

Received November 25, 2020, accepted December 8, 2020, date of publication December 21, 2020, date of current version January 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046131

# A Matrix Iteration Algorithm With Pruning for Pinpointing Multivariate Correlations From High Dimensional Data Sets

FUBO SHAO<sup>1,2,3</sup>, ZHIQIANG HOU<sup>4</sup>, LIMIN JIA<sup>1</sup>, AND ZHE ZHANG<sup>1</sup>

<sup>1</sup>State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China

<sup>2</sup>The Post-Doctor Station of the CRRC Corporation Limited, Beijing 100067, China

<sup>3</sup>CRRC Academy, Beijing 100070, China

<sup>4</sup>China Waterborne Transport Research Institute, Beijing 100088, China

Corresponding authors: Zhiqiang Hou (zhiqianghou@163.com) and Limin Jia (bhjx2009@163.com)


This work was supported in part by the Natural Science Foundation of Shandong Provincial, China, under Grant ZR2018MG003, and in part by the China Postdoctoral Science Foundation under Grant 2019M650981.

**ABSTRACT** There are a few dependent multivariate relationships among high dimensional data sets. Then how to identify these dependent variables from high dimensional data sets is an important issue for data analysis. Now, the most frequently used method is the enumeration method, that is all multivariate relationships in the high dimensional data sets are examined. However, the time complexity of the enumeration method is exponential ( $2^n$ ) and the calculation load is very heavy when the dimension is high. Aiming at solving this problem, the matrix iteration algorithm with pruning (MIP) is proposed for pinpointing multivariate dependent relationships in high dimensional data sets without examining all multivariate relationships. Some not dependent relationships are ignored without examined by the pruning process of the proposed MIP and the computing burden is reduced. The maximal information coefficient (MIC) is adopted as the measure of correlations in the proposed MIP algorithm due to the excellent properties, generality and equitability, of MIC. In the case of the data set with 5 variables, more than 50% multivariate relationships are pruned without examining. Numerical experiments also show that the calculating burden is greatly reduced. Compared to the enumeration method, 82.5% calculating time and 98.5% calculating times of multivariate relationships are saved for the data set with two dependent multivariate relationships among 30 variables in the experiment. The proposed MIP algorithm is effective for pinpointing multivariate dependent relationships from data sets with high dimensions.

**INDEX TERMS** Correlation, high dimensions, maximal information coefficient, pruning algorithm.

## I. INTRODUCTION

Pinpointing correlated stochastic quantities from high dimensional data sets is a very important issue. With the wide application of information technologies, various information can be obtained. The era of big data is approaching with large amount data emerging at the growing rate of fifty percent a year [1]. Besides of the properties of velocity (the speed of data in and out), variety (range of data types and resources) and veracity (an indication of data integrity and the ability for an organization to trust the data and be able to confidently use it to make crucial decisions) [2], [3],

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano .

the volume, the amount of data, is also an important property of which high dimension is an important feature of large volume.

In the exploratory data analysis of data sets with hundreds or thousands of dimensions (stochastic quantities, variables), the first step may be to identify promising bivariate or multivariate correlations for further research. In order to identify correlations among stochastic quantities, a natural approach, that is the enumeration method, is to compute a measure of dependence on all stochastic relationships, and then sort these stochastic relationships according to these measure values. The higher ranking relationships are what we want to identify. For bivariate correlations, the natural approach may be appropriate, that is all measures of dependence on

all bivariate stochastic quantities are computed. However, for multivariate correlations, if the interrelations among different multi stochastic quantities are ignored and all measures are also computed using the natural approach, the computation workload increases exponentially due to combinations of variables. For example, given  $k$  stochastic variables (the dimension of the data set is  $k$ ),  $C_k^2 = k(k-1)$  bivariate relationships are examined, while  $3C_k^3 + \dots + kC_k^k > 3(2^k - C_k^0 - C_k^1 - C_k^2)$  multivariate relationships should be examined. Then, if the natural approach of identifying bivariate correlations, that is calculating measures of the dependence of all bivariate relationships and ranking, is adopted, it is infeasible to identify multivariate correlations in the data set with hundreds or thousands of dimensions. The reason is that every multivariate relationship is examined even if there are independent variables in the multivariate relationships. Facing high dimensional data sets, how to avoid calculating the measures of independent multivariate relationships as much as possible and how to efficiently identify multivariate correlations from quite a lot of relationships is an important work.

There are many measures of correlations. The maximal information coefficient (MIC) is proposed by Reshef *et al.* [4] via employing mutual information [6], [8]. Compared to other measures, MIC has two excellent properties: generality and equitability. Generality means that, with sufficient sample size, MIC can discover a wide range of interesting functional and non-functional associations. Equitability means that MIC gives similar scores to equally noisy relationships of different types. However, the algorithm proposed by Reshef *et al.* [4], [5] can only identify bivariate correlations and the computation time is much longer. Aiming at these problems, Shao *et al.* [10] design a fast algorithm calculating the MIC of multi variables. In this paper, MIC is adopted as the measure of dependence of multi variables. Of course, other measures of dependence can also be adopted replacing MIC in the designed efficient matrix iteration algorithm with pruning.

However, if the fast algorithm is directly applied to identifying multivariate correlations in high dimensional data sets, that is the MIC of every multi-variable relationship is calculated using the proposed fast algorithm [10], the computation workload is also very heavy. Aiming at solving this problem, employing the interrelation of multivariate relationships, some nonabsolutely correlative multivariate relationships are filtered out without calculating measurement values and an efficient algorithm named as the matrix iteration algorithm with pruning (MIP) is designed for pinpointing multivariate correlations in high dimensional data sets. The contributions of this paper are as follows.

Firstly, without calculating measure values of all multivariate relationships, multivariate correlations can be pinpointed from quite a lot of multivariate relationships by employing the proposed MIP algorithm. Many nonabsolutely correlative multivariate relationships are filtered out (that is the pruning process). Multivariate correlations can be pinpointed out from

massive relationships and relatively less computation workload is needed.

Secondly, besides the pruning operation promoting computation efficiency, the proposed efficient algorithm MIP can also give the exact expression of the multivariate dependent relationship. Given  $k$  variables,  $x_1, x_2, \dots, x_k$ , there are  $k$  multivariate relationships which are relationships between variable  $x_i (i = 1, 2, \dots, k)$  and variables  $x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_k$ . MIP algorithm can give the specific form of the multivariate correlations, that is variable  $x_i$  can be determined by the algorithm MIP.

Thirdly, in this paper, if other measures of dependence replace MIC, the algorithm is also can be used for pinpointing multivariate correlations. Then the proposed efficient algorithm MIP can be regarded as an algorithmic framework which has good expansibility and transplantability.

The paper is organized as follows. In section II, some related research work is reviewed. In section III, employing the fast algorithm [10] calculating the MIC value of the multi-variable relationship, the efficient algorithm, named as the matrix iteration algorithm with pruning (MIP), for pinpointing multivariate correlations is proposed. A simple case of the MIP algorithm is given in Section IV. Section V gives some experimental results. Lastly, some conclusions and future work are given in section VI.

## II. LITERATURE REVIEW

Measures of variable dependence can be roughly divided into the following four categories: grid-based method, mutual information estimation, distance/kernel-based statistics and correlation-based methods [9]. These measures are summarized in Table 1.

Grid-based method. The grid-based method explores the space of all possible grids drawing on the sampled data, assigns a score to each grid, and aggregates these scores. Normally, the value of the correlation coefficient is equal to the maximal value of the aggregated scores. The maximal information coefficient (MIC) [4] is the maximal value of the metric of normalized mutual information scores. However, it is difficult to compute the MIC value efficiently over all grids. Then fast approximate algorithms computing the MIC values of bivariate and multivariate relationships is proposed by Shao *et al.* [10], [11]. Albanese *et al.* [12] develop a practical tool for detecting associations in big data sets. HHG [13] explores the three-by-three grids defined by pairwise and uses as its score Pearson's  $\chi^2$  test statistic computed on two-by-two contingency tables derived from the three-by-three grids. However, the HHG is not distribution free. The coefficient  $S^{DDP}$  [15] explores much more grids and its score is the normalized mutual information.

Mutual information estimation. Due to its information theoretic background, mutual information differs from the other measures of dependence of random variables. The theoretical advantages of mutual information derived from the reason that it is closely tying to Shannon entropy. Then the aim is to estimate mutual information only from the data set

**TABLE 1.** Summary of measures of dependence.

Methodological approach	Contents	Related work
Grid-based method	Explore the space of all possible grids drawing on the sampled data	The maximal information coefficient (MIC) [4], HHG [13], $S^{DDP}$ [15]
Mutual information estimation	Estimate mutual information only from the data set without knowing the densities of random variables	Kraskov <i>et al.</i> [6], Mlter <i>et al.</i> [7]
Distance/kernel-based statistics	Use the distance variance/covariance based on pairwise distances between points	The distance correlation (dCor) [16], Hilber-Schmidt information criterion (HSIC) [19]
Correlation-based methods	Search for arbitrary measurable functions such that the coefficient is maximized	The maximal correlation [20], randomized dependence coefficient (RDC) [22]

without knowing the densities of random variables (including joint density of variables). The easy and very crude approximations to mutual information is based on cumulate expansions [6]. However, these approximations are valid for distributions close to Gaussians alone. It is more robust for expressions obtained by entropy maximization using averages of functions of the data set [6]. The estimations based on explicit parametrization of densities are useful but are less efficient [7]. The promising methods is based on kernel density estimators [6], and Kraskov *et al.* estimate mutual information from  $k$ -nearest neighbor statistics [6].

Distance/kernel-based statistics. The distance correlation (dCor) [16], which is defined analogously to ordinary correlations, uses the distance variance/covariance based on pairwise distances between points. Going a step further, the distance covariance is developed as the metric spaces of negative type of which Euclidean spaces are a special case [17]. In addition to distance criteria, there are kernel-based measures which are formulated based on embedding of probability distributions into reproducing kernel Hilbert spaces [18]. Hilber-Schmidt information criterion (HSIC) is a more general statistic in kernel Hilbert spaces of which dCor is a special case [19].

Correlation-based methods. Pearson's correlation coefficient is the first coefficient. After that, many kinds of coefficients are proposed. The maximal correlation [20] may be the best-known correlation-based method which searches for arbitrary measurable functions such that the coefficient is maximized, and it is hard to be computed in general. However, the approximate method of alternating conditional expectations is widely used [21]. The recent method, randomized dependence coefficient (RDC) [22] which is the estimation of the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient (HGR), is the largest canonical correlation between random non-linear projections of their respective empirical copula-transformations.

Besides the above taxonomy, the statistical correlation coefficient can also divided as the two categories: bivariate correlation coefficient and multivariate coefficient. All above correlation coefficient can be as a measurement of dependence between two random variables. However, there are much more multivariate correlations in big data and there are relatively few correlation coefficient can be used to detect multivariate correlations [14], [23], [24].

Compared to other measures, the maximal information coefficient (MIC) has two excellent properties: generality and equitability. Generality means that with sufficient sample size, MIC can capture a wide range of interesting associations, not limited to specific functional types. Equitability means that MIC gives similar scores to equally noisy relationships of different types. Reshef *et al.* (2011) [4], [25] firstly proposed MIC which can only detect bivariate correlations. Shao *et al.* [10] further defined MIC of multi random variables and designed an approximate fast algorithm for calculating the MIC value of multi variables. However, if the fast algorithm is directly applied into identifying dependent multivariate relationships in data sets with thousands of variables or much more, that is the enumeration method is adopted, dependent multi-variable relationships cannot be effectively pinpointed from data sets.

