An estimation of causal structure based on Latent LiNGAM for mixed data

Mako Yamayoshi¹ · Jun Tsuchida² · Hiroshi Yadohisa¹

Received: 23 April 2019 / Accepted: 9 September 2019 / Published online: 19 September 2019
© The Author(s) 2019

Abstract

The linear non-gaussian acyclic model (LiNGAM) has been proposed as a method for estimating causal structures using structural equation modeling (SEM). LiNGAM is useful as an exploratory estimation method for a causal structure. However, the assumptions that all observed variables in LiNGAM are continuous is not applicable in case of mixed data (i.e., when discrete variables are also included in the dataset). Therefore, we propose the Latent LiNGAM (L-LiNGAM), where each variable corresponds to a continuous latent variable and is observed as data through transformation via a link function. In the numerical study, when mixing discrete variables, the estimation of causal structure using L-LiNGAM is proven useful in terms of sum of squared error and path recovery. Moreover, from real-world data applications, the causal structure estimated by L-LiNGAM is shown to be the best for evaluation under SEM. The model fit is also superior to that of existing methods.

Keywords SEM · Latent variable · Causal structure · Generalized linear model · Link function

1 Introduction

Recently, the estimated causal structure represented by directed acyclic graph (DAG) modeling became important in various research fields. Confirmatory structural equation modeling (SEM) and Bayesian Networks using conditional independence are proposed as conventional methods for estimating causal structures via DAG. Actually, management (Haigh and Bessler 2004), economics

Communicated by Shohei Shimizu.

✉ Jun Tsuchida
jtsuchida@rs.tus.ac.jp

¹ Doshisha University, 1-3 Tataramiyakodani, Kyotanabe, Kyoto, Japan

² Tokyo University of Science, 6-3-1, Nijjuku, Katsushika-ku, Tokyo, Japan

(Awokuse and Bessler 2003), and genetics (Goeman and Mansmann 2008) research using DAG is applied to real data.

Exploratory SEM is often used to estimate causal structures. However, when the number of observed variables is large, the number of candidates for the exploratory causal structure is also extremely large. Therefore, it is difficult for analysts to build a DAG model suitable to the data. To address this problem, the linear non-gaussian acyclic model (Shimizu et al. 2006; Shimizu 2014, LiNGAM) has been proposed for unique causal structures, as it can uniquely estimate DAG using non-Gaussianity and linear relationships between observed variables.

LiNGAM is thus useful as an exploratory causal structure estimation method, and is used in various fields, such as neuroscience (Xu et al. 2014) and psychology (Takahashi et al. 2012). However, some problems of LiNGAM have been pointed out by Shimizu and Bollen (2014) as well as Zhang and Hyvärinen (2010). One of them is the assumption that all observed variables are continuous. That is, if discrete variables, such as categorical variables, are included in the data (which are then called mixed data), the causal structure estimated by LiNGAM changes depending on the value of the categorical variable. The assumption that all relationships between variables are linear implies that the fitting of LiNGAM worsens, although we select a true causal relationship when discrete variables are introduced. The cause of this phenomenon is that the relationship between continuous and discrete variables is non-linear, such as with a sigmoid curve or a log–log curve in many cases.

When all observed variables are discrete, the Poisson DAG model (Park and Raskutti 2015), based on the Poisson model of overdispersion, and the Quadratic Variance Function (QVF) DAG model, which uses multivalued categorical data for overdispersion (Park and Raskutti 2017), have been proposed. Since the Poisson DAG and QVF DAG are methods for only discrete variables, they are inappropriate when handling mixed quantitative and qualitative data. Inazumi et al. (2014) proposed a LiNGAM-like algorithm for binary data, and Wiedermann and von Eye (2017) discussed log-linear models to test causal effect directionality, making use of the same latent variable assumption as was used in the present manuscript. Especially in fields such as biology (Goeman and Mansmann 2008; Huber et al. 2007), artificial intelligence (Tepeš et al. 2016), and economics (Castañeda and Guerrero 2018; Castañeda et al. 2018); these methods show information loss from discrete variables (Li and Shimizu 2018).

The nonlinear acyclic causal model (Zhang and Hyvärinen 2010) and post-nonlinear (PNL) causal model (Zhang and Hyvärinen 2009) have been proposed for estimating nonlinear relationships. The observed variable $x_{PA(i)}$ is the parent of the observed variable x_i , which is described through a nonlinear transformation. However, as in the case of LiNGAM, these models also assume that all observed variables are continuous, and thus they are not suitable for the estimation of causal structures with mixed data. As an extension of LiNGAM, a method using logistic regression as a nonlinear transformation of observed variables has been proposed, namely the hybrid causal model (HCM) (Li and Shimizu 2018), extended using a sigmoid function for LiNGAM as a causal structure estimation

method when mixing continuous and binary variables. However, this model still cannot deal with categorical data other than binary discrete data.

