Exogenous Matching: Learning Good Proposals for Tractable Counterfactual Estimation

Yikang Chen[1], Dehui Du[1,*], Lili Tian[1]

[1] Shanghai Key Laboratory of Trustworthy Computing, East China Normal University
* Corresponding author (dhdu@sei.ecnu.edu.cn)

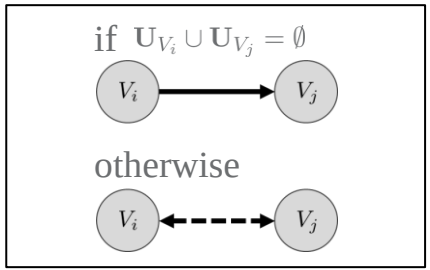
Preliminaries

Structural Causal Model (SCM) is a 4-tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P_{\mathbf{U}} \rangle$.

- $\mathbf{U}$ is a set of exogenous variables.
- $\mathbf{V}$ is a set of endogenous variables.
- $\mathcal{F}$ is a collection of functions describing causal mechanisms.
- $P_{\mathbf{U}}$ is a probability distribution over exogenous variables.

$$\mathcal{F} = \begin{cases} V_1 \leftarrow f_{V_1}(\mathbf{Pa}_{V_1}, \mathbf{U}_{V_1}) \\ \dots \\ V_n \leftarrow f_{V_n}(\mathbf{Pa}_{V_n}, \mathbf{U}_{V_n}) \end{cases},$$
$$\text{where } \forall i = 1 \dots n, \quad V_i \in \mathbf{V},$$
$$\text{and} \quad \mathbf{Pa}_{V_i} \subseteq \mathbf{V}, \mathbf{U}_{V_i} \subseteq \mathbf{U}.$$

The general form of $\mathcal{F}$.



Build causal graph $\mathcal{G}$ from $\mathcal{F}$.

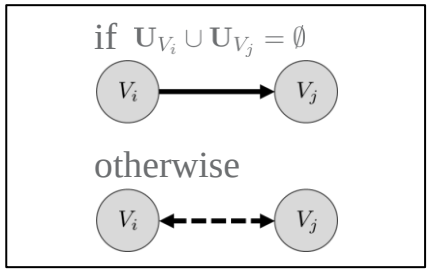
Preliminaries

Structural Causal Model (SCM) is a 4-tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P_{\mathbf{U}} \rangle$.

- $\mathbf{U}$ is a set of exogenous variables.
- $\mathbf{V}$ is a set of endogenous variables.
- $\mathcal{F}$ is a collection of functions describing causal mechanisms.
- $P_{\mathbf{U}}$ is a probability distribution over exogenous variables.

$$\mathcal{F} = \begin{cases} V_1 \leftarrow f_{V_1}(\mathbf{Pa}_{V_1}, \mathbf{U}_{V_1}) \\ \quad \dots \\ V_n \leftarrow f_{V_n}(\mathbf{Pa}_{V_n}, \mathbf{U}_{V_n}) \end{cases},$$
$$\text{where } \forall i = 1 \dots n, \quad V_i \in \mathbf{V},$$
$$\text{and} \quad \mathbf{Pa}_{V_i} \subseteq \mathbf{V}, \mathbf{U}_{V_i} \subseteq \mathbf{U}.$$

The general form of $\mathcal{F}$.



Build causal graph $\mathcal{G}$ from $\mathcal{F}$.

Recursiveness

There exists an order in $\mathcal{F}$ such that for any $f_i, f_j \in \mathcal{F}$, if $f_i < f_j$, then $V_j \notin \mathbf{Pa}_{V_i}$.

This implies that given values for exogenous variables $\mathbf{u} \in \Omega_{\mathbf{U}}$, all values of endogenous variables in $\mathbf{V}$ are also fixed (i.e., there exists a unique solution for endogenous variables w.r.t. $\mathbf{u}$), and the causal graph $\mathcal{G}$ is an acyclic directed mixed graph (ADMG).

Assumption on SCMs.

Pearl Causal Hierarchy (PCH) defines symbolic languages $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$.

