Joint estimation of linear non-Gaussian acyclic models

Shohei Shimizu
The Institute of Scientific and Industrial Research, Osaka University
Mihogaoka 8-1, Ibaraki, Osaka 567-0047, Japan

Abstract
A linear non-Gaussian structural equation model called LiNGAM is an identifiable model for exploratory causal analysis. Previous methods estimate a causal ordering of variables and their connection strengths based on a single dataset. However, in many application domains, data are obtained under different conditions, that is, multiple datasets are obtained rather than a single dataset. In this paper, we present a new method to jointly estimate multiple LiNGAMs under the assumption that the models share a causal ordering but may have different connection strengths and differently distributed variables. In simulations, the new method estimates the models more accurately than estimating them separately.

Keywords: structural equation models, non-Gaussianity, causal discovery, analysis of multiple groups, graphical models

1. Introduction
In causal discovery methods for continuous variables, it is typical [1] to assume i) the data generating process of observed variables \( x_i \) \((i = 1, \ldots, p)\) is graphically represented by a directed acyclic graph, that is, DAG; ii) the relations are linear; iii) there is no unobserved confounding variable. Let us denote by \( b_{ij} \) the connection strength from \( x_j \) to \( x_i \) and by \( k(i) \) a causal order of \( x_i \) in the DAG so that no later variable determines or has a directed path on any earlier variable. Without loss of generality, each observed variable \( x_i \)

---

Email address: sshimizu@ar.sanken.osaka-u.ac.jp (Shohei Shimizu)

Preprint submitted to Elsevier November 26, 2011
is assumed to have zero mean. Then we have

\[ x_i = \sum_{k(j) < k(i)} b_{ij} x_j + e_i, \]  

(1)

where \( e_i \) are external influences that are continuous variables with zero means and non-zero variances and are mutually independent. The independence assumption between \( e_i \) means that there is no latent confounding variable. In matrix form, the model (1) is written as

\[ \mathbf{x} = \mathbf{Bx} + \mathbf{e}, \]  

(2)

where the connection strength matrix or adjacency matrix \( \mathbf{B} \) collects \( b_{ij} \) and the vectors \( \mathbf{x} \) and \( \mathbf{e} \) collect \( x_i \) and \( e_i \), respectively. Note that the matrix \( \mathbf{B} \) can be permuted to be lower triangular with all zeros on the diagonal if simultaneous equal row and column permutations are made according to a causal ordering \( k(i) \) [2]. The problem of causal discovery is now to estimate the connection strength matrix \( \mathbf{B} \) based on data \( \mathbf{x} \) only. Each \( b_{ij} \) represents the direct causal effect of \( x_j \) on \( x_i \) and each \( a_{ij} \), the \((i, j)\)-th element of the matrix \( \mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1} \), the total causal effect of \( x_j \) on \( x_i \) [3]. In LiNGAM [4], external influences \( e_i \) are assumed to be non-Gaussian. This makes it possible to identify a causal ordering \( k(i) \) without using any background knowledge on the structure [4]. Once a causal ordering \( k(i) \) is identified, the connection strengths \( b_{ij} \) can be estimated by ordinary least squares regression [4]. Many works have been conducted to extend LiNGAM to more general cases [5, 6, 7] and apply LiNGAM and its extensions on real-world problems [8, 9, 10].

Previous methods for LiNGAM [4, 11] estimate a causal ordering \( k(i) \) based on a single dataset. However, in many applications, data are often obtained under different conditions. In neuroinformatics, fMRI (functional magnetic resonance imaging) signals are frequently recorded for multiple subjects [12, 13]. In bioinformatics, gene expression levels are measured under various experimental conditions [14]. That is, multiple datasets are obtained rather than a single dataset. Since such data are likely to have some common structure, it would be effective in estimation to exploit the similarities [15]. It is well known [16, 15] that it could cause serious bias in estimation to naively concatenate multiple datasets into a single dataset and analyze it. Many sophisticated methods for such situations have been proposed in other statistical models [17, 14, 18, 15, 19, 20, 16]. However, to our best knowledge, no such method is yet proposed in the literature of LiNGAM. Thus, in this
paper, we propose a method to jointly estimate such LiNGAMs that share a causal ordering but may have different connection strength matrices and differently distributed external influences.