### III. THE MATRIX ITERATION ALGORITHM WITH PRUNING

It is infeasible to examine all multivariate relationships in data sets with hundreds of or thousands of dimensions in limited time. Then in III-B, aiming at avoiding calculating some not dependent multivariate relationships (the pruning process), the efficient matrix iteration algorithm with pruning (MIP) is designed to reduce the computation load in the procedure of detecting multivariate correlations. In this paper, the maximal information coefficient (MIC) is selected as the measure of dependence of variables due to its two excellent properties: generality and equitability. In III-A, the definition and related algorithms of MIC are introduced.

#### A. INTRODUCTION TO THE MAXIMAL INFORMATION COEFFICIENT (MIC)

In this subsection, the definition of MIC [4], the approximate algorithm [4] calculating MIC values of bivariate relationships and the fast algorithm [10] calculating MIC values of multivariate relationships are introduced.

Given a finite data set  $D$  with two variables,  $x$ ,  $y$ , the  $x$ -value of  $D$  is partitioned into  $s$  bins and the  $y$ -value is partitioned into  $t$  bins. Then an  $s$ -by- $t$  grid  $G$  is obtained.  $D|_G$  is the distribution induced by the points of  $D$  in the cells of  $G$ . Given a fixed data set  $D$ , different grids with the same or different numbers of partitions of the  $x$ -value and  $y$ -value will result in different distributions. Specifically, in reference [4],

the maximal information coefficient of two random variables is given in definitions III.1-III.3.

*Definition III.1* [4] Given a finite data set  $D \subset R^2$  and positive integers  $s, t$ , define

$$I^*(D, s, t) = \max I(D|_G)$$

where the maximum is over all grids with  $s$  columns and  $t$  rows, that is the first variable in the data set  $D$  is divided into  $s$  partitions and the second one is divided into  $t$  parts.  $I(D|_G)$  is the mutual information of  $D|_G$ .

*Definition III.2* [4] The characteristic matrix  $M(D)$  of the data set  $D \subset R^2$  is a finite matrix with entities

$$M(D)_{s,t} = \frac{I^*(D, s, t)}{\log \min\{s, t\}}$$

*Definition III.3* [4] The MIC of the data set  $D \subset R^2$  with sample size  $n$  and grid size less than  $B(n)$  is given by

$$MIC(D) = \max_{st < B(n)} \{M(D)_{s,t}\}$$

where  $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$  for some  $0 < \varepsilon < 1$

From the above definitions, it can be found that the most important work is calculating the mutual information of the distribution  $D|_G$ . The space of grids that must be searched grows exponentially with the increasing of the number of data points. Then, considering the computational efficiency, an approximate algorithm [4] is designed to calculate the approximate MIC value via employing the dynamic programming algorithm. To be specific, intuitively, equal division of axes may lead to the maximal value of MIC [4]. Hence, the approximate algorithm firstly partitions a variable of the data set equally. For the other axis, some candidate partition clumps are given to reduce the computation load and then the partitions of the other axis are obtained by employing the dynamic program algorithm. With the partitions of the two axes, the grid on the scatter plot of the two variables is obtained. Another important problem in the above definitions is the maximal grid size  $B(n) = n^\alpha$ . If  $B(n)$  is set too high, that can lead to non-zero scores even for random data. While setting  $B(n)$  too low means that only simple patterns are searched for. To balance these competing considerations, the parameter  $\alpha$  is usually set to be 0.6 according to the practical experience [4].

The difficulty of directly applying the approximate algorithm [4] to detecting multivariate correlations is how to divide partitions of multi variables. Aiming at solving this problem, the MIC definition of multi variables is firstly given, and then the difficulty of dividing partitions of multi variables is solved via employing the clustering algorithm in the fast adaptive-MIC algorithm proposed by Shao et al. [10]. Compared to the approximate algorithm [4], the fast adaptive-MIC algorithm can calculate MIC values of bivariate and multi-variable relationships in a very short time.

However, if the fast adaptive-MIC algorithm [10] is directly applied to detecting multivariate correlations in data sets with hundreds of or thousands of dimensions, that is

the enumeration method is adopted, the computation time is much longer since the most time is spent on random relationships. Then the matrix iteration algorithm with pruning (MIP) is proposed for efficiently detecting multivariate correlations from high dimensional data sets.

## B. THE PROPOSED MATRIX ITERATION ALGORITHM WITH PRUNING (MIP)

In this paper, the matrix iteration algorithm with pruning (MIP) can precisely and efficiently identify the existent multivariate correlations, and it is proposed based on MIC and the fast adaptive-MIC algorithm [10]. Of course, if there are new better measures of dependence appear, MIC and the fast adaptive-MIC algorithm can be replaced by other measures and corresponding algorithms, respectively.

The fast adaptive-MIC algorithm [10] can calculate MIC values of relationships with  $VN$  ( $VN \geq 2$ ) variable in the data set  $D$ . Firstly, select a variable  $Y$  from the data set  $D$  and then the remaining  $VN - 1$  variables,  $X_1, X_2, \dots, X_{VN-1}$ , are as the whole. Secondly, the variables  $X_1, X_2, \dots, X_{VN-1}$  and variable  $Y$  are clustered into  $x, y$  partitions via employing the bisecting  $k$ -means clustering algorithm, respectively. Thirdly, according to the definition of MIC, calculate the MIC value of the  $VN$ -variable relationship.

There are at least two problems if the adaptive-MIC algorithm is directly applied to detecting multivariate correlations in big data with hundreds of thousands of variables (that is the enumeration method). Firstly, the number (more than  $3(2^k - C_k^0 - C_k^1 - C_k^2)$ ) of the examined multi-variable relationships exponentially increases with the increasing of the number ( $k$ ) of variables in the data set. As a consequence, the computation time is very long when the number of variables in the data set is large. Secondly, some multivariate correlations with relatively lower MIC values may be missed. The reason is as follows. Although MIC gives similar scores to equally noisy relationships of different types (the equitability property), there is a slightly difference among MIC values of relationships of different types and MIC values of multivariate relationships may be less than that of bivariate relationships. For example, the MIC value of the linear relationship might be higher than that of the sine/linear relationship for data sets with the same scale. Given two relationships  $y = \sin(4\pi x_1^2 x_3^2) + x_1 + x_3$  (sine/linear),  $x_4 = 5x_2 + 1$  (linear),  $x_1, x_2, x_3$  are mutually independent random variables generated in domain  $[0, 1]$ . The data set  $D$  with 5 variables,  $x_1, x_2, x_3, x_4, y$ , is generated according to definitions of these two relationships, and the scale of the data set  $D$  is 20000. If the enumeration method is adopted, that is the adaptive-MIC algorithm [10] is directly applied to detecting multivariate correlations in the data set  $D$ , the MIC values of all  $C_5^2 + 3C_5^3 + 4C_5^4 + 5C_5^5 = 65$  multi-variable relationships in the data set  $D$  are calculated. The highest MIC values of relationships,  $(x_2, x_4)$ ,  $(x_4, (x_1, x_2))$  and  $(y, (x_1, x_3))$ , are 0.9988, 0.9769 and 0.9178, respectively. According to these calculated MIC values, it is obvious that the multi-variable relationship  $(x_4, (x_1, x_2))$  is more important than the

relationship  $(y, (x_1, x_3))$ . However, the obvious conclusion is not correct. In fact, according to the data set generating process, the relationships  $(x_2, x_4)$  and  $(y, (x_1, x_3))$  are the most important and these two relationships should be identified. The relationship  $(x_4, (x_1, x_2))$  should be excluded. From this example, it can be found that, if only rank all relationships according to MIC values, some not important relationships may be paid more attention to and a lot of time is squandered on calculating MIC values of these not important relationships. An efficient algorithm is needed to exclude these not important relationships and to precisely identify the exist multivariate dependent correlations.

Aiming at solving these problems, the matrix iteration algorithm with pruning (MIP) is designed and the intuitions of the designed efficient MIP algorithm are as follows.

Firstly, pruning. The new relationship which is obtained by adding independent variables into the exist dependent relationship will be given a lower MIC value than that of the exist dependent relationship, and the new relationship which is obtained by adding variables into the independent relationship also is not the relationship we are looking for, even though the MIC of the new relationship is high. For example, if the variable  $y$  is dependent on the variables  $(x_1, x_3)$ , that is the relationship  $(y, (x_1, x_3))$  is dependent, the MIC value of the relationship  $(y, (x_1, x_3))$  is higher than these of the relationships  $(y, (x_1, x_2, x_3))$ ,  $(y, (x_1, x_3, x_4))$  and  $(y, (x_1, x_2, x_3, x_4))$ , where the variables  $x_2, x_4$  are independent of the variables  $y, x_1$  and  $x_3$ . The relationship  $(y, x_2)$  is the relationship of independent variables  $x_2, y$ . The relationship  $(y, (x_1, x_2))$  can be seen as adding the variable  $x_1$  into the relationship  $(y, x_2)$ .

Secondly, identifying dependent multivariate correlations. Given a multivariate relationship, new relationships can be obtained by adding one of the remaining variables into the given relationship. If these MIC values of the obtained new relationships are lower than that of the given multivariate relationship, the given multivariate relationship is regarded as a multivariate correlation.

For clearness, these symbols are introduced.  $(x_i, (x_j, x_{j+1}, \dots, x_{j+k-1}))$  is the multivariate relationship between the variable  $x_i$  and  $k$  variables  $x_j, x_{j+1}, \dots, x_{j+k-1}$ , and its corresponding MIC, which is calculated by the fast adaptive-MIC algorithm [10], is  $MIC_{(x_i, (x_j, x_{j+1}, \dots, x_{j+k-1}))}$ . The symbol  $MIC_{(x_1, x_2, \dots, x_k)} = \max_{1 \leq i \leq k} MIC_{(x_i, (x_1, x_{i-1}, x_{i+1}, \dots, x_k))}$ . For clarity, relationships with  $k$  (in this table,  $k = 2$ ) variables and corresponding MIC values are added above the designed matrix  $IA$ , and relationships are also added on the left of the matrix  $IA$ . Elements  $IA_{p,q}$  of the table body (the matrix  $IA$ ) are MIC values of relationships with  $k+1$  variables (see Table 2).

Based on the above intuitions, the designed matrix iteration algorithm with pruning (MIP) is presented in Algorithm 2. In MIP, the most important procedure is the pruning process which is presented in Algorithm 1.

