A short note on causality and its minimax framework

Yen-Chi Chen
University of Washington
June 25, 2020

In this note, we will briefly comment on a trending perspective of analyzing causality under a minimax framework. The main reference is the following paper:

[B2018] Bhlmann, P. (2018). Invariance, causality and robustness. arXiv preprint arXiv:1812.08233.

Consider a simple problem where we have an outcome variable Y and a set of covariates $X \in \mathbb{R}^p$ and both are random variables. For simplicity, we assume that both are continuous and has a joint PDF $p(x,y)$.

In a regular case, the causal effect from X on Y is encoded by the conditional density $p(y|x)$ (or the conditional distribution $P_{Y|X}$). $p(y|x)$ informs us how the change of X would affect the distribution of Y. This also implies an interesting and well-known result in terms of the prediction: the best predictor $f^*(x)$ of the response Y under the squared loss will be

$$f^*(x) = \mathbb{E}(Y|X=x) = \mathsf{argmin}_f \mathbb{E}\left((Y-f(X))^2\right). \tag{1}$$

We can decompose the joint PDF into $p(x,y) = p(y|x)p_x(x)$. Let $\mathcal{P}_{\mathsf{all}}$ be the collection of all possible marginal PDF of X. Then we can further rewrite the best predictor in equation (1) as the following *minimax framework:*

$$f^*(x) = \mathbb{E}(Y|X=x) = \mathsf{argmin}_f \sup_{p_x \in \mathcal{P}_{\mathsf{all}}} \mathbb{E}\left((Y-f(X))^2\right). \tag{2}$$

The fact that $f^*(x) = \mathbb{E}(Y|X=x) = \int y p(y|x) dy$ implies that $f^*(x)$ can be interpreted as the causal predictor of Y when $X=x$. Thus, the formulation in equation (2) connects the causality to a distributional robustness– if we discover the causal effect $\mathbb{E}(Y|X=x)$, this causal effect must has the distributional robustness against the change of marginal PDF p_x.

Note: suppose that we have a DAG (directed acyclic graph) of (X,Y) and let $S^* = \mathsf{PA}(Y) \subset \{1,2,\cdots,p\}$ be the parent nodes of Y. Then one can show that $f^*(x) = f^*(x^*_S)$, which again links back to the usual sense of causal effect from X on Y in a DAG.

1 The minimax framework

In reality, the optimization problem in equation (2) could be challenging because we have to search over all possible function $f(x)$. To reduce the complexity of the problem, we often restrict ourselves to a class $f \in \mathcal{F}$, leading to

$$f^\dagger_0(x) = \mathsf{argmin}_{f\in\mathcal{F}} \sup_{p_x \in \mathcal{P}_{\mathsf{all}}} \mathbb{E}\left((Y-f(X))^2\right). \tag{3}$$

The collection $\mathcal{P}_{\mathsf{all}}$ might be too large and in reality, we may only have data from a set of different conditions/environments. So we often restrict ourselves to a smaller subset $\mathcal{P} \subset \mathcal{P}_{\mathsf{all}}$. In some cases, we may have $\|\mathcal{P}'_X\| = k$, where $\|\mathcal{P}\|$ is the cardinality. In this scenario, we denote $\mathcal{P} = \{p_{x,1}, \cdots, p_{x,k}\}$ and equation (3) can be re-written as

$$f^\dagger(x) = \mathsf{argmin}_{f \in \mathcal{F}} \sup_{p_x \in \mathcal{P}} \mathbb{E}\left((Y - f(X))^2\right). \tag{4}$$

A causal inference scenario of equation (4) is that we have k different interventions on the system (k observational studies) that each of them changes the distribution of X but these interventions keep the same causal effect (conditional density $p(y|x)$). The set $\mathcal{P} = \{p_{x,1}, \cdots, p_{x,k}\}$ contains the marginal distributions that corresponds to each intervention. Note: the above scenario satisfies the 'ad-hoc' conditions in [B2018].

Example: linear causal parameter. One simple example is the case where $\mathcal{F} = \{f(x) = \beta^T x : \beta \in \mathbb{R}^p\}$, the class of linear function. In this case, the problem becomes finding the parameter $\beta^\dagger$ and equation (4) leads to

$$\beta^\dagger = \mathsf{argmin}_\beta \sup_{p_x \in \mathcal{P}} \mathbb{E}\left((Y - \beta^T X)^2\right). \tag{5}$$

The parameter $\beta^\dagger$ is called *causal parameter* in [B2018]. Note that $\beta^\dagger$ depends on the subclass $\mathcal{P}$ that we are considering when the true regression function is not linear.