- $\mathcal{L}_i$-expression **syntax**: (Boolean combinations of) inequalities between polynomials over terms $P(\alpha)$, where $P(\alpha)$ is an $\mathcal{L}_i$-term.
  - $\mathcal{L}_1$**-terms** are those of the form $P(\mathbf{Y})$
  - $\mathcal{L}_2$**-terms** additionally include $P(\mathbf{Y}_\mathbf{x})$
  - $\mathcal{L}_3$**-terms** encode conjunctions $P(\mathbf{Y}_\mathbf{x}, \ldots, \mathbf{Z}_\mathbf{w})$ ($\{\mathbf{Y}_\mathbf{x}, \ldots, \mathbf{Z}_\mathbf{w}\}$ will be abbreviated as $\mathbf{Y}_*$)

- $\mathcal{L}_i$-expression **semantics**: assign $\mathcal{L}_i$-term $P(\alpha)$

$$P(\mathbf{Y}_* \in \mathcal{Y}_*) = \int_{\Omega_\mathbf{U}} \mathbb{1}_{\Omega_\mathbf{U}(\mathcal{Y}_*)}(\mathbf{u})\, \mathrm{d}P_\mathbf{U}$$

called $\mathcal{L}_3$-valuation, while $\mathcal{L}_1$ and $\mathcal{L}_2$-valuations are special cases. $\mathcal{Y}_*$ is termed counterfactual event.
  - $\Omega_\mathbf{U}$ is the domain of exogenous variables
  - $\Omega_\mathbf{U}(\mathcal{Y}_*) = \{\mathbf{u} \mid \mathbf{Y}_*(\mathbf{u}) \in \mathcal{Y}_*\}$
  - $\mathbf{Y}_*(\mathbf{u})$ is called potential outcome, inferred from SCM

Variables: □, ○, △ Observations: ■, ●, ▲ Interventions: ↖, ↖, ↖
The submodel of SCM $\mathcal{M}$ under intervention ↖: $\mathcal{M}_{\nwarrow}$

SCM $\mathcal{M}$	Observational	Interventinal	Counterfactual
Human Behavior			
Natural Language	□ is ■?	○ would be ● under ↖?	if ○ were ● under ↖, △ would be ▲ under ↖?
Formal Language	$P(\square = \blacksquare)$	$P(\circ_{\nwarrow} = \bullet)$	$P(\triangle_{\nwarrow} = \blacktriangle \mid \circ_{\nwarrow} = \bullet)$
PCH	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$
SCM	$\mathcal{M}$	$\mathcal{M}_{\nwarrow}$	$\mathcal{M}_{\nwarrow}, \mathcal{M}_{\nwarrow}$

A brief illustration of SCM, PCH and counterfactual concepts.

The collection of observational, interventional, and counterfactual distributions induced by SCM, as delineated by the above syntax and semantics, is called **Pearl Causal Hierarchy (PCH)**.

[1] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. On Pearl's Hierarchy and the Foundations of Causal Inference

Counterfactual Estimation

Counterfactual Estimation is a subtask within counterfactual reasoning. From the perspective of PCH, it involves estimating the value of the $\mathcal{L}_3$-expression, which requires estimating the value of each term $P(\mathcal{Y}_*)$.

Estimating counterfactual expression of special forms, such as effect estimation, has been extensively discussed. However, estimating the general case of $P(\mathcal{Y}_*)$ is less frequently addressed due to several challenges:

1. According to definition, it requires the computation of a Lebesgue integral;
2. Exact inference in discrete settings has exponential complexity[1];
3. Marginal inference in SCMs is NP-hard[2].

Thanks to the recent advancements in **counterfactual generation**, a class of models called **neural proxy SCMs**, which leverage neural networks and simple exogenous distributions to simulate the underlying true SCMs, enables tractable approximations of $\mathcal{L}_3$-expressions on the true SCMs.

[1] Y. Han, Y. Chen, and A. Darwiche. On the complexity of counterfactual reasoning.
[2] M. Zečević, D. S. Dhami, and K. Kersting. Not all causal inference is the same.

Monte Carlo Integration

A widely used method for estimating the Lebesgue integral is **Monte Carlo integration.**

- Rejection Sampling

$$P(\mathcal{Y}_*) = \mathbb{E}_{\mathbf{u}\sim Q_\mathbf{U}}\left[\mathbb{1}_{\Omega_\mathbf{U}(\mathcal{Y}_*)}(\mathbf{u})\right] \approx \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\Omega_\mathbf{U}(\mathcal{Y}_*)}(\mathbf{u}^{(i)})$$

  ➔ The indicator limits the efficiency — many samples will be discarded.

