Causal Inference and Mechanism Clustering of A Mixture of Additive Noise Models

Shoubo Hu [1] Zhitang Chen [2] Vahid Partovi Nia [2] Laiwan Chan [1] Yanhui Geng [3]

[1] The Chinese University of Hong Kong [2] Huawei Noah's Ark Lab [3] Huawei Montréal Research Center

Motivation

Most causal models assume a single mapping from the cause to effect (causal mechanism) in a functional form, which makes them inapplicable in cases of data with complex distributions (see the illustration below).

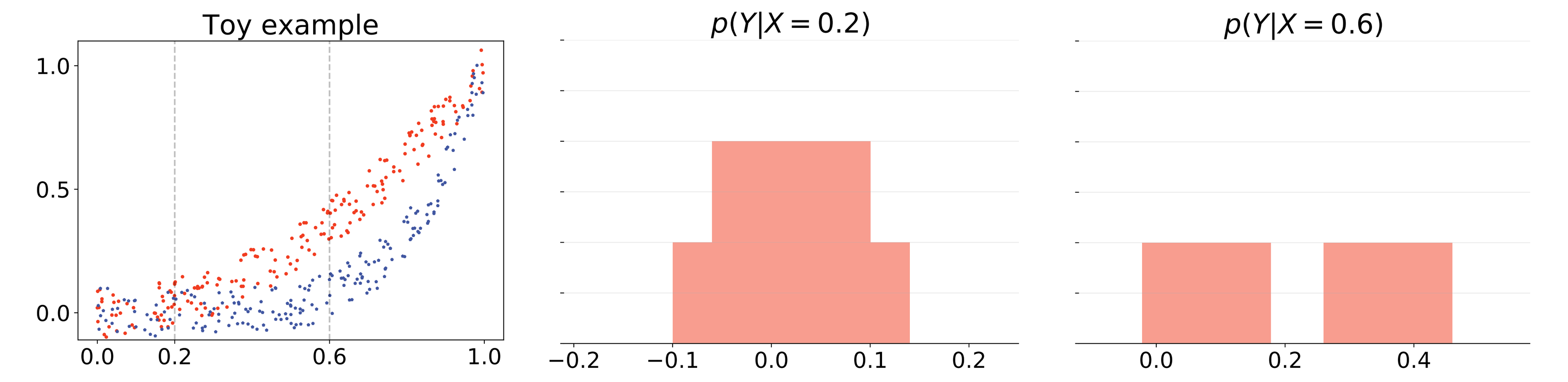

Figure: Example illustrating the failure of ANM on the inference of a mixture of ANMs (a) the distribution of data generated from $M_1 : Y = X^2 + \epsilon$ and $M_2 : Y = X^5 + \epsilon$, where $X \sim U(0,1)$ (x-axis) and $\epsilon \sim U(-0.1, 0.1)$; (b) Conditional $p(Y|X = 0.2)$; (c) Conditional $p(Y|X = 0.6)$. It is obvious that when the data is generated from a mixture of ANMs, the consistency of conditionals is likely to be violated which leads to the failure of ANM.

Objective

- **Causal inference:** infer the causal direction of data generated from a mixture of causal mechanisms;
- **Mechanism clustering:** cluster the data such that each cluster corresponds to a causal mechanism.

ANM Mixture Model (ANM-MM)

An ANM [2] Mixture Model is a set of causal models of the same causal direction between two continuous random variables X and Y. All causal models share the same function form given by the following ANM:

$$Y = f(X;\theta) + \epsilon,$$

where X denotes the cause, Y denotes the effect, f is a nonlinear function parameterized by θ and the noise $\epsilon \perp\!\!\!\perp X$. The differences between causal models in an ANM-MM stem only from different values of function parameter θ. In ANM-MM, θ is assumed to be drawn from a discrete distribution on a finite set $\Theta = \{\theta_1, \cdots, \theta_C\}$, i.e. $\theta \sim p_\theta(\theta) = \sum_{c=1}^{C} a_c \mathbf{1}_{\theta_c}(\cdot)$, where $a_c > 0$, $\sum_{c=1}^{C} a_c = 1$ and $\mathbf{1}_{\theta_c}(\cdot)$ is the indicator function of a single value θ_c.

