Causal Discovery with Multi-Domain LiNGAM for Latent Factors

Yan Zeng[1,2] , **Shohei Shimizu**[2,3] , **Ruichu Cai**[1] , **Feng Xie**[4] ,
Michio Yamamoto[2,5] and **Zhifeng Hao**[1,6]
[1]Guangdong University of Technology
[2]RIKEN
[3]Shiga University
[4]Peking University
[5]Okayama University
[6]Foshan University
yanazeng013@gmail.com, shohei-shimizu@biwako.shiga-u.ac.jp, {cairuichu, xiefeng009}@gmail.com,
m.yamamoto@okayama-u.ac.jp, zfhao@gdut.edu.cn

Abstract

Discovering causal structures among latent factors from observed data is a particularly 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 structure among latent factors of interest is shared for all domains, and we provide its identification results. 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 causal structure among shared 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. We show that the proposed method provides locally consistent estimators. Experimental results on both synthetic and real-world data demonstrate the efficacy and robustness of our approach.

1 Introduction

Learning causal relationships from observed data, termed as causal discovery, has been developed rapidly over the past decades [Pearl, 2009; Spirtes and Zhang, 2016; Peters *et al.*, 2017]. In many scenarios, including sociology, psychology, and educational research, the underlying causal relations are usually embedded between latent variables (or factors) that cannot be directly measured, e.g., anxiety, depression, or coping, etc [Silva *et al.*, 2006; Bartholomew *et al.*, 2008], in which scientists are often interested.

Some approaches have been developed to identify the causal structure among latent factors, which can be categorized into covariance-based 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 Ramsey, 2016], to ascertain how many latent factors as well as the structure of latent factors. However, these algorithms can only output structures up to the Markov equivalence class for latent factors. Non-Gaussianity-based methods address this indistinguishable identification problem by taking the best of 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 Linear, Non-Gaussian, Acyclic Models (LiNGAM) [Shimizu *et al.*, 2006]. They transformed the problem into the Noisy Independent Component Analysis (NICA). Recently, to avoid the local optima of the NICA, Cai *et al.* [2019] designed the so-called Triad constraints and Xie *et al.* [2020] developed the GIN condition. They both proposed a two-phase method to learn the structure or causal orderings among latent factors.

It is noteworthy that the above-mentioned methods all focus on the data which are originated from the same domain, i.e., single-domain data. However, in many real-world applications, data are often collected under distinct conditions. They may be originated from different domains, resulting in distinct distributions and/or various causal effects. For instance, in neuroinformatics, functional Magnetic Resonance Imaging (fMRI) signals are frequently extracted from multiple subjects or over time [Smith *et al.*, 2011]; in biology, a particular disease is measured by distinct medical equipment [Dhir and Lee, 2020], etc. Existing methods to handle multi-domain data in causal discovery are flourishing, e.g., Danks *et al.* [2009], Tillman and Spirtes [2011], Ghassami *et al.* [2018], Kocaoglu *et al.* [2019], Dhir and Lee [2020], Huang *et al.* [2020], Jaber *et al.* [2020], etc. Though there are some methods to handle multi-domain data that allow the existence of latent variables or confounders, no such method is yet proposed in the literature when learning the causal structure among latent factors with only observed data, to our best knowledge. Thus, it is desirable to perform causal discovery from multi-domain data to uncover the structure among latent factors.

When considering multi-domain instead of single-domain data in latent factor models, there may exist different causal structures among latent factors or with different causal effects in different domains. An important question then naturally raises, i.e., how to guarantee that factors in different domains are represented by the same factors of interest so that the underlying structure among latent factors of interest is uncovered. A solution may be to naively concatenate the multi-domain observed data, such that the multi-domain model can be regarded as a single-domain latent factor model. However, it may cause serious bias in estimating the causal structure among latent factors of interest [Shimizu, 2012]. 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. In addition, we propose an integrated two-phase approach to uniquely identify the underlying causal structure among latent factors (of interest). In particular, in the first phase, we locate the latent factors and estimate the factor loading matrix (relating the observed variables to its latent factors) for all domains, leveraging the ideas from Triad constraints and factor analysis [Cai *et al.*, 2019; Reilly and O'Brien, 1996]. In the second phase, we derive a score function to characterize the independence relations between external variables, and interestingly, we unify this function to characterize the dependence relations between latent factors from different domains and latent factors of interest. Then such unified function is enforced with acyclicity, sparsity, and elastic net constraints, with which our task is formulated as a purely continuous optimization problem. The method is (locally) consistent to produce feasible solutions.

Our contributions are mainly three-folded:

- This is the first effort to study the causal discovery problem of identifying causal structures between latent factors for multi-domain observed data.
- We propose an MD-LiNA model, which is a causal representation for both single and multi-domain data in latent factor models and offers a deep interpretation of dependencies between observed variables across domains, and show its identifiability results.
- We propose an integrated two-phase approach to uniquely estimate the underlying causal structure among latent factors, which simultaneously identifies causal directions and effects. It is capable of handling cases when the sample size is small or the latent factors are highly correlated. And the local consistency is also provided.