Example: ridge regression. Consider the linear model as equation (5) and we choose $\mathcal{P}$ to be a linear perturbation such that $\mathcal{P} = \mathcal{P}_\eta$ is the collection of PDF of $X' = X + \varepsilon$, where $\varepsilon \in \mathbb{R}$ is an independent random vector with variance $\mathsf{Var}(\varepsilon_j) \le \eta^2$ (and is independent from X and Y). In this case, equation (5) becomes

$$\begin{aligned}
\beta_\eta &= \mathsf{argmin}_\beta \sup_{\mathsf{Var}(\varepsilon_j) \le \eta^2} \mathbb{E}\left((Y - \beta^T(X+\varepsilon))^2\right) \\
&= \mathsf{argmin}_\beta \mathbb{E}\left((Y - \beta^T X)^2\right) + \sup_{\mathsf{Var}(\varepsilon_j) \le \eta^2} \sum_{j=1}^p \beta_j^2 \mathbb{E}(\varepsilon_j^2) \\
&= \mathsf{argmin}_\beta \mathbb{E}\left((Y - \beta^T X)^2\right) + \eta^2 \|\beta\|_2^2,
\end{aligned}$$

which is the ridge regression. Thus, the ridge regression can be interpreted as a causal parameter under the distribution perturbation

$$\mathcal{P} = \{p(x') : X' = X + \varepsilon, \mathsf{Var}(\varepsilon_j) \le \eta^2 \text{ for each } j = 1, \cdots, p\}.$$

2 Anchor regression and OLS, PA, IV methods

The anchor regression is a new regression method that trade-off between a PA (partialling out estimator), an OLS (ordinary least square), and an IV (instrumental variable) approach. Now we will show that in a linear structural equation model (SEM), the anchor regression can be derived under the minimax framework (equation (5)) with a suitable choice of $\mathcal{P}$. This implies that the PA, OLS, and IV estimators can all be casted in the minimax framework.

In addition to X, Y, we include an additional variable A, denoted as the anchor. Note that A may play a similar role as the IV. Let Π_A be a projection operator such that for any random variable W, $\Pi_A W = \mathbb{E}(W|A)$; also

let $\mathbf{I}_d$ be the identity operator such that $\mathbf{I}_d W = W$. This implies that $(\mathbf{I}_d - \Pi_A)W = W - \mathbb{E}(W|A)$ is the orthogonal projection. The anchor regression solves the following least square problem:

$$\beta_{AR,\gamma} = \text{argmin}_\beta \mathbb{E}\left((\mathbf{I}_d - \Pi_A)(Y - \beta^T X)^2\right) + \gamma \mathbb{E}\left(\Pi_A (Y - \beta^T X)^2\right). \tag{6}$$

One can see that when $\gamma = 0$, this reduces to the PA estimator, when $\gamma = 1$, this becomes the OLS, and when $\gamma = \infty$, this leads to the IV estimator. In the anchor regression γ is a tuning parameter that one can choose to tradeoff between these methods.

Consider a simple SEM that:

$$\begin{pmatrix} Y \\ X \end{pmatrix} = B \begin{pmatrix} Y \\ X \end{pmatrix} + \varepsilon + MA = B \begin{pmatrix} Y \\ X \end{pmatrix} + \varepsilon + V, \tag{7}$$

where $B \in \mathbb{R}^{(p+1)\times(p+1)}$ is an upper triangular matrix and $M \in \mathbb{R}^(p+1)$ is a deterministic vector and $A \in \mathbb{R}$ is the anchor variable and $\varepsilon \in \mathbb{R}^{(p+1)}$ is a random vector (that describes the distribution of X and the additive noise of Y) and $V = MA$. Note that the anchor variable may be multivariate; here we use univariate for simplicity.

Let $\mathcal{P}$ to be the collection of models under the SEM of equation (7) and the following constraints on V:

$$VV^T \leq \gamma M \mathbb{E}(A^2) M^T, \quad V \perp \varepsilon, \tag{8}$$

where $\leq$ means that $VV^T - \gamma M \mathbb{E}(A^2) M^T$ is negative definite matrix. We will show that under equations (7) and (8), the anchor regression estimator can be derived under the minimax framework (i.e., equation (5) implies (8)).

The following derivation is modified from the proof of Theorem 1 of the following paper:

> [RMBP2018] Rothenhäusler, D., Meinshausen, N., Bühlmann, P., & Peters, J. (2018). Anchor regression: heterogeneous data meets causality. arXiv preprint arXiv:1801.06229.