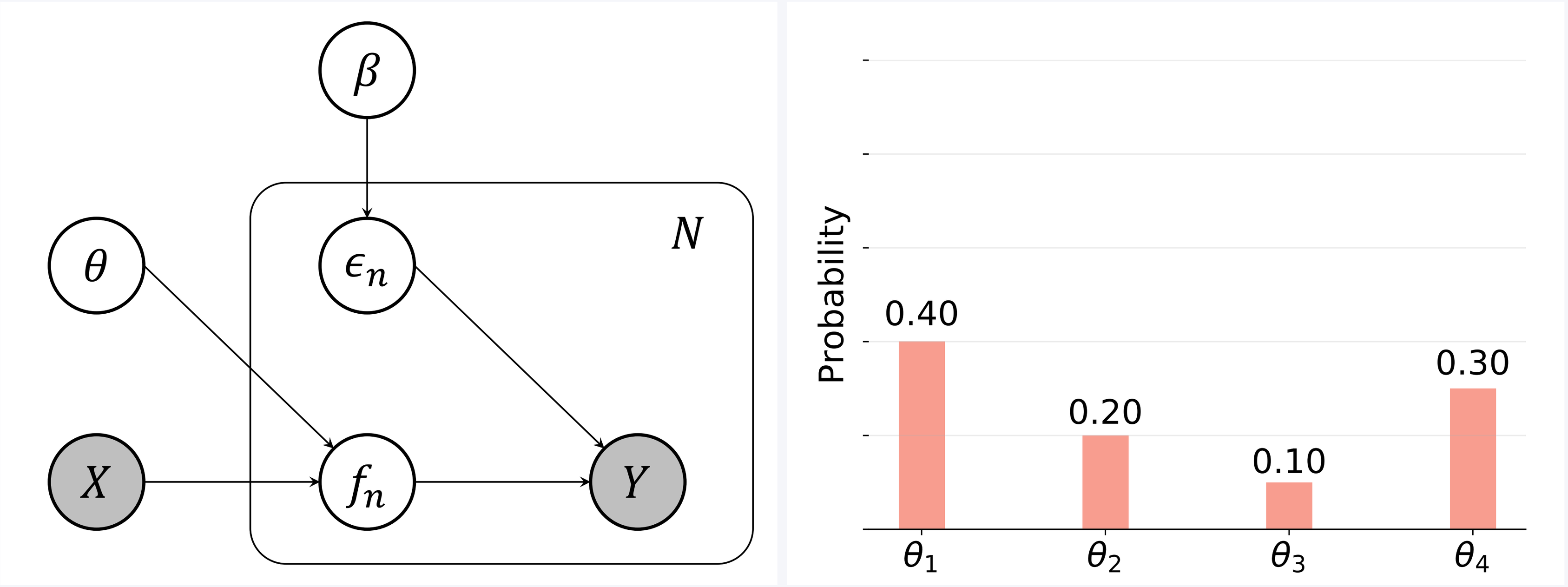

Figure: (a) The graphical representation of ANM mixture model. (b) An example of the distribution over θ.

Identifiability

Postulate 1. Independence of input and function [3]

If $X \to Y$, the distribution of X and the function f mapping X to Y are *independent* since they correspond to independent mechanisms of nature.

Theorem 1. Let $X \to Y$ and they follow an ANM-MM. If there exists a backward ANM-MM,

$$X = g(Y;\omega) + \tilde{\epsilon},$$

where $\omega \sim p_\omega(\omega) = \sum_{\tilde{c}=1}^{\tilde{C}} b_{\tilde{c}} \mathbf{1}_{\omega_{\tilde{c}}}(\cdot)$, $b_{\tilde{c}} > 0$, $\sum_{\tilde{c}=1}^{\tilde{C}} b_{\tilde{c}} = 1$ and $\tilde{\epsilon} \perp\!\!\!\perp Y$, in the anticausal direction, then $(p_X, p_\epsilon, f, p_\theta)$ should fulfill $\tilde{C}$ ordinary differential equations,

$$\xi''' - \frac{G^{(\tilde{c})}(X,Y)}{H^{(\tilde{c})}(X,Y)}\xi'' = \frac{G^{(\tilde{c})}(X,Y)V^{(\tilde{c})}(X,Y)}{U^{(\tilde{c})}(X,Y)} - H^{(\tilde{c})}(X,Y),\ \tilde{c} = 1, 2, \cdots, \tilde{C}, \tag{1}$$

where $\xi := \log p_X$, $G^{(\tilde{c})}(X,Y)$, $H^{(\tilde{c})}(X,Y)$, $U^{(\tilde{c})}(X,Y)$ and $V^{(\tilde{c})}(X,Y)$ are given in the supplementary.

