Natural Counterfactuals With Necessary Backtracking

Guang-Yuan Hao*, Jiji Zhang*, Biwei Huang, Hao Wang, Kun Zhang

Background

Consider a scenario: A sudden brake of a high-speed bus caused Tom (cat) to fall and injure Jerry (mouse).

Non-backtracking counterfactuals:
If **A** had been $\boldsymbol{a}^*$ what would the value of **B**? For example, if Tom had stood still (despite the sudden braking), then Jerry would not have been injured.

Figure 1. Motivation Example

The problem of non-backtracking counterfactuals:

- Surgical interventions are sometimes so removed from what are or can be. Preventing Tom's fall in a sudden braking scenario requires defying mechanisms that are difficult or even physically impossible to disrupt.
- As a result, there are likely to be no data points in the reservoir of observed scenarios that are consistent with a person standing still during a sudden braking.

Motivation: Naturalness

Our New Notion of **Natural Counterfactuals:** Allow a certain amount of backtracking, to keep the counterfactual scenario "natural" with respect to the available observations.

- **Non-Backtracking Counterfactuals** interpret a statement like "if **A** had been $\boldsymbol{a}^*$ …" as "if **A** had been $\boldsymbol{a}^*$, while keeping all upstream variables unchanged…" This implies that we impose the change on A without considering how earlier causes of **A** might need to be altered to accommodate this change. The focus is on an isolated alteration of **A**, holding all prior conditions fixed.
- **Natural Counterfactuals** interpret "if **A** had been $\boldsymbol{a}^*$ …" as "if **A** had been $\boldsymbol{a}^*$, while allowing small adjustments upstream to ensure naturalness." In this approach, changes to **A** are made along with slight modifications in its upstream causes, so that the scenario feels natural. *For the example above, a more natural counterfactual scenario to realize the change to not-falling would involve changing at the same time some causally preceding events, such as changing the sudden braking to gradually slowing down.*

Preliminaries

Structural Causal Model (SCM): A SCM $\mathcal{M} := \langle \mathbf{U}, \mathbf{V}, \mathbf{f}, p(\mathbf{U}) \rangle$

$$\mathbf{V}_i := f_i(\mathbf{PA}_i, \mathbf{U}_i), \quad i = 1, ..., N \tag{1}$$

Local Mechanisms: $p(\mathbf{V}_i | \mathbf{PA}_i)$ for $i = 1, ..., N$

Three-Step Procedure of Non-Backtracking Counterfactuals:.
A general counterfactual question takes the following form: given evidence **E = e**, what would the value of **B** have been if **A** had been a^*?
1. Update **p(U)** as **p(U|E=e)**;
2. Modify Causal Model as $\boldsymbol{M_A}$;
3. Do inference on < **p(U|E=e)**, $\boldsymbol{M_A}$ >

A Framework for Natural Counterfactuals

***Do(·)* and *Change(·)* Operators:** We use *Change*(**A**= $\boldsymbol{a}^*$) to indicate setting **A** to $\boldsymbol{a}^*$ in our natural counterfactuals.

Natural Counterfactuals:
1. **Minimal Change**: Counterfactual data point should be as close to the actual data point as possible.
2. **Necessary Backtracking**: Allow Necessary backtracking to achieve *Change*(**A**=$\boldsymbol{a}^*$), i.e., variables in A's causal upstream need to change together with **A**;
3. **Naturalness**: the counterfactual scenario is kept within the relevant support by necessary backtracking.

Feasible Intervention Optimization (FIO):

$$\begin{aligned} \underset{an(\mathbf{A})^*}{\text{minimize}} \quad & D(an(\mathbf{A}), an(\mathbf{A})^*) && \textbf{\textit{Distance for minimal change}} \\ \text{s.t.} \quad & \mathbf{A} = \mathbf{a}^*, && \textbf{\textit{Change}(A= }\boldsymbol{a}^*\textbf{)} \\ & g_n(an(\mathbf{A})^*) > \epsilon. && \textbf{Naturalness constraint} \end{aligned} \tag{2}$$

