Causality with Robust Machine Learning

清华大学 刘家硕

THU Trustworthy-AI Group

2020 年 11 月 11 日

1. Invariant Causal Prediction
2. Distributionally Robust Learning
3. Relationships
4. Conclusions

Invariant Causal Prediction

Paper: Causal inference using invariant prediction: identification and confidence intervals[1]

- Main idea: use the invariance of the causal relationships under different environments for causal inference.
- Problem setting: data (X^e, Y^e) from different environments $e \in \mathcal{E}$ with unknown experimental conditions or type of interventions.
- Invariance Assumption:

Assumption 1 (Invariant prediction) *There exists a vector of coefficients* $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^t$ *with support* $S^* := \{k : \gamma_k^* \neq 0\} \subseteq \{1, \ldots, p\}$ *that satisfies*

$$\begin{aligned} \text{for all } e \in \mathcal{E}: \quad & X^e \text{ has an arbitrary distribution and} \\ & Y^e = \mu + X^e \gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}, \end{aligned} \tag{3}$$

where $\mu \in \mathbb{R}$ *is an intercept term,* ε^e *is random noise with mean zero, finite variance and the same distribution* F_ε *across all* $e \in \mathcal{E}$.

  - Equivalent to $P(Y^e | X^e_{S*})$ identical for all $e \in \mathcal{E}$.
  - S^* corresponds to what kind of covariates?

[1] Peters, J. , Bühlmann, Peter, & Meinshausen, N. . (2015). Causal inference using invariant prediction: identification and confidence intervals. Stats, 78(5), 947-1012.

Invariant Causal Prediction

- Parents of Y satisfies the invariance assumption:

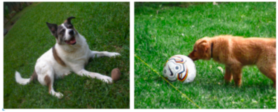
Proposition 1 *Consider a linear structural equation model, as formally defined in Section 4.1, for the variables $(X_1 = Y, X_2, \ldots, X_p, X_{p+1})$, with coefficients $(\beta_{jk})_{j,k=1,\ldots,p+1}$, whose structure is given by a directed acyclic graph. The independence assumption on the noise variables in Section 4.1 can here be replaced by the strictly weaker assumption that $\varepsilon_1^e \perp\!\!\!\perp \{\varepsilon_j^e; j \in \mathbf{AN}(1)\}$ for all environments $e \in \mathcal{E}$, where $\mathbf{AN}(1)$ are the ancestors of Y. Then Assumption 1 holds for the parents of Y, namely $S^* = \mathbf{PA}(1)$, and $\gamma^* = \beta_{1,\cdot}$ as defined in Section 4.1, under the following assumption:*

> *for each $e \in \mathcal{E}$: the experimental setting e arises by one or several interventions on variables from $\{X_2, \ldots, X_{p+1}\}$ but interventions on Y are not allowed; here, we allow for do-interventions [Pearl, 2009] (see also Section 4.2.1, and note that the assigned values can be random, too), or soft-interventions [Eberhardt and Scheines, 2007] (see also Sections 4.2.2 and 4.2.3).*

- Practically, S^* is not unique.

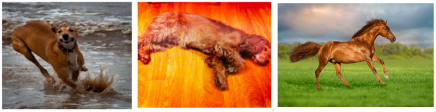
An Example for Demonstration

Env1



Env2

Env3



- When environment set $\mathcal{E}$ contains $Env1$ and $Env2$: grass is invariant.
- When environment set $\mathcal{E}$ contains $Env1$ and $Env3$: grass is variant.

Plausible causal predictors and coefficients

$$H_{0,\gamma,S}(\mathcal{E}): \quad \gamma_k = 0 \text{ if } k \notin S \quad \text{and} \quad \begin{cases} \exists F_\varepsilon \text{ such that for all } e \in \mathcal{E} \\ Y^e = X^e\gamma + \varepsilon^e, \text{ where } \varepsilon^e \perp\!\!\!\perp X_S^e \text{ and } \varepsilon^e \sim F_\varepsilon. \end{cases}$$

- Plausible causal predictors:

  (i) We call the variables $S \subseteq \{1, \ldots, p\}$ plausible causal predictors *under $\mathcal{E}$ if the following null hypothesis holds true:*

  $$H_{0,S}(\mathcal{E}): \quad \exists \gamma \in \mathbb{R}^p \text{ such that } H_{0,\gamma,S}(\mathcal{E}) \text{ is true.} \tag{5}$$