The main interest of this study is in addressing the problem caused by mixed data. With this as motivation, we propose the Latent LiNGAM (L-LiNGAM) method. Specifically, we assume that each observed variables generate from a latent variable via a one-to-one correspondence. This assumption is also used to compute polychoric and polyserial correlation coefficients (Drasgow 1986). The L-LiNGAM conducts DAG not by observed variables but by latent variables. Rather than estimating relationships between observed variables through the PNL Causal Model or HCM using nonlinear functions, a latent variable forms a background of continuous values in the data observed as discrete values. This is a natural extension to the likelihood of the DAG model through the latent variable transformation with link functions. Another extension method of LiNGAM (Hoyer et al. 2008) including an unobserved common cause has also been proposed, but it still assumes that all observed variables are continuous, making discrete values and nonlinear applications of the relationship difficult. L-LiNGAM aims to allow not only the use of continuous and binary variables, but also the estimation of causal structures with mixed data. It is assumed that L-LiNGAM has some expected values such as the generalized linear model and is obtained from a distribution having expected values (Molenaar and Bolsinova 2017). As such, the causal structure of variables observed as discrete values is estimated from the causal structure of latent variables. In this study, we do not focus on the theoretical results of DAG but on the algorithm, because the DAG model for mixed data is required in many research areas.

The article is structured as follows. Section 2 describes the L-LiNGAM and its relationship with previous works. Section 3 evaluates L-LiNGAM on some numerical and real dataset applications of L-LiNGAM and other existing methods. Section 4 summarizes the study and presents future research directions.

2 Latent LiNGAM

In this section, we relax the two assumptions from the LiNGAM model. Using latent variables and link functions, we develop a method of estimating causal structures, even with mixed data. First, we describe L-LiNGAM and then the estimation algorithm.

2.1 Model

In the causal structure assumed by L-LiNGAM, the latent variable forming LiNGAM exists as a background of the observed variables, and the causal structure among observed variables is estimated from the causal structure of the latent variables, which become the original measurement items. We assume that the observed variables, observed as discrete values, are measured from latent variables through link functions, as shown in Fig. 1.

Fig. 1 Causal structure of L-LiNGAM

L-LiNGAM assumes the causal structure described by DAG. With observed variable $x_i (\in \mathbb{R} \vee \in \mathbb{N})$, latent variable $f_i \in \mathbb{R}$, error variable $e_i \in \mathbb{R}$, coefficient b_{ij} ($j = 1, 2, \dots, p; j \neq i$), and the one-to-one correspondence function g_i , L-LiNGAM is expressed as:

$$\begin{cases} f_i = \sum_{j \neq i} b_{ij} f_j + e_i & (2.1) \\ x_i = g_i(f_i). & (2.2) \end{cases}$$

We assume that the error terms e_i are independent from each other. In L-LiNGAM, continuous latent variables are described by LiNGAM, and the mean structure of observed variables is described by the link function corresponding to the latent variable. Equation (2.1) is the same for LiNGAM and shows the structure among latent variables. Equation (2.2) expresses the relationship between observed and latent variables. Although g_i is used as a link function, it can be any differentiable one-to-one correspondence function. By these assumptions, we can fit structural equations and linear relationships when quantitative and qualitative data are considered, which was a problem for LiNGAM. When link functions g_i are identity functions, L-LiNGAM is equivalent to LiNGAM in terms of constructing a structural equation in which all functional forms are linear. In other words, the proposed model is an extension of LiNGAM, in the sense that it provides a larger functional system.

Next, we reconsider L-LiNGAM under the ICA framework. First, L-LiNGAM can be represented in matrix form as:

$$\mathbf{f} = \mathbf{B}\mathbf{f} + \mathbf{e}, \quad (2.3)$$

$$\mathbf{x} = \mathbf{g}(\mathbf{f}), \quad (2.4)$$

where $\mathbf{x} \in \mathbb{R}^{p \times 1}$ is an observed variable vector, $\mathbf{f} \in \mathbb{R}^{p \times 1}$ is a latent variable vector, $\mathbf{e} \in \mathbb{R}^{p \times 1}$ is an error variable vector, and $\mathbf{g}$ is a link function vector. Moreover, with $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$, Eqs. (2.3) and (2.4) are expressed as:

$$\mathbf{x} = \mathbf{g}(\mathbf{A}\mathbf{e}). \quad (2.5)$$

Equation (2.5) is the same model as the post non-linear ICA (Taleb and Jutten 1999). PNL ICA makes the same assumption as ICA that $\mathbf{A}$ has an inverse matrix and $\mathbf{e}$ has independent components that follow at most one Gaussian distribution. The assumption that $\mathbf{g}$ can be inverted is also added. $\mathbf{f}$ is an independent non-Gaussian component in L-LiNGAM, while $\mathbf{e}$ corresponds to $\mathbf{f}$ in Eq. (2.3) and Eq. (2.3) becomes an independent component. Consequently, the assumptions of L-LiNGAM can satisfy the assumption of PNL ICA and can be regarded as an ICA framework. That is, L-LiNGAM is represented by constrained PNL ICA. This relationship between L-LiNGAM and PNL ICA is the same as that between LiNGAM and ICA. We could select a distortion of $\mathbf{f}$ that satisfies PNL ICA assumption such as t -distribution.