Model Estimation

Gaussian Process Partially Observable Model (GPPOM)

As in standard GP-LVM, the log-likelihood of GPPOM is given by

$$\mathcal{L}(\boldsymbol{\Theta}|\mathbf{X}, \mathbf{Y}, \beta) = -\frac{DN}{2}\ln(2\pi) - \frac{D}{2}\ln\left(|\tilde{\mathbf{K}}|\right) - \frac{1}{2}\mathrm{tr}\left(\tilde{\mathbf{K}}^{-1}\mathbf{Y}\mathbf{Y}^T\right), \tag{2}$$

where $\mathbf{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_N]^T$ is the matrix collecting instances of the effect, $\tilde{\mathbf{K}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \boldsymbol{\beta}^{-1}\mathbf{I} = [\mathbf{X}, \boldsymbol{\Theta}][\mathbf{X}, \boldsymbol{\Theta}]^T + \boldsymbol{\beta}^{-1}\mathbf{I} = \mathbf{X}\mathbf{X}^\mathbf{T} + \boldsymbol{\Theta}\boldsymbol{\Theta}^\mathbf{T} + \boldsymbol{\beta}^{-1}\mathbf{I}$ is the covariance matrix after bringing in θ, $\mathbf{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]^T$ is the matrix collecting instances of the cause, and $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N]^T$ is the matrix collecting parameters associated with all instances.

Parameter estimation by independence enforcement

We include HSIC [1] in the objective to enforce X and θ to be independent. By incorporating HSIC term into the negative log-likelihood of GPPOM, the optimization objective reads

$$\underset{\boldsymbol{\Theta},\Omega}{\arg\min}\, \mathcal{J}(\boldsymbol{\Theta}) = \underset{\boldsymbol{\Theta},\Omega}{\arg\min}\, [-\mathcal{L}(\boldsymbol{\Theta}|\mathbf{X}, \mathbf{Y}, \Omega) + \lambda \log \mathrm{HSIC_b}(\mathbf{X}, \boldsymbol{\Theta})], \tag{3}$$

where λ is the parameter which controls the importance of the HSIC term and Ω is the set of all hyper parameters including β and all kernel parameters γ_d, $d = 1, \dots, D_x$.

Algorithms

Input: $\boldsymbol{\mathcal{D}} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ - the set of instances of two random variables;

λ - parameter of HSIC term;

C - the number of clusters

Causal inference **Output:** The causal direction	Mechanism clustering **Output:** The cluster labels
Standardize instances in $\boldsymbol{\mathcal{D}}$; Initialize β and kernel parameters; Optimize (3) in both directions; If $\mathrm{HSIC}_{X\to Y}$ < $\mathrm{HSIC}_{Y\to X}$ then $X \to Y$. Else if $\mathrm{HSIC}_{X\to Y}$ > $\mathrm{HSIC}_{Y\to X}$ then $Y \to X$. Else No decision made.	Standardize instances in $\boldsymbol{\mathcal{D}}$; Initialize β and kernel parameters; Find $\boldsymbol{\Theta}$ by optimizing (3) in causal direction; Apply k-means on $\boldsymbol{\theta}_n$ to obtain cluster labels.

Experiments

Causal Inference. The following elementary functions are adopted in the synthetic experiments: (a) $f_1 = \frac{1}{1.5+\theta_c X^2}$; (b) $f_2 = 2 \times X^{\theta_c - 0.25}$; (c) $f_3 = \exp(-\theta_c X)$; (d) $f_4 = \tanh(\theta_c X)$.