  (ii) The identifiable causal predictors *under interventions $\mathcal{E}$ are defined as the following subset of plausible causal predictors*

  $$S(\mathcal{E}) := \bigcap_{S: H_{0,S}(\mathcal{E}) \text{ is true}} S = \bigcap_{\gamma \in \Gamma(\mathcal{E})} \{k : \gamma_k \neq 0\}. \tag{6}$$

  - $S(\mathcal{E}) \subseteq S^*$
  - larger $\mathcal{E}$, the more $S(\mathcal{E})$
- Plausible causal coefficients

  Definition 2 (Plausible causal coefficients) *We define the set $\Gamma_S(\mathcal{E})$ of* plausible causal coefficients for the set *$S \subseteq \{1, \ldots, p\}$ and the global set $\Gamma(\mathcal{E})$ of* plausible causal coefficients *under $\mathcal{E}$ as*

  $$\Gamma_S(\mathcal{E}) := \{\gamma \in \mathbb{R}^p : \ H_{0,\gamma,S}(\mathcal{E}) \text{ is true}\}, \tag{7}$$

  $$\Gamma(\mathcal{E}) := \bigcup_{S \subseteq \{1,\ldots,p\}} \Gamma_S(\mathcal{E}). \tag{8}$$

Estimation of identifiable causal predictors

Generic method for invariant prediction

1) For each set $S \subseteq \{1, \ldots, p\}$, test whether $H_{0,S}(\mathcal{E})$ holds at level α (we will discuss later concrete examples).

2) Set $\hat{S}(\mathcal{E})$ as

$$\hat{S}(\mathcal{E}) := \bigcap_{S: H_{0,S}(\mathcal{E}) \text{ not rejected}} S. \tag{12}$$

3) For the confidence sets, define

$$\hat{\Gamma}(\mathcal{E}) := \bigcup_{S \subseteq \{1,\ldots,p\}} \hat{\Gamma}_S(\mathcal{E}), \tag{13}$$

where

$$\hat{\Gamma}_S(\mathcal{E}) := \begin{cases} \emptyset & H_{0,S}(\mathcal{E}) \text{ can be rejected at level } \alpha \\ \hat{C}(S) & \text{otherwise.} \end{cases} \tag{14}$$

Here, $\hat{C}(S)$ is a $(1-\alpha)$-confidence set for the regression vector $\beta^{\text{pred}}(S)$ that is obtained by pooling the data.

Identifiability results

Theorem 2 *Consider a (linear) Gaussian SEM as in (19) and (20) with interventions. Then, with $S(\mathcal{E})$ as in (6), all causal predictors are identifiable, that is*

$$S(\mathcal{E}) = \mathbf{PA}(Y) = \mathbf{PA}(1) \tag{22}$$

if one of the following three assumptions is satisfied:

i) The interventions are **do-interventions** *(Section 4.2.1) with $a_j^e \neq E(X_j^1)$ and there is at least one single intervention on each variable other than Y, that is for each $j \in \{2, \ldots, p+1\}$ there is an experiment e with $\mathcal{A}^e = \{j\}$.*

ii) The interventions are **noise interventions** *(Section 4.2.2) with $1 \neq E(A_j^e)^2 < \infty$, and again, there is at least one single intervention on each variable other than Y. If the interventions act additively rather than multiplicatively, we require $EC_j^e \neq 0$ or $0 < \mathrm{Var}\, C_j^e < \infty$.*

iii) The interventions are **simultaneous noise interventions** *(Section 4.2.3). This result still holds if we allow changing linear coefficients $\beta_{j,k}^{e=2} \neq \beta_{j,k}^{e=1}$ in (21) with (possibly random) coefficients $\beta_{j,k}^{e=2}$.*

The statements remain correct if we replace the null hypothesis (10) with its weaker version (16).

1 Invariant Causal Prediction

2 Distributionally Robust Learning

3 Relationships

4 Conclusions

Robust Learning

Robust learning takes the form:

$$\theta = \arg\min_{\theta \in \Theta} \sup_{P \in \mathcal{P}} \mathbb{E}_{X,Y \sim P}[\ell(\theta; X, Y)] \tag{1}$$

where $\mathcal{P}$ denotes the uncertainty set.