In Algorithm 1,  $RS$  is the set of relationships with  $i-1$  variables in the first line above the matrix  $IA$  and  $IV$  is the

**Algorithm 1** PrunPro( $RS, IV$ ): the pruning process of the proposed MIP algorithm

**Require:** The set of relationships  $RS$ , the set of independent relationships  $IV$ .

- 1: Select the  $p$ -th and  $q$ -th relationships,  $R_p, R_q$ , from  $RS$ .
- 2: **if** the number of different variables in  $R_p, R_q$  is equal to 0, or equal to or larger than 2 **then**
- 3:  $IA_{pq} = -$ , does not need to be calculated.
- 4: **else**
- 5: */\** There is only one different variable in relationships  $R_p = (x_{i_1}, x_{i_2}, \dots, x_{i_{k-1}}, x_{i_{k+1}})$  and  $R_q = (x_{i_1}, x_{i_2}, \dots, x_{i_k})$ . *\*/*  
**Case 1.** Completely prune.  $R_p \in IV$  and  $R_q \in IV$ ,  $IA_{pq} = -$ , does not need to be calculated.  
**Case 2.** Partly prune.  $R_p \notin IV \& R_q \in IV$  or  $R_p \in IV \& R_q \notin IV$ . If  $R_p \notin IV \& R_q \in IV$ ,  $IA_{pq} = MIC_{(x_{i_1}, x_{i_2}, \dots, x_{i_k}, x_{i_{k+1}})} = MIC_{(x_{i_{k+1}}, (x_{i_1}, \dots, x_{i_k}))}$ . If  $R_p \in IV \& R_q \notin IV$ ,  $IA_{pq} = MIC_{(x_{i_1}, x_{i_2}, \dots, x_{i_k}, x_{i_{k+1}})} = MIC_{(x_{i_k}, (x_{i_1}, \dots, x_{i_{k-1}}, x_{i_{k+1}}))}$ .  
**Case 3.** No pruning.  $R_p \notin IV \& R_q \notin IV$ .  $IA_{pq} = MIC_{(x_{i_1}, x_{i_2}, \dots, x_{i_k}, x_{i_{k+1}})} = \max\{MIC_{(x_{i_j}, (x_{i_1}, \dots, x_{i_{j-1}}, x_{i_{j+1}}, \dots, x_{i_k}, x_{i_{k+1}}))}, j = 1, 2, \dots, k\}$ .
- 6: **end if**
- 7: **return** the matrix  $IA$ .

set of dependent relationships with  $i-1$  variables. Select two relationships,  $R_p, R_q$ , from the set  $RS$ . If the number of different variables in  $R_p, R_q$  is equal to 0, or equal to or larger than 2, there is no new relationships generated according to  $R_p, R_q$ . The element  $IA_{pq} = -$  does not need to be calculated (line 2-3). If the number of different variables in  $R_p, R_q$  is equal to 1, new relationships can be generated according to  $R_p, R_q$ . There are three cases (line 5).

No loss of generality, the  $p$ -th and  $q$ -th relationships are  $(x_{i-1}, (x_1, x_2, \dots, x_{i-2}))$ ,  $(x_i, (x_1, x_2, \dots, x_{i-2}))$ , respectively.

Case 1. If  $(x_{i-1}, (x_1, x_2, \dots, x_{i-2})) \in IV$  and  $(x_i, (x_1, x_2, \dots, x_{i-2})) \in IV$ , completely prune. The MIC values of relationships  $(x_j, (x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_i))$ ,  $j = 1, 2, \dots, i$  do not need to be calculated and the element  $IA_{pq}$  does not need to be calculated.

Case 2. If  $(x_{i-1}, (x_1, x_2, \dots, x_{i-2})) \notin IV$  and  $(x_i, (x_1, x_2, \dots, x_{i-2})) \in IV$ , partly prune. Only calculate the MIC value of the relationship  $(x_{i-1}, (x_1, \dots, x_{i-2}, x_i))$  which is the value of the element  $IA_{pq}$ , and the other MIC values of the remained  $i-1$  relationships need not be calculated. That is,  $IA_{pq} = MIC_{(x_{i-1}, \dots, x_i)} = MIC_{(x_{i-1}, (x_1, \dots, x_{i-2}, x_i))}$ .

Case 3. If  $(x_{i-1}, (x_1, x_2, \dots, x_{i-2})) \notin IV$  and  $(x_i, (x_1, x_2, \dots, x_{i-2})) \notin IV$ , no pruning. Calculate MIC values of all relationships with  $i$  variables. Then the maximal MIC value, which is the value of the element  $IA_{pq}$ , is the MIC value of the  $i$ -variable relationship  $(x_1, x_2, \dots, x_{i-1}, x_i)$  and the corresponding relationship is the specific form of the  $i$  variables.

TABLE 2. The form of iteration matrix IA which is a symmetric matrix.

	$(x_1, x_2)$	...	$(x_1, x_v)$	$(x_2, x_3)$	...	$(x_2, x_v)$	...	$(x_{v-1}, x_v)$
	$MIC_{(x_1, x_2)}$	...	$MIC_{(x_1, x_v)}$	$MIC_{(x_2, x_3)}$	...	$MIC_{(x_2, x_v)}$	...	$MIC_{(x_{v-1}, x_v)}$
$(x_1, x_2)$	—	...	$MIC_{(x_1, x_2, x_v)}$	$MIC_{(x_1, x_2, x_3)}$	...	$MIC_{(x_1, x_2, x_v)}$	...	—
...	...	...	...	...	...	...	...	...
$(x_1, x_v)$	...	...	—	—	...	$MIC_{(x_1, x_2, x_v)}$	...	$MIC_{(x_1, x_{v-1}, x_v)}$
$(x_2, x_3)$	...	...	—	—	...	$MIC_{(x_2, x_3, x_v)}$	...	—
...	...	...	...	...	...	...	...	...
$(x_2, x_v)$	...	...	...	...	...	—	...	$MIC_{(x_2, x_{v-1}, x_v)}$
...	...	...	...	...	...	...	...	...
$(x_{v-1}, x_v)$	...	...	...	...	...	...	...	—

Algorithm 2 presents the whole MIP algorithm. The data set  $D$  is with  $n$  points and  $v$  variables.  $VR$  is the maximal number of variables in the detected relationships. The critical variable  $\epsilon$  is used for measuring the dependence of relationships. If the MIC value of the relationship is less than  $\epsilon$ , the relationship is independent.  $DR$  and  $IV$  are the set of identified dependent relationships and independent relationships, respectively. The count number  $i$  is the number of variables in the current relationships.

In Algorithm 2, the initialization process is presented in lines 1-7. All MIC values of bivariate relationships are calculated and independent bivariate relationships are identified. The  $RS$  and  $IV$  are obtained, and then the initialization matrix  $IA$  is obtained by employing the pruning process Prun-Pro( $RS, IV$ ). The specific procedure is as follows. Select two relationships, the  $p$ -th and  $q$ -th bivariate relationships  $R_p, R_q$ , from  $RS$ . If there is two different variables between  $R_p$  and  $R_q$ , the element  $IA_{pq}$  does not need to be calculated. If there is only one different variable between the  $p$ -th and  $q$ -th bivariate relationships in the set  $RS$ , such as  $(x_l, x_m)$  and  $(x_l, x_s)$ , this situation can be divided into three cases.

Case 1. If  $(x_l, x_m) \in IV$  and  $(x_l, x_s) \in IV$ , completely prune. The MIC values of relationships  $(x_l, (x_m, x_s))$ ,  $(x_m, (x_l, x_s))$  and  $(x_s, (x_l, x_m))$  do not need be calculated and the element  $IA_{pq}$  does not need to be calculated.

Case 2. If  $(x_l, x_m) \notin IV$  and  $(x_l, x_s) \in IV$ , partly prune. The MIC values of relationships  $(x_l, (x_m, x_s))$  and  $(x_s, (x_l, x_m))$  do not need be calculated and only the MIC value of the relationship  $(x_m, (x_l, x_s))$  is calculated. And  $IA_{pq} = MIC_{(x_l, x_m, x_s)} = MIC_{(x_m, (x_l, x_s))}$ .

Case 3. If  $(x_l, x_m) \notin IV$  and  $(x_l, x_s) \notin IV$ , no pruning. Calculate MIC values of all three-variable relationships  $(x_l, (x_m, x_s))$ ,  $(x_m, (x_l, x_s))$  and  $(x_s, (x_l, x_m))$ . And  $IA_{pq} = MIC_{(x_l, x_m, x_s)} = \max\{MIC_{(x_l, (x_m, x_s))}, MIC_{(x_m, (x_l, x_s))}, MIC_{(x_s, (x_l, x_m))}\}$ .

After the initialization procedure, the next step is the iterative loop for calculating the matrix  $IA$  in lines 8-25. The set  $DR$  of dependent relationships are updated in lines 8-9. And the  $k$ -th relationship  $R_k$  of the first row above  $IA$  is added into  $DR$  if the MIC of  $R_k$  is larger than all MIC values ( $IA_{.k}$ ) in the  $k$ -th column. In lines 12-17, the set  $IV$  of independent relationships is updated. With sets  $RS$  and  $IV$ , the Algorithm 2 is called to calculate the iteration matrix  $IA$ .

If the relationships above the new matrix  $IA$  is empty, stop and return the dependent relationships in  $DR$  and independent relationships in  $IV$ ; otherwise, calculate the element  $IA_{pq}$  of the matrix  $IA$  with pruning.

If the number of different variables between the  $p$ -th and the  $q$ -th relationships in the first row of the table header is equal to 0 or larger than 2, there is no need to calculate the element  $IA_{pq}$ . Otherwise (the number of different variables between the  $p$ -th and the  $q$ -th relationships is equal to 1), the element  $IA_{pq}$  is calculated with pruning which is similar to Step 1.4. No loss of generality, the  $p$ -th and  $q$ -th relationships are  $(x_{i-1}, (x_1, x_2, \dots, x_{i-2}))$ ,  $(x_i, (x_1, x_2, \dots, x_{i-2}))$ , respectively. There are also three cases.

Case 1. If  $(x_{i-1}, (x_1, x_2, \dots, x_{i-2})) \in IV$  and  $(x_i, (x_1, x_2, \dots, x_{i-2})) \in IV$ , completely prune. The MIC values of relationships  $(x_j, (x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_i))$ ,  $j = 1, 2, \dots, i$  do not need be calculated and the element  $IA_{pq}$  does not need to be calculated.

Case 2. If  $(x_{i-1}, (x_1, x_2, \dots, x_{i-2})) \notin IV$  and  $(x_i, (x_1, x_2, \dots, x_{i-2})) \in IV$ , partly prune. Only calculate the MIC value of the relationship  $(x_{i-1}, (x_1, \dots, x_{i-2}, x_i))$  which is the value of the element  $IA_{pq}$ , and the other MIC values of the remained  $i - 1$  relationships need not be calculated. That is,  $IA_{pq} = MIC_{(x_1, \dots, x_i)} = MIC_{(x_{i-1}, (x_1, \dots, x_{i-2}, x_i))}$ .

Case 3. If  $(x_{i-1}, (x_1, x_2, \dots, x_{i-2})) \notin IV$  and  $(x_i, (x_1, x_2, \dots, x_{i-2})) \notin IV$ , no pruning. Calculate MIC values of all relationships with  $i$  variables. Then the maximal MIC value, which is the value of the element  $IA_{pq}$ , is the MIC value of the  $i$ -variable relationship  $(x_1, x_2, \dots, x_{i-1}, x_i)$  and the corresponding relationship is the specific form of the  $i$  variables. That is,  $IA_{pq} = MIC_{(x_1, \dots, x_i)} = \max_{1 \leq j \leq i} MIC_{(x_j, (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_i))}$ .

There is the pruning process in the proposed MIP algorithm. The relationships, obtained by adding a variable into the exist dependent (independent) relationships in the set  $DR(IV)$ , are pruned. At last, the dependent multi-variable relationships and independent relationships are pinpointed from the data set  $D$ .