2 Problem Formalization

Suppose we have data from M domains. Let $\boldsymbol{x}^{(m)}$ and $\boldsymbol{f}^{(m)}$ $(m = 1, ..., M)$ be the random vectors that collect p_m observed variables and q_m latent factors in domain m, respectively. The total number of observed variables for all domains is $p = \sum_{m=1}^{M} p_m$ while that of latent factors is $q = \sum_{m=1}^{M} q_m$. In this study, we focus on linear models,

$$\begin{aligned} \boldsymbol{f}^{(m)} &= \boldsymbol{B}^{(m)} \boldsymbol{f}^{(m)} + \boldsymbol{\varepsilon}^{(m)}, \\ \boldsymbol{x}^{(m)} &= \boldsymbol{G}^{(m)} \boldsymbol{f}^{(m)} + \boldsymbol{e}^{(m)}, \end{aligned} \tag{1}$$

where $\boldsymbol{\varepsilon}^{(m)}$, and $\boldsymbol{e}^{(m)}$ are random vectors that collect external influences, and errors, respectively, and they are independent with each other. $\boldsymbol{B}^{(m)}$ is a matrix that collects causal effects $b_{ij}^{(m)}$ between $f_i^{(m)}$ and $f_j^{(m)}$, while $\boldsymbol{G}^{(m)}$ collects factor loadings $g_{ij}^{(m)}$ between $f_j^{(m)}$ and $x_i^{(m)}$. $f_i^{(m)}$ is assumed to have zero means and unit variances. Note that in a specific domain m, data are generated with the same $b_{ij}^{(m)}$ and $g_{ij}^{(m)}$. $\varepsilon_i^{(m)}$ and $e_i^{(m)}$ are sampled independently from the identical distributions. In different domains, $\boldsymbol{B}^{(m_1)}$ $(\boldsymbol{G}^{(m_1)})$ and $\boldsymbol{B}^{(m_2)}$ $(\boldsymbol{G}^{(m_2)})$ may be different, but they have a shared causal structure.

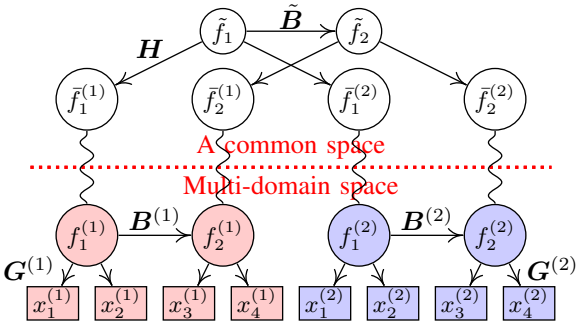


Figure 1: An MD-LiNA model. Variables in the same color (light red and light blue) are in the same domain. Observed variables $\boldsymbol{x}^{(m)}$ in domain m entail its latent factors $\boldsymbol{f}^{(m)}$. Augmented latent factors $\bar{\boldsymbol{f}}$ are obtained using the coding representation method, which are signified by the curved waved lines[1]. $\tilde{\boldsymbol{f}}$ are shared latent factors of interest, whose structure is shared by $\boldsymbol{f}$ from different domains.

Thereafter, we encounter two problems: how to integrate the multi-domain data effectively ; and how to guarantee factors $\boldsymbol{f}$ in different domains are represented by the same concepts (factors) of interest, with which we identify the underlying causal structure among latent factors of interest. To address the first problem, we leverage an idea of simple coding representation [Shimodaira, 2016], i.e., an observation is represented as an augmented one with p dimensions, where only p_m dimensions come from the original domain while the other $(p - p_m)$ dimensions are padded with zeros. Any augmented data are expressed with a bar, e.g., $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{f}}$. With such representations, we obtain,

$$\begin{aligned} \bar{\boldsymbol{f}} &= \bar{\boldsymbol{B}} \bar{\boldsymbol{f}} + \bar{\boldsymbol{\varepsilon}}, \\ \bar{\boldsymbol{x}} &= \bar{\boldsymbol{G}} \bar{\boldsymbol{f}} + \bar{\boldsymbol{e}}, \end{aligned} \tag{2}$$

where $\bar{\boldsymbol{B}} = \mathrm{Diag}(\boldsymbol{B}^{(1)}, ..., \boldsymbol{B}^{(M)})$, $\bar{\boldsymbol{\varepsilon}}$ and $\bar{\boldsymbol{e}}$ are independent, and $\bar{\boldsymbol{G}} = \mathrm{Diag}(\boldsymbol{G}^{(1)}, ..., \boldsymbol{G}^{(M)})$. A detailed explanation is presented in Supplementary Materials A (SM A). To address the second problem, we introduce factors of interest $\tilde{\boldsymbol{f}}$, which are embedded as different concepts with causal relations. As depicted in Figure 1, suppose $\bar{\boldsymbol{f}}$ are linearly generated by $\tilde{\boldsymbol{f}}$,

$$\bar{\boldsymbol{f}} = \boldsymbol{H} \tilde{\boldsymbol{f}}, \tag{3}$$

[1]These waved lines are used to emphasize the common space and they can be simply replaced by directed edges.

where $\boldsymbol{H} \in \mathbb{R}^{q \times \tilde{q}} (\tilde{q} \leq q)$ is a transformation matrix, and $\tilde{q}$ is the number of $\tilde{\boldsymbol{f}}$. The whole model is defined as follows.

Definition 1 (Multi-Domain LiNGAM for LAtent Factors (MD-LiNA))**.** *An MD-LiNA model satisfies the following assumptions:*

A1. $\boldsymbol{f}^{(m)}$ are generated linearly from a Directed Acyclic Graph (DAG) with non-Gaussian distributed external variables $\boldsymbol{\varepsilon}^{(m)}$, as in Eq.(1)*;*

A2. $\boldsymbol{x}^{(m)}$ are generated linearly from $\boldsymbol{f}^{(m)}$ plus Gaussian distributed errors $\boldsymbol{e}^{(m)}$, as in Eq.(1)*;*

A3. Each f_i has at least 2 pure measurement variables [2]*;*

A4. 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 $\bar{\boldsymbol{f}}$, as in Eq.(3)*.*

Note that assumption A4 implies that each row of $\boldsymbol{H}$ has only one non-zero element and each column has at least one non-zero element. It is reasonable since it is equal to the most interpretable structure in factor analysis. More plausibility of assumptions is discussed in SM B.