2. Model definition

We now define our model extending LiNGAM [4], that is, the model (2) with non-Gaussian external influences, to multiple group cases. We assume the following data generating processes:

$$x_i^{(g)} = \sum_{k(j) < k(i)} b_{ij}^{(g)} x_j^{(g)} + e_i^{(g)} \quad (g = 1, \ldots, c),$$

where $g$ is the group index, $c$ is the number of groups, $k(i)$ is such a causal ordering of variables that they form a DAG in each group $g$, $x_i^{(g)}$, $e_i^{(g)}$ and $b_{ij}^{(g)}$ are the observed variables, external influences, and connection strengths of Group $g$, respectively. External influences $e_i^{(g)}$ are continuous non-Gaussian variables with zero means and non-zero variances and are mutually independent. In matrix form, the model (3) is written as

$$x^{(g)} = B^{(g)} x^{(g)} + e^{(g)} \quad (g = 1, \ldots, c).$$

The connection strength matrix $B^{(g)}$ collects $b_{ij}^{(g)}$ and the vectors $x^{(g)}$ and $e^{(g)}$ collect $x_i^{(g)}$ and $e_i^{(g)}$, respectively. The assumption that the models share a causal ordering $k(i)$ might be rather weak since many recent methods [15, 19, 18, 17, 21] for conventional graphical models also assume a common causal ordering across groups. Real-world examples of cases where a common ordering is shared across groups but the DAG structures are different are discussed in [14, 22].

3. A direct estimation method for multiple LiNGAMs

First, we briefly review an estimation method for a single group in Subsection 3.1 and extend it to multiple group cases in Subsection 3.2.

3.1. Background: A direct method for a single group

In [11], a direct estimation method called DirectLiNGAM was proposed. DirectLiNGAM estimates causal orders one by one and eventually a causal ordering of all the variables. An exogenous variables is a variable with no
parents, and the corresponding row of $B$ has all zeros. Therefore, an exogenous variable can be at the top of such a causal ordering that makes $B$ lower triangular with zeros on the diagonal. Then we remove the effect of the exogenous variable from the other variables by regressing it out. We iterate this procedure until all the variables are ordered. Now the point is how we can find an exogenous variable. The following lemma of [11] shows how it is possible:

**Lemma 1.** Assume that the input data $x$ strictly follows LiNGAM, that is, the model (2) with non-Gaussian external influences. This means that we assume that all the model assumptions are met and the sample size is infinite. Denote by $r^{(j)}_i$ the residual when $x_i$ is regressed on $x_j$: $r^{(j)}_i = x_i - \frac{\text{cov}(x_i, x_j)}{\text{var}(x_j)} x_j$ $(i \neq j)$. Then a variable $x_j$ is exogenous if and only if $x_j$ is independent of its residuals $r^{(j)}_i$ for all $i \neq j$. □

To evaluate independence between a variable $x_j$ and its residuals $r^{(j)}_i$ $(i \neq j)$, we first evaluate pairwise independence between the variable and each of the residuals using a kernel-based estimator of mutual information called KGV [23], which we denote by $\text{KGV}(x_i; r^{(j)}_i)$, and subsequently compute the sum of the pairwise independence measures over the residuals. The non-negative estimator KGV asymptotically goes to zero if and only if the variables are independent [23]. Thus we obtain the following statistic to evaluate independence between a variable $x_j$ and its residuals $r^{(j)}_i$:

$$T_{\text{kernel}}(x_j; U) = \sum_{i \in U, i \neq j} \text{KGV}(x_j; r^{(j)}_i),$$

where $U$ denotes the set of the subscripts of variables $x_i$, that is, $U = \{1, \cdots, p\}$. The next lemma connects this measure to Lemma 1:

**Lemma 2.** Assume that the input data $x$ strictly follows the model (2). Then, $x_j$ is exogenous if and only if the measure $T_{\text{kernel}}(x_j; U)$ in Eq. (5) is zero. □

**Proof.** (i) Assume that $x_j$ is exogenous. Applying Lemma 1, $x_j$ is independent of its residuals $r^{(j)}_i$ for all $i \neq j$. Then all the $\text{KGV}(x_j, r^{(j)}_i)$ $(i \neq j)$ are zeros since the KGV estimator is zero for independent variables, which implies that $T_{\text{kernel}}(x_j; U)$ is zero. (ii) Assume that $T_{\text{kernel}}(x_j; U)$ is zero. Then, all the $\text{KGV}(x_j, r^{(j)}_i)$ $(i \neq j)$ must be zeros since they are non-negative by
definition [23]. This implies that \( x_j \) is independent of its residuals \( r_i^{(j)} \) for all \( i \neq j \) because the KGV estimator is zero only for independent variables. Then, applying Lemma 1, \( x_j \) is exogenous. From (i) and (ii), the lemma is proven. ■

3.2. A direct algorithm for multiple groups

In this subsection, we present a method to jointly estimate multiple LiNGAMs constraining their causal orderings to be identical. This would enable more accurate estimation of the LiNGAMs in Eq. (4) than estimating them separately since they share a causal ordering.