In the proposed MIP Algorithm 2, there are two important parameters  $i, \epsilon$ . The integer  $i$  is used to count the number of variables in the current calculated relationships. The integer

**Algorithm 2** The Matrix Iteration Algorithm With Pruning (MIP)

**Require:** The data set  $D_{n \times v}$  is with  $n$  points and  $v$  variables ( $x_i, i = 1, 2, \dots, v$ ).  $DR = \{\}$  is the set of dependent relationships and  $IV = \{\}$  is the set of independent relationships.  $VR$  is the maximal number of variables in detected multivariate relationships. The number  $i$  is the number of variables among current detected relationships. The critical variable  $\varepsilon$ .

```

/* Step 1 Initialization. */
1:  $i = 2$  (Step 1.1).
2: Calculate  $MIC_{(x_p, x_q)}(p, q = 1, 2, \dots, v, p \neq q)$ . (Step 1.2)
3: if  $MIC_{(x_p, x_q)} < \varepsilon$  (Step 1.3) then
4:    $IV = IV \cup \{(x_p, x_q)\}$ 
5: end if
6:  $i = i + 1$  (Step 1.4).
7: Calculate the initialization matrix. (Step 1.5)
   The set of relationships with two variables,  $RS = \{(x_l, x_m), 1 \leq l < m \leq v\}$ .
    $IA = PrunPro(RS, IV)$ .
8: while  $i < VR$  (Step 2) do
   /* Step 2.1 Exam the matrix  $IA_{mn}$  and update the set  $DR$ . */
9:   if the  $k$ -th MIC value of the matrix header is larger than  $\max\{IA_{jk}, j = 1, 2, \dots, m\}$  then
10:    The  $k$ -th relationship  $R_k$  of the matrix header is added into the set  $DR$ 
11:   end if
   /* Step 2.2 Update the set  $IV$ . */
12:    $IV^t = IV, IV = \{\}$ 
13:   Select two elements (relationships) from  $IV^t, R_p =, R_q =$ .
14:   if variables in the first part of  $R_p, R_q$  are the same and there is only one different variable between the second parts, for example,  $R_p = (x_{k_1}, (x_{k_2}, x_{k_3}, \dots, x_{k_i}))$ ,  $R_q = (x_{k_1}, (x_{k_2}, x_{k_3}, \dots, x_{k_{i-1}}, x_{k_{i+1}}))$  then
15:      $NR = (x_{k_1}, (x_{k_2}, x_{k_3}, \dots, x_{k_{i-1}}, x_{k_i}, x_{k_{i+1}}))$ .
16:      $IV = IV \cup \{NR\}$ 
17:   end if
18:    $i = i + 1$ . (Step 2.3)
   /* Step 2.4 Update the iteration matrix  $IA$  with pruning. */
19:    $RS$  is the set of relationships of the header of  $IA$  excluding relationships in  $IV$ .
20:   if  $RS == \{\}$  then
21:     Stop
22:   else
23:      $IA = PrunPro(RS, IV)$ .
24:   end if
25: end while
26: return (Step 3) The precisely identified relationships in the set  $DR$  and the independent relationships in the set  $IV$ .

```

$VR$  is the biggest number of variables in the relationships we want to detect. Another important parameter is  $\varepsilon$ . If  $\varepsilon$  is set too large, many not independent relationships are added into the independent relationship set  $IV$ . Then this will lead to that all relationships are independent and dependent relationships can not be identified. However, if  $\varepsilon$  is set too small, the independent relationship set  $IV$  will be empty at last, and all independent relationships are deemed to be a certain dependent. Then this will lead to that the MIP algorithm will have no pruning. In our opinion, the parameter  $\varepsilon$  should be set to be a little higher than the MIC value of two random variables under the same scale. And the reference values of  $\varepsilon$  under different scales are given in Table 3.

**TABLE 3.** The reference values of  $\varepsilon$ .

scale	$MIC_{(x_1, x_2)}$	$\varepsilon$
1000	0.033438	0.04
2000	0.021450	0.03
5000	0.016927	0.02
10000	0.012088	0.02
20000	0.008597	0.01

The time complexity of MIP is related to the number of independent relationships in the detected data set. If there are many dependent relationships, that is the ratio of the number of dependent relationships to that of independent relationships is relatively higher, the computational complexity of MIP is high and if all relationships are dependent in the data set, the complexity of MIP is equal to  $2^n$  which is the time complexity of enumeration method. However, if dependent relationships is sparse in the data set, that is the ratio of the number of dependent relationships to that of independent relationships is close to 0, the complexity of MIP is much lower than  $2^n$ . In reality, the number of dependent relationships is very small in data sets with high dimensions. And the proposed MIP algorithm is scalable and suitable for pinpointing dependent relationships from data sets.

The MIP algorithm has the following advantages. Firstly, the MIP algorithm has the procedure of pruning and MIC values of some not important relationships are avoided to be calculated. The calculation workload is reduced. Secondly, the MIP algorithm can precisely identify the dependent and independent relationships from data sets with many variables. And the relationships, which can be regarded as adding one or more variables into dependent (independent) relationships, are excluded although these MIC values of these relationships are relatively higher (lower). Thirdly, the dependent and independent relationships are identified at the same time.

**IV. CASE STUDY**

The validity of the proposed MIP algorithm is verified through a simple case in this section. The data set  $D$  has five variables (attributes),  $x_1, x_2, x_3, x_4, y$ , where  $x_1, x_2, x_3$  are mutually independent variables. There are two relationships

TABLE 4. The MIC of all relationships in the data set  $D$ .

Relationship	MIC	Relationship	MIC	Relationship	MIC
$(x_2, x_4)$	0.999741	$(x_3, (x_1, x_4, y))$	0.688464	$(x_1, x_4)$	0.009220
$(x_2, (x_1, x_4))$	0.999128	$(x_1, (x_2, x_3, x_4, y))$	0.682911	$(x_1, x_3)$	0.009217
$(x_2, (x_3, x_4))$	0.998641	$(x_3, (x_1, x_2, x_4, y))$	0.680027	$(x_3, x_4)$	0.009193
$(x_2, (x_1, x_3, x_4))$	0.998473	$(y, (x_1, x_3, x_4))$	0.657815	$(x_2, x_3)$	0.009175
$(x_2, (x_4, y))$	0.996968	$(y, (x_1, x_2, x_3, x_4))$	0.657767	$(y, (x_2, x_4))$	0.009171
$(x_2, (x_3, x_4, y))$	0.996276	$(x_3, y)$	0.374055	$(x_4, (x_1, x_3, y))$	0.009165
$(x_2, (x_1, x_3, x_4, y))$	0.996011	$(x_1, y)$	0.371880	$(x_2, (x_1, x_3, y))$	0.009129
$(x_2, (x_1, x_4, y))$	0.995776	$(x_3, (x_2, y))$	0.370785	$(x_2, (x_1, x_3))$	0.009117
$(x_4, (x_2, x_3))$	0.981239	$(x_1, (x_2, y))$	0.368703	$(x_1, (x_2, x_3))$	0.009102
$(x_4, (x_1, x_2))$	0.980502	$(x_1, (x_4, y))$	0.360307	$(x_2, (x_1, y))$	0.009092
$(x_4, (x_2, y))$	0.966387	$(x_3, (x_4, y))$	0.360291	$(x_4, (x_1, x_3))$	0.009056
$(x_4, (x_1, x_2, x_3))$	0.963928	$(x_1, (x_2, x_4, y))$	0.359547	$(x_4, (x_1, y))$	0.009051
$(x_4, (x_2, x_3, y))$	0.957025	$(x_3, (x_2, x_4, y))$	0.358830	$(x_4, (x_3, y))$	0.009046
$(x_4, (x_1, x_2, y))$	0.953667	$(y, (x_2, x_3))$	0.324844	$(x_1, (x_2, x_3, x_4))$	0.009043
$(x_4, (x_1, x_2, x_3, y))$	0.950137	$(y, (x_1, x_2))$	0.300628	$(x_3, (x_2, x_4))$	0.009027
$(y, (x_1, x_3))$	0.915383	$(y, (x_1, x_4))$	0.268222	$(x_2, y)$	0.009012
$(x_1, (x_3, y))$	0.912997	$(y, (x_1, x_2, x_4))$	0.267902	$(x_4, y)$	0.008995
$(x_3, (x_1, y))$	0.910849	$(y, (x_2, x_3, x_4))$	0.266694	$(x_1, x_2)$	0.008982
$(y, (x_1, x_2, x_3))$	0.794157	$(y, (x_3, x_4))$	0.264570	$(x_3, (x_1, x_2))$	0.008961
$(x_1, (x_2, x_3, y))$	0.790372	$(x_2, (x_3, y))$	0.009284	$(x_1, (x_2, x_4))$	0.008927
$(x_3, (x_1, x_2, y))$	0.790067	$(x_1, (x_3, x_4))$	0.009269	$(x_3, (x_1, x_2, x_4))$	0.008772
$(x_1, (x_3, x_4, y))$	0.689207	$(x_3, (x_1, x_4))$	0.009245		

in the data set  $D$ :  $x_4 = 5x_2 + 1$  (linear relationship) and  $y = \sin(4\pi x_1^2 x_3^2) + x_1 + x_3$  (sine/linear relationship). The scale of the data set  $D$  is 20000, that is  $n = |D| = 20000$ . Obviously, there are two dependent relationships,  $(x_2, x_4)$  and  $(y, (x_1, x_3))$ . And the aim of the MIP algorithm is to precisely identify these two dependent relationships from the data set  $D$ .

If the enumeration method is directly used to detect multi-variable relationships in the data set  $D$  via ranking MIC values of relationships, sort  $C_5^2 + 3C_5^3 + 4C_5^4 + 5C_5^5 = 10+30+20+5 = 65$  MIC values, which are shown in Table 4, of all relationships in the data set  $D$ . There are at least two problems. Firstly, every relationship is examined, even if this relationship is definitely not dependent. Secondly, because MIC values are slightly different for different types of relationships under the same noise level [4], the MIC values of some not dependent relationships by adding independent variables into dependent relationships may be higher than that of some dependent relationships. For example, the MIC value of  $(x_2, (x_1, x_4))$  is higher than that of  $(y, (x_1, x_3))$ . And if only according to MIC values in Table 4, these relationships  $(x_2, (x_1, x_4))$ ,  $(x_2, (x_3, x_4))$ ,  $(x_2, (x_1, x_3, x_4))$ , ...,  $(x_4, (x_1, x_2, x_3, y))$  are more important than the relationship  $(y, (x_1, x_3))$ . However, this result has big error with the fact.

Now, the MIP algorithm is applied to pinpoint the dependent relationships  $(x_2, x_4)$ ,  $(y, (x_1, x_3))$  in the data set  $D$ .

- Step 0: Initialization. There are five variables  $x_1, x_2, x_3, x_4, y$  in the data set  $D$ .  $V = 5, n = |D| = 20000, DR = \{\}, IV = \{\}, \epsilon = 0.01$  (refer to Table 3).
- Step 1: Initialization of the iteration matrix  $IA$ .

Step 1.1:  $i = 2$ .

Step 1.2:  $C_5^2$  MIC values of all bivariate relationships in the data set  $D$  are calculated which are listed in Table 5.

Step 1.3: Select independent pairs. If  $MIC_{(x_i, x_j)} < \epsilon$ ,  $(i, j = 1, 2, \dots, 5, i \neq j)$ ,  $(x_i, x_j)$  is added into the set  $IV$ .  $IV = \{(x_1, x_2), (x_1, x_3), (x_1, x_4), (x_2, x_3), (x_2, y), (x_3, x_4), (x_4, y)\}$ .

Step 1.4:  $i = i + 1 = 3$ .

Step 1.5: Calculate the initialization iteration matrix  $IA$ .

A  $C_5^2 \times C_5^2 (10 \times 10)$  matrix  $IA$  is created in Table 6. Two rows, these calculated bivariate relationships and corresponding MIC values, are added above  $IA$ , and the column of relationships is added on the left of  $IA$ .

