Alternative models for the modified form of ridge regularized linear model in discovering Markov boundary

Shu Yan1, 2, 3, Xiaobo Hu1 and Rujing Wang1

1 Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031 China
2 University of Science and Technology of China, Hefei 230026 China
3 Email: yanshu@iim.ac.cn

Abstract. It has been proved that the modified form of ridge regularized linear model (MRRLM) can discover a subset of Markov boundary of the target variable under some constrained conditions. However, MRRLM cannot be applied to the data sets with collinear variables due to covariance matrix is employed. To develop a suitable alternative model for MRRLM, we study the relationships of discovery performance of Markov boundary among MRRLM, ridge regression linear model (RRLM), and LASSO combining with permutation test through empirical method. In addition, we also proposed a new NVRRLM to discover Markov boundary of the target variable. The experimental results show that: (1) On the binary data sets, MRRLM has a basically similar performance with LASSO and RRLM; (2) On the continuous data sets, MRRLM has a basically similar discovery performance with LASSO but has higher discovery performance than RRLM; (3) The new NVRRLM can replace MRRLM on the data sets with collinear variables. The above experimental results demonstrate NVRRLM can effectively deal with variable collinearity problems.

1. Introduction

The causal relationships between variables are the foundation of all disciplines, such as computer science, medicine, statistics, economics, and social science [1-3]. Causal relationships have been widely accepted as an alternative to the best scheme by randomized controlled trials (RCTs) [4-6]. In most cases, RCTs are impractical to discover causal relationship from the observational data due to expensive, unethical or impossible [7-9]. As a directed acyclic graph model [3], a Bayesian Network can represent the causal relationships among all variables in the network. Specifically, for the target node Y in the network, it is only related to its parents, children, spouses, and independent with other nodes. The set of its parents, children, and spouses is called Markov boundary or Markov blanket (MB) of the variable Y. The property of MB is widely used as the feature selection algorithm for classification or regression in the fields of machine learning and data mining.

However, the global structure of Bayesian networks learnt from the high-dimension data sets has been proved to be NP-complete [10]. So researchers proposed many local structure search algorithms to address this challenging i.e. MB discover algorithms. According to incomplete statistics, there were as many as 17 representative algorithms from 1996 to 2013. In recent five years, more Markov boundary (blanket) learning algorithms have emerged [11-15]. These algorithms can be roughly divided into two categories: constraint-based algorithms and scoring-based algorithms. Constraint-based algorithms are based on condition independent test, while scoring-based algorithms are actually based on scoring and search strategy. However, the mainstream algorithms are mainly constraint-
based algorithms rather than scoring-based algorithms. The main reason is that the probability and topological structure information of Bayesian network can help to define constraint conditions effectively, but cannot help to establish the correlation between local scores and global scores [16].

In recent years, MB learning algorithms based on regularized linear model have also been reported. For example, Mark Schmidt [17] used BIC scoring mechanism to propose L1MB algorithm by virtue of LASSO in the process of building Bayesian network but did not give further theoretical proof. Strobl [18] proposed a modified form of ridge regularized linear model (MRRLM), and theoretically explored the relationship between the solution of the model and MB (subset) with permutation test, which was of great theoretical significance.

However, the covariance of MRRLM makes it unable to be applied to data sets with collinear variables. An instinctive thought is that: can ridge regression model replace MRRLM for MB discovery? If so, the above problem can be readily solved; if not, are there alternative regularized linear models? In order to answer the question, firstly, we study relationship of the discovery efficiency of MB among MRRLM, RRLM, and LASSO through empirical method, and tries to explore the applicability of the above regularized linear model on the data set. In addition, we also proposed a VRRLM for MB discovery on the data sets with collinear variables.

The structure of this article is as follows. The next section specifies some definitions and briefly reviews MRRLM. The third section reveal relationships between collinearity of variables and singularity of corresponding covariance matrix, and then introduces a new model NVRRLM that modified from MRRLM. Section four shows the experimental results and analysis. Section five concludes.

2. Definition and background

In order to introduce our developed work, some necessary definitions and concepts are presented as follows.