Once given multi-domain data, our goal is to estimate $\bar{\boldsymbol{G}}$ and $\tilde{\boldsymbol{B}}$, where $\tilde{\boldsymbol{B}}$ is a matrix that collects causal effects between shared latent factors of interest $\tilde{\boldsymbol{f}}$. $\tilde{\boldsymbol{B}}$ reflects the underlying causal structure shared by different $\boldsymbol{B}^{(m)}$. If there is only one single domain in the data, one just needs to estimate $\bar{\boldsymbol{G}} = \boldsymbol{G}^{(1)}$ and $\boldsymbol{B}^{(1)}$, where assumption A4 can be neglected and our model is simply called LiNGAM for LAtent Factors (LiNA). For simplicity, we use the **measurement model** to relate the structure from $\boldsymbol{x}^{(m)}$ to $\boldsymbol{f}^{(m)}$ while the **structure model** to record the causal relations among $\tilde{\boldsymbol{f}}$ or $\boldsymbol{f}^{(m)}$ [Silva *et al.*, 2006].

3 Model Identification

We state our identifiability results here. Note that the identifiability of LiNA has been provided by Shimizu *et al.* [2009], but with the assumption that each latent factor has at least 3 pure measurement variables. Below we show this identifiability can be strengthened to 2 pure measurement variables for each factor inspired by Triad constraints [Cai *et al.*, 2019].

Lemma 1. *Assume that the input data $\boldsymbol{X}$ strictly follow the LiNA model. Then the factor loading matrix $\boldsymbol{G}$ is identifiable up to permutation and scaling of columns and the causal effects matrix $\boldsymbol{B}$ is fully identifiable.*

The proof is in SM C. It relies on the corollary that Triad constraints also hold for our models, which helps find pure measurement variables and achieve LiNA's identifiability. Next, we show the identifiability of MD-LiNA in Theorem 1.

Theorem 1. *Assume that the input multi-domain data $\boldsymbol{X}$ with $\boldsymbol{X}^{(m)}$ of domain m, strictly follow the MD-LiNA model. Then the underlying factor loading matrix $\bar{\boldsymbol{G}}$ is identifiable up to permutation and scaling of columns and the causal effects matrix $\tilde{\boldsymbol{B}}$ is fully identifiable.*

We give a sketch of the proof below. For complete proofs of all theoretical results, please see SM C.

[2]Pure measurement variables are those which have only one single latent factor parent [Silva *et al.*, 2006].

Proof Sketch. Firstly in Eq.(2) and due to Lemma 1, $\bar{\boldsymbol{G}}$ is identifiable up to permutation and scaling of columns, since we estimate one measurement model for all domains simultaneously (mentioned in Section 4). Furthermore, combining Eqs.(2) and (3), we obtain $\tilde{\boldsymbol{f}} = \tilde{\boldsymbol{B}}\tilde{\boldsymbol{f}} + \tilde{\varepsilon}$, where $\tilde{\boldsymbol{B}} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\tilde{\boldsymbol{B}}\boldsymbol{H} \in \mathbb{R}^{\tilde{q}\times\tilde{q}}$ and $\tilde{\varepsilon} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\bar{\varepsilon} \in \mathbb{R}^{\tilde{q}}$. Note that the inverse matrix of $\boldsymbol{H}^T\boldsymbol{H}$ always exists since $\boldsymbol{H}$ is full column rank due to the assumption A4. To prove $\tilde{\boldsymbol{B}}$ is identifiable, we have to additionally ensure $\tilde{\boldsymbol{B}}$ can be permuted to a strictly lower triangular matrix and $\tilde{\varepsilon}$ are independent with each other. Fortunately, due to assumption A1, $\tilde{\boldsymbol{B}}$ satisfies the condition. Due to assumption A4, by virtue of the independence between $\bar{\varepsilon}$, its non-Gaussianity and the Darmois-Skitovich theorem [Kagan *et al.*, 1973], $\tilde{\varepsilon}$ are also independent with each other. Thus, $\tilde{\boldsymbol{B}}$ is fully identifiable, which implies the theorem is proved. □

4 Model Estimation

We exhibit a two-phase framework (measurement-model and structure-model phases) to estimate causal structures under latent factors in Algorithm 1, and provide its consistency.

4.1 MD-LiNA Algorithm

To learn measurement models, we have several approaches. Firstly, we can use the Confirmatory Factor Analysis (CFA) [Reilly and O'Brien, 1996], after employing Triad[3] to yield the structure between latent factors $\boldsymbol{f}^{(m)}$ and observed variables $\boldsymbol{x}^{(m)}$ [Cai *et al.*, 2019]. Secondly, more exploratory approaches are advocated, e.g., Exploratory Structural Equation Modeling (ESEM) [Asparouhov and Muthén, 2009], which enables us to use fewer restrictions on estimating factor loadings. Please see SM D for details. In our paper, we take the first approach, as illustrated in lines 1 to 3 of Algorithm 1, but we can use the second one as well in our framework.

To learn structure models, we introduce the log-likelihood function of LiNA, then unify it to MD-LiNA. For brevity, we omit the superscripts of all notations for LiNA. The log-likelihood function of LiNA is derived by characterizing the independence relations between $\boldsymbol{\varepsilon}$ from NICA models,

$$\mathcal{L}(\boldsymbol{B}, \hat{\boldsymbol{G}}) = \sum_{t=1}^{n} \left[\frac{1}{2} \left\| \boldsymbol{X}(t) - \hat{\boldsymbol{G}}\tilde{\hat{\boldsymbol{G}}}^T \boldsymbol{X}(t) \right\|_{\Sigma^{-1}}^2 + \sum_{i=1}^{q} \log \hat{p}_i(\boldsymbol{g}_i^T \boldsymbol{X}(t) - \boldsymbol{b}_i^T \tilde{\hat{\boldsymbol{G}}}^T \boldsymbol{X}(t)) \right] + C, \tag{4}$$