Identifiable Natural Counterfactuals

Theorem 4.1 (Identifiable Natural Counterfactuals)**.** *Given the causal graph and the joint distribution over* $\mathbf{V}$*, suppose* $\mathbf{V}_i$ *satisfies the following structural causal model:* $\mathbf{V}_i := f_i(\mathbf{PA}_i, \mathbf{U}_i)$ *for any* $\mathbf{V}_i \in \mathbf{V}$*, assume every* f_i*, though unknown, is smooth and strictly monotonic w.r.t.* $\mathbf{U}_i$ *for fixed values of* $\mathbf{PA}_i$*. Then, given an actual data point* $\mathbf{V} = \mathbf{v}$*, with a LBF intervention* $do(\mathbf{C} = \mathbf{c}^*)$ *(satisfying the criterion of* ϵ*-natural generation), the counterfactual instance* $\mathbf{V} = \mathbf{v}^*$ *is identifiable:* $\mathbf{V} = \mathbf{v}^* | do(\mathbf{C} = \mathbf{c}^*), \mathbf{V} = \mathbf{v}$.

Experiments

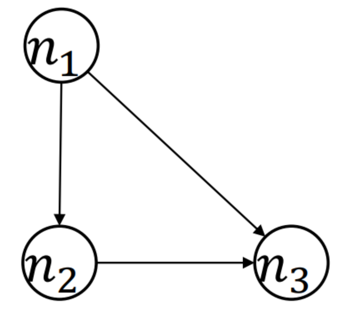
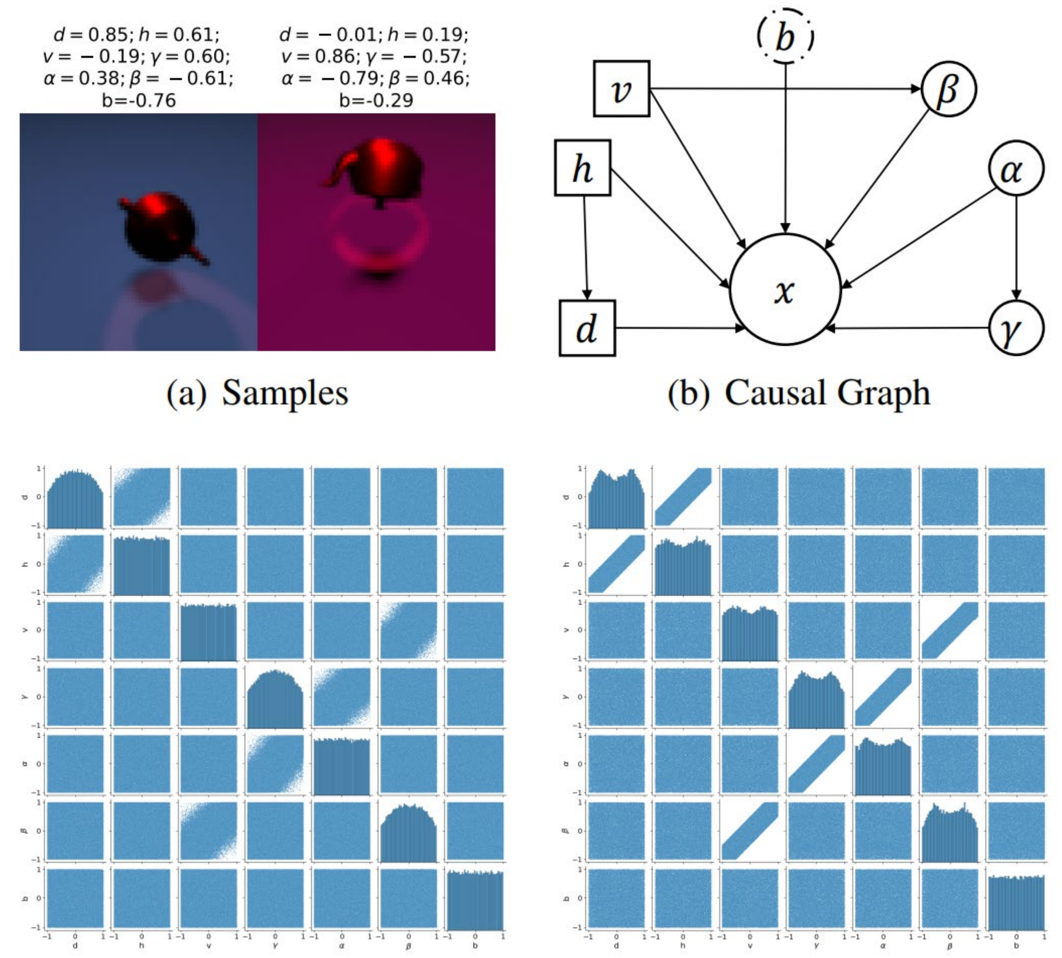


Figure 2. Causal Graph of *Toy 1*

Table 1: MAE Results on Toy 1 (Lower MAE is better). To save room, we also write "do" for "change" for natural counterfactuals.