- Importance Sampling

$$P(\mathcal{Y}_*) = \mathbb{E}_{\mathbf{u}\sim Q_\mathbf{U}}\left[\sigma_{\mathcal{Y}_*}(\mathbf{u})\right] \approx \frac{1}{n}\sum_{i=1}^{n}\sigma_{\mathcal{Y}_*}(\mathbf{u}^{(i)}), \quad \text{let } \sigma_{\mathcal{Y}_*}(\mathbf{u}) = \frac{p(\mathbf{u})}{q(\mathbf{u})}\mathbb{1}_{\Omega_\mathbf{U}(\mathcal{Y}_*)}(\mathbf{u})$$

  ➔ May allow more samples to contribute, but variance depends on the proposal distribution.

$$\mathbb{V}_{\mathbf{u}\sim Q_\mathbf{U}}\left[\sigma_{\mathcal{Y}_*}(\mathbf{u})\right] = n\cdot\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_{\mathcal{Y}_*}(\mathbf{u}^{(i)})\right] = \mathbb{E}_{\mathbf{u}\sim P_\mathbf{U}}\left[\sigma_{\mathcal{Y}_*}(\mathbf{u})\right] - P^2(\mathcal{Y}_*)$$

  ➔ Learning good proposal to reduce variance.
  - Optimize cross-entropy
  - Optimize χ^2 divergence
  - …

Exogenous Matching

A learned proposal is typically "one-off," meaning that a proposal used to estimate $P(\mathcal{Y}_*^{(1)})$ is not applicable to $P(\mathcal{Y}_*^{(2)})$, which is inefficient when estimating expressions involving multiple $P(\mathcal{Y}_*)$ terms. We are interested in exploring the following questions:

- ➔ **Q1.** Can we train a universal proposal for *different counterfactual events* $\mathcal{Y}_*^{(1)}, \mathcal{Y}_*^{(2)}, \cdots \subseteq \Omega_{\mathbf{Y}_*}$?
- ➔ **Q2.** Can we train a universal proposal even for *different counterfactual variables* $\mathbf{Y}_*^{(1)}, \mathbf{Y}_*^{(2)}, \ldots$?

- Exogenous Matching (EXOM)

$$\arg\min_{\textcircled{1}\ q} -\mathbb{E}_{s\sim Q_{\mathcal{S}}}\left[\mathbb{E}_{\mathbf{u}\sim P_{\mathbf{U}}}\left[\log q(\mathbf{u}\,|\,\mathbf{Y}_*^{(s)}(\mathbf{u}))\right]\right]$$

❶ the proposal distribution
❷ the exogenous distribution
❸ the prior state distribution
❹ potential outcomes

Exogenous Matching

A learned proposal is typically "one-off," meaning that a proposal used to estimate $P(\mathcal{Y}_*^{(1)})$ is not applicable to $P(\mathcal{Y}_*^{(2)})$, which is inefficient when estimating expressions involving multiple $P(\mathcal{Y}_*)$ terms. We are interested in exploring the following questions:

- ➔ **Q1.** Can we train a universal proposal for *different counterfactual events* $\mathcal{Y}_*^{(1)}, \mathcal{Y}_*^{(2)}, \cdots \subseteq \Omega_{\mathbf{Y}_*}$?
- ➔ **Q2.** Can we train a universal proposal even for *different counterfactual variables* $\mathbf{Y}_*^{(1)}, \mathbf{Y}_*^{(2)}, \ldots$?

- Exogenous Matching (EXOM)

$$\arg\min_{\textcircled{1}\ q} -\mathbb{E}_{s\sim Q_{\mathcal{S}}}\left[\mathbb{E}_{\mathbf{u}\sim P_{\mathbf{U}}}\left[\log q(\mathbf{u}\,|\,\mathbf{Y}_*^{(s)}(\mathbf{u}))\right]\right]$$

❶ the proposal distribution
❷ the exogenous distribution
❸ the prior state distribution
❹ potential outcomes