2.2 Estimation and algorithm

As in the Direct LiNGAM Algorithm (Shimizu et al. 2011), which is one of LiNGAM's estimation algorithms, the algorithm of L-LiNGAM is based on a structural equation model, in which the direction of the path between two variables is in opposition. The causal structure is estimated by comparing the mutual information on observed variables and residuals. In other words, observed variables that are independent of the residuals in the set of two variables are taken as the head of the causal structure. As with LiNGAM for an unobserved common cause (Shimizu and Bollen 2014), latent variables are generated as random numbers from an arbitrary non-Gaussian distribution.

We define an index $m(x_i, x_j)$, which compares the mutual information of observed variables x_i and x_j with their respective residuals by the following formula:

$$m(x_i, x_j) = NMI(x_j, r_i^{(j)}) - NMI(x_i, r_j^{(i)}),$$

where $r_i^{(j)} = g_i(b_{ij}f_j + e_i) - x_i$,

where $r_i^{(j)}$ is the residual when the regression is performed with x_i as objective variable and x_j as explanatory variable, where $NMI(x_j, r_i^{(j)})$ is the Normalized Mutual Information (NMI) (Strehl and Ghosh 2002), defined below with x_j and $r_i^{(j)}$. In L-LiNGAM, since the transformation is performed by different link functions for each observed variable, scaling is necessary to compare mutual information as NMI:

$$NMI(x_j, r_i^{(j)}) = \frac{I(x_j, r_i^{(j)})}{\sqrt{H(x_j)H(r_i^{(j)})}},$$

where $H(x) = \int_{-\infty}^{\infty} p(x) \log p(x) dx$ is the entropy of x . Moreover, $I(x_j, r_i^{(j)})$ is the mutual information of x_j and $r_i^{(j)}$. In practice, we need the estimator of the entropy and the mutual information. Among the many estimators available in the literature (e.g., Bach and Jordan 2002), we chose the histogram. That is, we chose the histogram as our probability model because of computational time. We could choose an alternative NMI among the many estimators for dependency, such as HSIC (Gretton

et al. 2005), because the algorithm does not depend on dependency measure. The advantage of NMI is that the adjustment between numerical and categorical examples is explicit and the range is independent from whether the example is categorical or numerical.

From the above, the observed variable with the largest $M(x_i; U)$ summarizing $m(x_i, x_j)$ is the beginning of the causal structure:

$$M(x_i; U) = - \sum_{j \in U} \min(0, m(x_i, x_j))^2. \quad (2.6)$$

U is a set of the observed variables to be analyzed. The algorithm for estimating L-LiNGAM is presented under Algorithm 1. For Algorithm 1 (e), the observed variable at the head is obtained, and the estimation of the causal structure is repeated using the residual matrix as the data matrix because we can eliminate the influence of the observed variable at the head on the causal structure.

ALGORITHM 1: Estimation algorithm for L-LiNGAM

Input: data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, ordered list of observed variables K

1. Generate p independently latent variable f_i from non-Gaussian distribution, such as t-distribution.
 2. Repeat the following (a) ~ (e) until $p - 1$ observed variable order is added to K .
 - (a) Perform regression with x_i as the objective variable and f_i as the explanatory variable.
 - if g_i is a link function
 - Perform regression via link function.
 - else g_i is an identity function
 - Perform regression via identity function.
 - (b) Update instance of f_i as predicted values by step (a)
 - (c) Calculate the residuals $r_i^{(j)}$ of regression of f_i on f_j
 - (d) Evaluate the independence of residuals $r_i^{(j)}$ and x_i obtained in (c) using Eq. (2.6).
 - (e) The residual matrix $\mathbf{R}^{(i)}$ is obtained by regression of all combinations when x_i is an explanatory variable set as the data matrix.
 3. Add the subscript of the last observed variable to the end of K .
 4. According to the order of K , perform regression of latent variable vector $\mathbf{f}$ and estimate the coefficient b_{ij} .
-

3 Application

In this section, we report numerical examples and real-world data application examples to compare L-LiNGAM to the other methods.

3.1 Numerical study

Here, L-LiNGAM, LiNGAM, HCM, ICA, and PNL ICA are compared using four variables (x_1, x_2, x_3, x_4) for a sample size of 100. To confirm the relationship between ICA and LiNGAM, LiNGAM uses an ICA-based algorithm to estimate causal structures.

