IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 11, 2019, accepted June 17, 2019, date of publication June 21, 2019, date of current version August 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2924341

Markov Boundary Discovery Based on Variant Ridge Regularized Linear Models

SHU YAN[1,2], CHAOYUAN CUI[1], BINGYU SUN[1], AND RUJING WANG[1]
[1]Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China
[2]University of Science and Technology of China, Hefei 230026, China

Corresponding author: Chaoyuan Cui (cycui@iim.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773360, and in part by the National Key Research and Development Program under Grant 2018YFD0700302.

ABSTRACT It has been proved that the modified form of ridge regularized linear models (MRRLMs) can get ''very close'' to identifying a subset of Markov boundary. However, it is assumed that the covariance matrix is non-singular, so MRRLMs cannot be applied to discover the Markov boundary (subset) from data sets when the covariance matrix is singular. The singularity of the covariance matrix means that there are some collinear variables in the data sets, and such data sets exist widely in the real world. In this paper, we present a novel variant of ridge regularized linear models (VRRLMs) to identify a subset of Markov boundary from data sets with collinear and non-collinear variables and, then, reveal the relationship between covariance matrix and collinearity of variables in the theory. In addition, we prove theoretically that the VRRLMs can identify a subset of Markov boundary under some reasonable assumptions and verify the theory on the four discrete data sets. The results show that VRRLMs outperform the MRRLMs in discovering a subset of Markov boundary on the data sets with collinear variables, while both of them have a similar discovery efficiency of the Markov boundary (subset) on the data sets with non-collinear variables.

INDEX TERMS CRP_δ, Markov boundary, Markov blanket, variant ridge regression models, linear regression models.

I. INTRODUCTION

Discovering causal relationships among variables from observation data sets is fundamental to the discipline, such as computer science, medicine, statistics, economics and social science [1]–[4]. Moreover, the causal relationships have been widely accepted as an alternative to randomized controlled trials(RCTs) [5]–[7]. In most cases, RCTs are impractical to discover causal relationship from the observational data due to expensive, unethical or impossible [8]–[10].

As a directed acyclic graph model [4], a Bayesian network can represent causal relationships among all nodes or variables in the network. Specifically, for a given target node Y in the graph, it is only related to its parents, children, and spouses of Y and independent with other nodes. The set of its parents, children, and spouses of Y is called Markov blanket (*MB*) of Y, which is different from Fuzzy Markov [11]. The property of *MB* is widely used the feature selection algorithm for classification or regression in the field of machine learning [12]–[14]. In 1996, Koller and Sahami from Stanford University, first associated Markov blanket and feature selection, proved theoretically that a Markov blanket of Y on a Bayesian network is an optimal feature subset of variables for Y on the corresponding data set [15], while feature selection as one of the important pre-processing methods in the field of machine learning has promoted the application and development of Markov blanket theory tremendously. Conversely, under some reasonable assumptions, an optimal feature subset for Y is also a Markov boundary (blanket) of Y on the corresponding Bayesian network. so it is another method for obtaining Markov boundary (blanket) by getting an optimal feature subset. This article focuses on this idea and does some work.

Over the past two decades, many algorithms were proposed for discovering Markov blanket from observational data sets. According to literature [16], at least 18 kinds of Markov blanket induction algorithms, such as the famous IAMB [17], HITON-MB [18], and PCMB [19], were mentioned from 1996 to 2013, and more *MB* induction algorithms were emerged in the recent five years [20]–[24]. However most of these algorithms are based on the conditional independent test, the main reason is that the probability and topological

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

structure information of Bayesian network are helpful to define the constraint condition effectively [16]. Other algorithms based on scoring, which widely used in Bayesian network structure learning, were rarely used in Markov blanket induction. Only two algorithms namely DMB and TPDMB [25] were proposed to identify Markov boundary in 2013.

In recent years, a new method based on regularized linear models was proposed to identify Markov blanket or boundary from data sets. At present, there are only two articles related to regularized linear models used to discover Markov boundary or blanket. Literature [26] focused on Bayesian network structure learning, while literature [27] theoretically proved that a subset of Markov boundary can get from a solution of MRRLMs. These algorithms have good performance in certain special conditions or on the special data sets.

The main contribution of MRRLMs recovers relationship between Markov boundary and solution of MRRLMs by combining ideas in Markov boundary and sufficient dimension reduction theory. One of the important assumptions of MRRLMs is that the covariance matrix is positive definite and non-singular, hence, MRRLMs can't be applied to the case where the covariance matrix is singular. The singularity of the covariance matrix means that there are some collinear variables in the data sets, and such data sets exist widely in the real world. The second deficiency of MRRLMs is that the response variable is multi-dimensional, but in practice, the response variable is usually one-dimensional. Hence, to tackle the challenges above, we propose new models (VRRLMs) based on regularized linear models, which is also our first motivation to write this article.

The second motivation is that using the regularized linear models to discover *MB* is still at stage of theoretical research, there are many unknowns that need to be explored and proved. For example, how to choose the appropriate expectation of number of variables within *MB*, and the simplification of covariance matrix operation, etc. We hope that more researchers will take part in this work in the future.

It should be emphasized that *MB* theory has a strong practical value in practice. For example, *M*B used for feature selection can reduce the data set dimension, reduce the search space dimension, and save time and storage cost for data collection, storage, transmission and preprocessing [16]. *MB* used for causality is often used to instead of randomized controlled trials (RCTs) to discover causality in data sets due to *MB* containing causal variables.

In this paper, we limit the response variable as one-dimensional variable and propose VRRLMs to identify a subset of Markov boundary from data sets with collinear and non-collinear variables.

In summary, our contributions of this paper are listed as follows:

1) We propose VRRLMs to tackle the problem of MRRLMs. Experiments on the data sets with collinear variables show that our proposed method is better than MRRLMs.
2) We prove VRRLMs in theory and demonstrate the performance of VRRLMs in discovering subset of *MB* from data sets with collinear and noncollinear variables.

The structure of this article is as follows. The next section briefly reviews the work related to *MB* and feature selection as well as relationships between *MB* and feature variables. The third section briefly introduces the theoretical foundation of MRRLMs. In the fourth section, we reveal relationships between collinearity of variables and singularity of the corresponding covariance matrix, and then propose VRRLMs and prove it theoretically under some assumptions. Section five shows the experimental results and analysis, and section six concludes.

II. RELATE WORK

Existing algorithms are generally divided into two categories: constraint and scoring, Constraint-based algorithms can be further divided into two categories: algorithms based on conditional independence test (ACIT) and algorithms based on topology structure information (ATSI) [16]. But the mainstream induction algorithms are still based on constraint, so constraint-based algorithms occupy a lot of space.

ACIT are directly constructed according to the definition of Markov blanket. This kind of algorithms has a simple search strategy, so that it has high time efficiency. Unfortunately, its data efficiency is not high, therefore it needs more samples. ACIT can be traced back to K&S [15] which not only requires the expectation of the number of variables within Markov boundary (blanket) but also requires the number of variables to be deleted in advance. To meet this challenge, Margaritis et al. provided a new *MB* induction algorithm GSMB [28] for constructing a complete Bayesian network through Markov blanket local properties, but the implementing process of GSMB is fundamentally different from that of K&S. The implementing process of K&S is to calculate distribution distance with the help of the information entropy theory; while the implementing process of GSMB is divided into two stages: growth and shrink. The greedy heuristic strategy in the growth stages makes a variable as a candidate variable within Markov blanket as long as the condition-independent testing holds. Such a process directly leads to some false positive variables added to the candidate Markov blanket, but the false positive variables are removed from the candidate Markov blanket in the shrinking stage. Tsamardinos et al. proposed IAMB [17] which optimized GSMB in the growth stage, and reduced the number of conditional independent test. However, IAMB inherited the problem of low data efficiency of GSMB and also needed more samples. Some researchers paid attention to the flaw of GSMB (or IAMB), and then proposed different improved algorithms of IAMB successively, such as inter-IAMB, IAMBnPC [17], fast-IAMB [29], k-IAMB [19], and λ-IAMB [30]. However, IAMB and its series of improved algorithms basically inherited the two-stage strategy of GSMB.

