

The Discovery of Generalized Causal Models with Mixed Variables using MML criterion

Gang Li*

Honghua Dai†

Abstract

One major difficulty frustrating the application of linear causal models is that they are not easily adapted to cope with discrete data. This is unfortunate since most real problems involve both continuous and discrete variables. In this paper, we consider a class of graphical models which allow both continuous and discrete variables, and propose the parameter estimation method and a structure discovery algorithm based on Minimum Message Length and parameter estimation. Experimental results are given to demonstrate the potential for the application of this method.

1 Introduction

There exist two kinds of *Graphical Models*, i.e., *Linear Causal Model* and *Bayesian Network*. While *Linear Causal Model* can only deal with continuous data, *Bayesian Network* can only deal with discrete data. However, most of the actual data set contains not only discrete data, but also continuous data. So how to combine them, is one of the problems in graphical models research. This paper extend the work of [1] in the way that both discrete and continuous variables are all allowed to exist in a model. A parameter estimation method and a model selection algorithm based on Minimum Message Length (MML) [2] are proposed, and experimental results are given to show the potential of this method.

The rest of this paper is organized as follows. In Section 2 we curtly describe a generalized linear causal model in which both continuous and discrete variables are allowed, and in Section 3 we develop an encoding scheme for this kind of generalized model using MML criterion. Experimental results are given and analyzed in Section 5. Finally, we conclude this paper in Section 6.

*School of Information Technology, Deakin University, Vic 3125, email: gangli@deakin.edu.au

†School of Information Technology, Deakin University, Vic 3125, email: hdai@deakin.edu.au

2 Generalized Linear Causal Model

Let $U = \{V_1, V_2, \dots, V_n\}$ be a set of random variables, which can be continuous or discrete. By definition, there exist two proper subsets of U , U_C and U_D , such that

- $U = U_C \cup U_D$ and $\emptyset = U_C \cap U_D$;
- U_C denotes the set of all continuous variables and U_D denotes the set of all discrete variables.

For any directed *Graphical Model* with a known *Directed Acyclic Graph* structure, we can define the conditional distribution of each node given its parents. According to the variables involved are pure discrete or continuous, or mixed discrete and continuous, we can have different kinds of local parameters.

2.1 Discrete Nodes with Discrete Parents For discrete nodes with discrete parents, suppose the local conditional probability distribution is a *Multinomial Distribution*.

For a *Generalized Linear Causal Model* with a structure S , when a discrete variable $V_i \in U_D$ has a set of discrete parent variables $Pa^S(V_i)$, the local parameter Θ_i is usually represented as conditional probability table at the variable V_i ,

$$(2.1) \quad \Theta_i = ((\theta_{ijk})_{k=1}^{r_i})_{j=1}^{q_i}$$

Where θ_{ijk} represents the conditional probability of variable V_i being in its k -th value, given that the parents of V_i is in its j -th value.

2.2 Continuous Nodes with Continuous Parents As for the case of continuous node with continuous parents, it can be defined similarly as in *Linear Causal Models*.

For a model with structure S , when a continuous variable $V_i \in U_C$ has a set of continuous parent variables $Pa^S(V_i) \subseteq U_C$, the relation between V_i and its parents $Pa^S(V_i)$ can be described by a linear function,

$$(2.2) \quad V_i = \sum_{k=1}^{K_i} \alpha_k \times Pa_k^S(V_i) + R_i$$

Where K_i is the number of parents for node V_i , $\{\alpha_1, \dots, \alpha_{K_i}\}$ are path coefficients, and R_i is assumed to be identically distributed following a *Gaussian* distribution with zero mean and a standard deviation that will also be treated as an adjustable parameters, that is $R_i \sim N(0, \sigma_i^2)$, so the set of local parameters Θ_i for a variable V_i with continuous parents is $\{\sigma_i^2, \alpha_1, \dots, \alpha_{K_i}\}$. This linear function can be written as a *Gaussian* distribution of variable V_i , conditioned on its parents:

$$(2.3) \quad p(V_i|Pa^S(V_i)) = N\left(\sum_{k=1}^{K_i} \alpha_k \times Pa_k^S(V_i), \sigma_i^2\right)$$

On the other hand, if the continuous variable V_i has no parents, we assume it as a random sample from a *Gaussian* distribution, $p(V_i) = N(\mu_i, \sigma_i^2)$, where μ_i is the expect value of node V_i , so the set of parameters Θ_i for a continuous variable V_i without parents is $\{\mu_i, \sigma_i^2\}$.