2.1. Notation

Except for specific annotations, the symbols and their relationships used are specified in Table 1.

| Symbol | Meanings |
|--------|----------|
| A, B, C | A, B, and C are a random variable |
| A, B, C | A, B, and C are a set of random variables |
| X | X is a set or matrix of explanatory variables |
| Y | Y is a response variable |
| (A; B) | A and B have the same number of columns, (A; B) denote that B is appended to the rows of A |
| (A, B) | A and B have the same number of rows, (A, B) denote that B is appended to the columns of A |
| A ⊥ B | A and B are independent given C |
| ΣX | covariance matrix of X |
| LASSO | Least absolute shrinkage and selection operator |
| MRRLM | Modified form of ridge regularized linear model |
| NVRRLM | New variant of ridge regression linear model |

2.2. A modified form of ridge regularized linear model

MRRLM is a variant of ridge regression model by modifying the penalty term, for the sake of knowledge completeness, we will introduce it briefly. Let K denote the number of the column dimensions of α and β, and let \( k \in [1, 2, ..., K] = \mathbb{K} \). For a linear regression model:

\[
Y = \alpha + \beta^T X + \epsilon
\]
where \( Y \in \mathbb{R}^{k \times n}, X \in \mathbb{R}^{p \times n}, \alpha \in \mathbb{R}^{k \times n}, \beta \in \mathbb{R}^{p \times K}, \epsilon \sim N(0, \sigma^2 I) \).

Following the reference [19], we make four assumptions as follows: 1) Global Markov condition holds; 2) Joint probability distribution of variables \((Y; X)\) satisfies the linear intersection property; 3) Covariance matrix \( \sum X \) is positive definite; 4) \( E(X|\eta^TX) \) is a linear function of \( \eta^TX \), when \( Y \perp X \mid \eta^TX (\eta \in \mathbb{R}^{p-d}) \).

If the above four assumptions are satisfied, then \( \beta \) is a non-zero matrix, and the solution \((\alpha^*, \beta^*)\) is to minimize the following optimization formula:

\[
\arg\min_{\alpha, \beta} \{ E(\alpha + \beta^TX, Y) \} + \lambda \text{tr}(\beta^T \sum X \beta)
\]

Then, \( S(\beta^*_k) \subseteq S(\eta) \) for all \( k \in \mathbb{K} \), where, \( S(\eta) \) is any subspace spanned by the columns of \( \eta \), \( \mu \) is convex function, and \( \lambda > 0 \) is a penalty parameter.

The above mentioned reveals the relationship between the solution of the model and the reduced dimension matrix, and the reason why the model can only find a subset of Markov boundary. We can recover a subset of the parents, children, and spouses of \( Y \) from the non-zero coefficients of \( \beta^* [19] \), however, in practice, the target variable \( Y \) is usually a one-dimensional variable, that is \( k = 1 \). Let \( \beta = (\sum X)^{-1/2} \gamma \), then the formula (2) is equivalent to the following ridge regression linear model:

\[
\arg\min_{\alpha, \beta} \{ E(\alpha + \beta^TZ, Y) \} + \lambda \gamma^T \gamma , \text{ where } Z = (\sum X)^{-1/2} X
\]

There are many off-the-shelf tools and software available for solving objective function (3). In this paper, the cvglmnet function in the Glmnet toolkit is used to get the model parameters and coefficients. Obviously, the covariance matrix should be non-singular. We will discussed this in Section 3.1, the above formula indicates that the model cannot be applied to the data set with collinear variables for Markov boundary discovery. And then we proposed a new model to address this challenge.

3. A new variant of ridge regularized linear model
This section first introduces the variables collinearity which leads to the singularity of covariance matrix, and then presents new variant of ridge regularized linear model.

3.1. Collinearity of variables

For the design matrix \( X = (X_1; X_2; \ldots; X_p) \), \( X_i \in \mathbb{R}^n \), if the variables \( X \) are collinear, then there exists a set of constants \( k_0, k_1, \ldots, k_p \) that is not zero, so that the following formula holds:

\[
k_1 X_1 + k_2 X_2 + \ldots + k_p X_p = k_0 \text{ or } k^T X = k_0 , \text{ where } k = (k_1, k_2, \ldots, k_p)^T. \tag{4}
\]

Obviously, \( \text{Var}(k^TX) = k^T \sum X k = 0 \), also consider \( k^T \sum X k = \sum_{i=1}^{p} (k^T v_i)^2 \lambda_i \), where \( \lambda_i \) (or \( v_i \)) is an eigenvalue (or eigenvector) of \( \sum X \). Then we have \( \sum_{i=1}^{p} (k^T v_i)^2 \lambda_i = 0 \), so there exist at least one \( \lambda_i = 0 \) for \( (k^T v_i)^2 \geq 0 \), \( \lambda_i \geq 0 \) and \( k \neq 0 \).