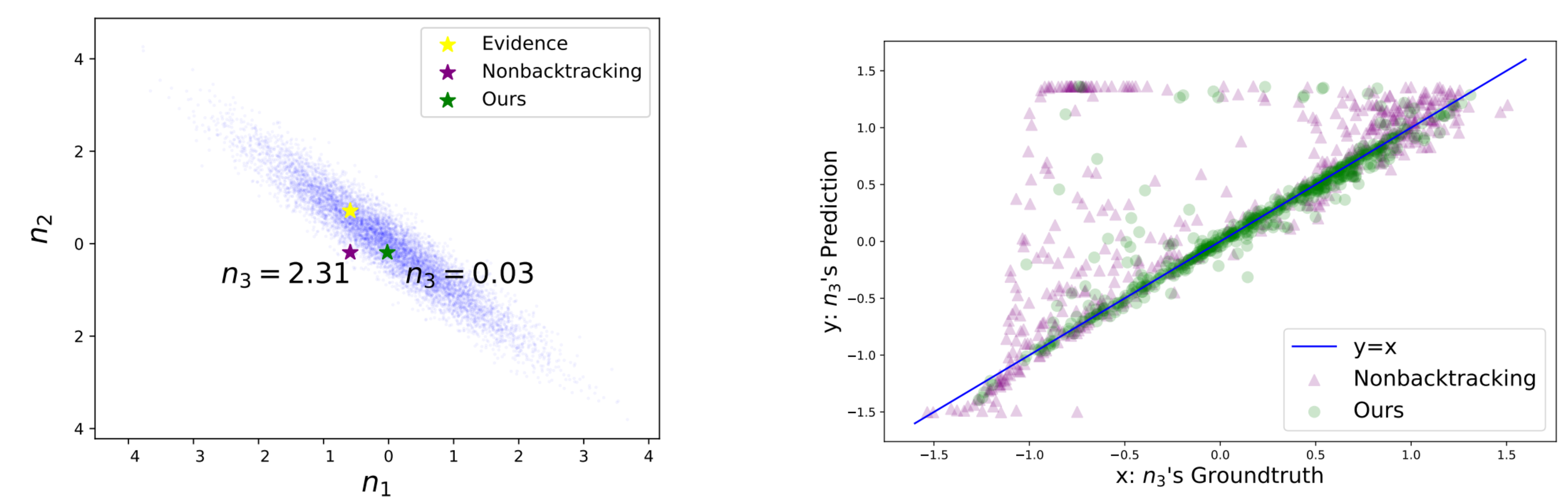
Dataset	*Toy 1*		
do or *change*	do(n_1)		do(n_2)
Outcome	n_2	n_3	n_3
Nonbacktracking	0.477	0.382	0.297
Ours	0.434	0.354	0.114



(a) Samples (b) Causal Graph

(c) Weak-3DIdent (d) Strong-3DIdent

Figure 4. 3DIdentBOX



(a) Outcome error on a single sample (b) Groudtruth-Prediction Scatter Plot

Figure 3. Visualization on *Toy 1*

Table 3: MAE Results on Weak-3DIdent and Strong-3DIdent (abbreviated as "Weak" "Strong" for simplicity). Lower MAE is better. For clarity, we use "Non" to denote Nonbacktracking.

Dataset	-	d	h	v	γ	α	β	b
Weak	Non	0.025	0.019	0.035	0.364	0.27	0.077	0.0042
	Ours	0.024	0.018	0.034	0.349	0.221	0.036	0.0041
Stong	Non	0.100	0.083	0.075	0.387	0.495	0.338	0.0048
	Ours	0.058	0.047	0.050	0.298	0.316	0.139	0.0047

Guang-Yuan will join virtually. For questions, contact guangyuanhao@outlook.com.