where $\|\boldsymbol{x}\|_{\Sigma^{-1}}^2 = \boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}$, $\boldsymbol{X}(t)$ is the t^{th} column (observation) of data $\boldsymbol{X}$. $\hat{\boldsymbol{G}}$ is the estimate of $\boldsymbol{G}$. $\tilde{\hat{\boldsymbol{G}}}^T = (\hat{\boldsymbol{G}}^T\hat{\boldsymbol{G}})^{-1}\hat{\boldsymbol{G}}^T$ relates to $\hat{\boldsymbol{G}}$ and $\boldsymbol{g}_i$ is the i^{th} column of $\tilde{\hat{\boldsymbol{G}}}$. The inverse matrix of $\hat{\boldsymbol{G}}^T\hat{\boldsymbol{G}}$ always exists due to assumption A3. $\boldsymbol{b}_i$ denotes the i^{th} column of $\boldsymbol{B}^T$. n is the sample size. C is a constant and $\hat{p}_i$ is their corresponding density function, which is specified to be Laplace distribution in estimation. Please see SM E.1 for detailed derivations.

[3]Triad constraints help locate latent factors and learn the causal structure between them, but they focus on single-domain data.

Algorithm 1 MD-LiNA Algorithm

Input: Data $\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(M)}$; M.
Output: Factor loadings $\bar{\boldsymbol{G}}$ ($\boldsymbol{G}$); effects matrix $\tilde{\boldsymbol{B}}$ ($\boldsymbol{B}$).
Phase I: Measurement models
Find the number of latent factors q_m and locate latent factors for each domain m by Triad constraints;
Get augmented data $\bar{\boldsymbol{X}} = \mathrm{Diag}(\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(M)})$ and estimate $\bar{\boldsymbol{G}}$ by CFA;
Estimate $\bar{\boldsymbol{f}} \doteq (\bar{\boldsymbol{G}}^T\bar{\boldsymbol{G}})^{-1}\bar{\boldsymbol{G}}^T\bar{\boldsymbol{X}}$;
Phase II: Structure models
if $M > 1$ **then**
  Optimize Eq.(7) iteratively for $\boldsymbol{H}$ and $\tilde{\boldsymbol{B}}$ until convergence using QPM (or ALM);
  Update $\tilde{\boldsymbol{B}}$ with regard to $\boldsymbol{H}$;
else
  Optimize Eq.(5) to get $\boldsymbol{B}$ with QPM (or ALM);
end if
return $\bar{\boldsymbol{G}}$ and $\tilde{\boldsymbol{B}}$ ($M > 1$); or $\boldsymbol{G} = \bar{\boldsymbol{G}}$ and $\boldsymbol{B}$ ($M = 1$).

Further, to strengthen the learning power in different cases, e.g., with small sample sizes or multicollinearity problem, we render $\boldsymbol{B}$ to satisfy an acyclicity constraint, adaptive ℓ_1 as well as ℓ_2 regularizations [Zheng *et al.*, 2018; Hyvärinen *et al.*, 2010; Zou and Hastie, 2005],

$$\begin{aligned} &\min_{\boldsymbol{B}} \ \mathcal{F}(\boldsymbol{B}, \hat{\boldsymbol{G}}), \quad \text{s.t.} \quad h(\boldsymbol{B}) = 0, \\ &\text{where} \quad \mathcal{F}(\boldsymbol{B}, \hat{\boldsymbol{G}}) = -\mathcal{L}(\boldsymbol{B}, \hat{\boldsymbol{G}}) + \lambda_1\|\boldsymbol{B}\|_{1*} + \lambda_2\|\boldsymbol{B}\|^2, \end{aligned} \tag{5}$$

$h(\boldsymbol{B}) = \mathrm{tr}(e^{\boldsymbol{B}\circ\boldsymbol{B}}) - q$ is the needed acyclicity constraint. $\circ$ is the Hadamard product, and $e^{\boldsymbol{B}}$ is the matrix exponential of $\boldsymbol{B}$. $\|\boldsymbol{B}\|_{1*} = \sum_{i=1}^{q}\sum_{j=1}^{q}|b_{ij}|/|\hat{b}_{ij}|$ represents the sparsity constraint where $\hat{b}_{ij}$ in $\hat{\boldsymbol{B}}$ is estimated by maximizing $\mathcal{L}(\boldsymbol{B}, \hat{\boldsymbol{G}})$. $\|\boldsymbol{B}\|^2$ is the ℓ_2 regularization. λ_1 and λ_2 are regularization parameters. This optimization function facilitates the simultaneous estimation of causal directions and effects between latent factors, without additional steps of permutation and rescaling, as required in Shimizu *et al.* [2009].

Thus, with estimated $\hat{\boldsymbol{G}}$, we leverage the Quadratic Penalty Method (QPM) (or Augmented Lagrangian Method, ALM) to optimize $\boldsymbol{B}$, transforming Eq. (5) into an unconstrained one,

$$\min_{\boldsymbol{B}} \ \mathcal{S}(\boldsymbol{B}), \tag{6}$$

in which $\mathcal{S}(\boldsymbol{B}) = \mathcal{F}(\boldsymbol{B}, \hat{\boldsymbol{G}}) + \frac{\rho}{2}h(\boldsymbol{B})^2$ is the quadratic penalty function. ρ is a regularization parameter. (For ALM, $\mathcal{S}(\boldsymbol{B}) = \mathcal{F}(\boldsymbol{B}, \hat{\boldsymbol{G}}) + \frac{\rho}{2}h(\boldsymbol{B})^2 + \alpha h(\boldsymbol{B})$, where α is a Lagrange multiplier.) Then Eq.(6) is solved by L-BFGS-B [Zhu *et al.*, 1997]. In case of avoiding numerical false discoveries from estimation, edges whose estimated effects are under a small threshold ϵ are ruled out [Zheng *et al.*, 2018].

