Algorithmic Recourse Under Incomplete Causal Graph

Algorithmic Recourse: from Counterfactual Explanations to Interventions

ACM 2021

Algorithmic Recourse

The systematic process of reversing unfavourable decisions by algorithms and bureaucracies across a range of counterfactual scenarios.

Problem Formulation

$$x^{*\mathrm{CFE}} \in \underset{x}{argmin} \quad dist(x, x^F) \quad s.t.\ h(x) \neq h(x^F), x \in \mathcal{P}$$

Contd.

Ustun et al.

$$\delta^* \in \underset{\delta}{argmin} \quad cost(\delta; x^F) \quad s.t.\ h(x^{\mathrm{CFE}}) \neq h(x^{\mathrm{F}}),$$

$$x^{\mathrm{CFE}} = x^{\mathrm{F}} + \delta,$$

$$x^{\mathrm{CFE}} \in \mathcal{P},\ \delta \in \mathcal{F}$$

Why does the above formulation fail?

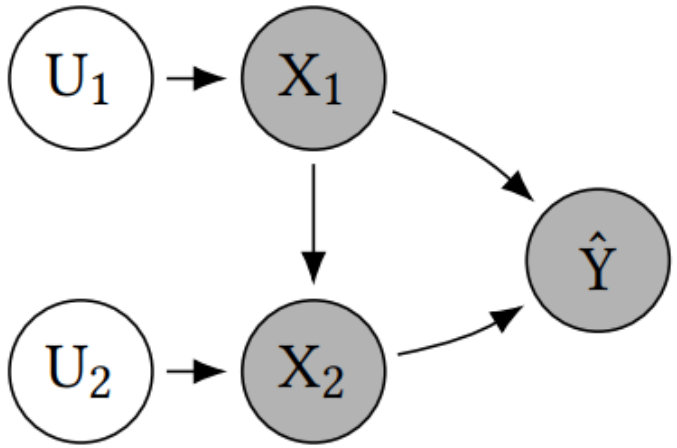
- Example: Consider an example where an individual has an annual salary of $75,000 and an account balance $25,000. Say the model is

$$h = sgn(X_1 + 5X_2 - \$225,000)$$

- The counterfactual explanation could be
    i. Annual salary: $100,000 or
    ii. Bank Balance: $30,000

In a world where home-seekers save 30% of their salary, a salary increase of 14% would result in positive decision of loan-granting algorithm.

Structural Causal Model



$$\left.\begin{aligned} X_1 &:= U_1 \\ X_2 &:= f_2(X_1) + U_2 \end{aligned}\right\} \mathcal{M}$$

$$\hat{Y} = h(X_1, X_2)$$

Proposition-1

A CFE-based action, $\mathbf{A}^{CFE}$, where $I = \{i | \delta_i^* \neq 0\}$, performed by individual x^F, in general results in the structural counterfactual, $x^{SCF} = x^{*CFE} := x^F + \delta^*$, if and only if, the set of descendants of the acted upon variables, determined by I, is the empty set.

Corollary-1

If the true world $\mathcal{M}$ is independent, i.e. all the observed features are root-nodes, then CFE-based actions always guarantee recourse.

Causal Persepective

$$
\begin{aligned}
\mathbf{A}^* \in \underset{A}{argmin} \quad & cost(\mathbf{A}; x^F) \\
s.t. \quad & h(x^{\mathrm{SCF}}) \neq h(x^{\mathrm{F}}), \\
& x^{\mathrm{SCF}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-\mathbb{1}}(\mathbb{x}^{\mathrm{F}})) \qquad x^{\mathrm{SCF}} \in \mathcal{P},\ \mathbf{A} \in \mathcal{F}
\end{aligned}
$$

Proposition-2

Given an individual x^F observed in world $\mathcal{M}$, a family of feasible actions $\mathcal{F}$. Assume that there exists CFE-based actions $A^{CFE} \in \mathcal{F}$ that achieves recourse, i.e., $h(x^F) \neq h(x^{*CFE})$. Then, $cost(A^*; x^F) \leq cost(A^{CFE}; x^F)$

Algorithm

1. **Abduction:** uniquely determines the values of all exogenous variables.
2. **Action:** modify the SCM according to the hypothetical interventions. ($\mathbb{F}_A$)
3. **Prediction:** Determine the values of of all endogenous variables.