The data generating process is as follows:

$$e_i \stackrel{\text{iid}}{\sim} t_{10}, f_i \stackrel{\text{iid}}{\sim} t_{10} \quad (i = 1, 2, 3, 4),$$

$$\mathbf{B} = (b_{ij}), \quad b_{ij} = \begin{cases} 2 & (i < j) \\ 0 & (\text{otherwise}) \end{cases}$$

We then set $\mathbf{f} = \mathbf{B}\mathbf{f}$. Both e_i and f_i were generated from a t-distribution with 10 degrees of freedom t_{10} , based on Shimizu and Bollen (2014). The number of iterations was 100, and the evaluation was made with the sum of square error (SSE) between predicted and actual values.

We build following four situations for the numerical case study.

Setting 1 Two variables are continuous and two are binary (g_i is an identity function and a logistic function). When generating binary data, logistic transformation is performed on the observed variables generated from t_{10} as follows:

$$x_i = \begin{cases} f_i + e_i & (i = 1, 3) \\ I\left(\frac{1}{1+\exp(-f_i+e_i)} < 0.5\right) & (i = 2, 4) \end{cases}$$

where $I(\cdot)$ is the indicator function. x_1 and x_3 are continuous data, while x_2 and x_4 are binary data.

Setting 2 All four variables are binary (g_i are all logistic functions). Binary data were generated by the same method as above.

Setting 3 All four variables are continuous (g_i are all identity functions). In this case, since LiNGAM, HCM, ICA, and PNL ICA are equivalent models, we compare LiNGAM, ICA, and L-LiNGAM. Continuous data were generated by the same method as above.

Setting 4 Two variables are five values and two variables are binary (g_i is an exponential and logistic function). When generating five-valued data, an exponential transformation is performed on the observed variable generated from t_{10} as follows:

$$x_i = 1 + \sum_{\ell=1}^4 I(\exp(f_i + e_i) > \ell) \quad (i = 1, 3).$$

Binary data were generated by the same method as above. In this case, LiNGAM, HCM, and ICA consider five-valued data as continuous.

We computed the sum of square error (SSE) between true x_i and the one estimated by L-LiNGAM, LiNGAM, HCM, ICA, and PNL ICA as follows:

$$\text{SSE}(\hat{x}_i) = \sum_{j=1}^{100} (x_{ji} - \hat{x}_{ji})^2.$$

We also computed the path recovery (PR) by L-LiNGAM, LiNGAM, and HCM as follows:

$$\text{PR}(\hat{B}) = \sum_{i=1}^4 \sum_{j=1}^4 (I(b_{ij} = 0 \wedge \hat{b}_{ij} = 0) + I(b_{ij} \neq 0 \wedge \hat{b}_{ij} \neq 0)).$$

In these settings, the maximum value of PR is 16.

Figures 2, 3, 4 and 5 are SSE of the corresponding settings. Tables 1 and 2 show mean and standard deviation of SSE and of PR under each setting, respectively. The dots in Tables 1 and 2 indicate no computation because the tables include the same results. The bold numbers in Tables 1 and 2 show the best result among settings.

From Table 1, the mean and standard deviation of SSE for L-LiNGAM are always stable for each variable, and there is no significant difference for the HCM with the highest accuracy in Setting 1. In Setting 2, L-LiNGAM has the best result with the exception of x_4 . The next most accurate model is L-LiNGAM, meaning there is no significant difference between HCM and L-LiNGAM. In Setting 3, the mean and standard deviation for the SSE of LiNGAM are the smallest, but there is no significant difference between LiNGAM and L-LiNGAM. The reason why we obtain different results for L-LiNGAM and LiNGAM is that we generated the

Fig. 2 SSE under mixed continuous and binary setting

Fig. 3 SSE for all variables are binary

Fig. 4 SSE for all variables are continuous

latent variable for all continuous observed variables; that is, we use different variables from observed variables for L-LiNGAM. In Setting 4, the mean and standard deviation of the SSE of LiNGAM are the smallest, but there is no significant difference between LiNGAM and L-LiNGAM. Compared to the SSE for five-value data,