ATSI is based on the topological structure information of the Bayesian network and checks the conditional set between the independent non-*MB* variables and the target variable which is often much smaller than *MB* of target variable, so the discovery efficiency of ATSI is relatively high, but more complex heuristic strategies also bring more computational costs. In addition, ATSI can derive more topological information than ACIT. they can not only get whether a node is a variable within *MB* but also distinguish the parents-children nodes and spouse nodes, part of children nodes and part of V structures. The typical representatives of ATSI are MMMB [31] and HITON-MB [18] proposed in 2003. The implementing process of both also is divided into two stages, but it is different from that of IAMB (or GSMB). The first stage is to learn parents-children nodes, the second stage is to learn spouse nodes, and finally get *MB*. In 2007, MMMB (or HITON) was proved to be unable to guarantee obtaining the correct Markov blanket [19], Pena *et al.* proposed improved PCMB [19] and Fu *et al.* proposed improved IPC-MB [32] respectively, but the two algorithms inherited implementing process of HITON (or MMMB). The difference between PCMB and IPC-MB algorithm is mainly in the implementing strategy of learning the parents-children nodes. PCMB chooses a forward strategy, while IPC-MB chooses a backward strategy. In addition, some researchers synthesized the advantages of the two or more algorithms above mentioned, and proposed some new algorithms, for example, MBOR [33] and DOS [34].

Score-based algorithms are actually strategy based on scoring and searching, and it is widely used in Bayesian network structure learning, but rarely used in Markov blanket learning. In 2013, DMB and RPDMB [25] based on scoring were reported for the first time, their experimental reports showed that RPDMB has competitive against PCMB in accuracy, but it required more time costs. Considering that the time efficiency of IPC-MB is much higher than that of PCMB, it can be reasonably predicted that IPC-MB has much better time efficiency than RPDMB. Even so, the two algorithms are an important attempt.

In recent years, *MB* induction algorithms based on regularized linear models were reported in the literature. Although there is not too much related work, it provides a new idea for obtaining *MB*. Mark Schmidt [26] used BIC scoring to propose an *MB* induction algorithm (L1MB) with lasso regression models for building a Bayesian network, but he did not give theoretical proof for L1MB. V.Strobl [27] used the modified form of ridge regularized linear models (MRRLMs) to explore the relationships between explanatory variables and response variable under certain conditions, and theoretically proved that the set of variables corresponding to the non-zero solution of the models is a subset of *MB*, and it has great significance in theory.

Feature selection, also known as variable selection or feature subset selection which is different from feature extraction [35], selects the minimum feature subset (feature variables) from the variables to satisfy the optimization of performance metric [36]. It is often used to improve the accuracy of the classifier or regression models and the explanatory ability of the data production process [37] in the field of machine learning. There are many methods available for feature selection, regularized linear model(s) is (are) one of the important methods, the main idea of this method is to compress coefficients, hence it also called as coefficient compression method [38], [39]. By adjusting penalty parameters until more coefficients are zero or tend to zero, users delete the corresponding variables to achieve the purpose of variable selection. LASSO and ridge regression based on regularized linear models are common feature selection methods. In this paper, ridge regression modes are used.

The relationships between feature variables and Markov blanket were proved theoretically in the literature [15], [40] (see definition 15 and theorem 7). Under the condition that $\mathbb{P}(Y|\boldsymbol{X})$ can be accurately estimated, the optimal predictor of the target variable is the Markov blanket of the target variable, where, the optimal predictor is the feature variable that satisfies some conditions. Literature [27] expanded the concept of the optimal predictor, and defined the minimal optimal predictor with the help of the optimal predictor. Moreover, it theoretically proved that a minimal optimal predictor is a Markov boundary. For MRRLMs, under some assumptions, the set of variables corresponding to the non-zero row coefficients (or non-zero coefficients) of the solution of the models is a subset of Markov boundary. However, in practice, the target variable is usually a one-dimensional variable and these conditions are hard to hold, therefore, MRRLMs finally become method based on ranking.

III. DEFINITION AND BACKGROUND

In order to introduce our proposed work, some necessary definitions and relevant contents are presented as follows.

A. NOTATIONS

In addition to specific annotations, the notations used in this paper and their related relationships are listed in Table 1:

TABLE 1. Notations.

Symbol	Mathematical meanings
A, B, C	A, B, and C are random variables respectively.
$\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$	$\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are a set of random variables respectively.
$\boldsymbol{X}$	$\boldsymbol{X}$ is a set or matrix of explanatory variables.
Y	Y is a response variable (a reserved word).
$(\boldsymbol{A};\boldsymbol{B})$	$\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of columns, $(\boldsymbol{A};\boldsymbol{B})$ denotes that $\boldsymbol{B}$ is appended to the rows of $\boldsymbol{A}$
$(\boldsymbol{A},\boldsymbol{B})$	$\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of rows, $(\boldsymbol{A},\boldsymbol{B})$ denotes that $\boldsymbol{B}$ is appended to the columns of $\boldsymbol{A}$
$\boldsymbol{A}\perp\boldsymbol{B}\|\boldsymbol{C}$	$\boldsymbol{A}$ and $\boldsymbol{B}$ are condition independent given $\boldsymbol{C}$.
$\sum \boldsymbol{X}$	Covariance matrix of $\boldsymbol{X}$.
MB or M	Markov boundary (blanket).
VRRLM	Variant ridge regression linear model.
MRRLM	Modified form of ridge regularized linear model.

B. MARKOV BOUNDARY THEORY

1) Markov boundary (blanket): A Markov blanket(*MB*) of a response variable Y in the joint probability distribution of variables $\boldsymbol{X}$ is a set of variables conditioned on which all other

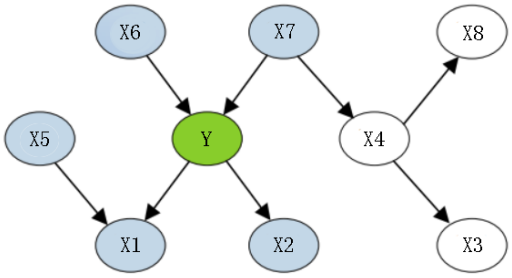

FIGURE 1. Markov blanket of *Y*.

variables are independent of Y, that is, for every $\boldsymbol{F}\subseteq(\boldsymbol{X}\backslash M)$, $Y\perp\boldsymbol{F}|M$, then M is a Markov blanket of Y; if no proper subset of M satisfies the definition of Markov blanket of Y, then M is a Markov boundary of Y. As shown in the Figure 1, $M = (X_1, X_2, X_5, X_6, X_7)$. PC $= (X_1, X_2, X_6, X_7)$ is set of causal variables of Y.

2) Intersection property: For the variables $\boldsymbol{X}$ with a joint probability distribution $\mathbb{P}$ and any subset ***A, B, C***, and ***D***, the joint probability distribution $\mathbb{P}$ was said to satisfy the intersection property if $\boldsymbol{A}\perp\boldsymbol{B}|(\boldsymbol{C}\cup\boldsymbol{D})$ and $\boldsymbol{A}\perp\boldsymbol{D}|(\boldsymbol{C}\cup\boldsymbol{B})\Rightarrow \boldsymbol{A}\perp(\boldsymbol{B}\cup\boldsymbol{D})|\boldsymbol{C}$.