2.3 Continuous Nodes with Mixed Parents For a continuous variable V_i with both continuous parents and discrete parents, we can specify a linear function between V_i and its continuous parents for each possible configuration of discrete parents. This is called *Conditional Gaussian Distribution* in related papers [3].

Let S be the structure of the *Generalized Linear Causal Model*, V_i is a continuous variable with continuous parents $Pa_C^S(V_i)$ and discrete parents $Pa_D^S(V_i)$. The distribution of V_i , conditioned on each state of its discrete parents $Pa_D^S(V_i)$, is a *Gaussian* distribution:

$$(2.4) \quad p(V_i|Pa^S(V_i)) = N\left(\sum_{k=1}^{K_i} \alpha_k (Pa_D^S(V_i))_k, \sigma_i^2 (Pa_D^S(V_i))\right) \times Pa_C^S(V_i)_k, \quad (3.6)$$

Where $Pa^S(V_i) = Pa_C^S(V_i) \cup Pa_D^S(V_i)$, and $Pa_C^S(V_i)_k$ is the k -th continuous parent of V_i .

Similarly, when the variable V_i has discrete parents $Pa_D^S(V_i)$ but without continuous parent, that is, $Pa^S(V_i) = Pa_D^S(V_i)$, we assume it to be a *Gaussian* distribution:

$$(2.5) \quad p(V_i|Pa^S(V_i)) = N(\mu_i(Pa_D^S(V_i)), \sigma_i^2(Pa_D^S(V_i)))$$

2.4 Summary of GLCM The *Generalized Linear Causal Model* allows both continuous and discrete variables, except that continuous variable can not be the parent of discrete parents. We can summarize different types of the relations between child node and its parents as in Table 1. Generally speaking, the effect from discrete variables to a continuous variable is captured by

a *Conditional Gaussian Distribution*. It is interesting that *Gaussian* distribution can be viewed as a special case of *Conditional Gaussian* distribution in which only one condition is allowed.

Table 1: Summary of Local Conditional Distribution

Parents / Child	Discrete	Continuous
Discrete	Multinomial	Conditional Gaussian
Continuous	not allowed	Gaussian
Mixed	not allowed	Conditional Gaussian

3 MML Discovery of Generalized Linear Causal Model

When discovering the *Generalized Linear Causal Model* from data set, there are two things to consider, namely specifying the model structure and specifying the local conditional probability distributions at each variable.

3.1 Minimum Message Length Criterion Minimum Message Length (MML) is an invariant Bayesian point estimation technique proposed by Wallace [2, 4]. According to the MML criterion [2], the shorter the encoding message length is, the better is the corresponding model.

Given data set D and d -dimensional parameters Θ , Wallace and Freeman [4] give us a formula which estimates the message length ¹:

$$MessLen(\Theta \& D) \approx \frac{d}{2} \log(\kappa_d) - \log(h(\Theta)) + \frac{1}{2} \log(\det(F(\Theta))) + L(\Theta) + \frac{d}{2}$$

where $h(\Theta)$ is a prior distribution over parameter values, $L(\Theta)$ is the negative log-likelihood function of the data set D , that is $L(\Theta) = -\log f(D|\Theta)$. $F(\Theta)$ is the *expected* Fisher Information matrix of expected second partial derivatives of the negative log-likelihood.

In the case of MML Discovery of *Generalized Linear Causal Models*, the whole encoded message consists of 3 segments,

1. message encoding the model structure S .

¹It is assumed that the prior probability of parameters is not rapidly changing, and the likelihood function satisfies some regularity conditions. A complete derivation of this formula can be found in [5].

2. message encoding parameters Θ_S , which includes a set of local parameters θ_i specifying the conditional probability at each node V_i in the model.
3. message encoding data set, under the assumption that the *Generalized Linear Causal Model* was the true model.

From *Information Theory*, the total message length can be approximated using the following formula 3.7²,

$$(3.7) \quad L = L(S) + \sum_{i=1}^n (L(\theta_i) + L(D_i|\theta_i))$$

Where n is the number of nodes, θ_i is the local parameters at node V_i , and D_i is the data set confined to node V_i .