Next, we show how Eq.(5) is unified to handle multi-domain cases ($M > 1$). Firstly, all $\boldsymbol{X}^{(m)}$ are projected into $\bar{\boldsymbol{X}}$ in a single common space so that $\bar{\boldsymbol{f}}$ are estimated. We then introduce the dependence relations between $\tilde{\boldsymbol{f}}$ and $\bar{\boldsymbol{f}}$ in Eq.(5), i.e., reconstruction errors $\mathcal{E}(\boldsymbol{H})$ and an adaptive sparsity $\|\boldsymbol{H}\|_{1*}$,

$$\begin{aligned} &\min_{\tilde{\boldsymbol{B}}, \boldsymbol{H}} \ \bar{\mathcal{F}}(\tilde{\boldsymbol{B}}, \boldsymbol{H}), \quad \text{s.t.} \quad h(\tilde{\boldsymbol{B}}) = 0, \\ &\text{where} \quad \bar{\mathcal{F}}(\tilde{\boldsymbol{B}}, \boldsymbol{H}) = \mathcal{F}(\tilde{\boldsymbol{B}}, \boldsymbol{H}) + \mathcal{E}(\boldsymbol{H}) + \lambda_3\|\boldsymbol{H}\|_{1*}, \end{aligned} \tag{7}$$

$\mathcal{E}(\boldsymbol{H}) = \|\bar{\boldsymbol{f}} - \boldsymbol{H}\tilde{\boldsymbol{f}}\|^2 = \|\bar{\boldsymbol{f}} - \boldsymbol{P}_{\boldsymbol{H}}\bar{\boldsymbol{f}}\|^2$, and λ_3 is a regularization parameter. $\boldsymbol{P}_{\boldsymbol{H}} = \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T$ is a projection matrix onto the column space of $\boldsymbol{H}$. $\mathcal{F}(\tilde{\boldsymbol{B}}, \boldsymbol{H})$ is the log-likelihood of MD-LiNA,

$$\begin{aligned} \mathcal{F}(\tilde{\boldsymbol{B}}, \boldsymbol{H}) &= -\mathcal{L}(\tilde{\boldsymbol{B}}, \boldsymbol{H}) + \lambda_1\|\tilde{\boldsymbol{B}}\|_{1*} + \lambda_2\|\tilde{\boldsymbol{B}}\|^2, \\ &= -\sum_{t=1}^{n}\left[\frac{1}{2}\left\|\bar{\boldsymbol{X}}(t) - \bar{\boldsymbol{G}}\tilde{\bar{\boldsymbol{G}}}^T\bar{\boldsymbol{X}}(t)\right\|^2_{\Sigma^{-1}}\right. \\ &\quad \left. + \sum_{i=1}^{\tilde{q}}\log\hat{p}_i(\boldsymbol{h}_i^T\bar{\boldsymbol{f}}(t) - \tilde{\boldsymbol{b}}_i^T\tilde{\hat{\boldsymbol{H}}}^T\bar{\boldsymbol{f}}(t))\right] - C \\ &\quad + \lambda_1\|\tilde{\boldsymbol{B}}\|_{1*} + \lambda_2\|\tilde{\boldsymbol{B}}\|^2, \end{aligned} \tag{8}$$

where $\tilde{\bar{\boldsymbol{G}}}^T = (\bar{\boldsymbol{G}}^T\bar{\boldsymbol{G}})^{-1}\bar{\boldsymbol{G}}^T$, $\tilde{\hat{\boldsymbol{H}}}^T = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T$, the inverse matrices of $\boldsymbol{H}^T\boldsymbol{H}$ and $\bar{\boldsymbol{G}}^T\bar{\boldsymbol{G}}$ always exist due to assumptions A3 and A4. $\boldsymbol{h}_i$ is the i^{th} column of $\tilde{\hat{\boldsymbol{H}}}$. $\tilde{\boldsymbol{b}}_i$ denotes the i^{th} column of $\tilde{\boldsymbol{B}}^T$. See SM E.2 for detailed derivations.

We iteratively optimize $\boldsymbol{H}$ and $\tilde{\boldsymbol{B}}$ using QPM (or ALM) until convergence. Specifically for QPM, i) to optimize $\boldsymbol{H}$, we find its descent direction to derive the next iteration for a given $\tilde{\boldsymbol{B}}$. Since there is no constraints for $\boldsymbol{H}$, it is an unconstrained problem. ii) to optimize $\tilde{\boldsymbol{B}}$, we compute its descent direction for the next iteration given $\boldsymbol{H}$, and update the penalty parameter ρ until the acyclicity constraint $h(\tilde{\boldsymbol{B}}) = 0$ is satisfied. Finally, we repeat steps i) and ii) until $\boldsymbol{H}$ and $\tilde{\boldsymbol{B}}$ are convergent. With $\boldsymbol{H}$, we can link the factors from multiple domains with high weights together to symbolize the same concepts so that we can decide which factors from different domains are represented by which factors of interest. Specifically, for $\tilde{f}_i$, those factors $\bar{\boldsymbol{f}}$ from different domains with the largest weights are considered to be represented by this $\tilde{f}_i$, where $\tilde{f}_i$ can also be named according to its corresponding observed measurement variables. Then according to these represented factors $\bar{\boldsymbol{f}}$ for each $\tilde{f}_i$, we obtain the causal ordering among $\tilde{\boldsymbol{f}}$, with which $\tilde{\boldsymbol{B}}$ can be updated. For the computational complexity, please see SM F.

4.2 Consistency Proofs

Here we prove that our methods could provide locally consistent estimators for MD-LiNA, including LiNA.

