Y_1, Y_2,…, Y_p. There is no sharing information between **X** and **Y**.

Under the PPC-LSE's heterogeneous model, we assume that the database T is vertically partitioned between two parties A and B. Party A has the matrix **X**, and party B has the matrix **Y**. Both parties know a vector **Z** of length n. Without either party disclosing its private matrix, they want to solve the linear function

$$Z = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_p X_p + \delta_{p+1} Y_1 + \delta_{p+2} Y_2 + \cdots + \delta_{p+q} Y_q$$

4 PROPOSED ALGORITHM

Concerning the privacy, the proposed algorithm detects multivariate linear relationship using factor analysis and least squares estimation. Factor analysis is a statistical technique that can be used to analyze interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (factors). In the proposed algorithm, factor analysis is employed at party A to find the common factors of its data partition **X**. Party A then sends $\mathbf{F}_x$, the factor scores matrix of **X**, to party B. Party B uses the least

square method to obtain the multivariate relationship Z' among its data partition **Y** and the factor scores matrix of **X**, and sends Z' to A. Finally party A recovers the multivariate relationship between **X** and **Y** by transforming Z' from the factor space back to the original variable space. Thus by factor analysis the problem of finding the multivariate relationship among original variables is transformed to finding the relationship among the common factors of party A's data **X** and the original variables of the party B's data **Y**. Moreover, factor analysis enables both parties to collaborate on the computation of the multivariate relationship without disclosing their own data.

In this section, we first briefly introduce factor analysis and classical linear regression model, and then we present the private-preserving algorithm to detect multivariate linear regression.

4.1 Factor Analysis

Factor analysis is used to uncover the latent structure (dimensions) of a set of variables. It has a variety of applications such as the assessment of underlying relationships or dimensions in the data, and the replacement of original variables with fewer, new variables. [21].

Definition 1

Let **X** be the observable random vector with p variables $X_1, X_2, \cdots, X_p$, that has sample mean vector μ and covariance matrix Σ. The factor model postulates that **X** is linearly dependent on a few unobservable random variables $\mathbf{F} = \{F_1, F_2, \cdots, F_m\}$, called common factors, and p additional sources of variation $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_p\}$, called errors, or specific factors. The factor analysis model is

$$\begin{aligned} X_1 - \bar{x}_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\ X_2 - \bar{x}_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \bar{x}_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p \end{aligned} \quad (1)$$

or, in matrix notion,

$$\mathbf{X} - \mu = \mathbf{LF} + \boldsymbol{\varepsilon} \quad (2)$$

where $\mathbf{L} = \{\ell_{ij}\}$ is the matrix of factor loadings. ℓ_{ij} is called the loading of the ith variable on the jth factor.

If we assume the unobservable random vectors **F** and $\boldsymbol{\varepsilon}$ satisfy the following conditions:

F and $\boldsymbol{\varepsilon}$ are independent

$E(F) = 0$, $Cov(F) = \mathrm{I}$

$E(\varepsilon) = 0$, $Cov(\varepsilon) = \Psi$, where

Ψ is a diagonal matrix

We get the covariance structure for the orthogonal factor model.

Property 1. The covariance structure for the orthogonal factor model is

1. $Cov(X) = LL' + \psi$ or

$Var(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2 + \psi_i$

$Cov(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km}$

2. $Cov(X, F) = L$ or

$Cov(X_i, F_j) = \ell_{ij}$

Given n observations on p correlated variables, the goal of factor analysis is to adequately represent the dataset $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$, $i = 1,2,\cdots,n$, with a small number of factors. The factor analysis can be performed via two approaches: the principal component method and maximum likelihood method. The principal component method assumes $\Sigma = \mathbf{LL}^T + 0 = \mathbf{LL}^T$. For simplicity, we only present the results from the principal component method here.

The principal component factor analysis of the sample covariance matrix **S** is specified in terms of its eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), \cdots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Let $m<p$ be the number of common factors. The matrix of estimated factor loadings $\{\hat{l}_{ij}\}$ is given by

$$\hat{\mathbf{L}} = [\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1, \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2, \cdots, \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m] \quad (3)$$

The estimated specific variance are provided by the diagonal elements of the matrix $\mathbf{S} = \hat{\mathbf{L}}\hat{\mathbf{L}}^T$, so

$$\hat{\boldsymbol{\Psi}} = \begin{bmatrix} \hat{\psi}_1 & 0 & \cdots & 0 \\ 0 & \hat{\psi}_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\psi}_p \end{bmatrix} \text{ with } \hat{\psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{l}_{ij}^2 \quad (4)$$

Communalities are estimated as

$$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \cdots + \hat{l}_{im}^2$$

Factor scores are estimated values of the unobserved random factor vector $\mathbf{F} = [F_1 F_2 \cdots F_p]^T$. That is, factor scores $\hat{\mathbf{f}}_j$ is equal to the estimate of the values $\mathbf{f}_j$ attained by **F** in jth case. Factor scores estimated by the principal component are generated using an un-weighted least squares procedure:

$$\hat{\mathbf{f}}_j = (\hat{\mathbf{L}}^T\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}^T(\mathbf{x}_j - \bar{\mathbf{x}}) \quad (5)$$

where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j$ is the sample mean.