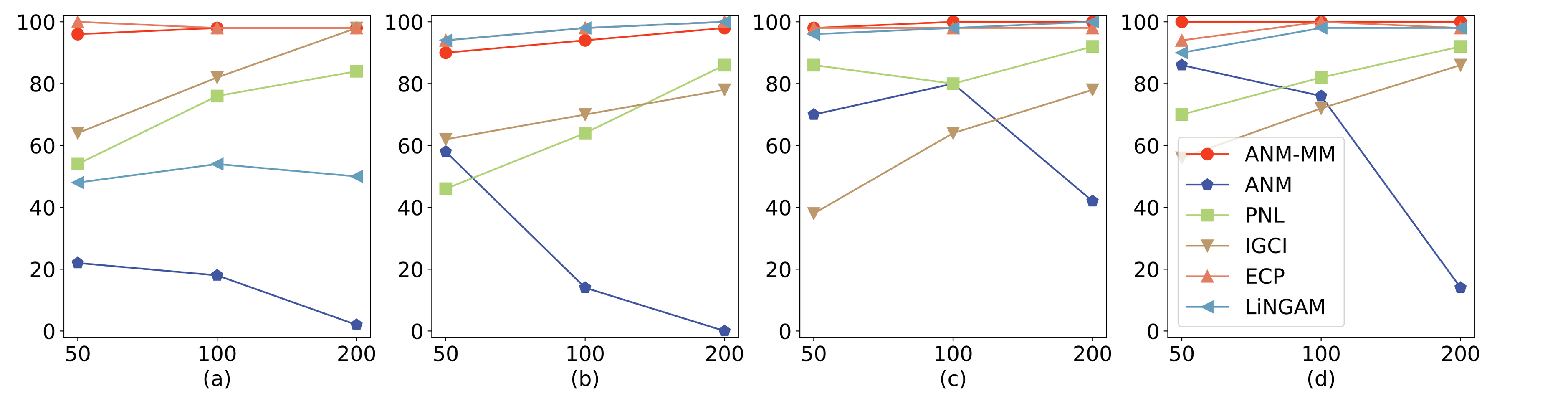

Figure: Accuracy (y-axis) versus sample size (x-axis) on different causal mechanisms: (a) f_1; (b) f_2; (c) f_3; (d) f_4.

Further experiments are conducted on (a) different number of causal mechanisms (C); (b) different noise standard deviations (σ); (c) different mixing proportions (a_c); (d) Tübingen cause-effect pairs.

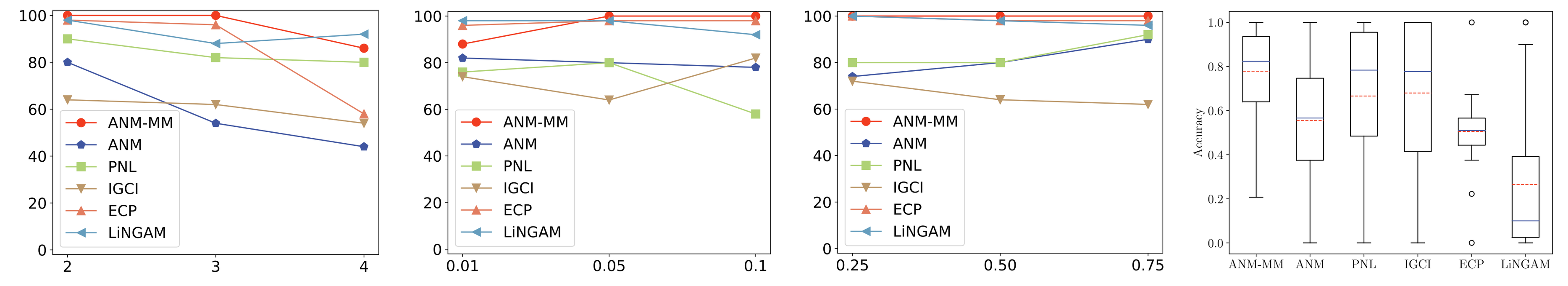

Figure: Accuracy (y-axis) versus (a) C; (b) σ; (c) a_1; on f_3 with $N = 100$. (d) Accuracy on real Tübingen cause-effect pairs.

Mechanisms clustering. Similar settings are used in clustering experiments. Average adjusted Rand index (avgARI $\in [-1, 1]$), which is the mean ARI over 100 experiments, are used for evaluation.

Table: avgARI of synthetic clustering experiments (Higher the better)

avgARI	(i) f				(ii) C		(iii) σ		(iv) a_1	
	f_1	f_2	f_3	f_4	3	4	0.01	0.1	0.25	0.75
ANM-MM	**0.393**	**0.660**	**0.777**	**0.682**	**0.610**	**0.447**	**0.798**	**0.608**	**0.604**	**0.867**
k-means	0.014	0.039	0.046	0.046	0.194	0.165	0.049	0.042	0.047	0.013
PCA-km	0.013	0.037	0.044	0.048	0.056	0.041	0.047	0.040	0.052	0.014
GMM	0.015	0.340	0.073	0.208	0.237	0.202	0.191	0.025	0.048	0.381
SpeClu	0.003	0.129	0.295	0.192	0.285	0.175	0.595	0.048	0.044	-0.008
DBSCAN	0.055	0.265	0.342	0.358	0.257	0.106	0.527	0.110	0.521	0.718

References

[1] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63--77. Springer, 2005.

[2] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689--696, 2009.

[3] Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168--5194, 2010.