3.2 Encoding the model structure The structure of a *Linear Causal Model* is a *Directed Acyclic Graph* (DAG), which can be described by specifying the parents set $Pa(V_i)$ for each node V_i of the DAG. This description consists of the number of parents, followed by the index of the set $Pa(V_i)$ in some enumeration of all sets of its cardinality. So the length for encoding the model structure can be calculated using the following formula:

$$(3.8) \quad L(S) = \sum_{i=1}^n \left(\ln n + \ln \binom{n}{m_i} \right)$$

Where n is the number of nodes, and m_i is the number of parents for node V_i . To avoid intensive computational time cost in calculating $\ln \binom{n}{m_i}$, we can use *Stirling's approximation* formula $x! = x^x e^{-x} \sqrt{2\pi x}$, thus, the length of encoding the model structure can be approximated by,

$$(3.9) \quad L(S) = \sum_{i=1}^n \left(\ln n + (n - m_i) \ln \left(\frac{n}{n - m_i} \right) + m_i \ln \left(\frac{n}{m_i} \right) \right)$$

3.3 Encoding the Multinomial Distribution For a discrete variable V_i with discrete parents, the conditional distribution at V_i is *Multinomial* for each configuration of the discrete parents.

Let r_i be the number of possible states of the discrete variable V_i , j be the configuration of its discrete parents, so $\{\theta_{ij1}, \theta_{ij2}, \dots, \theta_{ijr_i}\}$ is the parameters involved in this *Multinomial* distribution. Firstly,

²For convenience, we use natural logarithm through the paper, so calculate all message length in *nits*. If the logarithm is 2-based, the length will be calculated in *bits*.

a uniform prior, $h(\theta_{ij})$ is assumed over the $(r_i - 1)$ -dimensional space $\sum_{k=1}^{r_i} \theta_{ijk} = 1$.

Here we divide the data set D_i at variable V_i into q_i subset, $D_i = D_{i1} \cup D_{i2} \cup \dots \cup D_{iq_i}$, each subset corresponding to one possible configuration of discrete parents. First we focus on the configuration j of the discrete parents and the sub data set D_{ij} . Letting n_{jm} be the number of data items in which V_i is in state m and $N_j = n_{j1} + \dots + n_{jm}$, minimizing the message length formula gives that the MML estimates of θ_{ijk} is given by $\theta_{ijk} = \frac{n_{jk} + \frac{1}{2}}{N_j + \frac{r_i}{2}}$. This nominally gives rise to a minimum message length of

$$(3.10) \quad L(\theta_{ij}) + L(D_{ij}|\theta_{ij}) = \frac{r_i - 1}{2} \left(\log \frac{N_j}{12} + 1 \right) - \log(r_i - 1)! - \sum_k (n_{jk} + \frac{1}{2}) \log \theta_{ijk}$$

for both stating the parameter estimates and then encoding the things in light of these parameter estimates. So the total message length at variable V_i is

$$(3.11) \quad L(\theta_i) + L(D_i|\theta_i) = \sum_{j=1}^{q_i} \left\{ \log q_i + \frac{r_i - 1}{2} \left(\log \frac{N_j}{12} + 1 \right) - \log(r_i - 1)! - \sum_k (n_{jk} + \frac{1}{2}) \log \theta_{ijk} \right\}$$

3.4 Encoding the Gaussian Distribution For a continuous variable V_i , if none of its parents is discrete, the encoding length of the parameter and the confined data set can obtained similarly as in a *Linear Causal Model* [1].

When the variable V_i has no parent at all, we assume it is drawn from a *Gaussian* distribution $N(\mu_i, \sigma_i^2)$, each data measurement is given to an accuracy of $\pm \frac{\epsilon_i}{2}$. Assuming the prior for parameters $\theta_i = \{\mu_i, \sigma_i^2\}$ to be $h(\mu_i, \sigma_i^2) \sim \frac{1}{\sigma_i^2}$, the total length for parameters and confined data set can be calculated by,

$$(3.12) \quad L(\theta_i) + L(D_i|\theta_i) = \frac{1}{2} \ln \left(\frac{T^2}{2} \right) + \ln \kappa_2 + 1 + \frac{T}{2} \ln 2\pi + \frac{T-1}{2} \ln \sigma_i^2 + \sum_{t=1}^T \frac{(v_{it} - \mu_i)^2}{2\sigma_i^2} - T \ln \epsilon_i$$

where T is the size of the training data set D .