Theorem 2. *Assume the input single-domain data $\boldsymbol{X}$ strictly follow the LiNA model. Given that the sample size n, the number of observed variables p and the penalty coefficient ρ satisfy $n, p, \rho \to \infty$, then under conditions given in C0 & C1 (see SM C.3), our method using QPM with Eq.(5), is consistent and locally consistent to learn $\boldsymbol{G}$ and $\boldsymbol{B}$, respectively.*

Theorem 3. *Assume the input multi-domain data $\boldsymbol{X}$ with $\boldsymbol{X}^{(m)}$ strictly follow the MD-LiNA model. Given that the sample size n_m, the number of observed variables p_m of each domain m and the penalty coefficient ρ satisfy $n_m, p_m, \rho \to$*

∞, then under conditions given in C0-C5 (see SM C.4), our method using QPM with Eq.(7), is consistent to learn $\tilde{\boldsymbol{G}}$ and $\boldsymbol{H}$, and locally consistent to learn $\tilde{\boldsymbol{B}}$.

Having the identification results of LiNA/MD-LiNA, we present Theorems 2 and 3 to ensure that with our methods, asymptotically the resulting estimators $\tilde{\boldsymbol{G}}$ and $\boldsymbol{H}$ will be consistent to the true ones, while $\tilde{\boldsymbol{B}}$ will be locally consistent to its DAG solution. The proofs are shown in SM C.3 and C.4.

5 Experiments

We performed experiments on synthetic and real data, including multi-domain and single-domain ones. Due to the unstable performance of Triad, we assume the structure in measurement models is known a priori for all methods.

5.1 Synthetic Data

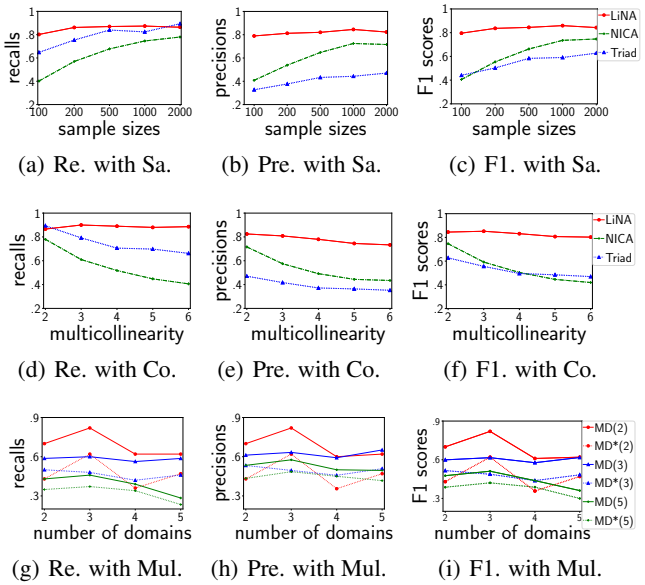
We generated the data according to Eq.(1). Unless specified, each structure in a domain has 5 latent factors and each factor has 2 pure measurement variables with sample size 1000. See SM G.1 for the details. We compared our LiNA with NICA [Shimizu *et al.*, 2009] and Triad [Cai *et al.*, 2019] for single-domain data. Since NICA and Triad do not focus on multi-domain data, we used our method that did not conduct line 6 of Algorithm 1 as the comparison (MD*). We did experiments with **i) different sample sizes**; **ii) highly-correlated latent variables**; **iii) different numbers of latent factors** and **iv) multi-domain data**. For other robustness performances, please see SM G.2.

i) Different sample sizes $n = 100, 200, 500, 1000, 2000$, to verify the capability in small-sample-size schemes. In Figure 2(a)-(c), we found our LiNA has the best performance, especially with smaller sample sizes. As sample sizes decrease, performances of other methods decrease whereas LiNA remains incredibly preponderant. Although Triad's recall is comparable to ours with enough sample sizes, its precision is the worst. The reason may be the sample sizes are not adequate to prune the directions, producing redundant edges.

ii) Highly-correlated variables through different effects of latent factors $[-i, -0.5] \cup [0.5, i]$, $i = 2,..., 6$. Their corresponding average Variable Inflation Factors (VIF) [4] are 22%, 47%, 69%, 78% and 84%, respectively, which measure the multicollinearity of variables and higher VIFs mean the heavier multicollinearity. In Figure 2(d)-(f), we found that as VIF increases, the accuracy of all methods declines in different degrees, but our method still outperforms the other comparisons, due to the employment of the elastic net regularization.

iii) Different numbers of latent factors $q = 2, 3, 5, 10$, to emphasize the capability of estimating causal effects, compared with NICA (Triad is not compared since it does not estimate effects). We applied the same method, CFA, to estimate $\hat{\boldsymbol{G}}$ as the NICA. Overall, in Figure 3, we found LiNA gives better performances in all cases. The accuracy decreases along with more latent factors. Specifically, NICA is comparable to ours when $q = 2$. However, as $q = 10$, accuracies both decrease, but LiNA decreases much more slowly than

[4] Average Variable Inflation Factors (VIF) are defined as the average VIF for all 100 independent trials of each range of weights.



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

(d) Re. with Co. (e) Pre. with Co. (f) F1. with Co.

(g) Re. with Mul. (h) Pre. with Mul. (i) F1. with Mul.

Figure 2: The recall (Re.), precision (Pre.) and F1 scores (F1.) of the recovered causal graphs between latent factors with different sample sizes (Sa.) i.e., $n = 100, 200, 500, 1000, 2000$ in (a), (b) and (c), whose noises of latent factors follow Laplace distributions; with different levels of multicollinearities (Co.) in (d), (e) and (f). In particular, in the x-axis, levels of multicollinearities of $i = 2, ..., 6$ are 22%, 47%, 69%, 78% and 84%, repectively, in the average VIF; and with different numbers of domains (Mul.) in (g), (h) and (i). Solid lines are from our MD-LiNA while dotted lines are from the comparison MD*. Higher F1 score represents higher accuracy.