3) Global Markov condition: The joint probability distribution $\mathbb{P}$ of variables satisfies the global Markov condition for a directed graph $\mathbf{G} = \langle \mathbf{H}, \mathbf{E}\rangle$ if and only if any three disjoint subsets ***A, B, C*** from $\mathbf{H}$, if $\boldsymbol{A}$ is d-separated from $\boldsymbol{B}$ given $\boldsymbol{C}$ in $\mathbf{G}$, then $\boldsymbol{A}$ is conditionally independent of $\boldsymbol{B}$ given $\boldsymbol{C}$ in $\mathbb{P}$.

Theorem 1: if a joint probability distribution $\mathbb{P}$ of variables $\boldsymbol{X}$ satisfies the intersection attribute, then for $\boldsymbol{V}\subseteq\boldsymbol{X}$, there exists a unique Markov boundary of $\boldsymbol{V}$ [41].

Theorem 2: if a joint probability distribution $\mathbb{P}$ of variables satisfies the global Markov condition for a directed graph $\mathbf{G}$, then the set of parents, children, and spouses of Y is a Markov blanket of Y [41].

According to the Theorem 2, a Markov boundary of Y is composed of parents, children, and spouses of Y. For the convenience of description, Markov boundary or Markov blanket is no longer distinguished in this paper.

C. SUFFICIENT DIMENSION REDUCTION THEORY

1) Sufficient Dimension Reduction (SDR): SDR attempts to find a Matrix $\eta \in \mathbb{R}^{p\times d}(d \le p)$, such that $Y\perp\boldsymbol{X}|\eta^T\boldsymbol{X}$, where $\boldsymbol{X} = (X_1; X_2 \ldots; X_p)$

2) Dimension Reduction Subspace (DRS): if $Y\perp\boldsymbol{X}|\eta^T\boldsymbol{X}$, the column space of η is called a DRS denoted as $S(\eta)$, where, the column space of η refers to the space spanned by the columns of η. A subspace $S_{Y|X}$ is called a central DRS for $Y|\boldsymbol{X}$ if $S_{Y|X}$ is a DRS and $S_{Y|X}\subseteq S_{DRS}$ for all DRS S_{DRS}

3) Linear intersection property: The distribution $\mathbb{P}$ of variables $\boldsymbol{Z}$ satisfies the linear intersection property if for any $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{R}^{(p+1)\times d}$. Where $d \le p+1$ such that $\alpha_1^T Z\perp\alpha_2^T Z|(\alpha_3^T Z, \alpha_4^T Z)$ and $\alpha_1^T Z\perp\alpha_4^T Z|(\alpha_3^T Z, \alpha_2^T Z)$, we also have $\alpha_1^T Z\perp(\alpha_2^T Z, \alpha_4^T Z)|\alpha_3^T Z$

Theorem 3: If a distribution $\mathbb{P}$ of variables satisfies the linear intersection property, then there exists a central dimension reduction subspace [27].

The relationships between Markov boundary and DRS is introduced below.

Suppose M is a Markov boundary of variables, p_M is denoted as the number of variables within M, $\eta = (\eta_M; \eta_{X\backslash M}) \in \mathbb{R}^{p\times d}(d \le p)$ is the dimension reduction matrix, where η_M has p_M, and $\eta_{X\backslash M}$ has $p - p_M$ rows.

Theorem 4: if a joint probability distribution of variables $(Y; \boldsymbol{X})$ satisfies the linear intersection property, then there exists a central DRS $S(\eta_M)$ for $Y|\ M$, and $S(\eta)$ is the central DRS for $Y|\ \boldsymbol{X}$, where $\eta_{X\backslash M}$ is a matrix of all zeros [27].

Theorem 5: if a joint probability distribution of variables $(Y; \boldsymbol{X})$ satisfy the linear intersection property, let d denote the number of column dimensions in η, then we have $\sum_{i=1}^{d} |\eta_{j,i}| > 0$, where the row j corresponds to a variable within M, and $\sum_{i=1}^{d} |\eta_{j,i}| = 0$, when the row j corresponds to a variable within $\boldsymbol{X}\backslash M$ [27].

We thus find that variables within Markov boundary are identified by discovering the central DRS and identifying any deviations from zero in the coefficients of η.

D. MODIFIED FORM OF RIDGE REGULARIZED LINEAR MODELS

Let K denote the number of the column dimensions of α and β, and let $k \in [1, 2, ..K] = \mathbb{K}$, Consider the following linear regression model:

$$Y = \alpha + \beta^T\boldsymbol{X} + \varepsilon, \textit{where}, \boldsymbol{X}\in\mathbb{R}^{p\times n},$$
$$\alpha \in \mathbb{R}^{K\times n}, \beta \in \mathbb{R}^{p\times K}, \varepsilon \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$$

If we have the following assumptions: (1)

1) The global Markov condition holds.
2) The joint probability distribution of $(Y;\boldsymbol{X})$ satisfies the linear intersection property.
3) The covariance matrix $\sum\boldsymbol{X}$ is positive definite.
4) $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is a linear function of $\eta^T\boldsymbol{X}$, when $Y\perp\boldsymbol{X}|\eta^T\boldsymbol{X}$.
5) The matrix β^* is a non-zero matrix, and the solution to the following optimization problem:

$$argminE\{\mu(\alpha + \beta^T\boldsymbol{X}, Y)\} + \lambda tr(\beta^T\sum\boldsymbol{X}\beta) \quad (1)$$

then $S(\beta_k^*) \subseteq S(\eta)$, for all $k \in \mathbb{K}$, where $S(\eta)$ is any DRS, μ is a convex function, λ is model's parameter, and $\lambda > 0$.

The above mentioned show the relationships between the solution β_k^* of MRRLMs and the dimension reduction matrix η. We can recover a subset of the parents, children, and spouses of Y from the non-zero coefficients of β^*. However, the target variable Y is usually a one-dimensional vector, that is, $K = 1$ in practice. Let $\beta = (\sum X)^{-1/2}\gamma$, then equation (1) is equivalent to the following normal ridge regression models.

$$argminE\{\mu(\alpha + \gamma^T\boldsymbol{Z}, Y)\} + \lambda\gamma^T\gamma \quad (2)$$

where $\boldsymbol{Z} = (\sum X)^{-1/2}\boldsymbol{X}$.

Obviously, the covariance matrix $\sum \boldsymbol{X}$ must be non-singular. The theoretical results show that MRRLMs cannot be applied to the case where $\left|\sum \boldsymbol{X}\right|$ is zero. The subsequent experimental results show that the closer $\left|\sum \boldsymbol{X}\right|$ be close to zero, the faster the discovery efficiency drops.

IV. VARIANT RIDGE REGULARIZED LINEAR MODELS (VRRLMs)

This section first introduces the variable collinearity which leads to the singularity of covariance matrix, and then presents VRRLMs as well as theoretical proof.

A. COLLINEARITY OF VARIABLES

For a $\boldsymbol{X} = (X_1; X_2; \ldots; X_p)$, $X_i \in \mathbb{R}^n$, $i = 1..p$, if variables $\boldsymbol{X}$ are collinear, then there exists a set of constants $k_0, k_1, k_2, \ldots, k_p$ not all zero, such that the following formula holds:

$$k^T \boldsymbol{X} = k_0, where, k = (k_1, k_2, \ldots, k_p)^T$$

So, we have

$$Var(k^T \boldsymbol{X}) = Var(\sum_{i=1}^{p} k_i X_i) = \sum_{i=1}^{p}\sum_{j=1}^{p} k_i k_j Cov(X_i, X_j)$$
$$= k^T \sum \boldsymbol{X} k = Var(k_0) = 0$$

Therefore, we have $k^T \sum \boldsymbol{X} k = 0$. Recall that covariance matrix $\sum \boldsymbol{X}$ is a very special matrix: it is square (P × P), symmetric and positive-definite, meaning that $k^T \sum \boldsymbol{X} k \geq 0$.