The assignment of structural counterfactual values can generally be written as

$$\begin{aligned} x_i^{\mathrm{SCF}} =& [i \in I] \cdot (x_i^{\mathrm{F}} + \delta_i) \\ &+ [i \notin I] \cdot (x_i^{\mathrm{F}} + f_i(\mathbf{pa}_i^{\mathrm{SCF}}) + f_i(\mathbf{pa}_i^{\mathrm{F}})) \end{aligned}$$

Limitations

1. The underlying causal model is rarely known in practice.
2. The assumption of one-to-one mapping from real world actions to interventions on endogenous variables may not hold.

Limitations

1. The underlying causal model is rarely known in practice.
2. The assumption of one-to-one mapping from real world actions to interventions on endogenous variables may not hold.

Algorithmic Recourse under imperfect causal knowledge: A probabilistic approach

NeurIPS 2020

Setting

The causal graph is known but the structural equations are not known.

No Recourse Guarantees for unknown structural equations

Example: *Consider the following two SCMs $\mathcal{M}_{\mathcal{A}}$ and $\mathcal{M}_{\mathcal{B}}$.*

$$\mathbf{S} = \left\{ \begin{array}{l} X_1 := f_1(U_1), \\ X_2 := f_2(X_1, U_2), \\ X_3 := f_3(X_1, X_2, U_3) \end{array} \right\}$$

$$P_{\mathbf{U}} = P_{U_1} \times P_{U_2} \times P_{U_3}$$

$$\text{(b) } \mathcal{M} = (\mathbf{S}, P_{\mathbf{U}})$$

Contd.

Choose $U_1, U_2 \sim Bernoulli(0.5)$ *and* $U_3 \sim Uniform(\{0, \ldots, K\})$ *independently in both* $\mathcal{M}_{\mathcal{A}}$ *and* $\mathcal{M}_{\mathcal{B}}$ *with structural equations*

$$
\begin{aligned}
X_1 &:= U_1, && \textit{in} \quad \{\mathcal{M}_A, \mathcal{M}_B\},\\
X_2 &:= X_1(1 - U_2), && \textit{in} \quad \{\mathcal{M}_A, \mathcal{M}_B\},\\
X_3 &:= \mathbb{I}_{X_1 \neq X_2}(\mathbb{I}_{U_3 > 0} X_1 + \mathbb{I}_{U_3 = 0} X_2) + \mathbb{I}_{X_1 = X_2} U_3, && \textit{in} \quad \mathcal{M}_A,\\
X_3 &:= \mathbb{I}_{X_1 \neq X_2}(\mathbb{I}_{U_3 > 0} X_1 + \mathbb{I}_{U_3 = 0} X_2) + \mathbb{I}_{X_1 = X_2}(K - U_3), && \textit{in} \quad \mathcal{M}_B.
\end{aligned}
$$

Then $\mathcal{M}_{\mathcal{A}}$ *and* $\mathcal{M}_{\mathcal{B}}$ *both imply exactly the same observational and interventional distributions, and thus indistiguishable from empirical data. However, having observed* $x^{\mathrm{F}} = (1, 0, 0)$*, they predict different counterfactuals had* X_1 *been 0, i.e.,* $x^{\mathrm{SCF}}(X_1 = 0) = (0, 0, 0)$ *and* $(0, 0, K)$*, respectively.*

Proposition

Unless the set of descendants of interevend-upon variables in empty, algorithmic recourse can in general, be guaranteed **only** if the true structural equations are known, irrespective of the amount and type of available data.

Method

The learned causal model could be imperfect due to finite sample of the observed data, or due to model misspecification.

The authors adopt a Bayesian approach to account for the uncertainty in the estimation of the structural equations. They assume additive gaussian noise and rely on probabilistic regression using a Gaussian process (GP) prior over the functions f_r.