NICA. The reasons may be 1) more measurement variables with the fixed sample size results in reduced power of CFA to estimate $\hat{\boldsymbol{G}}$, propagating errors to learn $\hat{\boldsymbol{B}}$; 2) the sparsity constraint deals with small sample sizes while NICA does not. And NICA does not estimate the causal directions and effects simultaneously, which may lack statistical efficiency.

iv) Multi-domain data $M = 2, 3, 4, 5$, through varying noises ε's distributions (sub-Gaussian or super-Gaussian). To obtain the true graph for evaluation, we generated the identical graphs of latent factors in each domain. We varied the number of latent factors, $q_m = 2, 3, 5$. In Figure 2(g)-(i), we found F1 scores of both methods tend to decrease with more domains or more 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. Further, though we experimented with only $q_m = 2, 3, 5$, the whole causal graph is much more complicated, which has totally Mq_m latent factors and $2Mq_m$ observed variables.

5.2 Real-World Applications

Yahoo stock indices dataset. We aimed to find the causal structure between different regions of the world, i.e., Asia, Europe, and the USA, each of which consisted of 2/3 stock indices. They were Asia := {N225, 000001.SS} from Japan and China, Europe := {BUK100P, FCHI, N100} from United Kingdom, France, and other European countries,

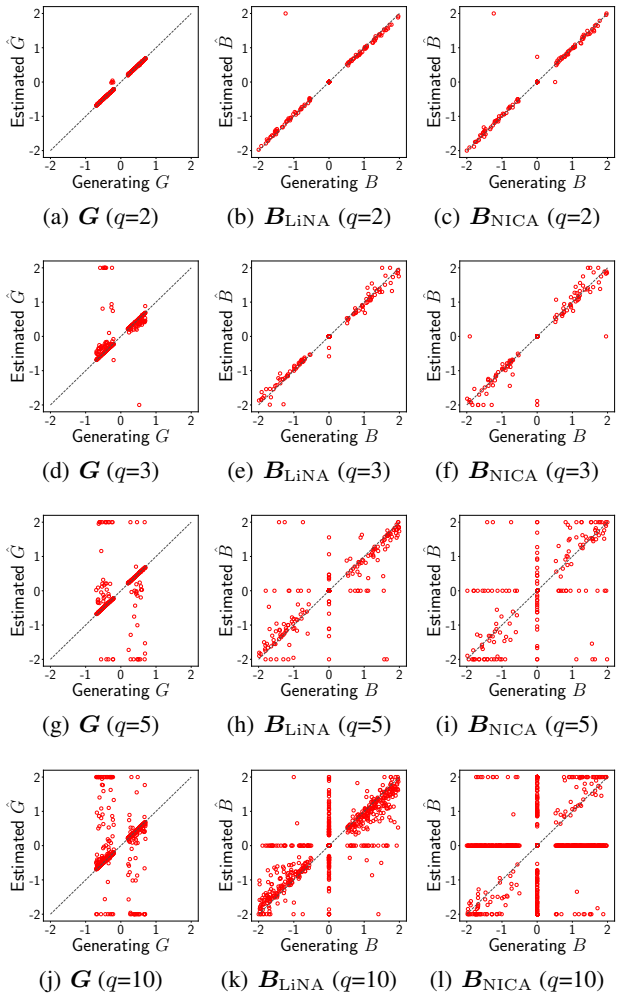


Figure 3: Scatter plots of estimated causal structures versus the true ones with different numbers of latent factors q. Note that estimates outside the interval [-2,2] are plotted at the edges of this interval. (a), (d), (g) and (j) are scatter plots of the estimated factor loading matrix versus the true ones. (b), (e), (h) and (k) are scatter plots of our method's estimated adjacency matrix versus the true ones while (c), (f), (i) and (l) are of NICA method's estimated matrix. The x-axis is the generating $\boldsymbol{G}$ or $\boldsymbol{B}$ while the y-axis is the estimated $\hat{\boldsymbol{G}}$ or $\hat{\boldsymbol{B}}$. Closer to the main diagonal means higher accuracy.

and USA := {DJI, GSPC, NYA} from the United States. We divided the data into two non-overlapping time segments such that their distributions varied across segments and are viewed as two different domains. We tested its multicollinearity and used 10-fold cross validation to select parameter values. Details are in SM H.1. Due to the different time zones, it is expected the ground truth is Asia $\rightarrow$ Europe $\rightarrow$ USA [Janzing *et al.*, 2010; Chen *et al.*, 2014], with which our recovered causal structure in Figure 4 was in accordance.

fMRI hippocampus dataset. We investigated causal relations between six brain regions of an individual: perirhinal cortex (PRC), parahippocampal cortex (PHC), entorhinal cortex (ERC), subiculum (Sub), CA1, and CA3/Dentate

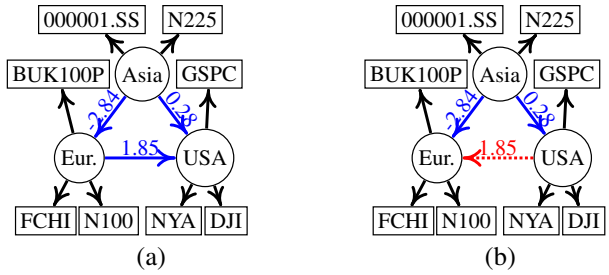


Figure 4: Estimated stock indices networks using the (a) MD-LiNA and (b) MD* methods. Solid blue lines denote consistent edges with the ground truth while densely dotted red lines are not.