Let $\lambda_i(v_i)$ is eigenvalue (eigenvector) of $\sum \boldsymbol{X}$, i = 1, 2, ..p. such that $\sum \boldsymbol{X} v_i = \lambda_i v_i$. Obviously, we have

$$k^T \sum \boldsymbol{X} k = \sum_{i=1}^{p} (k^T v_i)^2 \lambda_i = 0$$

Therefore, there exists at least one $\lambda_i = 0$.

Since $\sum \boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}}$, where $\boldsymbol{V}$ is the matrix and $\boldsymbol{D}$ is the diagonal matrix whose entries are $\lambda_1, \lambda_2, \ldots, \lambda_p$. So the covariance matrix $\sum \boldsymbol{X}$ is singular and vice versa.

B. VARIANT RIDGE REGULARIZED LINEAR MODELS

We add a matrix $\boldsymbol{I}$ to MRRLMs to solve the question about collinearity of variables. Let $\widetilde{\boldsymbol{X}} \approx \boldsymbol{X} + \delta\boldsymbol{\sigma}, \delta \in [-0.5, 0.5], \boldsymbol{\sigma} = (\sigma_1; \sigma_2; \ldots; \sigma_p), \sigma_i \in \mathbb{R}^n$, where σ_i is an independent identically distributed variable subject to the standard Gaussian distribution $\mathcal{N}(0, 1)$.

Suppose $Y \perp \widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}}$ holds ($\delta \in [-0.5, 0.5]$) when $Y \perp \boldsymbol{X} | \eta^T \boldsymbol{X}$, the hypothesis is completely reasonable, because the design matrix $\boldsymbol{X}$ is composed of observation data which has observation errors itself. Moreover, δ is very small number.

Theorem 6: $\sum \widetilde{\boldsymbol{X}} = \sum \boldsymbol{X} + \delta^2 \boldsymbol{I} (\delta \neq 0)$ is non-singular and positive definite.

Proof: Let λ_i (v_i) is an eigenvalue (eigenvector) of $\sum \boldsymbol{X}$, then $\sum \widetilde{\boldsymbol{X}} = \sum \boldsymbol{X} + \delta^2 \boldsymbol{I} = \boldsymbol{V}(\boldsymbol{D}\,\boldsymbol{\mathcal{C}}\,\boldsymbol{\delta^2 I})\boldsymbol{V}^{\boldsymbol{T}}$, $\lambda_i + \delta^2$ is the eigenvalue of $\sum \widetilde{\boldsymbol{X}}$. Obviously, $\sum \widetilde{\boldsymbol{X}}$ is non-singular for $\delta \neq 0$.

We know that $\sum \widetilde{\boldsymbol{X}}$ is a non-singular matrix which means $k^T \sum \widetilde{\boldsymbol{X}} k \neq 0$. As can be seen from section III.A.

$$Var(k^T \widetilde{\boldsymbol{X}}) = k^T \sum \widetilde{\boldsymbol{X}} k \geq 0$$

Therefore, we have $k^T \sum \widetilde{\boldsymbol{X}} k > 0 (k \neq 0)$, so $\sum \widetilde{\boldsymbol{X}}$ is positive definite.

The following is a revision of theorem 3.5.2 in the literature [27], and the conclusion also holds. Obviously, VRRLMs is equivalent to MRRLMs when $\delta = 0$ and $0K = 1$.

Theorem 7: Let $S(\eta)$ is any reduce dimension subspace, such that $E(\widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}})$ is a linear function of $\eta^T \widetilde{\boldsymbol{X}}$ when $Y \perp \widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}}$. The covariance matrix $\sum \widetilde{\boldsymbol{X}}$ is positive definite, and we have the assumption (1) and (2) in the section III.D. if the following formula (3) can be minimized for $\alpha^* \in \mathbb{R}$, $\beta^* \in \mathbb{R}^p$.

$$L(\alpha, \beta) = argminE\{\mu(\alpha + \beta^T \boldsymbol{X}, Y)\} + \lambda tr(\beta^T (\sum \widetilde{\boldsymbol{X}})\beta) \quad (3)$$

Then, $S(\beta^*) \subseteq S(\eta)$

Proof: Let $\widetilde{\boldsymbol{X}} \approx \boldsymbol{X} + \delta\boldsymbol{\sigma}$, following formula holds (see theorem 3.5.2 in the literature [27]).

$$E\{\mu(\alpha + \beta^T \boldsymbol{X}, Y)\} \approx E\{\mu(\alpha + \beta^T \widetilde{\boldsymbol{X}}, Y)\} \geq E_{Y, \eta^T \widetilde{\boldsymbol{X}}}[\mu(\alpha + \beta^T E(\widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}}), Y)] \quad (4)$$

Also consider,

$$Var(\beta^T \widetilde{\boldsymbol{X}}) = E(Var(\beta^T \widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}})) + Var(\beta^T E(\widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}})) \geq Var(\beta^T E(\widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}})) \quad (5)$$

Therefore, we apply proposition 4.2 of the literature [42], so that

$$L(\alpha, \beta) \geq L(\alpha, P_\eta(\sum \widetilde{\boldsymbol{X}})\beta)$$

where, $P_\eta(\sum \widetilde{\boldsymbol{X}}) = \eta(\eta^T \sum \widetilde{\boldsymbol{X}} \eta)^{-1} \eta^T \sum \widetilde{\boldsymbol{X}}$

Let $S(\eta)$ be minimum DRS such that $E(\widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}})$ is still linear function of $\eta^T \widetilde{\boldsymbol{X}}$. we have $E(\widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}}) = C^T \eta^T \widetilde{\boldsymbol{X}} (C \neq 0)$, then

$$Var(\beta^T E(\widetilde{\boldsymbol{X}} | \eta^T \widetilde{\boldsymbol{X}})) = Var(\beta^T C^T \eta^T \widetilde{\boldsymbol{X}}) = \beta^T C^T \eta^T (\sum \widetilde{\boldsymbol{X}}) \eta C \beta$$

We want to show that we have $[\beta^T C^T \eta^T] \sum \widetilde{\boldsymbol{X}} [\eta C \beta] > 0$. if $S(\beta) \supset S(\eta)$, so $S(\beta) \supset S(\eta C)$, and β is not orthogonal to a column in ηC, we have $\eta C \beta \neq 0$ by the definition of orthogonality. Therefore it implies that (5) is strict inequality, and such a β is impossible to minimize equation (3), so $S(\beta^*) \subseteq S(\eta)$.

C. ALGORITHM

We employ Covariance Ridge with Permutation test and parameter δ (CRP_δ) to solve object function (3) for recovering coefficient specific p-value. As shown in Algorithm 1, the input and output parameters are introduced first. $\boldsymbol{X}$ and Y are input data set, where $X = (X_1; X_2; \ldots; X_p)$, X_i is an n-dimensional row vector and Y is a one-dimensional column