GP-SCM

$$X_r := f_r(\mathbf{X}_{pa(r)}) + U_r$$
$$where \quad f_r \sim \mathcal{GP}(0, k_r), \quad U_r \sim \mathcal{N}(0, \sigma_r^2), \quad r \in [d]$$

GP-SCM noise posterior

$$u_r|\mathbf{X}_{pa(r)}, \mathbf{x}_r \sim \mathcal{N}\left(\sigma_r^2\left(\mathbf{K}+\sigma_r^2\mathbf{I}\right)^{-1}\mathbf{x}_r, \sigma_r^2\left(\mathbf{I}-\sigma_r^2\left(\mathbf{K}+\sigma_r^2\mathbf{I}\right)^{-1}\right)\right)$$

$$where \quad \mathbf{K} := \left(k_r\left(\mathbf{x}_{pa(r)}^i, \mathbf{x}_{pa(r)}^j,\right)\right)_{ij} denotes\ the\ Gram\ matrix$$

GP-SCM counterfactual distribution

$$X_r\left(\mathbf{X}_{pa(r)=\tilde{x}}\right)|\mathbf{x}^F, \{\mathbf{x}^i\}_{i=1}^n \sim \mathcal{N}\left(\mu^F + \tilde{\mathbf{k}}^T\left(\mathbf{K}+\sigma_r^2\mathbf{I}\right)^{-1}\mathbf{x}_r, s_r^F + \tilde{k} - \tilde{\mathbf{k}}^T\left(\mathbf{K}+\sigma_r^2\mathbf{I}\right)^{-1}\tilde{\mathbf{k}}\right)$$

$$where \quad \tilde{k} := k_r\left(\tilde{\mathbf{x}}_{pa(r)}, \tilde{\mathbf{x}}_{pa(r)}\right), \tilde{\mathbf{k}} := \left(k_r\left(\tilde{\mathbf{x}}_{pa(r)}, \mathbf{x}_{pa(r)}^1\right), \ldots, k_r\left(\tilde{\mathbf{x}}_{pa(r)}, \mathbf{x}_{pa(r)}^n\right)\right)$$

Probabilistic version of the individualised recourse

$$\min_{a=do(\mathbf{x}_{\mathcal{I}}=\theta)\in\mathbb{A}} \quad \text{cost}^{F}(a) \quad \text{subject to} \quad \mathbb{E}_{\mathbf{X}^{SCF}(a)}\left[h\left(\mathbf{X}^{SCF}(a)\right)\right] \geq \text{thresh(a)}$$

Algorithm for optimization

Brute-force approach

A way to solve the objective is

1. Iterate over $a \in \mathbb{A}^F$
2. Approximately evaluate the constraint via Monte Carlo
3. Select a minimum cost action amongst all evaluated candidates

Gradient-based approach

$$\mathcal{L}(\theta, \lambda) := \text{cost}^F(a) + \lambda \left(\text{thresh}(a) - \mathbb{E}_{\mathbf{X}_{d(\mathcal{I})|\theta}} \left[h \left(\mathbf{x}^F_{nd(\mathcal{I})}, \theta, \mathbf{X}_{d(\mathcal{I})} \right) \right] \right)$$

Since GP-SCM counterfactual admit reparametrisation trick we get

$$\nabla_\theta \mathbb{E}_{\mathbf{X}_{d(\mathcal{I})|\theta}} \left[h \left(\mathbf{x}^F_{nd(\mathcal{I})}, \theta, \mathbf{X}_{d(\mathcal{I})} \right) \right] = \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0,1)} \left[\nabla_\theta h \left(\mathbf{x}^F_{nd(\mathcal{I})}, \theta, \mathbf{x}_{d(\mathcal{I})}(\mathbf{z}) \right) \right]$$

Results

Table 1: Experimental results for the gradient-based approach on different 3-variable SCMs. We show average performance ± 1 standard deviation for $N_{\text{runs}} = 100$, $N_{\text{MC-samples}} = 100$, and $\gamma_{\text{LCB}} = 2$.