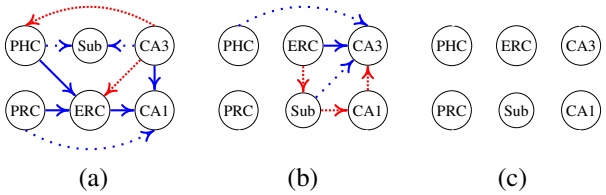


Figure 5: Causal structures of fMRI hippocampus data using (a) LiNA, (b) NICA, and (c) Triad methods. Note that Triad output an empty graph, which implied all brain regions are independent with each other. Solid blue lines are consistent edges with the anatomical connectivity while densely dotted red lines are spurious. Loosely dotted blue lines represent redundant edges.

Gyrus (DG), each of which had left and right sides and were treated as measurements [Poldrack *et al.*, 2015]. We used the anatomical connectivity between regions as a reference for evaluation [Bird and Burgess, 2008; Ghassami *et al.*, 2018]. From Figure 5, we see though our method estimated one more redundant edge, we obtained more consistent as well as less spurious edges than NICA, while Triad failed to learn the relations between these regions in this data. Besides, we found our results also coincide with some current findings, e.g., ERC $\rightarrow$ CA1 is supposed to correlate with memory loss [Kerchner *et al.*, 2012]. Please refer to SM H.2 for more details.

6 Conclusions

We proposed Multi-Domain Linear Non-Gaussian Acyclic Models for LAtent Factors (MD-LiNA) with its identification results, which gave deeper interpretation for latent factors that count. To discover the underlying causal structure for shared latent factors of interest, we proposed an integrated two-phase approach along with its local consistency. Our experimental results on simulated data and real-world applications validated the efficacy and robustness of the proposed algorithm.

Acknowledgments

This work was supported by the Grant ONR N00014-20-1-2501, the Natural Science Foundation of China (61876043, 61976052), Science and Technology Planning Project of Guangzhou (201902010058), and the Grant of China Scholarship Council. FX would like to acknowledge the support by China Postdoctoral Science Foundation (020M680225) and a research project of Huawei.

References

[Asparouhov and Muthén, 2009] Tihomir Asparouhov and Bengt Muthén. Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3):397–438, 2009.

[Bartholomew *et al.*, 2008] David Bartholomew, Fiona Steele, Ir Moustaki, and Jane Galbraith. *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Routledge (2 edition), 2008.

[Bird and Burgess, 2008] Chris M Bird and Neil Burgess. The hippocampus and memory: insights from spatial processing. *Nature Reviews Neuroscience*, 9(3):182–194, 2008.

[Cai *et al.*, 2019] 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.

[Chen *et al.*, 2014] Zhitang Chen, Kun Zhang, Laiwan Chan, and Bernhard Schölkopf. Causal discovery via reproducing kernel hilbert space embeddings. *Neural Computation*, 26(7):1484–1517, 2014.

[Danks *et al.*, 2009] David Danks, Clark Glymour, and Robert E Tillman. Integrating locally learned causal structures with overlapping variables. In *NIPS*, pages 1665–1672, 2009.

[Dhir and Lee, 2020] Anish Dhir and Ciarán M Lee. Integrating overlapping datasets using bivariate causal discovery. In *AAAI*, pages 3781–3790, 2020.

[Ghassami *et al.*, 2018] AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. In *NeurIPS*, pages 6266–6276, 2018.

[Huang *et al.*, 2020] Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery from multiple data sets with non-identical variable sets. In *AAAI*, pages 10153–10161, 2020.

[Hyvärinen *et al.*, 2010] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5):1709–1731, 2010.

[Jaber *et al.*, 2020] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *NeurIPS*, volume 33, pages 9551–9561, 2020.

[Janzing *et al.*, 2010] Dominik Janzing, Patrik O Hoyer, and Bernhard Schölkopf. Telling cause from effect based on high-dimensional observations. In *ICML*, pages 479–486. Omnipress, 2010.

[Kagan *et al.*, 1973] Abram M Kagan, Calyampudi Radhakrishna Rao, and Yurij Vladimirovich Linnik. *Characterization problems in mathematical statistics*. Wiley, 1973.

[Kerchner *et al.*, 2012] Geoffrey A Kerchner, Gayle K Deutsch, Michael Zeineh, Robert F Dougherty, Manojkumar Saranathan, and Brian K Rutt. Hippocampal ca1 apical neuropil atrophy and memory performance in alzheimer's disease. *Neuroimage*, 63(1):194–202, 2012.

[Kocaoglu *et al.*, 2019] Murat Kocaoglu, Karthikeyan Shanmugam, Amin Jaber, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. *NeurIPS*, 2019.

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

[Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition, 2009.

[Peters *et al.*, 2017] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

[Poldrack *et al.*, 2015] Russell A. Poldrack, Timothy O Laumann, Oluwasanmi Koyejo, Brenda Gregory, Ashleigh Hover, Mei-Yen Chen, Krzysztof J Gorgolewski, Jeffrey Luci, Sung Jun Joo, Ryan L. Boyd, et al. Long-term neural and physiological phenotyping of a single human. *Nature Communications*, 6(1):1–15, 2015.

[Reilly and O'Brien, 1996] 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.

[Shimizu *et al.*, 2006] 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.

[Shimizu *et al.*, 2009] 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.

[Shimizu, 2012] Shohei Shimizu. Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81:104–107, 2012.

[Shimodaira, 2016] Hidetoshi Shimodaira. Cross-validation of matching correlation analysis by resampling matching weights. *Neural Networks*, 75:126–140, 2016.

[Silva *et al.*, 2006] 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.

[Smith *et al.*, 2011] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

[Spirtes and Zhang, 2016] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3, 2016.

[Tillman and Spirtes, 2011] Robert Tillman and Peter Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *AISTATS*, pages 3–15, 2011.

[Xie *et al.*, 2020] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *NeurIPS*, 33, 2020.

[Zheng *et al.*, 2018] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *NeurIPS*, pages 9472–9483, 2018.

[Zhu *et al.*, 1997] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.