Algorithm 1 CRP_δ

Require: $\boldsymbol{X}$, Y, *ref _mb_num*, δ, *numPermute*.
Ensure: *crp_mb*
1: Calculate covariance matrix $\sum \boldsymbol{X}$
2: $\sum \boldsymbol{X} = \sum \boldsymbol{X} + \delta^2 * \boldsymbol{I}$
3: Data transformation: $\boldsymbol{X} = \sum \boldsymbol{X}^{-0.5} * \boldsymbol{X}$
4: Calculate γ_0 and λ with ridge regression(cvglmnet).
5: Calculate original $\beta_0 = \sum \boldsymbol{X}^{-0.5} * \gamma_0$.
6: Calculate number of the row of $\boldsymbol{X}$: p;
7: Set the matrix *mat_p*(p, *numPermute*),
8: **for** $k = 0$ to *numPermute* **do**
9: Random perm Y.
10: Calculate γ with ridge regression(glmnet)
11: Calculate original $\beta = \sum \boldsymbol{X}^{-0.5} * \gamma$
12: $mat\_p(:, k) = (abs(\beta) \geq abs(\beta_0))$
13: **end for**
14: $p\_Value = (sum(mat\_p^T) + 1)./(numPermute + 1)$
15: Calculate index *p_value_index* from small to larger
16: Calculate *crp_mb* : *p_value_index*(1 : *ref _mb_num*)
17: **return** *crp_mb*

vector; *ref _mb_num* refers to the number of variables within *MB* returned by a reference algorithm or expected value given in advance; δ is a regulatory parameter, NumPermute is the number of duplicates in permutation test and *crp_mb* is the *MB* returned by CRP_δ. From the first row to the third row of the algorithm, the covariance matrix of $\boldsymbol{X}$ denoted as $\sum \boldsymbol{X}$ is calculated first, then the $\sum \boldsymbol{X}$ is modified by applying the method shown in Theorem 7 in this paper, and then the variable replacement is made as shown in formula (2). At this moment, VRRLMs become normal ridge regression models. The fourth row is a question for the solution of the normal ridge regression models. There are many off-the-shelf tools and software available, in this paper, the cvglmnet function in the *Glmnet* toolkit is used to get the model parameters, while the parameter λ is selected by 10-fold cross-validation. Lines 8 to 15 are the process of computing the p-value by using the permutation test. Line 15 is the sequence of the $\boldsymbol{X}$ variable obtained by sorting the p-value from smallest to largest. Since permutation test [43] is not the focus of this paper and is omitted here. Line 16 returns the index set *crp_mb* of the front *ref _mb_num* variables in the sequence of p-values. According to Theorems 4, Theorems 5, and Theorems 7, *crp_mb* is a subset of *MB*.

V. SIMULATION

To verify the theory above, we evaluate the algorithms using Precision Rate, Recall Rate, F-Score and running time. Experimental data are composed of four discrete data sets: Alarm(10), Child(10), Gene, and Insurance (downloaded: http://pages.mtu.edu/lebrown/supplements/mmhc_paper/mmhc_index.html). The attribute of data sets is listed as follows.

TABLE 2. Attribute of data sets.

Data sets	Alarm	Alarm10	Child
Nodes	37	370	20
Max/Min(*MB*)	8/1	8/1	8/1
Data sets	Child10	Insurance	Gene
Nodes	200	27	801
Max/Min(*MB*)	8/1	10/1	15/0

TABLE 3. Selection of reference algorithm (F-Score).

Data set Name	HITON	IAMB	interIAMBnPC	GSMB
Alarm1_s500_v1	0.819	0.752	0.769	0.681
Alarm10_s500_V1	0.605	0.567	0.616	0.552
Alarm10_s1000_V1	0.709	0.614	0.689	0.556
Child10_s500_v9	0.579	0.674	0.715	0.660
Child_s500_v2	0.833	0.799	0.791	0.786
Gene_s500_v1	0.667	0.710	0.696	0.658
Insurance_s500_v1	0.668	0.553	0.616	0.545
Insurance_s500_v9	0.610	0.626	0.612	0.544
Insurance_s1000_v1	0.752	0.706	0.706	0.656
Mean	**0.694**	0.667	0.690	0.626

To realize the performance comparison between several classical algorithms and the regularized linear models, we first choose several classical algorithms to compare with each other on the same data set in advance (see Table 3), and then choose the optimal algorithm (HITON: G^2 test and $\alpha = 0.05$) to compare with regularized linear models (see Table 4 and Table 5). The reasons are that the algorithms based on regularized linear models are essentially sorting algorithms, and they need the expectation of the number of variables within *MB* which is provided by HITON in advance. It should be emphasized that the main purpose of the experiment is to compare the discovery performance of MRRLMs and VRRLMs on the data sets with collinear variables.

A. ENVIRONMENT AND PARAMETER

We use the *Causal Explorer* and *Glmnet* package for matlab, and call HITON-MB, IAMB, interIAMBnPC and GSMB algorithm in Matlab2014a. The running environment of the experiment is that the processor is Inter(R) Core(TM) i7-6700 CPU @3.4g, the memory is 16GB, and the operating system is 16-bit windows 7. The CRP-δ algorithm is implemented by Matlab language, and cvglmnet and glmnet function in the *Glmnet* package is directly called.

The experiment specifically investigated the discovery efficiency of VRRLMs when $\delta = (0, 0.2, 0.3, 0.4, 0.5)$. The experimental parameters are agreed as follows:

1) Zero eigenvalue problem. An eigenvalue λ_i is considered equal to zero when the absolute value of eigenvalue is less than 1*E-12. Obviously, the covariance matrix is singular when $\lambda_i = 0$; the covariance matrix is non-singular when $\lambda_i \neq 0$.

2) The number of samples. Considering the high calculation costs of covariance operation and permutation test, the sample size of data set we uniformly considered is 500 or 1000.

3) The number of target node. Considering the computational time costs, In the experiment, fifteen target nodes randomly selected are adapted to evaluate *MB* discovery

TABLE 4. MRRLMs ($\eta = 0$) and VRRLMs ($\eta \neq 0$) on the data sets with collinear variables.

Alarm10 (Alarm10_s500_v1.txt)						
Item	HITON-MB	MRRLMs(δ=0)	VRRMs(δ=0.2)	VRRLMs(δ=0.3)	VRRLMs(δ=0.4)	VRRLMs(δ=0.5)
F-Score	0.662(0.063)	0.214(0.067)	0.452(0.059)	0.470(0.067)	0.482(0.052)	0.478(0.077)
Precision rate	0.621(0.067)	0.193(0.057)	0.421(0.065)	0.434(0.071)	0.444(0.06)	0.439(0.082)
Recall rate	0.714(0.083)	0.243(0.084)	0.493(0.068)	0.517(0.080)	0.532(0.067)	0.530(0.086)
Running time	0.396(0.037)	15.39(0.073)	18.22(0.102)	19.14(0.115)	19.79(0.120)	20.33(0.129)
Insurance (Insurance_s1000_v1.txt)						
Item	HITON-MB	MRRLMs(δ=0)	VRRMs(δ=0.2)	VRRLMs(δ=0.3)	VRRLMs(δ=0.4)	VRRLMs(δ=0.5)
F-Score	0.747(0.037)	0.380(0.050)	0.644(0.030)	0.608(0.035)	0.643(0.030)	0.621(0.028)
Precision rate	0.826(0.047)	0.401(0.063)	0.705(0.038)	0.660(0.033)	0.705(0.037)	0.6759(0.037)
Recall rate	0.685(0.053)	0.364(0.050)	0.595(0.047)	0.567(0.056)	0.595(0.049)	0.5783(0.049)
Running time	0.109(0.007)	0.322(0.003)	0.323(0.003)	0.323(0.003)	0.324(0.003)	0.328(0.003)
Insurance (Insurance_s500_v1.txt)						
Item	HITON-MB	MRRLMs(δ=0)	VRRMs(δ=0.2)	VRRLMs(δ=0.3)	VRRLMs(δ=0.4)	VRRLMs(δ=0.5)
F-Score	0.647(0.035)	0.408(0.080)	0.563(0.047)	0.571(0.044)	0.567(0.045)	0.580(0.036)
Precision rate	0.638(0.042)	0.390(0.076)	0.553(0.052)	0.560(0.057)	0.557(0.052)	0.575(0.046)
Recall rate	0.658(0.045)	0.429(0.088)	0.576(0.054)	0.586(0.047)	0.581(0.050)	0.588(0.460)
Running time	0.079(0.008)	0.260(0.002)	0.263(0.005)	0.263(0.002)	0.265(0.002)	0.268(0.002)
Gene(Gene_s500_v1.txt)						
Item	HITON-MB	MRRLMs(δ=0)	VRRMs(δ=0.2)	VRRLMs(δ=0.3)	VRRLMs(δ=0.4)	VRRLMs(δ=0.5)
F-Score	0.727(0.068)	0.151(0.052)	0.532(0.061)	0.536(0.074)	0.520(0.040)	0.502(0.055)
Precision rate	0.599(0.073)	0.119(0.042)	0.430(0.062)	0.432(0.069)	0.421(0.049)	0.404(0.054)
Recall rate	0.928(0.057)	0.216(0.086)	0.700(0.057)	0.708(0.086)	0.684(0.038)	0.666(0.070)
Running time	0.889(0.069)	30.48(0.018)	30.35(0.018)	30.35(0.019)	30.37(0.022)	30.39(0.021)

TABLE 5. MRRLMs ($\eta = 0$) and VRRLMs ($\eta \neq 0$) on the data sets with non-collinear variables.

