Causal Discovery with Heterogeneous Observational Data

Fangting Zhou [1,2] Kejun He [2] Yang Ni [1]

[1]Department of Statistics, Texas A&M University, College Station, Texas, USA [2]Institute of Statistics and Big Data, Renmin University of China, Beijing, China

Introduction

Causal discovery is a central task in various fields including social science, artificial intelligence, and systems biology. Discovering causality from purely observational data is of vital importance.

- One prominent approach in presenting and learning causality is to use the structural equation model (SEM) and the associated causal graph.
- The recursive linear Gaussian SEM is among the most popular ones although the associated causal directed acyclic graph is only identifiable up to Markov equivalence classes.
- In order to uniquely identify causal structures with observational data, additional distributional assumptions have been made in prior works including the linear non-Gaussian model, the non-linear additive noise model, and the linear Gaussian model with equal error variances.
- None of these methods allow for both cycles and confounders.
- Recently, some methods are proposed to explicitly address the heterogeneity issue by incorporating covariates (environments), exploiting invariance, or using a latent mixture model.

We propose a novel method for Causal discovery with Heterogeneous Observational Data (CHOD). We do not restrict our model to be acyclic and do not assume causal sufficiency. By exploiting the data heterogeneity via exogenous covariates, we provide sufficient conditions under which CHOD is structurally identifiable. Our method is among the first model-based causal discovery methods to identify causal graphs with both cycles and confounders in purely observational settings without prior domain knowledge.

Model

Our key idea to discover causality is to take advantage of the data heterogeneity, which we assume can be explained by some exogenous covariates $Z \in \mathbb{R}$. The exogenous covariates may be observed (e.g., biomarkers in cancer genomic data) or latent. Alternatively, latent covariates can be learned simultaneously with our model. Given Z, we model X as a conditionally linear Gaussian SEM,

$$X = B(Z)X + E, E \sim N(0, S),$$

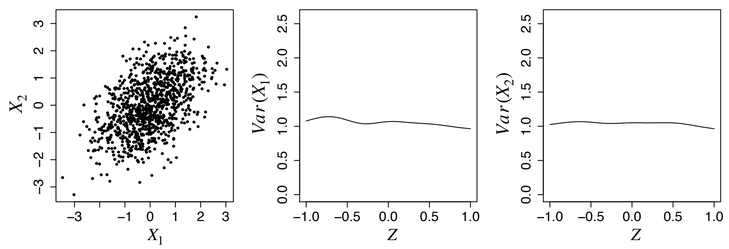
where $B(Z) = [b_{j\ell}(Z)] : \mathbb{R} \mapsto \mathbb{R}^{p\times p}$ is a matrix-valued function of Z, which characterizes the changes of the direct causal effects with respect to Z. Because each observation potentially has a different value of covariate Z, the direct causal effects $B(Z)$ are heterogeneous and observation-specific. When $B(Z)$ is constant in Z, model is reduced to an ordinary linear Gaussian SEM and hence its underlying causal graph $\mathcal{G}$ is not identifiable.

A bivariate toy example

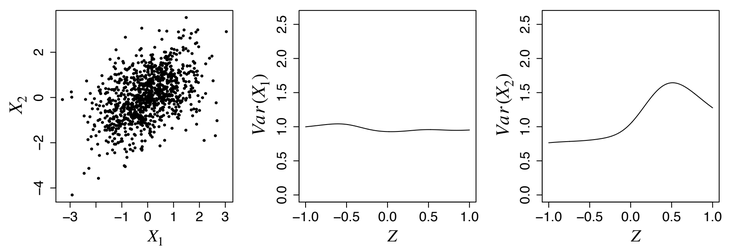
We illustrate the identifiability with simulated data from graphs $X_1 \cdots X_2$, $X_1 \to X_2$, and $X_1 \leftrightarrow X_2$. The $n = 1000$ data points as well as the marginal variances estimated by kernel method of the two nodes as functions of Z are depicted in the figure below for these 3 cases, from which the causal relationships between X_1 and X_2 are intuitively identifiable in the presence of both confounders and cycles: in Figure (a), both Var(X_1) and Var(X_2) are constant in Z indicating no direct causal link; in Figure (b), Var(X_1) is constant but Var(X_2) is not constant in Z indicating a direct causal link $X_1 \to X_2$; and in Figure (c), neither Var(X_1) nor Var(X_2) is constant in Z indicating a cyclic causal link $X_1 \rightleftarrows X_2$.