Variance Upper Bound

Theorem 1 (Variance Upper Bound). Let $\sigma_{\mathcal{Y}_*}(\mathbf{u}) = (p(\mathbf{u})/q(\mathbf{u}\,|\,\mathbf{y}_*))\,\mathbb{1}_{\Omega_{\mathbf{U}}(\mathcal{Y}_*)}(\mathbf{u})$, where $q(\mathbf{u}\,|\,\mathbf{y}_*)$ denote the density of the proposal distribution $Q_{\mathbf{U}|\mathbf{y}_*}$, and let $\mathbf{Y}_*(\mathbf{u})$ be the potential response w.r.t. $\mathbf{u}$. If for any $\mathbf{y}_* \in \mathcal{Y}_*$, there exists $\kappa \geq 1$ such that $1/\kappa \leq p(\mathbf{u})/q(\mathbf{u}\,|\,\mathbf{y}_*) \leq \kappa$ holds almost surely on the support $\Omega_{\mathbf{U}}(\mathcal{Y}_*)$, then for any $Q_{\mathbf{U}|\mathbf{y}_*}$ where $\mathbf{y}_* \in \mathcal{Y}_*$:

$$\mathbb{V}_{\mathbf{u}\sim Q_{\mathbf{U}|\mathbf{y}_*}}\left[\sigma_{\mathcal{Y}_*}(\mathbf{u})\right] \leq -\mathbb{E}_{\mathbf{u}\sim P_{\mathbf{U}}}\left[\log q(\mathbf{u}\,|\,\mathbf{Y}_*(\mathbf{u}))\right] + c, \tag{6}$$

where the constant c is solely dependent on κ and $P_{\mathbf{U}}$.

- Use conditional distribution $Q_{\mathbf{U}|\mathbf{y}_*}$ where $\mathbf{y}_* \in \mathcal{Y}_*$, as proposal for different counterfactual events.
- If importance weights are bounded, then for any $\mathbf{y}_* \in \mathcal{Y}_*$, any estimator using $Q_{\mathbf{U}|\mathbf{y}_*}$ as proposal will share a common variance upper bound independent of $\mathbf{y}_*$.

Expected Variance Upper Bound

Corollary 1 (Expected Variance Upper Bound). Let $Q_{\mathcal{S}}$ denote an arbitrary distribution defined over the state space $\mathcal{S}$ of a stochastic counterfactual process, and let $P_{\mathcal{Y}_*}^{(s)}$ denote an arbitrary distribution defined over the σ-algebra $\Sigma_{\mathbf{Y}_*}^{(s)}$ corresponding to a set of counterfactual variables $\mathbf{Y}_*^{(s)}$ given a state $s \in \mathcal{S}$. $P_{\mathcal{Y}_*}$ is the joint distribution induced by $P_{\mathcal{Y}_*}^{(s)}$ and $Q_{\mathcal{S}}$. If the conditions in Thm. 1 are met for any $\mathcal{Y}_*^{(s)}$ and any $s \in \mathcal{S}$, then for any $Q_{\mathbf{U}|\mathbf{y}_*}$ where $\mathbf{y}_* \in \mathcal{Y}_*^{(s)}$ and any $s \in \mathcal{S}$:

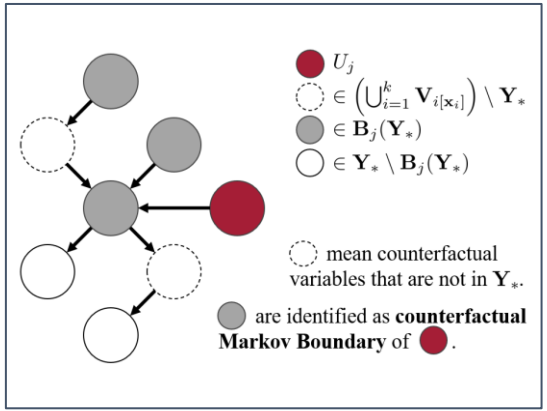
$$\mathbb{E}_{\mathcal{Y}_*^{(s)}\sim P_{\mathcal{Y}_*}}\left[\mathbb{V}_{\mathbf{u}\sim Q_{\mathbf{U}|\mathbf{y}_*}}\left[\sigma_{\mathcal{Y}_*}(\mathbf{u})\right]\right] \leq -\mathbb{E}_{s\sim Q_{\mathcal{S}}}\left[\mathbb{E}_{\mathbf{u}\sim P_{\mathbf{U}}}\left[\log q(\mathbf{u}\,|\,\mathbf{Y}_*^{(s)}(\mathbf{u}))\right]\right] + c, \tag{7}$$

where the constant c is the same as in Thm. 1.