First, we define a measure to evaluate independence between an observed variable \( x_j \) and its residuals in multiple-group cases as follows:

\[
T(j; U) = \sum_{g=1}^{c} T_{\text{kernel}}(x_j^{(g)}; U) = \sum_{g=1}^{c} \sum_{i \in U, i \neq j} KGV(x_j^{(g)}, r_i^{(j)}), \tag{6}
\]

where \( r_i^{(j)} \) is the residual when \( x_i^{(g)} \) is regressed on \( x_j^{(g)} \):

\[
r_i^{(j)} = x_i^{(g)} - \frac{\text{cov}(x_i^{(g)}, x_j^{(g)})}{\text{var}(x_j^{(g)})} x_j^{(g)} \quad (i \neq j). \tag{7}
\]

Next, we extend Lemma 2 to multiple-group cases:

**Lemma 3.** Assume that the input data \( x^{(g)} \) (\( g = 1, \ldots, c \)) strictly follow the model (4). Then a variable \( x_j^{(g)} \) is exogenous for any \( g \) if and only if the measure \( T(j; U) \) in Eq. (6) is zero. □

**Proof.** (i) Assume that \( x_j^{(g)} \) is exogenous for any \( g \). Applying Lemma 2 on the LiNGAM of each group \( g \), all the \( T_{\text{kernel}}(x_j^{(g)}; U) \) (\( g = 1, \ldots, c \)) are zeros. Therefore, \( T(j; U) \) is zero. (ii) Assume that \( T(j; U) \) is zero. Then, all the \( T_{\text{kernel}}(x_j^{(g)}, U) \) (\( g = 1, \ldots, c \)) must be zeros since they are non-negative by definition. Further, applying Lemma 2, \( x_j^{(g)} \) is exogenous for any \( g \). From (i) and (ii), the lemma is proven. ■

Thus, we can estimate a variable that is exogenous in every group by finding the variable subscript of such a variable that minimizes the independence measure in Eq. (6).
Then in each group we subtract the effect of the exogenous variable found from the other variables using least squares regression. Similarly to the single group case, we can find all the causal orders by iterating this.

This is a deflation approach and enables estimation of sub-orderings instead of the whole ordering. It would be a more practical goal to estimate the first $q$ ($q < p$ and $q < \min(n^{(1)}, \ldots, n^{(c)})$) orders in a causal ordering rather than all the orders when the sample size is not very large compared to the number of variables [10, 24] since the number of causal orders to be estimated gets smaller. Moreover, it is necessary to stop the algorithm before the covariance matrices of residuals get singular when any of the sample sizes is smaller than the number of variables.

We now present a direct algorithm to estimate a causal ordering and the connection strengths in the multiple LiNGAMs (4):

1. Given a set of $p$-dimensional random vectors $x^{(g)} (g = 1, \ldots, c)$, a set of its variable subscripts $U$ and $p \times n^{(g)}$ data matrices of the random vectors as $X^{(g)}$ respectively, initialize an ordered list of variables $K := \emptyset$ and $m := 1$ and choose the number of causal orders to be estimated $q (q = 1, \ldots, p)$.

2. Remove the mean from the data.

3. Repeat until $\min(q, p - 1)$ subscripts are appended to $K$:
   (a) For each group $g$, perform least squares regressions of $x^{(g)}_i$ on $x^{(g)}_j$ for all $i \in U \setminus K$ ($i \neq j$) and compute the residual vectors $r^{(g)}_{ij}$ and the residual data matrix $R^{(g)}_{(j)}$ from the data matrix $X^{(g)}$ for all $j \in U \setminus K$. Find the variable subscript $m$ of such a variable that is most independent of its residuals:

   $m = \arg \min_{j \in U \setminus K} T(j; U \setminus K), \quad (8)$

   where $T$ is the independence measure defined in Eq. (6).

   (b) Append $m$ to the end of $K$.

   (c) Let $X^{(g)} := R^{(g)}_{(m)}$ and replace $x^{(g)}$ by $r^{(g)}_{(m)}$, that is, $x^{(g)} := r^{(g)}_{(m)} (g = 1, \ldots, c)$.

4. Append the remaining variable to the end of $K$ if $q = p$.

5. For each group $g$, construct a $q \times q$ strictly lower triangular matrix $B^{(g)}_q$, that is, a sub-matrix of $B^{(g)}$ by following the order in $K$, and estimate
the connection strengths by using least squares regression on the original random vectors $x^{(g)}$ and the original data matrix $X^{(g)}$.

