Causal Discovery with Multi-Domain LiNGAM for Latent Factors*

Yan Zeng YANAZENG013@GMAIL.COM
Guangdong University of Technology

Shohei Shimizu SHOHEI-SHIMIZU@BIWAKO.SHIGA-U.AC.JP
Shiga University; RIKEN

Ruichu Cai CAIRUICHU@GMAIL.COM
Guangdong University of Technology

Feng Xie XIEFENG009@GMAIL.COM
Peking University

Michio Yamamoto M.YAMAMOTO@OKAYAMA-U.AC.JP
Okayama University; RIKEN

Zhifeng Hao ZFHAO@GDUT.EDU.CN
Foshan University; Guangdong University of Technology

Editor: Sisi Ma, Erich Kummerfeld

Abstract

Discovering causal structures among latent factors from observed data is particularly significant yet challenging problem. Despite some efforts for this problem, existing methods focus on the single-domain data only. In this paper, we propose Multi-Domain Linear Non-Gaussian Acyclic Models for LAtent Factors (MD-LiNA), where the causal structures from different domains may be different, but they have a shared causal structure among latent factors of interest. The model enriches the causal representation for multi-domain data. We propose an integrated two-phase algorithm to estimate the model. In particular, we first locate the latent factors and estimate the factor loading matrix. Then to uncover the shared causal structure among latent factors of interest, we derive a score function based on the characterization of independence relations between external influences and the dependence relations between multi-domain latent factors and latent factors of interest. Experimental results on synthetic data demonstrate the efficacy of our approach.

Keywords: Causal Discovery; LiNGAM; Multi-Domain Data; Latent Factors.

1. Introduction

Causal relations usually lie between latent variables (factors) that cannot be directly measured in many real-world applications, e.g., anxiety, depression, or coping, etc (Silva et al., 2006). The methods to discover the causal structures among latent factors roughly fall into two categories, namely covariance-based methods and non-Gaussianity-based ones. Covariance-based methods employ the covariance structure of data alone, e.g., BuildPureClusters algorithm (Silva et al., 2006), or FindOneFactorClusters algorithm (Kummerfeld and

* This work will appear in the 30th International Joint Conference on Artificial Intelligence (IJCAI-21).

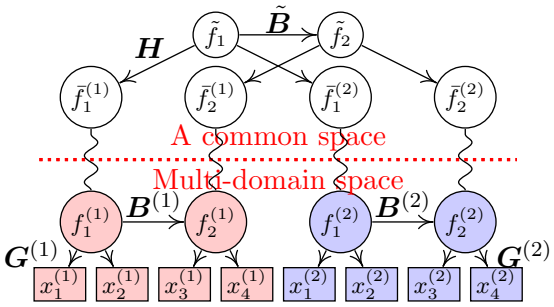


Figure 1: An MD-LiNA model. Variables in the same color (light red and light blue) are in the same domain. Observed variables $\boldsymbol{x}^{(m)}$ in domain m entail its latent factors $\boldsymbol{f}^{(m)}$. $\bar{\boldsymbol{f}}$ are augmented latent factors and $\tilde{\boldsymbol{f}}$ are latent factors of interest, whose structure $\tilde{\boldsymbol{B}}$ is shared by $\boldsymbol{f}$ from different domains.

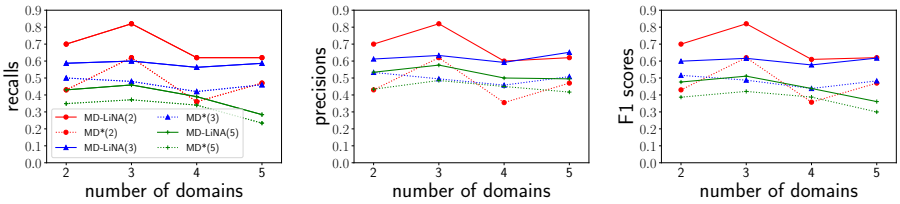
Ramsey, 2016), etc. However, they can only output structures up to the Markov equivalence class for latent factors. Non-Gaussianity-based methods address this indistinguishable problem by the non-Gaussianity of data. Specifically, Shimizu et al. (2009) leveraged non-Gaussianity and firstly achieved identifying a unique causal structure between latent factors based on the LiNGAM (Shimizu et al., 2006). Cai et al. (2019) designed the so-called Triad constraints and proposed a two-phase method to learn the structure among latent factors.

