Discovering Linear Non-Gaussian Models for All Categories of Missing Data (Student abstract)

Matteo Ceriscioli[1, 2], **Shohei Shimizu**[2, 3, 4], **Karthika Mohan**[1]
[1]Oregon State University, Corvallis, USA
[2]RIKEN, Tokyo, Japan
[3]The University of Osaka, Osaka, Japan
[4]Shiga University, Hikone, Japan
{ceriscim, karthika.mohan}@oregonstate.edu, shohei-shimizu@ds.sanken.osaka-u.ac.jp

Abstract

Causal discovery is the task of learning causal models, encoding causal relationships, from a source of information, such as a dataset containing observational data. While many algorithms have been developed to discover causal models under varied sets of assumptions, the case in which the dataset is affected by missing data remains significantly underexplored. Naively applying standard causal discovery algorithms to listwise, test-wise, or regression-wise deleted datasets, or imputing the missing data, can introduce spurious associations between variables and bias function estimation in functional causal models. This issue arises when the data is missing at random or not at random. It ultimately invalidates the theoretical guarantees of these algorithms and prevents finding the true underlying causal model, even in the large-sample limit. An established family of causal models is the Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu et al. 2006), which assumes linear functional relationships and non-Gaussian independent noise terms. We propose a new causal discovery algorithm for the LiNGAM model, capable of identifying the true underlying causal structure and providing unbiased estimates of the model's parameters, even when the data is affected by MNAR missingness.

Preliminaries

Representing missingness. Missingess can be represented using graphical models called missingness graph (m-graph) (Mohan, Pearl, and Tian 2013). An m-graph is a causal Directed Acyclic Graph (DAG) $G = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ corresponds to the set of variables and can be partitioned as follows: $\mathbf{V} = \mathbf{V}_o \cup \mathbf{V}_m \cup \mathbf{U} \cup \mathbf{V}^* \cup \mathbf{R}$ where $\mathbf{V}_o$ contains all variables that are fully observed, and $\mathbf{V}_m$ contains partially observed variables, i.e. variables affected by missing data. For each variable $X \in \mathbf{V}_m$ there exists a corresponding binary variable called missingness mechanism $R_X \in \mathbf{R}$ and a proxy variable $X^* \in \mathbf{V}^*$ such that $X^* := X$ if $R_X = 0$ and $X^* := m$ if $R_X = 1$, where m is a special value indicating missingness. $\mathbf{U}$ is the set of latent variables. The columns in a dataset generated by G correspond to variables in $\mathbf{V}_o$ and $\mathbf{V}^*$, and the values of $\mathbf{R}$ follow from those of $\mathbf{V}^*$. Variables in $\mathbf{V}_m$ and $\mathbf{U}$ are unobserved. The type of missingness can be read from the independence relations encoded in the m-graph: it is Missing Completely At Random (MCAR) if $\mathbf{V}_m \cup \mathbf{V}_o \cup \mathbf{U} \perp\!\!\!\perp \mathbf{R}$, Missing At Random (MAR) if $\mathbf{V}_m \cup \mathbf{U} \perp\!\!\!\perp \mathbf{R} \mid \mathbf{V}_o$, and Missing Not At Random (MNAR) otherwise (Mohan and Pearl 2021). Figure 1 shows an example of an m-graph with MAR missingness.

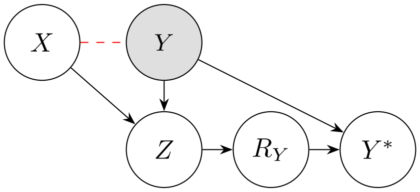

Figure 1: An m-graph entailing MAR missingness. $\mathbf{V}_o = \{X, Z\}, \mathbf{V}_m = \{Y\}, \mathbf{U} = \emptyset, \mathbf{R} = \{R_Y\}, \mathbf{V}^* = \{Y^*\}$. The red dashed edge is not part of the m-graph, it highlights that naive handling of the missingness problem (e.g. listwise deletion: $R_Y = 0$) induces collider bias between X and Y, since R_Y is a child of the collider Z.

LiNGAM and LiM. Given continuous random variables $\{X_i\}_{i=1}^n$, a causal order $i \in \{1, \ldots, n\} \mapsto k(i)$, and independent non-Gaussian error terms $\{E_i\}_{i=1}^n$, the LiNGAM model (Shimizu et al. 2006) assumes the data is generated according to the following assignments:

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i, \qquad e_i \sim \text{Non-Gaussian}(\cdot) \quad (1)$$

where $B = \{b_{ij}\}_{i,j=1}^n$ is the weight matrix, corresponding to the adjacency matrix of the causal DAG.