\* Because there are two different variables between  $(x_1, x_2)$  and  $(x_3, x_4)$ , then the corresponding element  $IA_{81}$  is ignored. Similarly, elements  $IA_{18}, IA_{19}, IA_{1,10}, IA_{26}, IA_{27}, IA_{2,10}, IA_{35}, IA_{36}, IA_{39}, IA_{45}, IA_{46}, IA_{48}, IA_{53}, IA_{54}, IA_{5,10}, IA_{62}, IA_{64}, IA_{69}, IA_{72}, IA_{73}, IA_{78}, IA_{81}, IA_{84}, IA_{87}, IA_{91}, IA_{93}, IA_{96}, IA_{10,1}, IA_{10,2}, IA_{10,5}$  are also ignored. Elements  $IA_{ii}, i = 1, \dots, 10$  are also ignored, because the variables of the corresponding two bivariate relationships are the same.

\* There are only one different variable between the corresponding two relationships of the element  $IA_{pq}$  in the matrix  $IA$ , and then it can be divided into three cases.

Case 1. Because  $(x_1, x_2) \in IV, (x_1, x_3) \in IV$ , completely prune. The MIC of three-variable relationships  $(x_1, (x_2, x_3)), (x_2, (x_1, x_3))$  and  $(x_3, (x_1, x_2))$  need not to be calculated. That is, the element  $IA_{12}$  needs not to be calculated. Similarly, elements  $IA_{13}, IA_{15}, IA_{17}, IA_{21}, IA_{23}, IA_{25}, IA_{28}, IA_{31}, IA_{32}, IA_{38}, IA_{31}, IA_{51}, IA_{52}, IA_{57}, IA_{58}, IA_{71}, IA_{75}, IA_{7,10}, IA_{82}, IA_{83}, IA_{85}, IA_{8,10}, IA_{10,3}, IA_{10,7}, IA_{10,8}$  needs not to be calculated.

Case 2. Because  $(x_1, x_2) \in IV$  and  $(x_1, y) \notin IV$ , partly prune. The MIC of relationships  $(x_1, (x_2, y)), (x_2, (x_1, y))$



TABLE 5. The MIC values of bivariate relationships in the data set D.

Relationship	MIC	Relationship	MIC	Relationship	MIC
$(x_1, x_2)$	0.008982	$(x_2, x_3)$	0.009175	$(x_3, y)$	0.374055
$(x_1, x_3)$	0.009217	$(x_2, x_4)$	0.999741	$(x_4, y)$	0.008995
$(x_1, x_4)$	0.009220	$(x_2, y)$	0.009012		
$(x_1, y)$	0.371880	$(x_3, x_4)$	0.009193		

TABLE 6. Iteration matrix ( $i = 3$ ) in the data set D.

	$(x_1, x_2)$	$(x_1, x_3)$	$(x_1, x_4)$	$(x_1, y)$	$(x_2, x_3)$	$(x_2, x_4)$	$(x_2, y)$	$(x_3, x_4)$	$(x_3, y)$	$(x_4, y)$
	0.008982	0.009217	0.009220	0.37188	0.009175	0.999741	0.009012	0.009193	0.374055	0.008995
$(x_1, x_2)$	—	—	—	0.300628	—	0.980502	—	—	—	—
$(x_1, x_3)$	—	—	—	0.915383	—	—	—	—	0.915383	—
$(x_1, x_4)$	—	—	—	0.268222	—	0.999128	—	—	—	—
$(x_1, y)$	0.300628	0.915383	0.268222	—	—	—	0.368703	—	0.915383	0.360307
$(x_2, x_3)$	—	—	—	—	—	0.981239	—	—	0.324844	—
$(x_2, x_4)$	0.980502	—	0.991281	—	0.981239	—	0.966387	0.998641	—	0.996968
$(x_2, y)$	—	—	—	0.368703	—	0.966387	—	—	0.370785	—
$(x_3, x_4)$	—	—	—	—	—	0.998641	—	—	0.264570	—
$(x_3, y)$	—	0.915383	—	0.915383	0.324844	—	0.370785	0.26457	—	0.360291
$(x_4, y)$	—	—	—	0.360371	—	0.996968	—	—	0.3602914	—

TABLE 7. Iteration matrix ( $i = 4$ ) in the data set D.

	$(x_1, x_2, y)$	$(x_1, x_3, y)$	$(x_1, x_4, y)$	$(x_2, x_3, y)$	$(x_3, x_4, y)$
	0.368703	0.915383	0.360307	0.370785	0.360291
$(x_1, x_2, y)$	—	c1	—	c1	—
$(x_1, x_3, y)$	c1	—	c3	c1	c3
$(x_1, x_4, y)$	—	c3	—	—	c3
$(x_2, x_3, y)$	c1	c1	—	—	—
$(x_3, x_4, y)$	—	c3	c3	—	—

TABLE 8. P values of t-test of deviations of calculating time of the MIP algorithm under different dimensions.

Dimensions 1	Dimensions 2	P value	Dimensions 1	Dimensions 2	P value
10	15	1.47E-13	15	25	1.34E-13
10	20	3.28E-12	15	30	7.65E-12
10	25	2.57E-12	20	25	5.12E-14
10	30	5.92E-13	20	30	8.52E-12
15	20	3.89E-13	25	30	4.91E-13

need not be calculated and only calculate the MIC of relationship  $(y, (x_1, x_2))$ .  $MIC_{(y,(x_1,x_2))} = 0.300628$  and  $IA_{14} = MIC_{(x_1,x_2,y)} = MIC_{(y,(x_1,x_2))} = 0.300628$ . Similarly, elements  $IA_{16}, IA_{24}, IA_{29}, IA_{34}, IA_{36}, IA_{3,10}, IA_{29}, IA_{41}, IA_{42}, IA_{43}, IA_{47}, IA_{4,10}, IA_{56}, IA_{59}, IA_{61}, IA_{63}, IA_{29}, IA_{65}, IA_{67}, IA_{68}, IA_{6,10}, IA_{74}, IA_{76}, IA_{79}, IA_{86}, IA_{89}, IA_{92}, IA_{94}, IA_{95}, IA_{97}, IA_{98}, IA_{9,10}, IA_{10,4}, IA_{10,6}, IA_{10,9}$  are also calculated.

Case 3. Because  $(x_1, y) \notin IV$  and  $(x_3, y) \notin IV$ , no prune.  $IA_{49} = MIC_{(x_1,x_3,y)} = \max\{MIC_{(x_1,(x_3,y))} = 0.912997, MIC_{(x_3,(x_1,y))} = 0.910849, MIC_{(y,(x_1,x_3))} = 0.915383\} = 0.915383$ .

- Step 2(1):  $i = 3$  and  $i < VR = 5$ .

Step 2.1(1): Exam the iteration matrix  $IA$  and update the set  $DR$ .

Exam every column of the matrix  $IA$

in Table 6. The MIC of the relationship  $(x_2, x_4)$ ,  $MIC_{(x_2,x_4)} = 0.999741$ , is larger than any element of the 6-th column of the matrix  $IA$ . Then  $DR = DR \cup \{(x_2, x_4)\}$ .

Step 2.2(1) : Update the set  $IV$ .

$IV = IV \cup \{(x_1, (x_2, x_3)), (x_1, (x_2, x_4)), (x_2, (x_1, x_3)), (x_2, (x_1, y)), (x_1, (x_3, x_4)), (x_3, (x_1, x_2)), (x_3, (x_1, x_4)), (x_4, (x_1, x_3)), (x_4, (x_1, y)), (x_2, (x_3, y)), (x_3, (x_2, y)), (x_3, (x_2, x_4)), (y, (x_2, x_4)), (x_4, (x_3, y))\}$ .

Step 2.3(1):  $i = i + 1 = 4$ .

Step 2.4(1): Update the iteration matrix  $IA$  with pruning.

All of the relationships are  $(x_1, (x_2, y)), (y, (x_1, x_2)), (x_1, (x_4, y)), (y, (x_1, x_4)), (x_3, (x_4, y)), (y, (x_3, x_4)), (y, (x_1, x_3)), (x_1, (x_3, y)), (x_3, (x_1, y)), (y, (x_2, x_3)),$

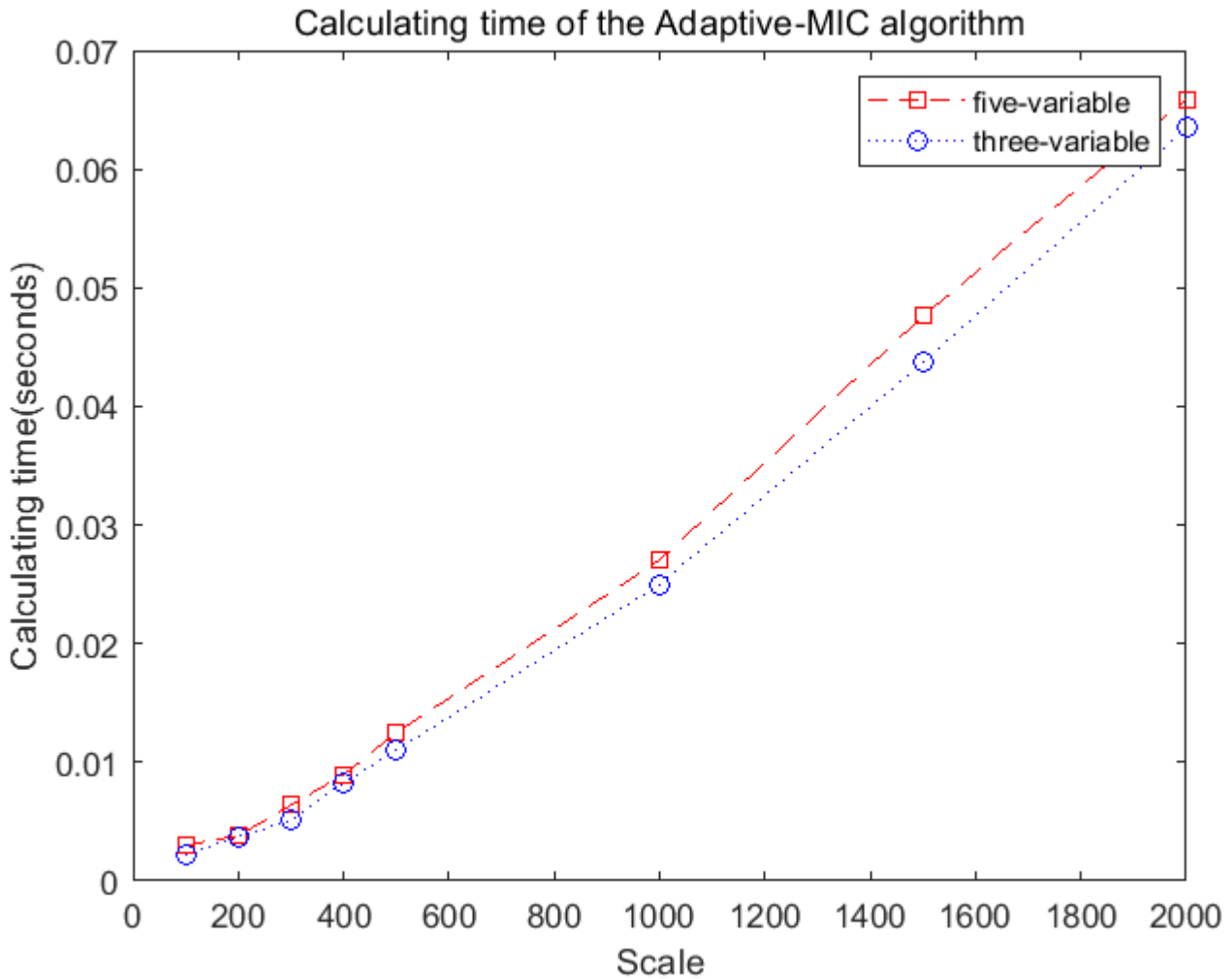


FIGURE 1. The calculating time of the fast Adaptive-MIC algorithm.