Fig. 5 SSE for five-value and binary data

Table 1 Mean and standard deviation of SSE under each setting

Setting		L-LiNGAM	LiNGAM	HCM	ICA	PNL ICA
1	x_1	4.81 (6.52)	14.64 (76.63)	1.01 (1.46)	10.3 (20.99)	11.07 (25.75)
	x_2	4.74 (4.12)	1.01 (5.24)	1.19 (1.89)	5.85 (13.67)	5.91 (13.33)
	x_3	75.08 (109.11)	236.57 (1242.8)	165.06 (292.84)	72.27 (114.08)	74.46 (122.13)
	x_4	12.69 (16.53)	1.58 (6.67)	2.76 (4.19)	6.26 (21.74)	5.3 (15.17)
2	x_1	4.10 (1.41)	12.61 (32.68)	7.80 (10.32)	43.54 (95.61)	43.05 (51.47)
	x_2	4.74 (4.12)	28.01 (44.01)	10.41 (13.74)	56.97 (100.54)	46.75 (52.59)
	x_3	7.48 (8.53)	34.03 (46.43)	10.00 (12.70)	65.59 (99.59)	49.84 (57.13)
	x_4	12.69 (16.53)	37.35 (47.27)	9.89 (11.90)	68.47 (108.52)	52.63 (59.36)
3	x_1	4.81 (6.52)	5.76 (13.71)	—	32.78 (72.00)	—
	x_2	26.71 (3.93)	21.22 (3.28)	—	24.72 (3.46)	—
	x_3	75.08 (109.11)	86.17 (231.53)	—	93.64 (156.37)	—
	x_4	212.89 (314.29)	281.01 (726.58)	—	240.79 (387.36)	—
4	x_1	4.55 (7.14)	11.25 (26.05)	5.43 (4.93)	8.73 (11.91)	9.12 (11.58)
	x_2	1.55 (2.52)	0.55 (1.25)	1.01 (1.55)	0.84 (1.62)	0.96 (1.62)
	x_3	3.71 (4.98)	18.88 (34.66)	14.90 (19.60)	12.92 (12.75)	13.08 (13.01)
	x_4	1.06 (1.78)	0.60 (1.06)	1.78 (3.08)	1.03 (1.99)	1.11 (1.94)

the SSE of binary data is clearly smaller, and the overall accuracy of L-LiNGAM is superior. From Figs. 2, 3, 4 and 5, the tendency for variance of SSE of each method is different between binary or continuous. The variance of SSE under MCH tends to be small when the variables are binary. On the other hand, when the variables are

Table 2 Mean and standard deviation of PR

	L-LiNGAM	LiNGAM	HCM
Setting 1	13.24 (1.01)	12.00 (0.00)	12.20 (2.11)
Setting 2	11.64 (0.84)	9.80 (0.57)	12.20 (4.40)
Setting 3	10.40 (1.97)	9.80 (1.67)	–
Setting 4	12.44 (3.28)	11.60 (4.69)	11.80 (1.98)

binary, the variance SSE under LiNGAM tends to be large. In all of the analyzed cases, L-LiNGAM has the smallest sum of SSE for each variable in the entire causal structure.

From the results of SSE, ICA and PNL ICA showed a tendency to become local solutions due to ICA influence. Similarly, the ICA-based algorithm, which is one of the causal structure estimation algorithms of LiNGAM, is also found to be easily affected by local solutions. On the other hand, L-LiNGAM, based on the Direct LiNGAM Algorithm, always converges by repeating the regression and the independence evaluation by the number of observed variables. Thus, it is hardly affected by the local solution.

From Table 2, L-LiNGAM exhibits the best performance except in setting 2. In this case, where all variables are binary, HCM is the best result. However, HCM's standard deviation of PR is higher than that of other methods. Therefore, the structure estimation by L-LiNGAM is more stable than by HCM, in the scene of PR.

HCM has the best SSE when binary and continuous variables are mixed. However, the mean and standard deviation of the SSE and the PR of L-LiNGAM are generally better than those of HCM. From the above, it emerges that L-LiNGAM can prevent loss of discrete variable information with respect to discrete variables such as binary and five-valued. Although the SSE when continuous values are included becomes somewhat inferior to other methods, any variable can be reduced to the most stable SSE.

3.2 Real-world data application

Here, we report the setting and results when applying L-LiNGAM, LiNGAM, and HCM to real-world data. In this application example, the comparison is made as SEM. The real-world data are a dataset on low birth weights, birthwt, in the MASS package in R. The observed variables are shown in Table 3. Similar to the numerical example, the latent variables for L-LiNGAM were generated from t_{10} . We set logit function as link function for binary variables because logit function is used for generalized linear model.

Additionally, the evaluation as SEM is based on the Comparative Fit Index (CFI) (Bentler 1990), Tucker–Lewis Index (TLI) (Tucker and Lewis 1973), Root Mean Square Error of Approximation (RMSEA) (Steiger 1980), and standard root mean square residual (SRMR) (Hu and Bentler 1999, SRMR):

$$\text{CFI} = 1 - \frac{\max [(\chi^2_t - \text{df}_t), 0]}{\max [(\chi^2_t - \text{df}_t), (\chi^2_i - \text{df}_i), 0]}.$$

Table 3 Dataset ($n = 189$)

Variables	Details
low	Indicator of birth weight less than 2.5 kg (binary)
age	Mother's age in years
lwt	Mother's weight in pounds at last menstrual period (<i>lb</i>)
smoke	Smoking status during pregnancy (binary)
ptl	Number of previous premature labors
ht	History of hypertension (binary)
ui	Presence of uterine irritability (binary)
ftv	Number of physician visits during the first trimester