- Use conditional distribution $Q_{\mathbf{U}|\mathbf{y}_*}$, where $\mathbf{y}_* \in \mathcal{Y}_*^{(s)}$ and $s \in \mathcal{S}$, as proposal for different counterfactual events and variables.
- Same conclusion extended to different counterfactual variables.

Counterfactual Markov Boundary

- To model the conditional distribution $Q_{\mathbf{U}|\mathbf{y}_*}$ we aim to infer information about the exogenous variables $\mathbf{U}$ from the counterfactual variables $\mathbf{Y}_*$.
- We seek to identify which variables in the counterfactual variables $\mathbf{Y}_*$ are informative to the exogenous variable U_j. The smallest subset $\mathbf{B}_j(\mathbf{Y}_*) \subseteq \mathbf{Y}_*$ encoding complete information is referred to as the **counterfactual Markov boundary**.
- Masking out information from variables that are not part of the Markov boundary, akin to feature selection, intuitively improves performance.



An example of counterfactual Markov boundary.

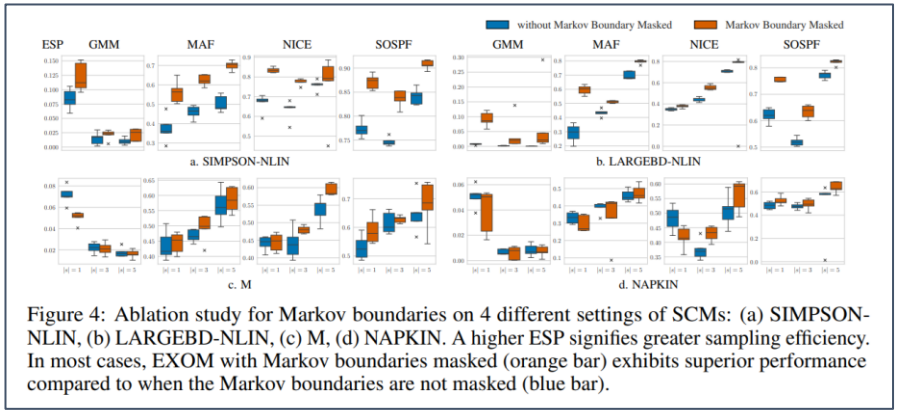


Figure 4: Ablation study for Markov boundaries on 4 different settings of SCMs: (a) SIMPSON-NLIN, (b) LARGEBD-NLIN, (c) M, (d) NAPKIN. A higher ESP signifies greater sampling efficiency. In most cases, EXOM with Markov boundaries masked (orange bar) exhibits superior performance compared to when the Markov boundaries are not masked (blue bar).

Ablation study on counterfactual Markov boundary for EXOM.

Results

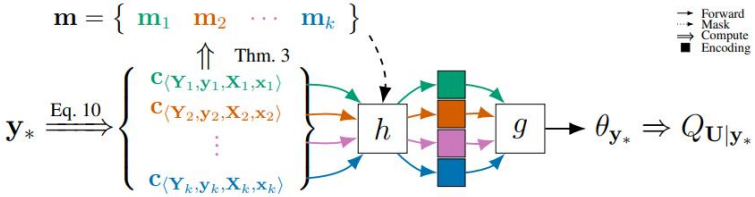


Figure 2: Overview of the conditioning and masking process. $\mathbf{y}_*$ serves as the input to the entire process, $\mathbf{m}$ represents the inferred mask, and the vectorized parameters $\theta_{\mathbf{y}_*}$ of the proposal distribution $Q_{\mathbf{U}|\mathbf{y}_*}$ are the output. Different colors represent information from different submodels. Both h and g represent neural networks.

Overview of the conditioning and masking process.

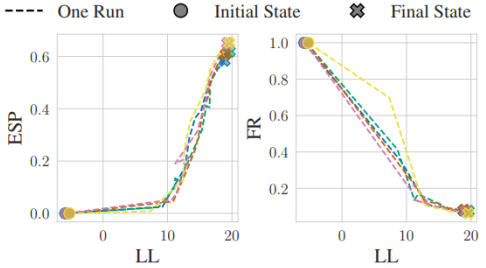


Figure 3: LL (negative Eq. 8) and ESP, FR on SIMPSON-NLIN. As LL increases, ESP increases while FR decreases, until convergence.