$(x_3, (x_2, y)), (x_2, (x_3, x_4)), (x_4, (x_2, x_3)),$   
 $(x_2, (x_4, y)), (x_4, (x_2, y)), (x_4, (x_1, x_2))$   
 in the last iteration matrix  $IA$ . Because the relationship  $(x_2, x_4)$  is added into the set  $DR$ , then the relationships in the column of  $(x_2, x_4)$ , that is  $(x_4, (x_2, x_3)), (x_2, (x_4, y)), (x_4, (x_2, y)), (x_4, (x_1, x_2))$ , in the last iteration matrix  $IA$  are excluded from all relationships in the last iteration matrix  $IA$ . Then the remained relationships are  $\{(x_1, (x_2, y)), (y, (x_1, x_2)), (x_1, (x_4, y)), (y, (x_1, x_4)), (x_3, (x_4, y)), (y, (x_3, x_4)), (y, (x_1, x_3)), (x_1, (x_3, y)), (x_3, (x_1, y)), (y, (x_2, x_3)), (x_3, (x_2, y)), (x_2, (x_3, x_4)), (x_4, (x_2, x_3))\}$ . And  $MIC_{(x_1, x_2, y)} = \max\{MIC_{(x_1, (x_2, y))}, MIC_{(y, (x_1, x_2))}\}$ .  $MIC_{(x_1, x_4, y)} = \max\{MIC_{(x_1, (x_4, y))}, MIC_{(y, (x_1, x_4))}\}$ .  $MIC_{(x_3, x_4, y)} = \max\{MIC_{(x_3, (x_4, y))}, MIC_{(y, (x_3, x_4))}\}$ .  $MIC_{(x_1, x_3, y)} = \max\{MIC_{(x_1, (x_3, y))}, MIC_{(y, (x_1, x_3))}\}$ .

$MIC_{(x_3, (x_1, y))}$ .  $MIC_{(x_2, x_3, y)} = \max\{MIC_{(x_3, (x_2, y))}, MIC_{(y, (x_2, x_3))}\}$ . Then the first row above  $IA$  and the column on left of  $IA$  are relationships  $(x_1, x_2, y), (x_1, x_4, y), (x_3, x_4, y), (x_1, x_3, y), (x_2, x_3, y)$  and the second row above  $IA$  are these corresponding MIC values.

The updated matrix  $IA$  is shown in Table 7.  $c1 = MIC_{(y, (x_1, x_2, x_3))} = 0.794157$ ,  $c3 = MIC_{(y, (x_1, x_3, x_4))} = 0.657815$ . Because  $(x_1, (x_4, y)) \in IV$ , the MIC value of the relationship  $(x_1, (x_2, x_4, y))$  need not be calculated. Similarly, these MIC values of relationships  $(x_2, (x_1, x_4, y)), (x_4, (x_1, x_2, y)), (y, (x_1, x_2, x_4))$  also need not be calculated. Then the element  $IA_{31}$  is ignored. Similarly, elements  $IA_{13}, IA_{15}, IA_{34}, IA_{43}, IA_{45}, IA_{51}, IA_{54}$  are also ignored. Because variables in the corresponding two relationships of  $IA_{ii}, i = 1, 2, \dots, 5$ , are the same, these

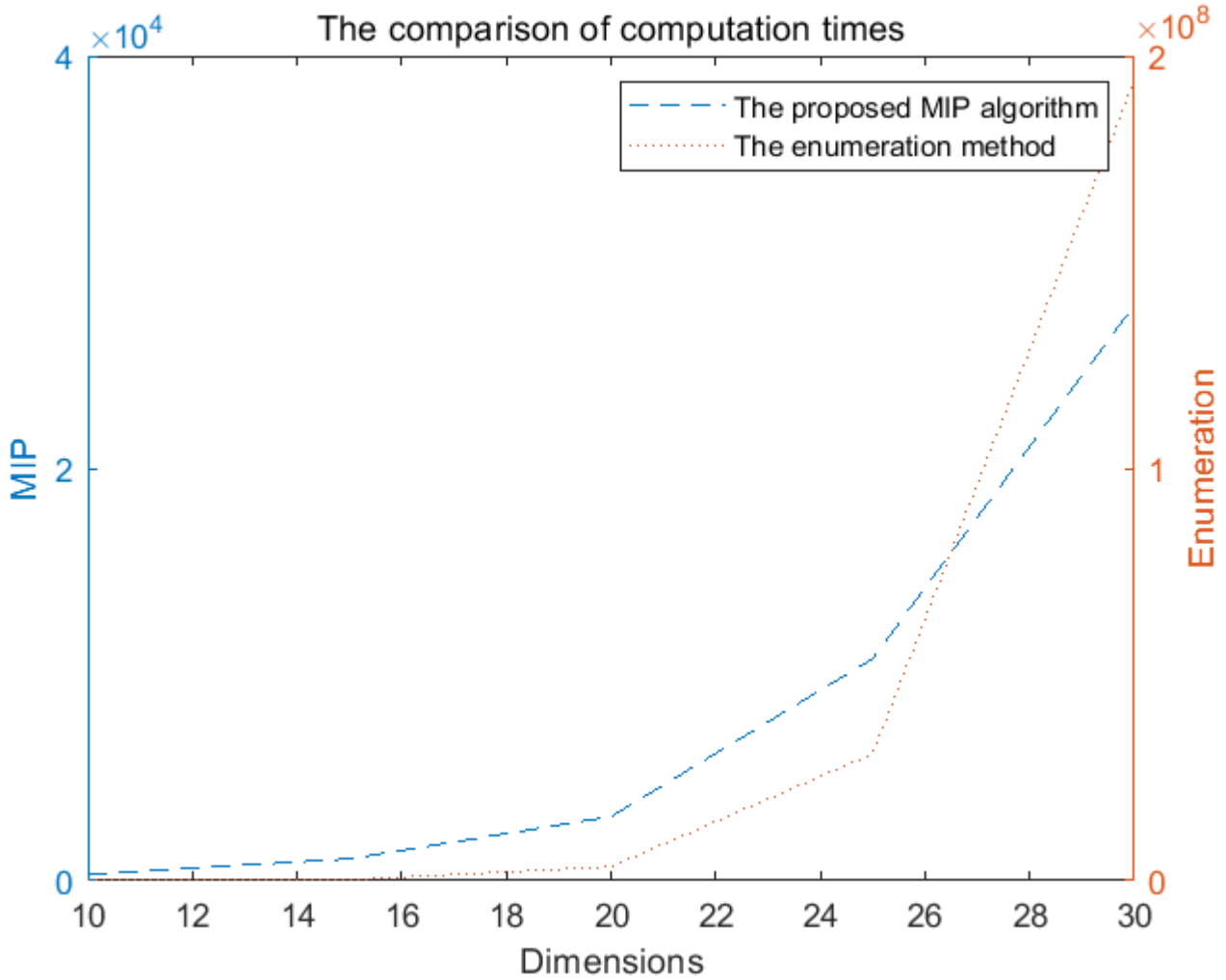


FIGURE 2. The comparison of calculation times between the proposed MIP algorithm and the enumeration method.

elements  $IA_{ij}$ ,  $i = 1, 2, \dots, 5$ , are also ignored.

- Step 2(2):  $i = 4$  and  $i < VR = 5$ .

Step 2.1(2): Exam the iteration matrix  $IA$  and update the set  $DR$ .

Exam every column of the matrix  $IA$ . The MIC value of the relationship  $(y, (x_1, x_3))$ ,  $MIC_{(y, (x_1, x_3))} = 0.915383$ , is larger than any element of the 2-th (the corresponding) column of the matrix  $IA$ .  $DR = DR \cup \{(y, (x_1, x_3))\} = \{(x_2, x_4), (y, (x_1, x_3))\}$ .

Step 2.2(2): Update the set  $IV$ .

$IV = IV \cup \{(x_1, (x_2, x_3, x_4)), (x_2, (x_1, x_3, y)), (x_3, (x_1, x_2, x_4)), (x_4, (x_1, x_3, y))\}$ .

Step 2.3(2):  $i = i + 1 = 5$ .

Step 2.4(2): Update the iteration matrix  $IA$  with pruning. In the Step 2.3(1), the MIC of relationships  $(y, (x_1, x_2, x_3))$  and  $(y, (x_1, x_3, x_4))$  are calculated. But  $(y, (x_1, x_3)) \in DR$ , the relationships  $(y, (x_1, x_2, x_3))$  and  $(y, (x_1, x_3, x_4))$

are ignored. And the iteration matrix  $IA$  is blank, that is  $IA = [ ]$ .

Goto Step 3.

- Step 3: Output.

The dependent relationships in the data set  $D$  are relationships  $(x_2, x_4)$  and  $(y, (x_1, x_3))$  in the set  $DR$ . The independent relationships are in the set  $IV$ . Stop.

The pinpointed dependent relationships are in the set  $DR$ ,  $DR = \{(x_2, x_4), (y, (x_1, x_3))\}$ . The other MIC values of relationships calculated in the MIP algorithm can be stored as a reference. The results agree well with the reality.

Besides, the calculation workload is reduced. The number of relationships, of which MIC values are calculated, is  $C_5^2 + 17 + 2 = 29$ . However, the number of relationships is 65 if the enumeration method is employed. More than 50% calculation is saved. The proposed MIP algorithm not only can pinpoint the exist dependent relationships in big data, but also can reduce calculation workload compared to the enumeration method.