Child(Child_s500_v2.txt)						
Item	HITON-MB	MRRLMs(δ=0)	VRRMs(δ=0.2)	VRRLMs(δ=0.3)	VRRLMs(δ=0.4)	VRRLMs(δ=0.5)
F-Score	0.827(0.017)	0.787(0.033)	0.773(0.035)	0.767(0.043)	0.787(0.035)	0.760(0.038)
Precision rate	0.771(0.034)	0.741(0.047)	0.727(0.048)	0.721(0.054)	0.741(0.046)	0.714(0.047)
Recall rate	0.892(0.013)	0.841(0.025)	0.827(0.031)	0.821(0.038)	0.841(0.035)	0.814(0.039)
Running time	0.066(0.007)	0.197(0.002)	0.210(0.001)	0.211(0.002)	0.213(0.002)	0.213(0.001)
Alarm10(Alarm10_s1000_v1.txt)						
Item	HITON-MB	MRRLMs(δ=0)	VRRMs(δ=0.2)	VRRLMs(δ=0.3)	VRRLMs(δ=0.4)	VRRLMs(δ=0.5)
F-Score	0.663(0.077)	0.507(0.083)	0.520(0.080)	0.535(0.072)	0.539(0.075)	0.526(0.084)
Precision rate	0.651(0.089)	0.491(0.087)	0.505(0.082)	0.518(0.072)	0.523(0.074)	0.509(0.086)
Recall rate	0.683(0.093)	0.528(0.092)	0.541(0.093)	0.558(0.094)	0.562(0.097)	0.552(0.100)
Running time	0.373(0.016)	20.35(0.054)	24.30(0.066)	24.83(0.071)	25.28(0.076)	25.65(0.086)
Alarm(Alarm1_s500_v1.txt)						
Item	HITON-MB	MRRLMs(δ=0)	VRRMs(δ=0.2)	VRRLMs(δ=0.3)	VRRLMs(δ=0.4)	VRRLMs(δ=0.5)
F-Score	0.820(0.047)	0.723(0.067)	0.756(0.043)	0.752(0.052)	0.721(0.049)	0.682(0.061)
Precision rate	0.817(0.068)	0.712(0.075)	0.740(0.049)	0.736(0.069)	0.703(0.073)	0.666(0.079)
Recall rate	0.826(0.043)	0.737(0.076)	0.777(0.062)	0.773(0.058)	0.744(0.047)	0.703(0.061)
Running time	0.061(0.006)	0.317(0.002)	0.360(0.003)	0.364(0.003)	0.368(0.003)	0.373(0.002)
Child10(Child10_s500_v9.txt)						
Item	HITON-MB	MRRLMs(δ=0)	VRRMs(δ=0.2)	VRRLMs(δ=0.3)	VRRLMs(δ=0.4)	VRRLMs(δ=0.5)
F-Score	0.605(0.044)	0.473(0.041)	0.485(0.057)	0.502(0.044)	0.504(0.039)	0.521(0.037)
Precision rate	0.472(0.052)	0.360(0.035)	0.370(0.047)	0.386(0.041)	0.387(0.036)	0.401(0.034)
Recall rate	0.853(0.058)	0.695(0.083)	0.708(0.096)	0.725(0.080)	0.728(0.078)	0.749(0.082)
Running time	0.298(0.041)	3.656(0.005)	4.791(0.005)	4.928(0.005)	5.069(0.007)	5.186(0.008)

efficiency of the data set, and the average of the discovery efficiency of fifteen target nodes was considered as the discovery efficiency on the data set.

4) The number of permutation test and penalty parameter. Theoretically, the larger the number of permutation test is, the more accurate the result will be. Considering the time costs, the number of permutation test in this experiment is set to 199. The same penalty parameter λ is used when the estimation parameters are repeatedly calculated.

B. DATA RESULTS AND ANALYSIS

1) Results on the data sets with collinear variables.

By calculating the eigenvalue of the covariance matrix of the data sets in advance, we find collinearity of variables in the data sets shown in the Table 4, repeat 10 times and calculate the average discovery efficiency of Markov boundary, and the results are also shown in Table 4 (the standard deviation is also shown in brackets).

As can be seen from Table 4, in terms of precision, recall rate or F-Score, the performance of VRRLMs significantly higher than that of MRRLMs on the data set with collinear variables, which also indicates the correctness of VRRLMs in theory. More specifically, taking the F-Score as an example, discovery performance of VRRLMs can be improved by 3.54 times at the highest than that of MRRLMs (i.e. from 0.151 to 0.536 in Gene_s500_v1 dataset) and 1.42 times at the lowest (i.e. from 0.408 to 0.58 in insurance_s500_v1 dataset). Note the size of the data set's dimension, we seem to get

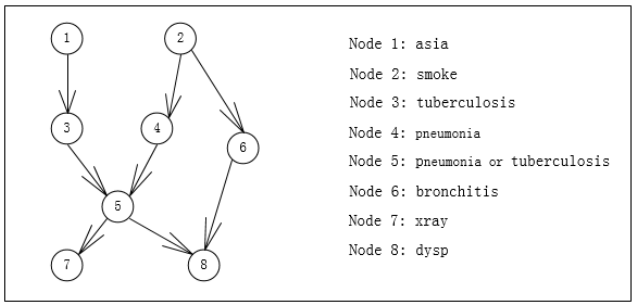

FIGURE 2. Small Bayesian network.

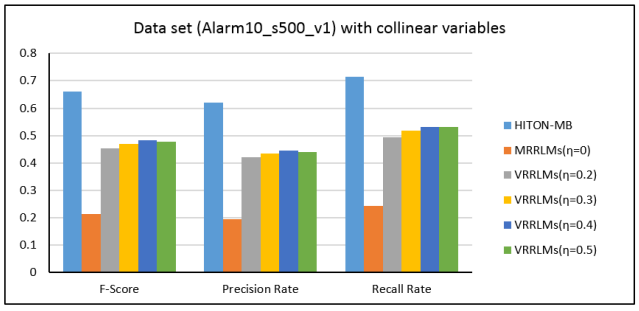

FIGURE 3. Discovery efficiency on the data set (Alarm10_s500_v1) with collinear variables.

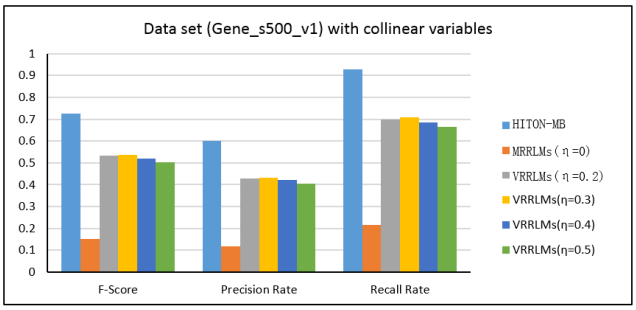

FIGURE 4. Discovery efficiency on the data set (Gene_s500_v1) with collinear variables.