Theorem: Causally insufficient bivariate cyclic graphs

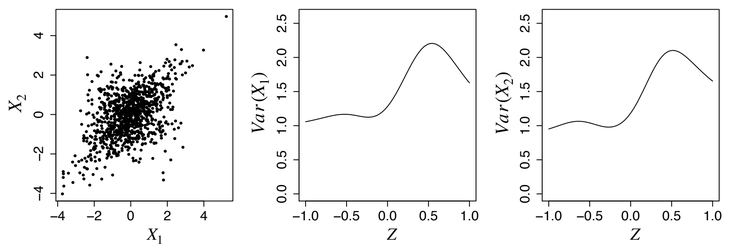
Consider bivariate CHOD models with direct causal effects $[b_{12}(Z), b_{21}(Z)]$ and $[b'_{12}(Z), b'_{21}(Z)]$, respectively. Assume $b_{j\ell}(Z)$ and $b'_{j\ell}(Z)$ are either zero or non-constant functions for all $j \neq \ell \in \{1, 2\}$. Then if the two CHOD models are distribution equivalent, we must have $\mathcal{G} = \mathcal{G}'$.



(a) $x_1 \nrightarrow x_2$ and $x_2 \nrightarrow x_1$



(b) $x_1 \to x_2$ and $x_2 \nrightarrow x_1$



(c) $x_1 \to x_2$ and $x_2 \to x_1$

Theorem: Causally insufficient multivariate acyclic graphs

Consider the CHOD model restricted to acyclic causal graphs. Assume without loss of generality $(1, \ldots, p)$ is a true causal ordering (i.e., $\ell \not\to j$ if $\ell > j$). If for any node j, and any set $S = \{1, \ldots, m\}$ such that $pa(j) \not\subset S$, we have Var$(X_j|X_S)$ is a non-constant function of the covariate Z, then the causal ordering is identifiable. Moreover, if $pa(j) \cap ds(j) = \emptyset, \forall j$, then the causal graph is identifiable.

Theorem: Causally sufficient multivariate cyclic graphs

Consider the CHOD model where there are no unmeasured confounders and all cycles are disjoint. The causal graph is generally identifiable.

Proposition: Multivariate latent exogenous covariates

Assume the vector $m(Z)$ that stacks the non-zero elements of $B(Z)$ is continuous and injective, and $(m, S) \mapsto \mathbb{P}(X|m, S)$ is continuous and injective in m given $\mathcal{G}$. Then the latent exogenous covariates are identifiable up to a monotone transformation.

Bayesian structure learning

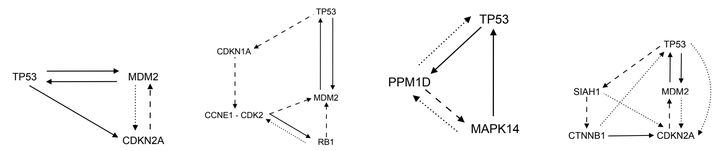
We learn the causal structure through a Bayesian approach by assigning priors on the space of graphs and model parameters. The posterior distribution is not analytically available and we use Markov chain Monte Carlo (MCMC) to approximate it. We can assess the credibility of inferred edges via posterior inference: edges that have nearly constant causal effects (e.g., if 95% credible bands of $b_{j\ell}(Z)$, which can be computed from Monte Carlo samples, cover constant functions) are deemed less reliable.

Experiment

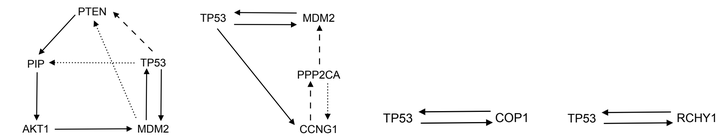
We conducted extensive simulation experiments to illustrate the performance of the proposed method.

- We considered sample size $n \in \{125, 250, 500, 1000\}$ and the number of nodes $p \in \{10, 25, 50\}$.
- We generated data from cyclic graphs with confounders, cyclic graphs without confounders, and acyclic graphs with (without) confounders.
- We compared with various state-of-the-art algorithms: RFCI, RICA, LiNG, ANM, CAM, GDS, RESIT, IGCI, EMD, bQCD, NOTEARS, and DAG-GNN.
- We considered various model misspecifications in terms of non-Gaussian errors, different confounding effects, varying degrees of heterogeneity, and unobserved covariates.
- Our method performed better than others throughout.

Application. We demonstrate the capability of CHOD in identifying gene feedback loops using breast cancer gene expression data downloaded from the Cancer Genome Atlas (https://www.cancer.gov/tcga). Breast cancer is a well-known extremely heterogeneous genetic disease. We focused on 8 feedback loops involving gene TP53. Gene expressions were log-transformed and we learned a one-dimensional embedding using UMAP as an input covariate and regressed out the effects of the covariate on the mean gene expression.



(a) Network A. (b) Network B. (c) Network C. (d) Network D.



(e) Network E. (f) Network F. (g) Network G. (h) Network H.

Future work

A natural future direction is to extend this paper to nonlinear and non-Gaussian models via e.g., basis expansion and mixture of Gaussian error distributions. Second, we plan to prove the most general case with causally insufficient multivariate cyclic graphs in the future. As was done in many previous works, one may first fix the skeleton or learn some partial structures with common structure learning algorithms, and then orient indeterminate edges by applying the bivariate causal discovery method to identify the full structure.

fangtingzhou@tamu.edu