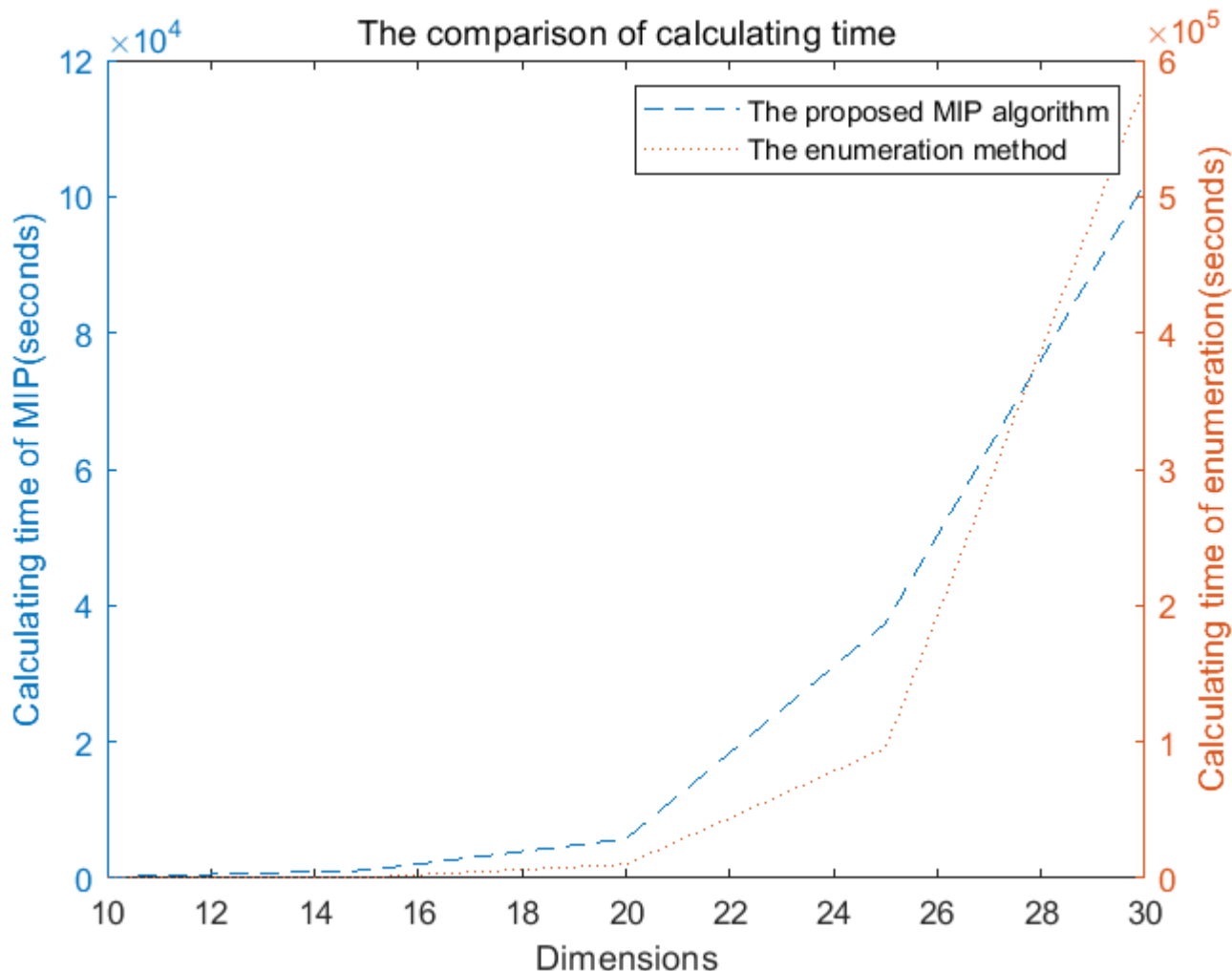


FIGURE 3. The comparison of calculating time between the MIP algorithm and the enumeration method.

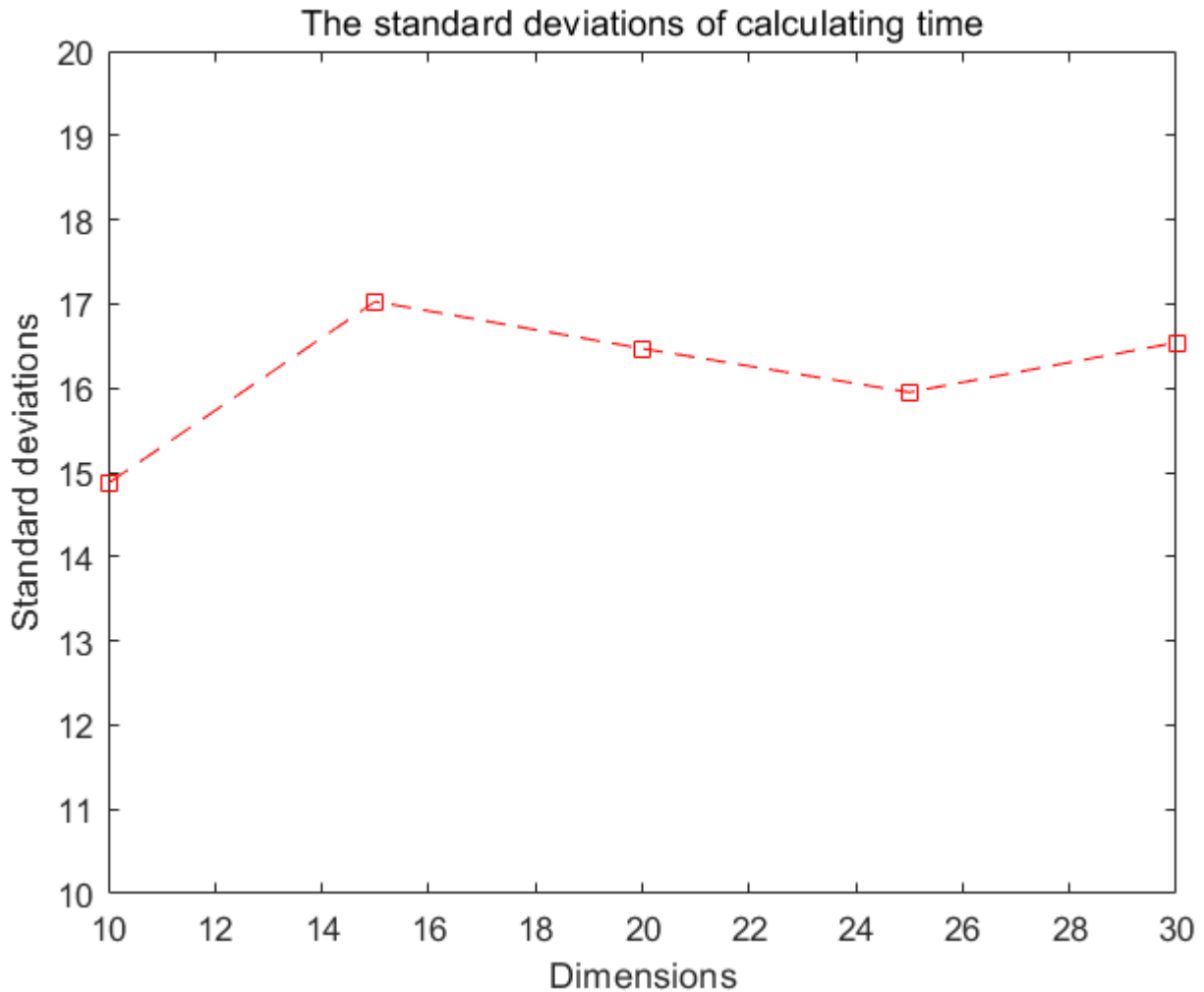
**V. EXPERIMENTAL RESULTS**

In the enumeration method, MIC values of relationships are calculated by the Adaptive-MIC algorithm proposed by Shao *et al.* [10] which is implemented in C programming language and parameters are default, that is,  $\alpha = 0.6$  and  $C = 15$ . The proposed MIP algorithm is also implemented in C programming language. The computing platform of the two C programs is personal notebook computer and its configuration is as following. Win 8 Operating System; CPU: Intel(R) Core(TM) i7 – 4510U, 2.60GHz; RAM: 8.00GB.

In order to compare calculating time between the proposed MIP algorithm and the enumeration method, the time of the fast Adaptive-MIC algorithm, which is used in the MIP algorithm and the enumeration method, calculating MIC values of three-variable and five-variable relationships under different scales [10], is firstly shown in Fig. 1. With the increasing of the scale, the calculating time increases for three-variable and five-variable relationships. When scales are the same, the calculating time of three-variable relationships is slightly lower than that of five-variable relationships

overall. However, the difference of calculating time between three-variable relationships and five-variable relationships is extremely tiny under the same scale. The calculating time of the Adaptive-MIC algorithm for relationships with different variables is almost equal under the same scale. Then, in the enumeration method, the time of calculating an MIC value of a relationship, no matter the number of variables in the relationship, is approximated as the time of calculating the MIC value of the relationship with five variables.

From the simple example in Section IV, it can be found that the proposed MIP algorithm can significantly reduce calculation times of MIC values of multivariate relationships in the procedure of detecting dependent multivariate relationships in data sets. In all data sets of experiments in Fig. 2, there are two dependent multivariate relationships,  $x_3 = 3x_2 + 5x_1$ ,  $x_4 = x_5^2 - 6$ , and the other variables are independent. With the increasing of the number of variables in the data set, the calculation times of MIC values of multivariate relationships are increasing. However, the calculation times of MIC values of the proposed MIP algorithm is considerably lower than



**FIGURE 4.** The standard deviations of calculating time of the MIP algorithm.

that of the enumeration method, and the increasing speed of calculation times of the proposed MIP algorithm is also much slower than that of the enumeration method. The calculation times of the proposed MIP algorithm is much lower than that of the enumeration method. If the dimension of the data set is lower than 15, the calculating time of MIP is longer than that of the enumeration method. And if the dimension of the data set is equal to or higher than 20, the calculating time of MIP is much shorter than that of the enumeration method. When the dimension of the data set is high, for example larger than 30, about 1 order of magnitude of time cost is reduced by the MIP algorithm compared to the enumeration method. If there are 30 variables in the data set, compared to the enumeration method, 4 orders of magnitude of calculation times are reduced in the proposed MIP algorithm. The MIP algorithm can significantly reduce the calculation times of MIC values in the procedure of detecting dependent relationships in data sets and the MIP algorithm is suitable for detecting multivariate dependent relationships in high dimensional data sets.

In Fig. 3, the calculating time of MIP algorithm is the average time. In order to further investigate the variation of calculating time, the standard deviations of calculating time is given in Fig. 4. From Fig. 4, it can be found that the variation of deviations of the calculating time varies very small and the proposed MIP algorithm is relatively stable in the aspect of calculating time. And the further investigate is analyzed by the t-test. In the t-test, the null hypothesis is that the two deviations are equal. The p-value of the t-test of any two deviations is shown in Tab. 8. From Tab. 8, it can be found that these p-values are all less than confidence level 0.01. Then it can be concluded that there is no significant differences between any two deviations. That is all deviations of calculating time are equal under different dimensions. And the variation of calculating time of the MIP algorithm is the same for different dimensions.

In Fig. 2, the comparison of calculation times of MIC values between the MIP and enumeration method is shown. Because there are some other operations in the MIP algorithm besides calculating MIC values of multivariate relationships