Since \( \sum X = VD V^T \), \( VV^T = E \) (unit matrix), where, \( V \) is the matrix whose \( i \)th column is \( v_i \), \( i = 1, \ldots, p \), \( D \) is the diagonal matrix whose entries are \( \lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_p \). So there exists at least \( \lambda_i = 0 \) for collinearity of variables, therefore, the covariance matrix \( \sum X \) is singular and vice versa. Hence, we present a new model to address the issue.

3.2. A new variant of ridge regularized linear model

We add a unit matrix \( E \) to MRRLM for solving the collinear problem among variables.

Let \( \tilde{X} = X + \delta \epsilon, \delta \in [-1, 1], \Sigma = (\sigma_1; \sigma_2; \ldots; \sigma_p), \sigma_i \in \mathbb{R}^n \) \( \tilde{X} \)

\[
\text{Where } \sigma_i \text{ is an independent identically distributed pseudo-variable subject to the standard Gaussian distribution } N(0, 1). \tag{5}
\]
Suppose $Y \perp \tilde{X} | \eta^T \tilde{X}$ holds when $Y \perp X | \eta^T X$. This hypothesis is completely reasonable, because the design matrix is composed of observational data which has observed errors itself. Moreover, $\delta$ is very small number. Then, we have the following theorem:

**Theorem 1:** $\sum \tilde{X} = \sum X + \delta^2 E$ ($\delta \neq 0$) is non-singular and positive definite.

**Proof:** Let $\lambda_i$ is an eigenvalue of $\sum X$, and $v_i$ is an eigenvector of $\sum X$, then, $\sum \tilde{X} = \sum X + \delta^2 E = V(D + \delta^2 E) V^T$ for $\sum X = V D V^T$, $\lambda_i + \delta^2 > 0 (\forall i)$ is an eigenvalue of $\sum \tilde{X}$, so $\sum \tilde{X}$ is non-singular when $\delta \neq 0$.

We know that $\sum \tilde{X}$ is a nonsingular matrix which means variables without collinear. As can be seen from section 3.1:

$\text{Var}(k^T \tilde{X}) = \text{Var}(k^T \tilde{X}) = k^T \sum \tilde{X} k \neq 0$, and $\text{Var}(k^T \tilde{X}) \geq 0$, therefore, $k^T \sum \tilde{X} k > 0 (k \neq 0)$. So $\sum \tilde{X}$ is positive definite.

Next, we rewrite section 2.2 and have following conclusion, but prove it in the future. We further make new four assumptions as follows: 1) Global Markov condition holds; 2) Joint probability distribution of variables $(Y; X)$ satisfies the linear intersection property; 3) Covariance matrix $\sum \tilde{X}$ is positive definite; 4) $E (X | \eta^T \tilde{X})$ is a linear function of $\eta^T \tilde{X}$, when $Y \perp \tilde{X} | \eta^T \tilde{X}$ ($\eta \in \mathbb{R}^{p \times d}$).

If the new four assumptions are satisfied and the following formula can be minimized for solving $\alpha' \in \mathbb{R}, \beta' \in \mathbb{R}$

$$\arg\min E\{u(\alpha + \beta^T X, Y)\} + \lambda \text{tr}(\beta'(\sum X + \delta^2)\beta)$$

Then, $S(\beta') \subseteq S(\eta)$. Where $S(\eta)$ is any subspace spanned by the columns of $\eta$, $u$ is a convex function. Obviously, this model is equivalent to MRRLM when $\delta = 0$ and $k = 1$.

### 4. Simulation and analysis

Covariance matrix restricts the application of MRRLM, therefore, we are looking for alternate models of MRRLM to discover Markov boundary (subset) on the data sets with collinear variables. In this section, first, we compared MRRLM with RRLM and LASSO, and then MRRLM with the new model NVRRLM.