χ^2_i and χ^2_i are the χ^2 values of the model and independent model, respectively, and df_i , df_i is the degree of freedom in the independent model. CFI is an indicator that the proposed model is located between the independent and saturated models, and it is preferable it is closer to 1.00.

$$TLI = \frac{(\chi^2_i/df_i) - (\chi^2_i/df_i)}{(\chi^2_i/df_i) - 1}.$$

TLI shows how much the proposed model has improved the fitness of the data over the independent model. It is preferable that it is closer to 1.00.

$$RMSEA = \sqrt{\frac{\max[(\chi^2_i - df)/(n - 1), 0]}{df}}.$$

$df = \frac{p(p+1)}{2} - q$ represents the degree of freedom of the model and q is the number of parameters to estimate. RMSEA shows how much the dispersion covariance matrix in the population and the variance covariance matrix estimated by the proposed model diverge. It is desirable this value is below 0.05.

$$SRMR = \sqrt{\frac{2}{p(p+1)} \sum_{i=1}^p \sum_{j=1}^i ((s_{ij} - \hat{\sigma}_{ij})/s_{ii}s_{jj})^2}.$$

s_{ij} and $\hat{\sigma}_{ij}$ are the (i, j) elements of the variance-covariance matrix of the population estimated by the data covariance matrix and proposed model, respectively. SRMR shows the divergence of the variance covariance matrix of the data and the variance covariance matrix of the population estimated by the model. It is desirable to be below 0.05.

Figure 6 and Table 4 show the results of real-world data application. HCM (a) is the causal structure with the smallest BIC among all DAG models, and (b) is the causal structure with the smallest BIC among the DAG models using all observed variables.

Table 4 Index evaluation

	LiNGAM	L-LiNGAM	HCM (a)	HCM (b)
CFI	0.31	0.82	1.00	0.66
TLI	0.08	0.71	1.00	0.13
RMSEA	0.12	0.04	0.00	0.09
SRMR	0.19	0.11	0.00	0.06

Fig. 6 Estimated causal structures

From Fig. 6, due to smoking during the pregnancy and the number of consultations by the doctor in early pregnancy (ftv), there is a tendency to give birth low-birth weight infants in the causal structure identified by LiNGAM. On the other hand, in the causal structure of L-LiNGAM, all variables other than ftv were causes of low birth weight. Additionally, since the coefficients on the mother's high blood pressure (ht) and maternal uterine nerve sensitivity (ui) are negative and the coefficients on the other variables are positive, if mothers have high blood pressure and the presence or absence of uterine nerve hypersensitivity, other symptoms and causes may be difficult to occur simultaneously. In the causal structure identified by the HCM, all variables except for the presence/absence of low body weight (low) lead to low-birth-weight infants, and the mother's latest weight (lwt) and maternal hypertension (ht) were confirmed as causes for a low weight infant. However, since the values of the coefficients are small, these two variables can be ignored as causes. Additionally, it can be considered that there are few mothers who have experienced multiple premature births because the coefficient on ptl has a negative value.

From Table 4, HCM (a) has the best values for all evaluation indexes, but it shows only the relationship between smoking and low birth weight, so that it is impossible to examine its relationship with other variables. On the other hand, in the causal structure using all observed variables included in the dataset, the causal structure identified by L-LiNGAM has the highest fitness. Next is HCM (b), but the somewhat distinguished causal structure becomes more complex. The reason is considered to be the HCM estimation method: approximation accuracy is poor when observed variables have non-Gaussianity.

In addition de Bernabé et al. (2004), the age of the mother, smoking, premature birth experience, hypertension during pregnancy, and diseases related to the uterus were mentioned as risk factors. These risk factors are considered to be similar to the variables of age, smoke, ptl, ht, and ui included in our dataset. Additionally, risk factors such as physician's number of consultations (ftv) were not mentioned. In view of the above, it is desirable that variables other than the number of doctor's visits (ftv) are considered as predictors of low-birth weight infants than low-birth weight infants or not. Only the causal structures identified by L-LiNGAM and HCM(b) satisfy this, and L-LiNGAM has the highest evaluation, as SEM represents the causal structure of low birth weight and its risk factors.

We regard the variable set before low birth weight (low) as the cause group and the subsequent variable set as the result group. In the causal structure identified by L-LiNGAM, the number of doctor's visits (ftv), which was not listed as a risk factor, is not included as a cause and it is possible to consider the relationship among the variables in the cause group rather than the HCM(b).

4 Discussion

In this study, we proposed a method to estimate DAG using latent variables and link functions for quantitative mixed data. As a result of the numerical simulation, the square error of the predicted value can be kept small by assuming a DAG model transformed by the link function from the latent variable to the discrete value.