Method	LINEAR SCM			NON-LINEAR ANM			NON-ADDITIVE SCM		
	Valid$_\star$ (%)	LCB	Cost (%)	Valid$_\star$ (%)	LCB	Cost (%)	Valid$_\star$ (%)	LCB	Cost (%)
$\mathcal{M}_\star$	100	-	10.9±7.9	100	-	20.1±12.3	100	-	13.2±11.0
$\mathcal{M}_{\text{LIN}}$	100	-	11.0±7.0	54	-	20.6±11.0	98	-	14.0±13.5
$\mathcal{M}_{\text{KR}}$	90	-	10.7±6.5	91	-	20.6±12.5	70	-	13.2±11.6
$\mathcal{M}_{\text{GP}}$	100	.55±.04	12.2±8.3	100	.54±.03	21.9±12.9	95	.52±.04	13.4±12.8
$\mathcal{M}_{\text{CVAE}}$	100	.55±.07	11.8±7.7	97	.54±.05	22.6±12.3	95	.51±.01	13.4±12.2
CATE$_\star$	90	.56±.07	11.9±9.2	97	.55±.05	26.3±21.4	100	.52±.02	13.5±13.0
CATE$_{\text{GP}}$	93	.56±.05	12.2±8.4	94	.55±.06	25.0±14.8	94	.52±.03	13.2±13.1
CATE$_{\text{CVAE}}$	89	.56±.08	12.1±8.9	98	.54±.05	26.0±14.3	100	.52±.05	13.6±12.9

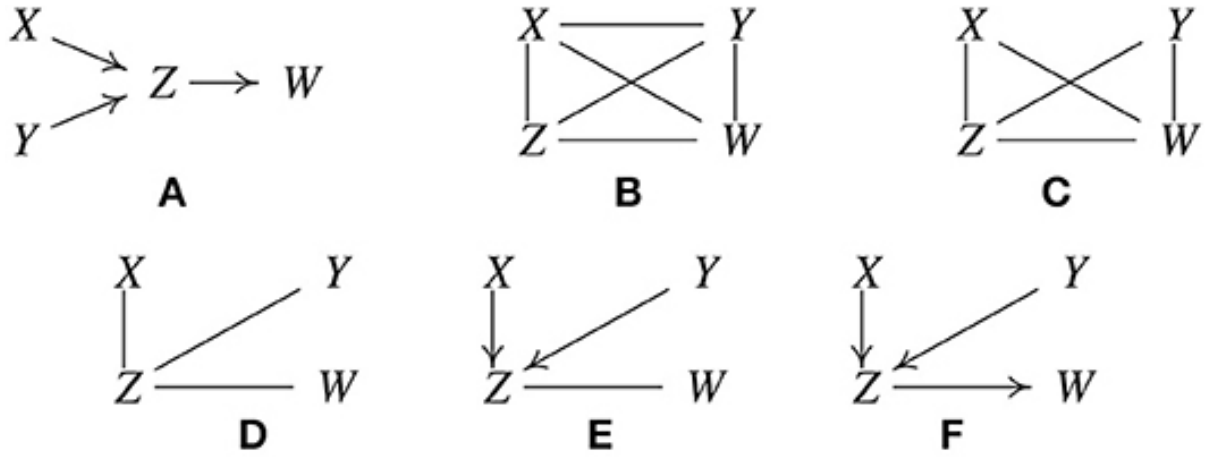
Limitations

1. Complete causal graph should be known.
2. The assumption of one-to-one mapping from real world actions to interventions on endogenous variables may not hold.

Towards Incomplete Causal Graph

PC Algorithm

The key idea is that two statistically independent variables are not causally linked.



1. (B) Start with fully connected graph
2. (C), (D) remove edges b/w statistically independent variables.
3. (E) Orient Colliders
4. (F) Additional Constraints.

Greedy Equivalence Search (GES)

The Greedy Equivalence Search (GES) algorithm uses this trick. GES starts with an empty graph and iteratively adds directed edges to maximize the improvement in a model fitness measure (i.e. score). An example score is the Bayesian Information Criterion (BIC)

Linear Non-Gaussian Assumption

All structural equation (causal mechanisms that generate the data) are of the following form:

$$Y := f(X) + U$$

where f is a linear function, $X \perp\!\!\!\perp U$, and U is distributed as some non-Gaussian

Nonlinear Additive Noise Setting

$$X_i := f_i(pa_i) + U_i \quad \text{where } f_i \text{ is nonlinear}$$

Theorem (Hoyer et al. 2008): Under the Markov assumption, causal sufficiency, acyclicity, the nonlinear additive noise assumption, and a technical condition from Hoyer et al. 2008, we can identify the causal graph.

Post-Nonlinear Setting

Nonlinear additive noise setting: $Y := f(x) + U, \quad X \perp\!\!\!\perp U$

Post-nonlinear:

$$Y := g(f(X) + U), \quad X \perp\!\!\!\perp U$$

References