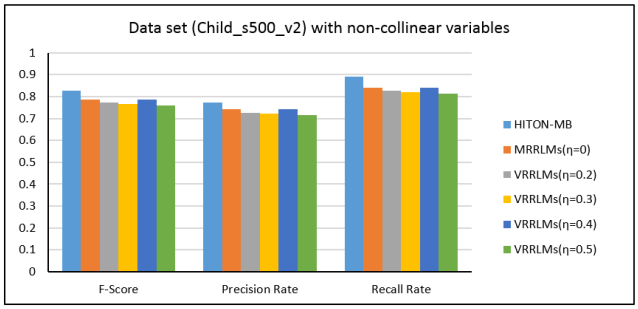

FIGURE 5. Discovery efficiency on the data set (Child_s500_v2) with non-collinear variables.

a rule: The discovery performance of *MB* of VRRLMs is more obvious than that of MRRLMs on the data sets with higher dimension of data sets or more samples. But in terms of running time costs, the opposite is true. That is, the higher the number of variables or the larger the sample size, the longer the running time will be. Note that MRRLMs can't be applied in theory when the covariance is singular, but in practice, when the absolute value of the eigenvalue is less than 1*E-12, the determinant value of the covariance matrix is close to zero, and the discovery efficiency of Markov boundary decreases when $\delta = 0$. We can also more intuitively understand the performance advantages of VRRLMs from Figure 3 and Figure 4. In addition, both of them have lower discovery efficiency of *MB* than that of HITON-MB which is also consistent with the literature [27].

We can also see that HITON-MB is significantly higher than MRRLMs and VRRLMs in identifying *MB* from Table 4. The reason is that MRRLM and VRRLM can get "very close" to identifying a subset of *MB*; while HITON-MB based on conditional independent test can approximately find all of the *MB*. However, the above results do not negate the superiority of VRRLM in finding the key causal variables affecting a target variable. The implementation process of Algorithm 1 says the VRRLM is actually a sequence of ranking from small to large based on p-value (i.e. causal correlation from strong to weak), we can easily select the most primary or secondary variable from the p-value sequence, such a result has great significance in practice. For example, when studying the effect of drug therapy, we mainly focus on the primary or secondary factors; while HITON algorithm selects a group of influential variables with strong causal correlation from variables, especially when there are many influencing variables, people cannot obtain the primary or secondary variable at all.

Figure 2 shows a Bayesian network with 8 nodes. (http://www.bnlearn.com/bnrepository/discrete-small.html). For target node 5 (tuberculosis or pneumonia), the result obtained by HITTON algorithm is {3,4,6,7,8}, and the result of VRRLM is {3,4,7,8,6}. Unfortunately, both algorithms have the same the primary and secondary causal variables (tuberculosis, pneumonia), so we look at the third factor. VRRLM outputs node 7 (i.e. X-ray); while HITTON outputs node 6 (i.e. bronchitis). However, for patients with tuberculosis or pneumonia, it is obvious that the causality of X-ray is stronger than bronchitis, and the superiority of VRRLM is reflected.

The VRRLMs proposed in this paper are actually an extension of MRRLMs, which also provides the theoretical basis for the improvement of the *MB* discovery algorithms based on regularized linear models in the future.

For δ, we suggest that δ should be consistent with the assumptions in the paper, which is to select values between -0.5 and 0.5 (except 0). Our experience is that when δ is between -0.5 and 0.5 (except 0), the performance is found to be relatively stable. Otherwise, there is a tendency for performance to decline on the some data sets. This conclusion is also applicable to VRRLMs on the data set with non-collinear variables.

2) Result on the data sets with non-collinear variables.

Similarly, the variables within the data sets shown in the Table 5, were found to be non-collinear. The calculation results of the two models are also shown in Table 5.

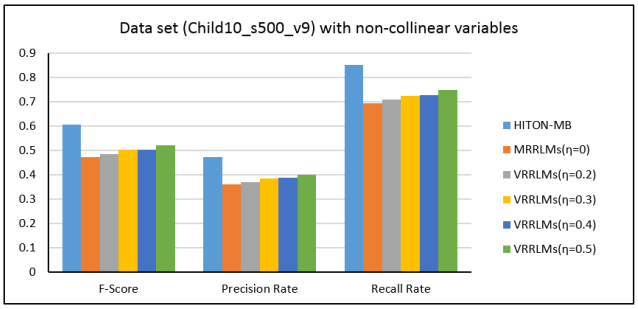

FIGURE 6. Discovery efficiency on the data set (Child_s500_v9) with non-collinear variables.

The Table 5 says that there is little difference in the discovery efficiency of the two models ($\delta = 0$ and $\delta \neq 0$), therefore, MRRLMs can be replaced by VRRLMs. We unify models which can get better discovery efficiency on the data sets with collinear and non-collinear variables. The same conclusion can be drawn more easily from Figure 5 and Figure 6.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose VRRLMs modified from MRRLMs to identify a subset of Markov boundary on the data sets with collinear and non-collinear variables. Theoretical and experimental results show that the *MB* discovery efficiency of VRRLMs is significantly improved than that of MRRLMs on the data sets with collinear variables, while both of them are basically similar discovery efficiency on the data sets with non-collinear variables, therefore, VRRLMs can completely replace MRRLMs and is fully applicable to discover *MB* on the data sets with collinear and non-collinear variables.

The literature [27] theoretically assumes that MRRLMs has the ability of variables selection, in fact, neither the ridge regression models nor MRRLMs has the ability of variables selection, both of them need to use other techniques for implementing variable selection. The most common method is used to set the threshold of p-value (such as 0.05), but the results vary with different thresholds, hence, it is difficult to get the optimal solution in practice, this is why we do not directly use the p-value threshold but adopt the p-value ranking method. By using p-value ranking method, the example in the literature [27] shows that MRRLMs are competitive against the traditional algorithm in discovering part of the *MB* on the NOTCH1 and RELA gene data sets, we reasonably infer that VRRLMs have similar conclusions.

VRRLMs can significantly improve the discovery performance of Markov blanket on data sets with collinear variables. However using regularized linear models to get Markov blanket is still at stage of the theoretical research and practical exploration, so there are still some limitations of application, for example, ridge regularized linear models are worse than traditional algorithms such as HITON on our low dimensional data sets. Moreover, high-dimensional covariance matrix and permutation test need too much computing power. However, regularized linear models as an important method of feature selection, such as ridge regression and LASSO, have been widely applied in the fields of machine learning in high-dimension small sample data sets and achieved good performance, therefore, further work can be carried out in the following aspects in the future: (1) covariance matrix shrinkage algorithm. Under the condition of no loss or little loss of matrix information, the optimization algorithm of covariance matrix calculation is studied to reduce the computing costs. (2) Study the new computing p-value method for improving discovery performance. The permutation test used in this paper is essentially a method to computing the P value. Compared with the traditional p-value method, and the permutation test can significantly improve the discovery performance of *MB*(subset), but it takes more computing time.

ACKNOWLEDGMENT

In closing, the author would like to appreciate Mr. Li, Mrs. Liu, Mr. Jerry and Mr. Yu from the University of South Australia, they taught me a lot in scientific research during my study at the University of South Australia. The idea of the paper comes from their research team, and they also provided some valuable opinions when I wrote this article. Most importantly, the authors thank all reviewers who played a crucial role in the peer-review process for their professional comments. Finally, I would like to thank the editors of IEEE ACCESS for their support and encouragement.