To minimize the total encoding length, we examine its partial derivatives with respect to μ_i and σ_i^2 , and

obtain the estimate of them to be $\mu_i = \frac{\sum_{t=1}^T v_{it}}{T}$ and $\sigma_i^2 = \frac{\sum_{t=1}^T (v_{it} - \mu_i)^2}{T-1}$.

When the variable V_i has K continuous parents (without discrete parents), the combined message length for the parameter and the confined data set D_i is

$$(3.13) \quad \begin{aligned} L(\theta_i) + L(D_i|\theta_i) &= \frac{1}{2} \ln\left(\frac{T}{2}\right) + \frac{1}{2} \ln |A| + \frac{K+1}{2} \ln \kappa_{K+1} \\ &+ \frac{K+1}{2} + \frac{T}{2} \ln 2\pi + \frac{T-K}{2} \ln \sigma_i^2 \\ &+ \sum_{t=1}^T \frac{(v_{it} - \sum_k \alpha_k P a_k(v_{it}))^2}{2\sigma_i^2} \\ &- T \ln \epsilon_r \end{aligned}$$

The coefficients of the linear regression, $\{\alpha_1, \dots, \alpha_K\}$ can be estimates by least squares estimation, and the estimation for σ_i^2 is $\sigma_i^2 = \frac{\sum_{t=1}^T (v_{it} - \sum_k \alpha_k P a_k(v_{it}))^2}{T-K}$.

3.5 Encoding the Conditional Gaussian distribution For a continuous variable V_i with both continuous and discrete parents, the conditional distribution at V_i is a conditional Gaussian, that is, for each configuration of its discrete parents, the relation between V_i and its continuous parents is a Gaussian distribution.

Let q_i be the number of configurations of the discrete parents, and the confined data set D_i can be divided into q_i subset, $D_i = D_{i1} \cup D_{i2} \cup \dots \cup D_{iq_i}$, each corresponding to one possible configuration of discrete parents.

If we focus on the configuration j of the discrete parents and its corresponding sub data set D_{ij} . Since the relation between V_i and its continuous parents can be captured by a Gaussian distribution, the encoding length for these linear coefficients and D_{ij} can be calculated in the same way as in formula 3.12 or 3.13. Considering that some message segment need to be used to distinguish different configurations of discrete parents, the total message length at variable V_i is

$$(3.14) \quad L(\theta_i) + L(D_i|\theta_i) = \sum_{j=1}^{q_i} \left\{ \log q_i + \left(L(\theta_{ij}) + L(D_{ij}|\theta_{ij}) \right) \right\}$$

4 Search Strategy

For a given sample data set, the number of possible model structures which may fit the data is exponential in the number of variables. To find out the best model structure from this huge space, an efficient search strategy is highly demanded [6].

In this paper we use the Message Length based Greedy Search (MLGS) [7] algorithm: Starts with a directed acyclic graph provided by user or a null graph without any edge, *Message Length Based Greedy search* runs through each pair of nodes attempting to add an edge if there is none or to delete or to reverse it if there already is one. Such adding, deleting or reversing is done only if such changes result in a decrease of the total message length of the new structure. If the new structure is better, it is kept and then try another change. This process continues until no better structure is found within a given number of search steps, or the search from the whole structure space is completed.

5 Empirical Results

In this section, we report the results of discovering generalized linear causal models from several models using the method proposed in this paper. Intuitively, if a causal discovery algorithm is working perfectly, it should reproduce exactly the model used to generate the data. In practice, sampling errors will result in deviations from the original model, but algorithm which can reproduce a model structure similar to the original structure should be considered to be better than those do not. Two experiments were conducted with synthetic data sampled from known models. In experiment 1, the proposed algorithm was compared with BIC-based algorithm [8] on 3 Bayesian networks, and with the algorithm in [1] on 3 linear causal models. In experiment 2, the proposed algorithm was called to induce model from generated data sampled from the *waste incineration plant* model [9].

5.1 Experiment 1 Firstly, we want to evaluate the ability of discovering Bayesian network, and Linear causal models. In the experiment, we selected 3 Bayesian networks, which contain only discrete variables, and 3 linear causal models which contain only continuous variables, The summary information of these models are described in Table 2.