- Adversarial Robustness:

$$\min_{\theta} \mathbb{E}_{X,Y \sim P_{tr}}[\sup_{\|\Delta\| \leq \rho} \ell(\theta; X + \Delta, Y)] \tag{2}$$

- Distributional Robustness:

$$\min_{\theta} \sup_{P: D(P, P_{tr}) \leq \rho} \mathbb{E}_{X,Y \sim P}[\ell(\theta; X, Y)] \tag{3}$$

- Remarks:
  - Adversarial robustness is a special case of distributional robustness.
  - The choice of the uncertainty set is important.

Distributionally Robust Optimization(DRO)

Recently, there are mainly two kinds of DRO based on the chosen distance metric:

- f-divergence DRO[2]:

$$D_f(P\|P_{tr}) = \int f(\frac{dP}{dP_{tr}})dP_{tr} \tag{4}$$

- Wasserstein DRO[3]:

$$W_c(P;P_{tr}) = \inf_{M\in\Pi(P;P_{tr})} \mathbb{E}_{Z_1,Z_2\sim M}[c(Z_1,Z_2)] \tag{5}$$

[2] Duchi, J. C. . (2018). Learning models with uniform performance via distributionally robust optimization.

[3] Sinha, A. , Namkoong, H. , Volpi, R. , & Duchi, J. . (2017). Certifying some distributional robustness with principled adversarial training.

f-divergence DRO

Paper: Learning Models with Uniform Performance via Distributionally Robust Optimization[4]

- Objective function:

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta\in\Theta}\left\{\mathcal{R}_f(\theta;\widehat{P}_n) := \sup_{Q\ll\widehat{P}_n}\left\{\mathbb{E}_Q[\ell(\theta;X)] : D_f(Q\|\widehat{P}_n)\leq\rho\right\}\right\}.$$

- Optimization:

Proposition 1. *Let P be an arbitrary probability measure on $(\mathcal{X},\mathcal{A})$. Then, for any $\rho>0$, we have for all $\theta\in\Theta$*

$$\sup_{Q\ll P}\left\{\mathbb{E}_Q[\ell(\theta;x)] : D_f(Q\|P)\leq\rho\right\} = \inf_{\lambda\geq 0,\eta\in\mathbb{R}}\left\{\mathbb{E}_P\left[\lambda f^*\left(\frac{\ell(\theta;X)-\eta}{\lambda}\right)\right]+\lambda\rho+\eta\right\}. \tag{5}$$

Moreover, if the supremum on the left hand side is finite, there are finite $\lambda(\theta)\geq 0$ and $\eta(\theta)\in\mathbb{R}$ attaining the infimum on the right hand side.

- Simplified dual formulation for the Cressie-Read family:

Lemma 1. *Let P be an arbitrary probability measure on $(\mathcal{X},\mathcal{A})$. Then, for $k\in(1,\infty)$ and $k_*=k/(k-1)\in(1,\infty)$, and any $\rho>0$, we have for all $\theta\in\Theta$*

$$\mathcal{R}_k(\theta;P) = \inf_{\eta\in\mathbb{R}}\left\{c_k(\rho)\mathbb{E}_P\left[(\ell(\theta;X)-\eta)_+^{k_*}\right]^{\frac{1}{k_*}}+\eta\right\}. \tag{8}$$

where $c_k(\rho) := (k(k-1)\rho+1)^{\frac{1}{k}}$.

[4] Duchi, J. C. . (2018). Learning models with uniform performance via distributionally robust optimization.

Wasserstein DRO

Paper: Certifying Some Distributional Robustness with Principled Adversarial Training[5]

- Why Wasserstein distance?
  - More flexible: do not require the same support.
  - More difficult to optimize.
- Objective function:

$$\min_{\theta} \sup_{P:W_c(P,P_{tr})\leq\rho} \mathbb{E}_{X,Y\sim P}[\ell(\theta;X,Y)] \tag{6}$$

- Optimization:

Proposition 1. *Let $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ and $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be continuous. Let $\phi_\gamma(\theta; z_0) = \sup_{z\in\mathcal{Z}} \{\ell(\theta; z) - \gamma c(z, z_0)\}$ be the robust surrogate (2b). For any distribution Q and any $\rho > 0$,*

$$\sup_{P:W_c(P,Q)\leq\rho} \mathbb{E}_P[\ell(\theta;Z)] = \inf_{\gamma\geq 0} \{\gamma\rho + \mathbb{E}_Q[\phi_\gamma(\theta;Z)]\}, \tag{5}$$

and for any $\gamma \geq 0$, we have

$$\sup_P \{\mathbb{E}_P[\ell(\theta;Z)] - \gamma W_c(P,Q)\} = \mathbb{E}_Q[\phi_\gamma(\theta;Z)]. \tag{6}$$

[5] Sinha, A. , Namkoong, H. , Volpi, R. , & Duchi, J. . (2017). Certifying some distributional robustness with principled adversarial training.