REFERENCES

[1] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, B. Sun, and S. Ma, "From observational studies to causal rule mining," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 2, p. 14, Jan. 2015. doi: 10.1145/2746410.
[2] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, no. 5659, pp. 799–805, Feb. 2004.
[3] D. Koller, N. Friedman, and F. Bach, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.
[4] R. E. Neapolitan, *Learning Bayesian Networks*, vol. 38. Upper Saddle River, NJ, USA: Prentice-Hall, 2004.
[5] J. Li, L. Liu, and T. D. Le, *Practical Approaches to Causal Relationship Exploration*. Springer, 2015.
[6] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *J. Amer. Stat. Assoc.*, vol. 100, no. 469, pp. 322–331, 2005.
[7] K. Yu, X. Wu, W. Ding, and H. Wang, "Exploring causal relationships with streaming features," *Comput. J.*, vol. 55, no. 9, pp. 1103–1117, Sep. 2012. doi: 10.1093/comjnl/bxs032.
[8] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Educ. Psychol.*, vol. 66, no. 5, p. 688, Oct. 1974.
[9] A. Hyttinen, F. Eberhardt, and P. O. Hoyer, "Experiment selection for causal discovery," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3041–3071, Oct. 2013.
[10] A. Nichols, "Causal inference with observational data," *Stata J., Promoting Commun. Statist. Stata*, vol. 7, no. 4, p. 507, Dec. 2007.
[11] M. Zhang, P. Shi, L. Ma, J. Cai, and H. Su, "Quantized feedback control of fuzzy Markov jump systems," *IEEE Trans. Cybern.*, vol. 49, no. 9, pp. 3375–3384, Sep. 2019.
[12] M. S. Mahmoud and Y. Xia, "Improved exponential stability analysis for delayed recurrent neural networks," *J. Franklin Inst.*, vol. 348, no. 2, pp. 201–211, Mar. 2011.
[13] J. Wang, Q.-L. Han, and F. Yang, "Event-triggered dissipative control of networked interconnected stochastic systems," in *Proc. IEEE Int. Symp. Ind. Electron.*, May 2013, pp. 1–6.
[14] L. Wu, Y. Gao, J. Liu, and H. Li, "Event-triggered sliding mode control of stochastic systems via output feedback," *Automatica*, vol. 82, pp. 79–92, Aug. 2017.

[15] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. 13th Int. Conf. Int. Conf. Mach. Learn.* San Mateo, CA, USA: Morgan Kaufmann, 1996, pp. 284–292.
[16] S. Fu and M. Sein, "A review of Markov blanket induction algorithms," *Appl. Res. Comput.*, vol. 29, no. 1, Jan. 2011.
[17] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale Markov blanket discovery," in *Proc. FLAIRS Conf.*, vol. 2, May 2003, pp. 376–380.
[18] C. F. Aliferis, I. Tsamardinos, and A. Statnikov, "HITON: A novel Markov blanket algorithm for optimal variable selection," in *Proc. AMIA Annu. Symp. Proc.*, 2003, p. 21.
[19] J. M. Peña, R. Nilsson, and J. Björkegren, and J. Tegnér, "Towards scalable and data efficient learning of Markov boundaries," *Int. J. Approx. Reasoning*, vol. 45, no. 2, pp. 211–232, Jul. 2007.
[20] T. Gao and Q. Ji, "Efficient Markov blanket discovery and its application," *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1169–1179, May 2017.
[21] X. Liu and X. Liu, "Swamping and masking in Markov boundary discovery," *Mach. Learn.*, vol. 104, no. 1, pp. 25–54, Jul. 2016.
[22] E. V. Strobl and S. Visweswaran, "Markov blanket ranking using kernel-based conditional dependence measures," 2014, *arXiv:1402.0108*. [Online]. Available: https://arxiv.org/abs/1402.0108
[23] K. Yu, L. Liu, J. Li, and H. Chen, "Mining Markov blankets without causal sufficiency," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6333–6347, Dec. 2018.
[24] S. Fu, Z. Su, and M. Sein, "Accelerating the recovery of Markov blanket using topology information," *Comput. Sci.*, vol. 42, no. 11A, Nov. 2015.
[25] S. Acid, L. M. de Campos, and M. Fernández, "Score-based methods for learning Markov boundaries by searching in constrained spaces," *Data Mining Knowl. Discovery*, vol. 26, no. 1, pp. 174–212, Jan. 2013.
[26] M. Schmidt, A. Niculescu-Mizil, and K. Murphy, "Learning graphical model structure using l1-regularization paths," in *Proc. AAAI*, vol. 7, Jul. 2007, pp. 1278–1283.
[27] E. V. Strobl and S. Visweswaran, "Markov boundary discovery with ridge regularized linear models," *J. Causal Inference*, vol. 4, no. 1, pp. 31–48, Mar. 2015.
[28] D. Margaritis and S. Thrun, "Bayesian network induction via local neighborhoods," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 505–511.
[29] S. Yaramakala and D. Margaritis, "Speculative Markov blanket discovery for optimal feature selection," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2005, p. 4.
[30] Y. Zhang, Z. Zhang, K. Liu, and G. Qian, "An improved iamb algorithm for Markov blanket discovery," in *Proc. JCP*, vol. 5, no. 11, pp. 1755–1761, Nov. 2010.
[31] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, "Time and sample efficient discovery of Markov blankets and direct causal relations," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2003, pp. 673–678.
[32] S. Fu and M. Desmarais, "Local learning algorithm for Markov blanket discovery," in *Proc. Australas. Joint Conf. Artif. Intell.* Springer, 2007, pp. 68–79.
[33] S. R. de Morais and A. Aussem, "A novel scalable and data efficient feature subset selection algorithm," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer, 2008, pp. 298–312.
[34] Y. Zeng, X. He, Y. Xiang, and H. Mao, "Dynamic ordering-based search algorithm for Markov blanket discovery," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Springer, 2011, pp. 420–431.
[35] L. Fei, G. Lu, W. Jia, S. Teng, and D. Zhang, "Feature extraction methods for palmprint recognition: A survey and evaluation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 2, pp. 346–363, Feb. 2019.
[36] H. Liu and H. Motoda, *Computational Methods of Feature Selection*. Boca Raton, FL, USA: CRC Press, 2007.
[37] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, pp. 1157–1182, Mar. 2003.
[38] D. R. Wang and Z. Z. Zhang, "Variable selection for linear regression models: A survey," *J. Appl. Statist. Manage.*, vol. 29, no. 4, pp. 615–627, Jul. 2010.
[39] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, Nov. 1995.
[40] A. Statnikov, N. I. Lytkin, J. Lemeire, and C. F. Aliferis, "Algorithms for discovery of multiple Markov boundaries," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 499–566, Feb. 2013.
[41] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 1988.
[42] R. D. Cook, *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Hoboken, NJ, USA: Wiley, 1998.
[43] M. A. Long, K. J. Berry, and P. W. Mielke, Jr., "A note on permutation tests of significance for multiple regression coefficients," *Psychol. Rep.*, vol. 100, no. 2, pp. 339–345, Apr. 2007.

SHU YAN received the master's degree from the Hefei University of Technology, in 2006. He is currently pursuing the Ph.D. degree with the University of Science and Technology of China. He is an Engineer with the Institute of Intelligent Machines, Chinese Academy of Sciences. His research interests include causal inference, pattern recognition and intelligent systems, the agriculture Internet of Things, and decision support systems.

CHAOYUAN CUI received the Ph.D. degree in computer science from the University of Tsukuba, in 2005. He is currently a Professor with the Institute of Intelligent Machines, Chinese Academy of Sciences. His research interests include system virtualization, computer architectures, information and communication security, and machine learning.

BINGYU SUN received the Ph.D. degree from the University of Science and Technology of China, in 2004. In 2005, he was a Research Assistant with the Chinese University of Hong Kong. He is currently a Doctoral Supervisor and a Research Fellow with the Institute of Intelligent Machines, Chinese Academy of Sciences. He engages mainly in machine learning and intelligent decision. His research interests include data mining, knowledge discovery, big data, and the agriculture Internet of Things.

RUJING WANG is currently the Director, a Doctoral Supervisor, and a Research Fellow with the Institute of Intelligent Machines, Chinese Academy of Sciences. His research interests include the agriculture Internet of Things, ontology knowledge representation and visualization, automatic and semi-automatic knowledge acquisition, master-slave reasoning and decision fusion, and decision support systems development.