The Linear Mixed model (LiM) (Zeng et al. 2022) is an extension of LiNGAM that handles both continuos and binary data. In this setting the set of variables X can be partitioned in $X = X_{con} \cup X_{dis}$ where X_{con} is the set of continuous variables and X_{dis} is the set of discrete (binary) variables. Given a causal order $i \in \{1, \ldots, n\} \mapsto k(i)$, X_{con} follow the same assignments as in Equation 1 while X_{dis} follow:

$$x_i = \mathbf{1}[\![\sum_{k(j)<k(i)} b_{ij} x_j + e_i > 0]\!], \; e_i \sim \text{Logistic}(0, 1) \quad (2)$$

Here, $1[\![\cdot]\!]$ denotes the Iverson bracket, which evaluates to 1 if the condition inside holds and 0 otherwise. Both models are identifiable from observational data.

Missingness-LiNGAM

We denote the set of parents of X as Pa_X and the set of ancestors of X as Anc_X.

Definition 1. Given an m-graph $G = (\mathbf{V}, \mathbf{E})$, a *Missingness-LiNGAM (m-LiNGAM)* is a causal model over $\mathbf{V} = \mathbf{V}_o \cup \mathbf{V}_m \cup \mathbf{U} \cup \mathbf{V}^* \cup \mathbf{R}$ where the variables in the induced subgraph $G[\mathbf{V}_o \cup \mathbf{V}_m]$ follow a LiNGAM.

Note that the definition above implies that there is no latent confounding involving $\mathbf{V}_o \cup \mathbf{V}_m$, and that missingness mechanisms cannot be ancestors of of these variables.
We also make the following assumptions:
A1. No causal interactions between missingness mechanisms, i.e. $\forall R_i, R_j \in \mathbf{R}.\ R_i \notin Pa(R_j)$.
A2. No direct self-masking ($\forall X \in \mathbf{V}_m,\ X \notin Pa(R_X)$).
Even if we do not allow direct self-masking, it is possible for a variable $X \in \mathbf{V}_m$ to indirectly cause R_X, i.e. X can be in Anc_{R_X}.
A3. The missingness mechanisms follow Equation 2.
This implies that $G[\mathbf{V}_o \cup \mathbf{V}_m \cup \mathbf{R}]$ forms a LiM. Since every missingness mechanism $R_i \in \mathbf{R}$ corresponds to a partially observed variable, it follows that $P(R_i = 1) > 0$, and since, according to A3, each R_i is generated using Equation 2, as the support of the logistic distribution is $\mathbb{R}$, it also follows that $P(R_i = 1) < 1$.
Consider the following identifiability result:

Theorem 1. *Let X be a $p \times n$ matrix of observational data over $\mathbf{V}_o \cup \mathbf{V}^*$ generated from an m-LiNGAM with graph $G = (\mathbf{V}, \mathbf{E})$ where $p = |\mathbf{V}_o \cup \mathbf{V}^*|$ and n is the sample size. If Assumptions A1-3 are satisfied, in the large-sample limit the partition $\mathbf{V} = \mathbf{V}_o \cup \mathbf{V}_m \cup \mathbf{U} \cup \mathbf{V}^* \cup \mathbf{R}$, the structure of the causal graph $\mathbf{E}$, and the parameters B of the LiM on $G[\mathbf{V}_o \cup \mathbf{V}_m \cup \mathbf{R}]$ can be identified from X.*

Proof. (Sketch of proof) In the large-sample limit, the columns of X can be mapped to the sets in the partition $\mathbf{V}$: columns corresponding to fully observed variables belong to $\mathbf{V}_o$, columns with missing data correspond to $\mathbf{V}^*$, and for each proxy $V_i^* \in \mathbf{V}^*$ there is a partially observed variable $V_i \in \mathbf{V}_m$ and a missingness mechanism $R_i \in \mathbf{R}$. We can completely reconstruct the values of $\mathbf{R}$ for every observation because $\forall R_i \in \mathbf{R}$ if the corresponding $V_i^* \in \mathbf{V}^*$ takes value m then $R_i = 1$ and $R_i = 0$ otherwise.
Let $Pa^o_{R_i} := Pa_{R_i} \cap \mathbf{V}_o$ and $Pa^m_{R_i} := Pa_{R_i} \cap \mathbf{V}_m$. Then, for each R_i we can estimate $P(R_i \mid \mathbf{V}_o, \mathbf{V}_m)$ since by A1 it is equal to $P(R_i = r_i \mid Pa^o_{R_i}, Pa^m_{R_i}, R_{pa^m_{R_i}} = 0)$. Let $V_i \in \mathbf{V}_m$ be the partially observable node corresponding to R_i, by A2 $V_i \notin Pa^m_{R_i} \cup Pa^o_{R_i}$, where $Pa^m_{R_i} \subseteq \mathbf{V}_m$, $Pa^o_{R_i} \subseteq \mathbf{V}_o$ are the parents of R_i. Using A3, in the large-sample limit we are guaranteed to identify the parent set for each R_i and also have samples to estimate $P(R_i = r_i \mid Pa^o_{R_i}, Pa^m_{R_i}, R_{pa^m_{R_i}} = 0)$ for both $r_i = 0$ and $r_i = 1$.
Next, given A1, A2, and $U = \emptyset$ it is possible to apply Theorem 2 in (Mohan, Pearl, and Tian 2013) and thus rewrite the joint distribution of $\mathbf{V}_o, \mathbf{V}_m$ as follows:

$$P(\mathbf{V}_o, \mathbf{V}_m) = \frac{P(\mathbf{V}_o, \mathbf{V}^*, \mathbf{R} = 0)}{\prod_i P(R_i = 0 \mid Pa^o_{R_i}, Pa^m_{R_i}, R_{Pa^m_{R_i}} = 0)} \tag{3}$$

Since the terms on the r.h.s. are known, we can reconstruct the joint distribution of $\mathbf{V}_o, \mathbf{V}_m$. From LiNGAM identifiability results (Shimizu et al. 2011), we can then reconstruct the subgraph $G[\mathbf{V}_o, \mathbf{V}_m]$ and estimate the parameters.
As we know the parent set of each $R_i \in \mathbf{R}$ and the exact functional relation of each $V_i^* \in \mathbf{V}^*$, we can reconstruct the remaining part of the graph. □

Table 1: Comparison of m-LiNGAM and DirectLiNGAM using SHD across different sample sizes (n).

n	LD-LiNGAM	RD-LiNGAM	m-LiNGAM
50	4.30±1.92	3.52±2.02	**3.13**±**1.93**
100	2.26±1.95	1.61±1.48	**1.58**±**1.49**
300	1.28±1.27	1.35±1.53	**0.95**±**1.45**
500	1.55±1.49	1.18±1.20	**0.69**±**0.99**
1000	1.63±1.23	1.40±1.50	**0.79**±**1.20**
3000	2.37±1.43	2.01±1.72	**0.79**±**1.10**
5000	2.69±1.46	2.40±1.82	**0.59**±**1.02**

Experiments

Baseline. We compare the performance of m-LiNGAM with DirectLiNGAM (Shimizu et al. 2011) on 100 Erdős–Rényi graphs with five variables and a sparsity parameter of 0.3. Two variables in each graph are affected by missingness, with their missingness categories assigned randomly, resulting on average in 30% MNAR, 46% MAR, and 24% MCAR. Since DirectLiNGAM can only be applied to datasets without missing values, we consider two variants: one using listwise deletion and another using regression-wise deletion, in which we remove only the rows that have missing values for the variables involved in each regression.

Metric. To evaluate the algorithms, we measure the Structural Hamming Distance (SHD) between the true and estimated adjacency matrices. SHD quantifies structural discrepancies between two graphs as the total number of missing, extra, and reversed edges. A smaller SHD indicates that the estimated graph more closely matches the true one.

Results. The results (Table 1) indicate that ignoring missingness leads to biased estimates, while m-LiNGAM achieves higher accuracy even with small sample sizes. Notably, as the sample size increases, the error for DirectLiNGAM grows because the bias induced by missingness becomes more pronounced and detectable in the observed data.

Conclusions

We introduced missingness-LiNGAM, a family of causal models for settings with missing data based on LiNGAM. We formulated assumptions enabling identification of the causal graph and unbiased parameter estimation from observational data under MCAR, MAR, or MNAR missingness. This work provides a foundation for future extensions to nonlinear models and for enabling more robust causal discovery in real-world datasets with missing values.

References

Mohan, K.; and Pearl, J. 2021. Graphical Models for Processing Missing Data. *Journal of the American Statistical Association*, 116(534): 1023–1037.

Mohan, K.; Pearl, J.; and Tian, J. 2013. Graphical Models for Inference with Missing Data. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and 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, P. O.; and Bollen, K. 2011. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *J. Mach. Learn. Res.*, 12: 1225–1248.

Zeng, Y.; Shimizu, S.; Matsui, H.; and Sun, F. 2022. Causal Discovery for Linear Mixed Data. In Schölkopf, B.; Uhler, C.; and Zhang, K., eds., *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, 994–1009. PMLR.