Property 2. The contribution to total variance from the j-th common factor is

$$\frac{\lambda_j}{s_{11} + s_{22} + \cdots + s_{pp}}$$

Then, the number of common factors m can be selected based on the estimated eigenvalues. It is generally set to the number of positive eigenvalues of **S**.

4.2 Classical Linear Regression Model Introduction

Given n independent observations on r predictor variables $z_1, z_2, \cdots, z_r$ and the associated response variable **Y**, the classical linear regression model is defined as [22]

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1r} \\ 1 & z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nr} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_r \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (6)$$

or,

$$Y = Z\beta + \varepsilon$$

The least squares estimate of β in (6) is given by

$$\beta = (Z^T Z)^{-1} Z^T Y \quad (7)$$

4.3 Algorithm

The intuition behind the algorithm is that the representation of an original data matrix in the factor space is transferred for computation. By doing this, it is possible to reduce the size of data being transferred if only a small number of common factors are used. In addition, it is infeasible to retrieve the original data matrix without the knowledge of common factors. The least square method seeks the linear model based on the common factors. Finally, the linear function is obtained via transformation.

Step 1 Party A calculates eigenvalue-eigenvector pairs for its sample covariance matrices via scanning private matrices once [23]. Assume **X** is an $n \times p$ matrix, **Y** is an $n \times q$ matrix. Let (λ_i^X, e_i^X) $(i = 1,2,\cdots, p)$ be the eigenvalue-eigenvector pairs of the positive definite matrix $\mathbf{S_x}$, which is the sample covariance matrix of X.

Step 2 Party A chooses h ($h \le p$) common factors, and computes the factor loading matrix $\mathbf{L}^X{}_{p\times h}$, specific error vector $\boldsymbol{\varepsilon}^X$, and the factor scores matrix $\mathbf{F}^X{}_{n\times h} = \{f_{ij}^X\}$, $i = 1,2,\cdots,n,\ j = 1,2,\cdots,h$ respectively.

Step 3 Party A sends its common factor matrix $\mathbf{F}^X$ to party B.

Step 4 Let the predictor be

$$\mathbf{W} = \begin{pmatrix} 1 & f_{11}^X & f_{12}^X & \cdots & f_{1h}^X & y_{11} & y_{12} & \cdots & y_{1q} \\ 1 & f_{21}^X & f_{22}^X & \cdots & f_{2h}^X & y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & f_{n1}^X & f_{n2}^X & \cdots & f_{nh}^X & y_{n1} & y_{n2} & \cdots & y_{nq} \end{pmatrix} \quad (7')$$

the associated response be $\mathbf{Z}^T = (z_1 \quad z_2 \quad \cdots \quad z_n)$. Party B calculates the $h + k + 1$ coefficients $\boldsymbol{\beta} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z}$, then obtains the linear function

$$\begin{aligned} Z = F\beta = \beta_0 &+ \beta_1 F_1^X + \beta_2 F_2^X + \cdots \\ &+ \beta_h F_h^X + \beta_{h+1} Y_1 + \beta_{h+2} Y_2 + \cdots + \beta_{h+q} Y_q \end{aligned} \quad (8)$$

Step 5 Party B sends the linear function (8) to party A.

Step 6 Party A replaces $\mathbf{F}_i^X$, $i = 1,2,\cdots,h$, in (8) by $\mathbf{X} = (X_1, X_2, \cdots, X_p)$. Finally, the linear function (9) is obtained.

$$Z = \delta_0 + \delta_1 X_1 + \cdots + \delta_p X_p + \delta_{p+1} Y_1 + \cdots + \delta_{p+q} Y_q \quad (9)$$

Step 7 Party A sends the equation (9) to party B.

5 Accuracy of the Algorithm

Suppose the population is normally distributed on either party. To assess the adequacy of the model obtained by the proposed algorithm, it involves the following two kinds of estimation analysis.

- Test the adequacy of factor model [21].

  Suppose the number of common factors is m. Testing the adequacy of the *m* common factors model (1) is equivalent to testing

  H_0: $\mathbf{S} = \mathbf{L}\mathbf{L}^{\mathbf{T}} + \boldsymbol{\psi}$ *vs* H_1: any other positive definite matrix.