#### 4.1. Experimentation

From Figure 1, we know that IAMB algorithm has the best overall performance on the continuous data set, while HITON-MB has the best overall performance on the binary data set. Therefore, IAMB and HITON-MB are selected as the reference algorithms for the regularized linear model. The experimentation is as follows: 1) Select the appropriate model. According to the trial destination, MRRLM, RRLM, and LASSO were selected respectively; 2) Determine the number of target variables. The rule adopted in this experiment is: if the number of variables (dimension) in the data set is greater than 15, 15 target nodes will be randomly selected; otherwise, all variables will be extracted as target variables; 3) Determine the permutation number and model parameters. Considering of time costs, the permutation number in this experiment is set as 199. For the same reason, the same penalty parameter is used when the estimation parameters are repeatedly calculated; 4) Calculate parameter p-value. the p-value is calculated by using permutation test; 5) Determine the evaluation index of the model. The output results of this experiment include F-score and operation time. After the above process is completed, the discovery performance of each data set is summarized and averaged, and then used as the discovery performance of each model on the low-dimensional data sets.
Table 2. Data sets properties.

| Data set     | Alarm | Child | Insurance | Barley | Water |
|--------------|-------|-------|-----------|--------|-------|
| Nodes        | 37    | 20    | 27        | 48     | 32    |
| MB (Max/Min) | 8/1   | 8/1   | 10/1      | 13/2   | 13/1  |
| Data set     | Carpo | Chain | Factors   | Hailfinder | Mildew |
| Nodes        | 61    | 7     | 27        | 56     | 35    |
| MB (Max/Min) | 18/0  | 2/1   | 26/1      | 17/1   | 9/1   |

Figure 1. Comparison among three traditional algorithms, the solid line represents the continuous data sets, and the dotted line represents the binary data sets.

4.2. Result analysis

4.2.1. Comparison between regularized linear models and traditional algorithms. According to the trial goal, considering the high operating costs of permutation test and the representativeness of data set, we selected from the standard data sets of ten industries, whose number of dimensions less than 100. DAGlearn [18] is used to generate 10 continuous data sets and 10 binary discrete data sets with data samples of \{300, 600, 900, 1200, 1500\}, equivalent to 100 data sets. Data set properties are shown in Table 2.

As can be seen in Figure 2, first, traditional algorithms have slight higher discover efficiency than regularized linear model on the data sets, second, on the binary data sets, MRRLM has basically similar discovery performance than RRLM and LASSO in term of precision rate and overall performance (F-Score), while on the continuous data sets, MRRLM is much higher discovery performance than RRLM but MRRLM has basically similar discovery performance than LASSO. In terms of running time, the traditional algorithm takes the least time, while the regularized linear model algorithm generally takes too long, but the running time on the low-dimensional data set is still within the acceptable range.
4.2.2. Comparison between NVRRLM and MRRLM. This section is to verify the discovery performance of Markov blanket for NVRRLM on the data sets. Therefore, four discrete data sets with collinear variables and four discrete data sets with non-collinear variables were selected. Data set and evaluation index results are shown in Table 3.

Table 3. Discovery performance between NVRRLM and MRRLM.

| Data set Name | Collinearity (Y/N) | Source file     | MRRLM | NVRRLM |
|---------------|--------------------|-----------------|-------|--------|
|               |                    |                 | F-Score | Precision | F-Score | Precision |
| Alarm10       | Y                  | Alarm10_s500_v1 | 0.2473 | 0.2235   | 0.4587 | 0.3969   |
| Insurance     | Y                  | Insurance_s1000_v1 | 0.4830 | 0.5084   | 0.6508 | 0.6974   |
| Insurance     | Y                  | Insurance_s500_v1 | 0.3998 | 0.3571   | 0.5579 | 0.5260   |
| Gene          | Y                  | Gene_s500_v1    | 0.1310 | 0.1060   | 0.4812 | 0.3831   |
| Child         | N                  | Child_s500_v2   | 0.7783 | 0.7270   | 0.7983 | 0.7470   |
| Alarm10       | N                  | Alarm10_s1000_v1 | 0.4300 | 0.4359   | 0.4321 | 0.4327   |
| Alarm         | N                  | Alarm_s500_v1   | 0.6933 | 0.6933   | 0.6972 | 0.6604   |
| Child10       | N                  | Child10_s500_v9 | 0.4745 | 0.3626   | 0.4775 | 0.3665   |