Other works in DRO

- Group DRO: consider the group-level DRO[6]
- Marginal DRO: consider the DRO on the marginal distributions of covariates[7]
- f-divergence DRO with varaince regularizer[8]
- WDRO for logistic regression [9]
- WDRO for linear regression [10]

[6] Shiori Sagawa et al. DISTRIBUTIONALLY ROBUST NEURAL NETWORKS FOR GROUP SHIFTS: ON THE IMPORTANCE OF REGULARIZATION FOR WORST-CASE GENERALIZATION

[7] John C. Duchi et al. Distributionally Robust Losses Against Mixture Covariate Shifts

[8] Hongseok Namkoong et al. Variance-based Regularization with Convex Objectives

[9] Soroosh Shafieezadeh-Abadeh et al. Distributionally Robust Logistic Regression

[10] Ruidi Chen et al. A Robust Learning Approach for Regression Models Based on Distributionally Robust Optimization

1 Invariant Causal Prediction

2 Distributionally Robust Learning

3 Relationships

4 Conclusions

Relationships between Causality and DRO

Causality can be viewed as robust learning with certain uncertainty set[11]:

$$\theta^{causal} = \arg\min_{\theta} \sup_{Q \in \mathcal{Q}^{do}} \mathbb{E}_{X,Y\sim Q}[\ell(\theta; X, Y)] \tag{7}$$

Understanding:

- $\rightarrow$: θ^{causal} obviously minimizes the worst case loss
- $\leftarrow$: consider the converse-negative proposition:

$$\text{Not } \theta^{causal} \rightarrow \text{Not } \arg\min_{\theta} \sup_{Q \in \mathcal{Q}^{do}} \tag{8}$$

- $\leftarrow$: another perspective:

$$\sup_{Q \in \mathcal{Q}^{do}} \mathbb{E}_{X,Y\sim Q}[\ell(\theta; X, Y)] = \begin{cases} \infty & \theta \neq \theta^{causal} \\ \mathrm{Var}(\epsilon_y) & \theta = \theta^{causal} \end{cases} \tag{9}$$

[11] Meinshausen, N. . (2018). Causality from a distributional robustness point of view. 6-10.

A Toy Illustration of $\leftarrow$

Consider a simple data generation mechanism：

$$S \rightarrow Y \dashrightarrow V \tag{10}$$

or in equation:

$$S \leftarrow \mathcal{N}(0, \epsilon_S) \tag{11}$$

$$Y \leftarrow S + \epsilon_Y \tag{12}$$

$$V \leftarrow \alpha_e Y + \epsilon_V \tag{13}$$

where α_e is changing across environments/distributions, which means V should not be used to predict Y. Then $\mathcal{Q}^{do}$ contains:

$$P^{do}_{S=a_1}, \ldots, P^{do}_{S=a_m}, P^{do}_{V=b_1}, \ldots, P^{do}_{V=b_k} \tag{14}$$

Actually, from $P^{do}_{V=b}$, we can know that V is not the direct cause of Y.

Difference between Causality, Robustness and Distributional Robustness

The goals are different:

- Causality: to estimate causal coefficients, to identify causal covaraites
- Traditional Robustness: the predictive accuracy for a reference distribution is optimal:

$$\arg\min_{\theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{X,Y \sim P^{obs}}[\ell(\theta; X, Y)] \tag{15}$$

- Distributional Robustness: robust over a set of distributions

$$\arg\min_{\theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{X,Y \sim Q}[\ell(\theta; X, Y)] \tag{16}$$

Difference between Causality and DRO:

- Causality: seeks for the true causal parameters
- DRO: only seeks for robustness over distributions but does not care the true parameters

1 Invariant Causal Prediction

2 Distributionally Robust Learning

3 Relationships

4 Conclusions

Conclusion

- Distributionally robust learning is closely related to causality.
- How to approach causality by the way of DRO?
- How to build the uncertainty set with the assistance of causality?

QA