  We reject H_0 at the α level of significance if

  $$(n - 1 - (2p + 4m + 5)/6) \ln \frac{|LL' + \Psi|}{|S|} > \chi^2_{[(p-m)^2 - p - m]/2, \alpha} \quad (10)$$

  provided that *n* and $n - p$ are large. This condition can be guaranteed in large databases.

  Because the number of degrees of freedom $[(p-m)^2 - p - m]/2$ in (10) must be positive, then

  $$m < \tfrac{1}{2}(2p + 1 - \sqrt{8p + 1}) \quad (11)$$

  Suppose party A has α_1 level of significance for the *h* common factors model via step 2.

- Test the linear regression model [22].

  After fitting a multiple regression model (6), the next step is to determine which predictor variables have statistically significant effects on the response variable. This can be done by testing the hypotheses

  H_0: $\beta_1 = \cdots = \beta_r = 0$ *vs* H_1: at least one $\beta_j \neq 0$

  Before giving statistics, we introduce some definitions.

  Definition 2

  Let y_i be observed values, $\hat{y}_i = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r$ be fitted values $(i = 1,2,\cdots,n)$, and $\bar{y}$ be the mean of the observed values.

  The error sum of squares $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

  The total sum of squares $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

  The regression sum of squares $SSR = SST - SSE$.

  The coefficient of multiple determination $c^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

  We reject H_0 at the α level of significance if

  $$F = \frac{c^2\{n - (r+1)\}}{r(1 - c^2)} > f_{r, n-(r+1), \alpha} \quad (12)$$

  Where statistic *F* is an increasing function of c^2, follows an *F*-distribution with *r* and *n*-(*r*+*1*) degree freedom.

Without considering the error from step 2, we assume party B has α_2 level of significance for the multivariate linear model via step 4.

Finally, taking both errors from step 2 and step 4 into account, the algorithm has α_{all} level of significance for the multivariate linear model via step 6.

$$\alpha_{all} = 1-(1-\alpha_1)(1-\alpha_2) \quad (13)$$

6 Security and Communication/Computation Analysis

6.1 Security Analysis

The goal of the work is to create a practical, efficient method to compute multivariate linear regression model without disclosing entity values. This does not require a complete zero-knowledge solution. In this section, we discuss what must be disclosed and what is not disclosed.

In the PPC-LSE problem, each party knows its own data and learns the global multivariate linear relationship. It naturally brings some disclosure. For example, if we have 90% level of significance for a multivariate linear model

$$Z = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_p X_p + \delta_{p+1} Y_1 + \delta_{p+2} Y_2 + \cdots + \delta_{p+q} Y_q,$$

and party A has exactly 90% level of significance for the linear model $Z' = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_p X_p$, party A knows that party B has 90% % level of significance for the linear model $Z - Z' = \delta_{p+1} Y_1 + \delta_{p+2} Y_2 + \cdots + \delta_{p+q} Y_q$. Further, if party B has only one variable in the final linear model, party A can guess the value in party B with 90% confidence.

The security of the algorithm is based on the inability of either side to solve *n* equations in more than *n* unknowns. Using factor analysis, matrix X, with *n* observations on *p* variables, is replaced by *m* common factors. In step 3, party A sends an $n \times m$ $(m < p)$ matrix *F* to party B. With the *n* equations, party B is not able to solve $(n+m+1)p$ unknowns. On the other hand, without the matrix of loading factors, the meaning of the m items cannot be interpreted. Hence, party A's private individual information is not disclosed. Party B's privacy is also kept because it never sends its private individual information to party A.

It is impossible that specific individual data values and private constraints will be disclosed with certainty by this method.

6.2 Communication/Computation Analysis

For the PPC-LSE problem, assume the Gaussian elimination is used in both the general solution [13] and the PPC-LSE [10]. The former one costs $O(n \times (m+q)^2 \times d^2)$ to conduct Gaussian elimination, where *d* is the maximum length to represent a number. The latter one, based on 1-out-of-N Oblivious Transfer protocol [12], costs $O(\mu \times n \times (p+q))$, where μ is security parameter.

In the proposed algorithm, party A sends *n* messages, each with $m < p$ values; party B replies with one message; finally, party A sends results. Thus, there are total three rounds of communication, and the total cost is $O(n \times m)$.

The proposed algorithm requires party A to do a single pass, $O(n \times p^2)$ operations, to compute its covariance matrix, and then $O(p^3)$ operations to calculate the eigensystem. $O(p^3)$ is negligible compared with $O(n \times p^2)$ because $n >> p$ holds in large database. In the similar manner, party B needs $O(n \times (m+q)^2)$ operations to compute the linear coefficients β. The total computation cost of the algorithm is less than $O(n \times (p+q)^2)$.