Table 2: Models used in Experiment 1

Model	# Nodes	# Edges
<i>Asia</i>	8(d)	8
<i>Burglary</i>	5(d)	4
<i>Cancer</i>	5(d)	5
<i>Blau</i>	6(c)	9
<i>Case9</i>	9(c)	12
<i>Case10</i>	10(c)	10

For Bayesian networks, we compared our method

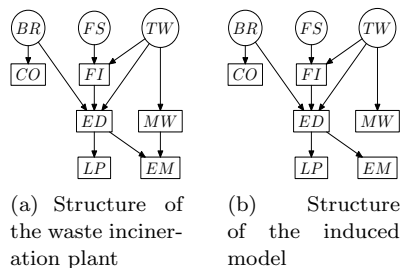


Figure 1: the known and the discovered WIP model

with the BIC criterion, both using a greedy search. For linear causal models, we compare our method with the MML-CI II algorithm [1]. The results are given in Table 3, in which we give the number of incorrect edges for the results by different algorithms. *The Number of Incorrect Edges* is decomposed into a triple-tuple: $[m, a, r]$, in which m is the number of missing edges, a is the number of added edges, r is the number of reversed edges.

Table 3: Comparison between MML and BIC

Model	Old method	New Method
<i>Asia</i>	[1,0,1]	[0,0,1]
<i>Burglary</i>	[0,0,1]	[0,0,1]
<i>Cancer</i>	[0,0,1]	[0,0,1]
<i>Blau</i>	[0,1,2]	[0,1,2]
<i>Case9</i>	[0,0,0]	[0,0,0]
<i>Case10</i>	[0,0,0]	[0,0,0]

From this table, we can see that the new method can keep the ability of both BIC criterion and the MML criterion used before.

5.2 Experiment 2 In order to evaluate the performance of learning graphical models from mixed variables, we use the *waste incineration plant* (WIP) model as described in [9]. This model is not intended to be a complete description of a waste incineration plant, but is limited to moderate size in the interest of simplicity. Its structure is given in figure 1(a).

In this figure, circles represent discrete variables, and squares represent continuous variables. In the interest of simplicity, we don't give the parameters of this model here. Based on this model, we generate one data set of 2000 items using BNT toolbox [10].

The result induced by our method is given in figure 1(b). Comparing the result with the original model structure, we can see that there is only one difference between them. The estimated parameters are also very close to the parameters in the original models,

except in the node *EM*.

6 Conclusion

One major difficulty frustrate the application of learning algorithms for graphical model is their ability to deal with mixed variables. This is unfortunate because most real problems involve both continuous and discrete variables. In this paper, we focused on graphical models which allow both discrete and continuous variables, and proposed a Minimum Message Length-based learning algorithm. Our experimental results indicate that this method is capable of recovering mixed model which is quite close to the original model, while preserve the discovering ability for Bayesian network and Linear causal model.

Further work includes how to deal with the case in which a discrete variable with continuous parents, and how to deal with missing data in these mixed data sets.

References

- [1] Li, G., Dai, H., Tu, Y.: Linear causal model discovery using MML criterion. In: Proceedings of 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, IEEE Computer Society (2002) 274–281
- [2] Wallace, C., Boulton, D.: An information measure for classification. *Computer Journal* **11** (1968) 185–194
- [3] Olesen, K.G.: Causal probabilistic networks with both discrete and continuous variables. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15** (1993) 275–279
- [4] Wallace, C., Freeman, P.: Estimation and inference by compact coding. *Journal of the Royal Statistical Society* **B,49** (1987) 240–252
- [5] Oliver, J.J., Baxter, R.A.: MML and Bayesianism: Similarities and differences. Technical Reports TR94/206, Department of Computer Science, Monash University, Clayton, Victoria 3168, Australia (1994)
- [6] Jordan, M.I.: *Learning in Graphical Models*. 1 edn. MIT Press, Cambridge, MA (1998)
- [7] Dai, H., Li, G., Tu, Y.: An empirical study of encoding schemes and search strategies in discovering causal networks. In: Proceedings of 13th European Conference on Machine Learning (Machine Learning: ECML 2002), Helsinki, Finland, Springer (2002) 48–59
- [8] Heckerman, D., Geiger, D.: Learning Bayesian networks. Technical Report MSR-TR-95-02, Microsoft Research, Redmond, WA (1995)
- [9] Lauritzen, S.: Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistics Association* **87** (1992) 1098–1108
- [10] Murphy, K.: *The Bayes Net Toolbox for Matlab*. *Computer Science and Statistics* **33** (2001) 331–351