From the Table 3, we can see that the discovery performance of NVRRLM is generally higher than that of MRRLM on the data sets with collinear variables, and the improvement of performance is obvious in some data sets. For example, on the Gene data set, the F-score can go from 0.1310 to 0.4812, while the discovery performance of the two models is basically equal on the non-collinear data sets. Note that in theory, MRRLM cannot be applied to data set with collinear variable. However, in practice, it is only found that the discovery performance is reduced. The reason is that the value "zero" is related to the number of decimal places defined in computer numerical storage, which is usually zero when the value is less than 10-12. But the discovery performance decreases with the degree of determinant of the covariance matrix approaching zero.

5. Conclusions
In this paper, we reveal the relationship of discovery performance among MRRLM, RRLM and LASSO. We also proposed a NVRRLM verified on the data sets. Experimental results show that on the low dimensional binary data sets, RRLM and LASSO can replace with MRRLM in discovering Markov boundary, while on the continuous data sets, LASSO can replace with MRRLM in
discovering Markov boundary. In addition, the results also show that NVRRLM can completely replace MRRLM in discovering Markov boundary on the data sets no matter which are collinear or non-collinear.

Acknowledgments
This research was financially supported by the National Key Research and Development Program (grant 2018YFD0700302). Study on key technologies of intelligent crop phenotype detection (grant Y821F42).

References
[1] Li J, Duy T, Liu, L, Liu J, Jin Z, Sun B and Ma S 2015 Acm Transactions on intelligent Systems &Technology 7(2) 1-27
[2] Friedman N 2004 inferring cellular networks using probabilistic graphical models J. Science. 303 799-805
[3] Neapolitan R E 2007 Learning Bayesian networks Acm Sigkdd national Conference on Knowledge Discovery & Data Mining
[4] Li J, Liu L and Duy T 2015 Practical approaches to causal relationship exploration (Springer)
[5] Rubin D B 2005 Causal inference using potential outcomes: Design, modeling decisions Journal of the American Statistical Association pp 322-331
[6] Yu K, Wu X, Ding W and Wang H 2012 Exploring causal relationships with streaming features Computer Journal 55(9) 1103-1117
[7] Rubin D B 1974 Estimating causal effects of treatments in randomized and nonrandomized studies Journal of educational Psychology 66(5) 688
[8] Rubin D B 2013 Experiment selection for causal discovery Journal of Machine Learning Research 14(1) 3041-71
[9] Nichols A et al. 2008 Causal inference with observational data Stata Journal 7(4) 507
[10] Chickering D M , Geiger D, Hecherman D 1996 Learning Bayesian Network is NP-Complete Networks 112(2) 121-130
[11] Fu S,Sein M 2011 A review of Markov blanket induction algorithms Application Research of Computers 29(1)
[12] Liu X, Liu X 2016 Swamping and masking in Markov boundary discovery J. Mach. Learn 104 25-54
[13] Gao T and Ji Q 2016 Efficient Markov Blanket Discovery and its Application IEEE transactions on Cybernetics 47(5) 1-11
[14] Strobl V E and Visweswaran S 2013 Markov Blanket Ranking using Kernel-based Conditional Dependence Measures NIPS Workshop on Causality
[15] Yu K, Liu L and Li J 2018 Mining Markov Blankets without Causal Sufficiency IEEE Transactions on Neural Networks and Learning Systems 29(12) 6333-47
[16] Fu S, Su Z and Sein M 2015 Accelerating the recovery of Markov blanket using topology information, Computer Science 42(11A)
[17] Schmidt M et al. 2007 learning graphical model structure using L1-regularization paths Proceedings of the 22nd national conference on Artificial intelligence 7 1278-83
[18] Strobl V E and Visweswaran S 2016 Markov Boundary Discovery with Ridge Regularized Linear Models Journal of Causal inference 4(1) 31-48
[19] Long M , Berry K P and Mielke P W 2007 A note on permutation tests of significance for multiple regression coefficients Psychological Rep. 100 339-45