The analysis shows that, when compared with other PPC_LSE algorithms, the proposed algorithm not only significantly reduces the communication cost, but also avoids the random matrix generation of either party to hide private information.

7 Conclusion and Future Work

In this paper, we propose a privacy-preserving linear-relationship mining algorithm and give the linear model's confidence level. Using factor analysis, the algorithm can effectively compress data and protect private information. It significantly reduces the communication cost, avoids transferring lots of random matrix and extra computation. Another advantage is that the algorithm can be easily extended to calculating a multiple multivariate linear regression model.

There are several directions for future research, such as how to improve the efficiency of computing eigensystem, and how to handle multiple parties, especially, if we consider collusion between parties as well.

8 References

[1] R. Agrawal. Data Mining: Crossing the chasm. **In 5th International Conference on Knowledge Discovery in Databases and Data Mining,** San Diego, California, August 1999.

[2] R. Agrawal and R. Srikant. Privacy Preserving Data Mining. **In Proceedings of ACM SIGMOD'00**, pages 439-450, 2000.

[3] Evfimievski, R. Srikant, R. Agrawal and J. Gehrke. Privacy Preserving Mining of Association Rules. **In Proceedings of ACM SIGKDD'02**, Edmonton, Canada, July 2002.

[4] D. Agrawal and C. C. Aggarwal. On The Design and Quantification of Privacy Data Mining Algorithms. **In Proceedings of PODS'01**, Santa Barbara, California, USA, May 2001.

[5] C. Clifton and D. Marks. Security and Privacy Implications of Data Mining. **In Proceedings of the 1996 ACM SIGMOD Workshop on Data Mining and Knowledge Discovery**, Montreal, Canada, June 2, 1996.

[6] C. Clifton. Tutorial Privacy, Security, and Data Mining, presented at the combined conference **13th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)**, Helsinki, Finland, 19-23

August, 2002.

[7] C. Clifton, M. Kantarcioglu, X. Lin, J. Vaidya and M. Y. Zhu. Tools for Privacy Preserving Distributed Data Mining. **SIGMOD Explorations**, 2003.

[8] M. Kantarcioglu and C. Clifton. Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. **The ACM SIGMOD Workshop on Research Issues In Data Mining and Knowledge Discovery (DMKD'2002)**, Madison, Wisconsin, June 2, 2002.

[9] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. **In Proceedings of ACM SIGKDD'02**, Edmonton, Alberta, CA, July 2002.

[10] W. Du and M. J. Atallah. Privacy-Preserving Cooperative Scientific Computations. **In Proceedings of The 14TH IEEE Computer Security Foundations Workshop**, Pages 273-282, Nova Scotia, Canada, June 11-13, 2001.

[11] W. Du and M. J. Atallah. Secure Multi-Party Computation Problems and their Applications: A Review and Open Problems. **In New Security Paradigms Workshop 2001**, pages 11-20. Cloudcroft, New Mexico, USA, Sept. 2001.

[12] M. Naor and B. Pinkas. Oblivious transfer and polynomial evaluation (extended abstract). **In Proceedings of the 31st ACM Symposium on Theory of Computing**, pages 245-254, Atanta, GA, USA, May 1999.

[13] C. Yao. Protocols for Secure Computations. **In Proceedings of the 23rd Annual IEEE Symposium on Foundation of Computer Science**, 1982.

[14] Goldreich, S. Micali and A. Wigderson. How to play any mental game. **In Proceedings of the 19th Annual ACM Symposium on Theory of Computing**, pages 218-229, 1987.

[15] V. Estivll-Castro and L. Brankovic. Data Swapping: Balancing Privacy Against Precision In Mining for Logic Rule. **In Proceedings of DaWak'99**, pages 389-398, 1999.

[16] C. K. Liew, U. J. Choi, and C. J. Liew. A Data Distortion by Probability Distribution. **ACM TODS**, 10(3):395-411, 1985.

[17] D. Dobkin, A. K. Jones, and R. J. Lipton. Secure Databases: Protection Against User Influence. **ACM TODS**, 4(1):97-106, March 1979.

[18] L. H. Cox Suppression Methodology and Statistical Disclosure Control. **J. Am. Stat. Assoc**., 75(370):377-395, April 1980.

[19] Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. **In Proceedings of CRYPTO'00**, pages 36-54, 2000.

[20] D. E. Denning. Cryptography and Data Security. **Addison-Wesley**, 1982.

[21] R. A. Johnson and D. W. Wichern. Applied Multivariate Statistical Analysis. **Prentice-Hall**, 1998.

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

[23] F. Korn, A. Labrinidis, Y. Kotidis and C. Faloutsos. Quantifiable Data Mining Using Ratio Rules. **The VLDB Journal** 8(3:4), February 2000.