However, the above methods all focus on the single-domain data which are originated from the same domain. In many scenarios, data are often collected under distinct conditions, resulting in distinct distributions and/or various causal effects. Thus, in this paper, we propose Multi-Domain Linear Non-Gaussian Acyclic Models for LAtent Factors (MD-LiNA) to represent the causal mechanism of latent factors, which tackles not only single-domain data but multi-domain ones. We propose an integrated two-phase approach to uniquely identify the underlying causal structure among latent factors of interest.

2. Methodology

We propose MD-LiNA in Fig. 1. It assumes (1) $\boldsymbol{f}^{(m)}$ are generated linearly from a Directed Acyclic Graph (DAG) with non-Gaussian distributed external variables $\boldsymbol{\varepsilon}^{(m)}$; (2) $\boldsymbol{x}^{(m)}$ are generated linearly from $\boldsymbol{f}^{(m)}$ plus Gaussian distributed errors $\boldsymbol{e}^{(m)}$; (3) Each f_i has at least 2 pure measurement variables [1]; (4) Each $\bar{f}_i^{(m)}$ is linearly generated by only one latent in $\tilde{\boldsymbol{f}}$ and each $\tilde{f}_i$ generates at least one latent in $\bar{\boldsymbol{f}}$. To estimate MD-LiNA, firstly, we locate the latent factors using Triad (Cai et al., 2019) [2]. Then we estimate the factor loading matrix $\boldsymbol{G}^{(m)}$ by the factor analysis (Reilly and O'Brien, 1996). Secondly, to estimate $\tilde{\boldsymbol{B}}$ and $\boldsymbol{H}$, we derive a score function to characterize the independence relations between external variables, and unify it to characterize the dependence relations between latent factors from different domains and latent factors of interest. Then such function is enforced with acyclicity constraints, with which our task is formulated as a purely continuous optimization problem. Finally, we update $\tilde{\boldsymbol{B}}$ according to the estimated $\boldsymbol{H}$.

1. Pure measurement variables are those which have only one single latent factor parent (Silva et al., 2006).
2. Triad constraints are based on the independence with "pseudo residuals", which help locate the latent factors. For details, please see Cai et al. (2019).



(a) Re. with Mul. (b) Pre. with Mul. (c) F1. with Mul.

Figure 2: The recall (Re.), precision (Pre.) and F1 scores (F1.) of the recovered causal graphs between latent factors with different numbers of domains (Mul.).

3. Results and Conclusions

We generated the data by Fig. 1, with different numbers of domains ($M = 2, 3, 4, 5$). In each domain, we used the identical number of latent factors, and the identical causal graphs of latent factors. We varied the number of latent factors in each domain, i.e., $q_m = 2, 3, 5$. Since no methods focus on multi-domain data in latent factor models, we used our method which did not conduct the last update step as the comparison (MD*). As is shown in Fig. 2, overall we found F1 scores of both methods tend to decrease as the number of domains or the number of latent factors increases. Specifically, in all cases MD-LiNA gives a better performance compared with MD*, in that MD* did neglect the problem that factors from different domains are represented by which factors of interest.

We proposed Multi-Domain Linear Non-Gaussian Acyclic Models for LAtent Factors (MD-LiNA), which gave deeper interpretation for latent factors that count. To discover the underlying shared causal structure for latent factors of interest, we proposed an integrated two-phase method for estimation.

References

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. In *NeurIPS*, pages 12863–12872, 2019.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *KDD*, pages 1655–1664, 2016.

Terence Reilly and Robert M. O'Brien. Identification of confirmatory factor analysis models of arbitrary complexity: The side-by-side rule. *Sociological Methods & Research*, 24(4): 473–491, 1996.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (10):2003–2030, 2006.

Shohei Shimizu, Patrik O Hoyer, and Aapo Hyvärinen. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027, 2009.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(2):191–246, 2006.