Furthermore, from the real-world data application example, the DAG via the latent variable using L-LiNGAM is superior to the conventional DAG model identification method of LiNGAM or HCM as SEM. Consequently, when discrete values are included in the data, assuming the latent variables forming LiNGAM exist as the background of the variables observed as discrete values is better for estimating causal structures. Even when all observed variables are continuous, since the loss of continuous information by L-LiNGAM is about the same as that of LiNGAM, it is also possible to identify the causal structure using L-LiNGAM. Especially when dealing with discrete data in the fields of psychology, medicine, or sociology, the DAG model estimated by the proposed method is more useful as a natural causal structure which considers the backgrounds of observed variables suitable for the data. When we regard the proposed method as one SEM, we consider multi-group L-LiNGAM as an extension from L-LiNGAM. When researchers are interested in the difference between indicator variables of a multi-group, such as gender, multi-group L-LiNGAM is required. On the other hand, although truly categorical variables, that is, not assuming continuous latent variables, such as gender and race, exist in datasets, the proposed method is applied to a dataset if the relationship between truly categorical variables and other variables is linear; that is, when we can select identity function as a link function for truly categorical variables, we assume no latent variables for truly categorical variables.

However, L-LiNGAM has the problem that the latent variable generation and link function transformation methods become arbitrary. When we change latent variable generation and link function, the result is different from another. In this paper, we set latent variable as t-distribution. The assumption of latent variables is non-Gaussian. Hence, we can also select various distributions for latent variables, including exponential distribution, skew t-distribution, and logistic distribution. If we use the link function of an exponential family, the calculation method of coefficients is the same as that of the generalized linear model. On the other hand, we should select an appropriate link function for the data. One manner for selection of latent variable distribution and link function is comparing the fitting of models such as SEM. As such, in future work, the simulation of various situations concerning the setting of latent variables and link functions is important.

We did not focus on theoretical results of DAG. For this reason, there are some unclear properties of the proposed method. In particular, the identification and consistency of L-LiNGAM are unclear, although the relationship between L-LiNGAM and the PNL Causal Model is clear. The identification and consistency of the PNL Causal Model are satisfied under general conditions. Hence, L-LiNGAM seems to exhibit identification and consistency properties, and this is suggested by our simulation results. We note that numerical examples contain some implicit assumptions. One of these is homoscedasticity of error terms. Another is that we select true link functions. We should carefully discuss identification and consistency under these implicit assumptions. Moreover, we should evaluate the effects of the number of objects using numerical simulations.

It is also useful to identify important variables included in the cause group within the causal structure by pruning directed edges in the graph through the sparse estimation method. Furthermore, since the proposed method may be susceptible to

errors when discriminating the causal structure in case of high-dimensional data, it is also important to perform a causal structure search at low dimensions first using the dimensionality reduction method. Similarly to Wiedermann et al. (2018), the estimation of causal structures considering measurement errors is also a useful extension.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Awokuse TO, Bessler DA (2003) Vector autoregressions, policy analysis, and directed acyclic graphs: an application to the U.S. economy. *J Appl Econ* 6(1):1–24
- Bach F, Jordan MI (2002) Kernel independent component analysis. *J Mach Learn Res* 3:1–48
- Bentler PM (1990) Comparative fit indexes in structural models. *Psychol Bull* 107(2):238–246. <https://doi.org/10.1037/0033-2909.107.2.238>
- Castañeda G, Guerrero OA (2018) The resilience of public policies in economic development. *Complexity* 2018:1–15. <https://doi.org/10.2139/ssrn.3221081>
- Castañeda G, Chávez-Juárez F, Guerrero OA (2018) How do governments determine policy priorities? Studying development strategies through spillover networks. *J Econ Behav Org* 154:335–361. <https://doi.org/10.1016/j.jebo.2018.07.017>
- de Bernabé JV, Soriano T, Albaladejo R, Juarranz M, Calle ME, Martínez D, Domínguez-Rojas V (2004) Risk factors for low birth weight: a review. *Eur J Obst Gynecol Reprod Biol* 116(1):3–15. <https://doi.org/10.1016/j.ejogrb.2004.03.007>
- Drasgow F (1986) Polychoric and polyserial correlations. In: Kotzm S, Johnson NL (eds) *Encyclopedia of statistical sciences, proceedings of machine learning research*, Wiley, New York, vol 7, pp 68–74
- Goeman JJ, Mansmann U (2008) Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* 24(4):537–544. <https://doi.org/10.1093/bioinformatics/btm628>
- Gretton A, Bousquet O, Smola A, Schölkopf B (2005) Measuring statistical dependence with hilbert-schmidt norms. In: *International conference on algorithmic learning theory*, Springer, New York, pp 63–77
- Haigh MS, Bessler DA (2004) Causality and price discovery: an application of directed acyclic graphs. *J Bus* 77(4):1099–1121. <https://doi.org/10.1086/422632>
- Hoyer PO, Shimizu S, Kerminen AJ, Palviainen M (2008) Estimation of causal effects using linear non-gaussian causal models with hidden variables. *Int J Approx Reason* 49(2):362–378. <https://doi.org/10.1016/j.ijar.2008.02.006>
- Hu LT, Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model* 6(1):1–55. <https://doi.org/10.1080/10705519909540118>
- Huber W, Carey VJ, Long L, Falcon S, Gentleman R (2007) Graphs in molecular biology. *BMC Bioinf* 8(6):S8. <https://doi.org/10.1186/1471-2105-8-s6-s8>
- Inazumi T, Washio T, Shimizu S, Suzuki J, Yamamoto A, Kawahara Y (2014) Causal discovery in a binary exclusive-or skew acyclic model: Bexsam. [arXiv:1401.5636](https://arxiv.org/abs/1401.5636)
- Li C, Shimizu S (2018) Combining linear non-gaussian acyclic model with logistic regression model for estimating causal structure from mixed continuous and discrete data. [arXiv:1802.05889](https://arxiv.org/abs/1802.05889)
- Molenaar D, Bolsinova M (2017) A heteroscedastic generalized linear model with a non-normal speed factor for responses and response times. *Br J Math Stat Psychol* 70:297–316. <https://doi.org/10.1111/bmsp.12087>
- Park G, Raskutti G (2015) Learning large-scale poisson dag models based on overdispersion scoring. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) *Advances in neural information processing systems*, vol 28, Curran Associates, Inc., pp 631–639