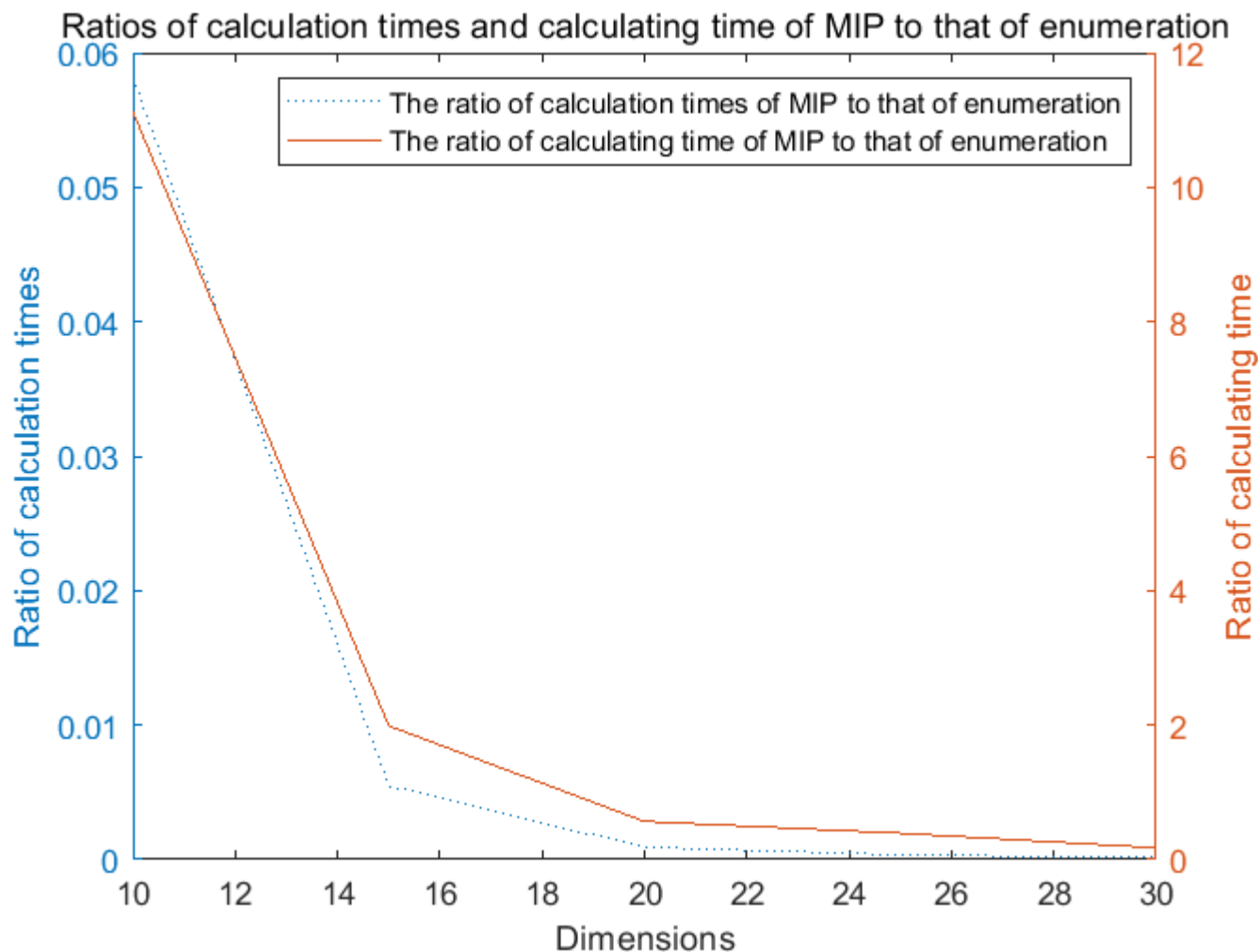


FIGURE 5. The ratio of the calculation times of MIP to that of enumeration and the ratio of the calculating time of MIP to that of enumeration.

by employing the Adaptive-MIC algorithm, the comparison of calculating time between the MIP and enumeration method is also shown in Fig. 3. The scale of the data set in experiments is 100 and the calculating time of the enumeration method is equal to the product of calculation times and the time of calculating an MIC value of five-variable relationships with 100 data points. The calculating time of the MIP algorithm is the CPU time. When the dimension is low (10, 15), the calculating time of the MIP algorithm (168.69, 1154.51 seconds) is higher than that of the enumeration method (15.16, 580.64 seconds). However, when the dimension is high (20), the calculating time of MIP (5776.9 seconds) is much lower than that of the enumeration method (10185.33 seconds), and if the dimension is higher than 20, about 1 order of magnitude of the calculating time is reduced in the MIP algorithm compared to the enumeration method. The proposed MIP algorithm is suitable for detecting dependent relationships in high dimensional data set and much calculating time can be saved.

In order to make the comparison of calculation times and calculating time more intuitively, the ratio of the calculation

times of the MIP algorithm to that of the enumeration method and the ratio of the calculating time of the MIP algorithm to that of the enumeration method are displayed in Fig. 5. Both the ratio of the calculation times of the MIP algorithm to that of the enumeration method and the ratio of the calculating time of the MIP algorithm to that of the enumeration method decrease with dimension. The ratio of the calculation times of the MIP algorithm to that of the enumeration method is very small even when the dimension is very low. And the ratio of calculation times declines with dimensions. However, the ratio of calculating time of the MIP algorithm to that of the enumeration method is very large when the dimension is very low. With the increasing of dimension, the ratio of calculating time rapidly declines to less than 1 and the ratio of calculating time is close to zero when the dimension is large. When dimension is large in the data set, both ratios are very low which is close to zero. That is, when dimension is large, the the calculating time and calculation times of the MIP algorithm are much lower than the calculating time and calculation times of the enumeration method, respectively.

However, there is a big difference between the ratio of the calculation times and the ratio of the calculating time. No matter how large the dimension, the ratio of the calculation times of the MIP algorithm to that of the enumeration method is very low. That is, the calculation times of the MIP algorithm is much lower than that of the enumeration method. When the dimension is lower than 20, the ratio is larger than 1. That is, the calculating time of the MIP algorithm is much longer than that of the enumeration method. When the dimension is equal to 10, the ratio of the calculating time is equal to 11.12, that is, the calculating time of the MIP algorithm is about 11 times of that of the enumeration method. There is a big difference of the calculating time between the low dimension and the large dimension. The reason is as follows. When the dimension is low, the saved time of calculation of MIC of multivariate relationships is not longer than the time of increased processes. And when the dimension is large, the saved time of calculation of MIC is much longer than the time of increased processes. The proposed MIP is suitable for detecting dependent multivariate relationships in high dimensional data sets.

## VI. CONCLUSION

The matrix iteration with pruning algorithm (MIP) is proposed for pinpointing a small amount of multivariate dependent relationships from high dimensional data sets. The MIP algorithm can also be seen as a framework, in which other excellent coefficients can replace the maximal information coefficient (MIC) as the measure of correlations in future. In MIP, there is a pruning process by which some not dependent relationships with relatively higher correlation values are discarded. Then the calculation load is significantly reduced. With the increasing of sparsity, that is the ratio of the number of variables in dependent relationships to the number of all variables in data sets, of data sets, the ratio of the number of correlation values calculated by the MIP algorithm to that of the enumeration method is decreasing and the calculating time of the MIP algorithm is greatly reduced. Without calculating all correlation values of all multivariate relationships, the proposed MIP algorithm can pinpoint correlations among high dimensional data sets.

## REFERENCES

- [1] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014.
- [2] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How we Live, Work, and Think*. New York, NY, USA: Houghton Mifflin Harcourt, 2013, pp. 15–64.
- [3] F. Shao, S. Yang, B. Sun, L. Jia, Y. Dong, and D. Wang, "The big data analysis of rail equipment accidents based on the maximal information coefficient," *J. Transp. Saf. Secur.*, vol. 12, no. 7, pp. 959–976, Aug. 2020.
- [4] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011.
- [5] D. N. Reshef, Y. A. Reshef, P. C. Sabeti, and M. Mitzenmacher, "An empirical study of the maximal and total information coefficients and leading measures of dependence," *Ann. Appl. Statist.*, vol. 12, no. 1, pp. 123–155, Mar. 2018.
- [6] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138.
- [7] J. Miter and G. J. Goodhill, "Limitations to estimating mutual information in large neural populations," *Entropy*, vol. 22, no. 4, p. 290, 2020.
- [8] C. E. Shannon, "A mathematical theory of communication," *ACM SIG-MOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
- [9] T. F. Móri and G. J. Székely, "Four simple axioms of dependence measures," *Metrika*, vol. 82, no. 1, pp. 1–16, Jan. 2019.
- [10] F. Shao, K. Li, and Y. Dong, "Identifying multi-variable relationships based on the maximal information coefficient," *Intell. Data Anal.*, vol. 21, no. 1, pp. 151–166, Jan. 2017.
- [11] F. Shao, K. Li, and X. Xu, "Railway accidents analysis based on the improved algorithm of the maximal information coefficient," *Intell. Data Anal.*, vol. 20, no. 3, pp. 597–613, Apr. 2016.
- [12] D. Albanese, S. Riccadonna, C. Donati, and P. Franceschi, "A practical tool for maximal information coefficient analysis," *GigaScience*, vol. 7, no. 4, Apr. 2018, Art. no. gyy032.
- [13] R. Heller, Y. Heller, and M. Gorfine, "A consistent multivariate test of association based on ranks of distances," *Biometrika*, vol. 100, no. 2, pp. 503–510, Jun. 2013.
- [14] H. Rootzén, J. Segers, and J. L. Wadsworth, "Multivariate generalized Pareto distributions: Parametrizations, representations, and properties," *J. Multivariate Anal.*, vol. 165, pp. 117–131, May 2018.
- [15] R. Heller, Y. Heller, S. Kaufman, B. Brill, and M. Gorfine, "Consistent distribution-free  $K$ -sample and independence tests for univariate random variables," *J. Mach. Learn. Res.*, vol. 17, no. 29, pp. 1–54, 2016.
- [16] G. J. Székely and M. L. Rizzo, "Brownian distance covariance," *Ann. Appl. Statist.*, vol. 3, no. 4, pp. 1236–1265, 2009.
- [17] R. Lyons, "Distance covariance in metric spaces," *Ann. Probab.*, vol. 41, no. 5, pp. 3284–3305, Sep. 2013.
- [18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.
- [19] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *Ann. Statist.*, vol. 41, no. 5, pp. 2263–2291, Oct. 2013.
- [20] A. Rényi, "On measures of dependence," *Acta Math. Acad. Sci. Hungarica*, vol. 10, nos. 3–4, pp. 441–451, 1959.
- [21] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Stat. Assoc.*, vol. 80, no. 391, pp. 580–598, Sep. 1985.
- [22] D. Lopezpaz, P. Hennig, and S. Bernhard, "The randomized dependence coefficient," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 1–9.
- [23] H. V. Nguyen, E. Müller, J. Vreeken, P. Efron, and K. Böhm, "Multivariate maximal correlation analysis," in *Pro. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 775–783.
- [24] E. Domanovitz and U. Erez, "On the importance of asymmetry and monotonicity constraints in maximal correlation analysis," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 3112–3116.
- [25] D. Reshef, Y. Reshef, M. Mitzenmacher, and P. Sabeti, "Equitability analysis of the maximal information coefficient, with comparisons," 2013, *arXiv:1301.6314*. [Online]. Available: <http://arxiv.org/abs/1301.6314>



**FUBO SHAO** received the B.S. degree in information management and information system and the M.S. degree in operational research from the Shandong University of Science and Technology (SDUST), China, in 2006 and 2009, respectively, and the Ph.D. degree in system science from Beijing Jiaotong University, Beijing, in 2017. He is currently the Postdoctoral Researcher with CRRC Corporation Limited. His research interests include statistical learning, artificial intelligence, association rules, and their applications in railway transportation and safety.



**ZHIQIANG HOU** was born in Zhangjiakou, Hebei, China, in 1978. He received the B.S. and M.S. degrees in port, waterway and coastal engineering from Southeast University, Nanjing, China, in 2004, and the Ph.D. degree in communication and transportation engineering from Beijing Jiaotong University, Beijing, in 2017.

Since 2017, he has been a Researcher with the China Waterborne Transport Research Institute. He is the author of two books, more than 40 articles, ten norms, 11 technical awards, and five patents. His research interests include accident causation and risk assessment of engineering.



**ZHE ZHANG** was born in 1988. He received the Ph.D. degree in transportation engineering from Beijing Jiaotong University, in 2017. He is currently a Researcher with the State Key Laboratory of Rail Traffic Control, Beijing Jiaotong University. His research interests include pedestrian flow simulation, crowd control, and facility optimization in terminals.

...



**LIMIN JIA** received the Ph.D. degree in automation and control in transportation from the China Academy of Railway Sciences, Beijing, China, in 1991. He is currently a Professor with Beijing Jiaotong University, and the Chief Scientist of the National Center of Collaborative Innovation Center for Rail Safety and the State Key Laboratory of Rail Traffic Control and Safety. He is the first batch of Millions of Leading Engineering Talents Project. His research interests include intelligent

transportation systems, computational intelligence, and rail traffic control and safety.