This essentially reduces to the original DirectLiNGAM [11] if the number of groups is one, that is, $c = 1$. Matlab codes are available at http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/code/jointDlingamcode.html.

The computational complexity of the proposed algorithm is $O(cnp^2qM^2 + cp^3qM^3)$ that is $c$ (the number of groups) times larger than that of DirectLiNGAM [11], where $n$ is the maximum sample size and $M$ is the maximal rank considered by the low-rank decomposition used to compute the kernel-based independence measure $KGV$ [23] in the group with the sample size $n$.

4. Simulations

As a sanity check of our method, we performed two experiments with simulated data. The first experiment consisted of 100 trials. In each trial, we generated 10 datasets with dimension $p = 10$ and sample sizes $n^{(1)} = \cdots = n^{(5)} = 50$ and $n^{(6)} = \cdots = n^{(10)} = 100$ and applied our joint estimation method on the data. For comparison, we also tested two methods; 1) a separate estimation method that applies DirectLiNGAM [11] on the datasets separately and 2) a naive method that creates a single dataset concatenating all the datasets and applies DirectLiNGAM on it. The sample sizes would not be very sufficient to reliably estimate the models separately [25]. Each dataset was created as follows:

1. We constructed the $p \times p$ connection strength matrix with all zeros and replaced every element in the lower-triangular part by independent realizations of Bernoulli random variables with success probability $s$ similarly to [26]. The probability $s$ determines the sparseness of the model. We set the sparseness $s$ so that the expected number of adjacent variables was half of the dimension.

2. We replaced each non-zero entry in the connection strength matrix by a value randomly chosen from the interval $[-1.5, -0.5] \cup [0.5, 1.5]$ and selected variances of the external influences from the interval $[1, 3]$. The resulting matrix was used as the data-generating connection strength matrix $B^{(g)}$. 


3. We generated data with sample size $n^{(g)}$ by independently drawing the external influence variables $e^{(g)}_i$ from various 18 non-Gaussian distributions used in [23] including super- and sub-Gaussian distributions and symmetric and asymmetric distributions [11]. Then, we generated constants following the Gaussian distribution $N(0, 4)$ and added them to the external influence variables as their means.

4. The values of the observed variables $x^{(g)}_i$ were generated using the connection strength matrix $B^{(g)}$ and external influences $e^{(g)}_i$. Finally, we permuted the variables according to a random ordering shared by all the groups.

We computed the percentages of datasets where all the causal orders were correctly learned in the total datasets, that is, 1,000 datasets (= 10 datasets × 100 trials). The percentages for the joint estimation method, the separate estimation method and the naive method were 93.1%, 44.9% and 0.4%, respectively. We also computed the percentages of actually absent direct edges in estimated absent directed edges, that is, false negatives. The percentages of false negatives for the three methods were 0.16%, 2.41% and 19.4%, respectively. Further, we computed the average squared errors between true connection strengths and estimated ones. The average squared errors for the three methods were 0.03, 0.09 and 0.53, respectively.

In the second experiment, we considered a situation with more variables than observations, that is, $p > n^{(g)}$. We generated 10 datasets with dimension $p = 40$ and sample size $n^{(1)} = \cdots = n^{(5)} = 10$ and $n^{(6)} = \cdots = n^{(10)} = 20$ in the same manner as above and only estimated the first five variables in a causal ordering ($q = 5$). We repeated this procedure 100 times. The percentages of datasets for which the joint, separate and naive methods correctly estimated such first five variables were 90.4%, 43.1% and 44.4%, respectively. The percentages of false negatives for the three methods were 0.05%, 0.66% and 0.44%, respectively. The average squared errors for the three methods were 0.12, 0.31 and 0.35, respectively.

To sum up, our joint estimation method worked much better than the other two methods in the simulations as summarized in Fig. 1.

5. Conclusions

We proposed a new joint estimation method for multiple LiNGAMs. The new method assumes that such LiNGAMs share a causal ordering, but the
connection strengths of variables and the distributions of external influences may be different between the models. Future works would include i) development of statistical tests to detect violations of the model assumptions including the common-ordering assumption, e.g., test of independence between observed variables and their residuals using a non-parametric method [27]; ii) evaluation of statistical reliability of estimation results using bootstrapping as in [28] and iii) empirical comparison of our method and related algorithms on artificial and real-world data sets.

Acknowledgements

We would like to thank Yoshinobu Kawahara, Takashi Washio, Satoshi Hara and Aapo Hyvärinen for interesting discussions and the anonymous reviewers for helpful comments. This work was supported by MEXT Grant-in-Aid for Young Scientists #21700302.

References