- Park G, Raskutti G (2017) Learning Quadratic Variance Function (QVF) DAG models via OverDispersion Scoring (ODS). [arXiv:170408783](https://arxiv.org/abs/1704.08783)
- Shimizu S (2014) Lingam: non-gaussian methods for estimating causal structures. *Behaviormetrika* 41(1):65–98
- Shimizu S, Bollen K (2014) Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *J Mach Learn Res* 15(1):2629–2652
- Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A (2006) A linear non-gaussian acyclic model for causal discovery. *J Mach Learn Res* 7:2003–2030
- Shimizu S, Inazumi T, Sogawa Y, Hyvärinen A, Kawahara Y, Washio T, Hoyer PO, Bollen K (2011) DirectLiNGAM: a direct method for learning a linear non-gaussian structural equation model. *J Mach Learn Res* 12:1225–1248
- Steiger JH (1980) Statistically based tests for the number of common factors. In: Paper presented at the annual meeting of the Psychometric Society Iowa City, IA
- Strehl A, Ghosh J (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* 3:583–617
- Takahashi Y, Ozaki K, Roberts BW, Ando J (2012) Can low behavioral activation system predict depressive mood? An application of non-normal structural equation modeling. *Jpn Psychol Res* 54(2):170–181. <https://doi.org/10.1111/j.1468-5884.2011.00492.x>
- Taleb A, Jutten C (1999) Source separation in post-nonlinear mixtures. *IEEE Trans Signal Process* 47(10):2807–2820. <https://doi.org/10.1109/78.790661>
- Tepeš B, Lešin G, Hrkač A, Tepeš K (2016) Causal bayes model of mathematical competence in kindergarten. *J Syst Cybern Inf* 14(3):14–17
- Tucker LR, Lewis C (1973) A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 38(1):1–10. <https://doi.org/10.1007/bf02291170>
- Wiedermann W, von Eye A (2017) Log-linear models to evaluate direction of effect in binary variables. *Stat Pap*. <https://doi.org/10.1007/s00362-017-0936-2>
- Wiedermann W, Merkle EC, von Eye A (2018) Direction of dependence in measurement error models. *Br J Math Stat Psychol* 71:117–145. <https://doi.org/10.1111/bmsp.12111>
- Xu L, Fan T, Wu X, Chen K, Guo X, Zhang J, Yao L (2014) A pooling-LiNGAM algorithm for effective connectivity analysis of fMRI data. *Front Comput Neurosci* 8:125. <https://doi.org/10.3389/fncom.2014.00125>
- Zhang K, Hyvärinen A (2009) On the identifiability of the post-nonlinear causal model. In: Bilmes JA, Ng AY (eds) Proceedings of the 25th conference on uncertainty in artificial intelligence, AUAI Press, Arlington, Virginia, United States, UAI '09, pp 647–655
- Zhang K, Hyvärinen A (2010) Distinguishing causes from effects using nonlinear acyclic causal models. In: Guyon I, Janzing D, Schölkopf B (eds) Proceedings of workshop on causality: objectives and assessment at NIPS 2008, PMLR, Whistler, Proceedings of machine learning research, Canada, vol 6, pp 157–164

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.