An example of the convergence.

Results

Table 1: Comparison of RS, CEIS, NIS, and EXOM (ours) across 3 different SCMs, with $|s| = 1, 3, 5$. Among the three SCMs, SIMPSON-NLIN is Markovian diffeomorphic, NAPKIN is Semi-Markovian continuous, and FAIRNESS-XW is Regional canonical. The results are averaged over 5 runs. Higher ESP and lower FR indicate better performance.

		SIMPSON-NLIN		NAPKIN		FAIRNESS-XW	
$\|s\|$	Model	ESP ↑	FR ↓	ESP ↑	FR ↓	ESP ↑	FR ↓
1	RS	0.008	0.971	0.006	0.967	0.156	0.065
	CEIS	0.037	0.809	0.017	0.911	0.255	0.532
	NIS	0.008	0.961	0.007	0.965	0.193	0.527
	EXOM[GMM]	0.117	0.245	0.039	0.476	**0.358**	0.009
	EXOM[MAF]	**0.581**	**0.005**	**0.306**	**0.066**	0.339	**0.007**
3	RS	0.000	1.000	0.000	1.000	0.038	0.281
	CEIS	0.000	1.000	0.000	1.000	0.122	0.704
	NIS	0.000	1.000	0.000	1.000	0.116	0.686
	EXOM[GMM]	0.024	0.500	0.011	0.613	0.247	0.056
	EXOM[MAF]	**0.606**	**0.075**	**0.397**	**0.183**	**0.261**	**0.043**
5	RS	0.000	1.000	0.000	1.000	0.030	0.368
	CEIS	0.000	1.000	0.000	1.000	0.106	0.727
	NIS	0.000	1.000	0.000	1.000	0.094	0.724
	EXOM[GMM]	0.020	0.497	0.009	0.644	0.231	0.084
	EXOM[MAF]	**0.698**	**0.031**	**0.482**	**0.094**	**0.237**	**0.070**

Comparison of EXOM with other methods.

Table 2: Estimation of counterfactual densities on CausalNF and counterfactual effects on NCM. Here, "O" represents the original SCM, and "P" represents the proxy SCM. For SIMPSON-NLIN, the proxy SCM is CausalNF, and the metric used is FR ("-" indicates FR equals 1); whereas for FAIRNESS, the proxy SCM is NCM, and the metric used is the average bias w.r.t. the ground truth. The subscript denotes the 95% CI error bound over 5 trials. For more details, see App. C.9

		SIMPSON-NLIN			FAIRNESS			
Method	SCM	$\|s\| = 1$	$\|s\| = 3$	$\|s\| = 5$	ATE	ETT	NDE	CtfDE
RS	O	$0.89_{0.014}$	-	-	$0.01_{0.013}$	$0.01_{0.018}$	$0.01_{0.015}$	$0.01_{0.020}$
	P	$0.90_{0.012}$	-	-	$0.01_{0.013}$	$0.01_{0.021}$	$0.01_{0.014}$	$0.01_{0.023}$
EXOM[MAF]	O	$0.00_{0.005}$	$0.02_{0.015}$	$0.01_{0.021}$	$0.04_{0.095}$	$0.06_{0.168}$	$0.04_{0.131}$	$0.07_{0.236}$
	P	$0.01_{0.005}$	$0.40_{0.012}$	$0.61_{0.016}$	$0.01_{0.013}$	$0.01_{0.024}$	$0.01_{0.013}$	$0.01_{0.044}$
EXOM[NICE]	O	$0.00_{0.006}$	$0.02_{0.016}$	$0.01_{0.046}$	$0.01_{0.018}$	$0.01_{0.030}$	$0.01_{0.020}$	$0.01_{0.039}$
	P	$0.01_{0.005}$	$0.40_{0.013}$	$0.61_{0.018}$	$0.01_{0.017}$	$0.01_{0.021}$	$0.01_{0.012}$	$0.01_{0.022}$

Results of EXOM on synthetic datasets and proxy SCMs.

THANKS

Code is available: https://github.com/cyisk/exom

Yikang Chen
https://cyisk.github.io
51265902150@stu.ecnu.edu.cn