This SEM implies that

$$\begin{pmatrix} Y \\ X \end{pmatrix} = (I - B)^{-1}(\varepsilon + V)$$

and we can further rewrite X and Y as

$$\begin{aligned} Y &= [(I-B)^{-1}]^T_{1,+}(\varepsilon + V) \\ X &= [(I-B)^{-1}]^T_{-1,+}(\varepsilon + V), \end{aligned}$$

where $[(I-B)^{-1}]_{1,+}$ is the first row of $(I-B)^{-1}$ and $[(I-B)^{-1}]_{-1,+}$ is the matrix of $(I-B)^{-1}$ without the first row[1]. As a result, we can rewrite

$$Y - \beta^T X = \underbrace{\left([(I-B)^{-1}]^T_{1,+} - \beta^T [(I-B)^{-1}]^T_{-11,+}\right)}_{\omega^T_\beta} (\varepsilon + V) \tag{9}$$

[1]The main reason for rewriting it as such additive form is that later in the derivation of least square criterion, the additive form gives an elegant decomposition.

so the least square criterion becomes

$$\mathbb{E}((Y-\beta^T X)^2) = \mathbb{E}((\omega_\beta^T(\varepsilon+V))^2) = \omega_\beta^T \mathbb{E}(\varepsilon\varepsilon^T)\omega_\beta + \omega_\beta^T \mathbb{E}(VV^T)\omega_\beta.$$

Note that we use the fact that ε, V are independent so the product term disappears. Recall that $\mathcal{P}$ only changes V in the additive form so the first part remains the same. Thus,

$$\begin{aligned}\sup_{p_X\in\mathcal{P}} \mathbb{E}((Y-\beta^T X)^2) &= \omega_\beta^T \mathbb{E}(\varepsilon\varepsilon^T)\omega_\beta + \sup_{V:\mathbb{E}(VV^T)\le\gamma M\mathbb{E}(A^2)M^T} \omega_\beta^T \mathbb{E}(VV^T)\omega_\beta \\ &= \omega_\beta^T \mathbb{E}(\varepsilon\varepsilon^T)\omega_\beta + \gamma\omega_\beta^T M\mathbb{E}(A^2)M^T\omega_\beta.\end{aligned} \tag{10}$$

Now we connect ε, V to the projection operator Π_A. Using the fact that $V = MA$, so $V \perp \varepsilon$ implies that $A \perp \varepsilon$. Thus,

$$\begin{aligned}\Pi_A(Y-\beta^T X) &= \mathbb{E}(Y-\beta^T X|A) = \mathbb{E}(\omega_\beta^T(\varepsilon+V)|A) = \omega_\beta^T MA \\ (\mathbf{I}_d-\Pi_A)(Y-\beta^T X) &= Y-\beta^T X - \mathbb{E}(Y-\beta^T X|A) = \omega_\beta^T\varepsilon\end{aligned}$$

so

$$\begin{aligned}\mathbb{E}(\Pi_A(Y-\beta^T X)^2) &= \mathbb{E}((\omega_\beta^T MA)^2) = \omega_\beta^T M\mathbb{E}(A^2)M^T\omega_\beta, \\ \mathbb{E}((\mathbf{I}_d-\Pi_A)(Y-\beta^T X)^2) &= \omega_\beta^T\mathbb{E}(\varepsilon\varepsilon^T)\omega_\beta\end{aligned}$$

are the two terms in the last equality of equation (10). Putting them into equation (10) and return to the equation (5), we conclude that

$$\begin{aligned}\beta_\gamma^\dagger &= \mathsf{argmin}_\beta \sup_{p_X\in\mathcal{P}} \mathbb{E}((Y-\beta^T X)^2) \\ &= \mathsf{argmin}_\beta \omega_\beta^T\mathbb{E}(\varepsilon\varepsilon^T)\omega_\beta + \gamma\omega_\beta^T M\mathbb{E}(A^2)M^T\omega_\beta \\ &= \mathsf{argmin}_\beta \mathbb{E}((\mathbf{I}_d-\Pi_A)(Y-\beta^T X)^2) + \gamma\mathbb{E}(\Pi_A(Y-\beta^T X)^2) \\ &= \beta_{RA,\gamma},\end{aligned}$$

which is the solution of equation (8).

The above analysis shows that the anchor regression can be written under the minimax framework. Since different choice of γ leads to different regression estimators (PA, OLS, and IV), this result also implies that PA, OLS, and IV regressions can all be written under such minimax framework.

3 Remarks

Generalization to other loss function. We may apply the same minimax framework to other loss function. Suppose that the function $f = f_\theta$ is indexed by some parameter $\theta \in \Theta$. This leads to the following estimator:

$$\theta^* = \mathsf{argmin}_{\theta\in\Theta} \sup_{p_x\in\mathcal{P}} L(f_\theta(X), Y)$$

for some suitable class $\mathcal{P}$. The corresponding model f_{θ^*} is robust against distributional perturbation. See the following papers for more details and examples of this procedure:

[M2018] Meinshausen, N. (2018, June). Causality from a distributional robustness point of view. In 2018 IEEE Data Science Workshop (DSW) (pp. 6-10). IEEE.

Robust regression. One may find the minimax framework similar to the robust regression problem. Both are attempting to estimate the parameter that are robust against to some distributional perturbations. Indeed, the two problems are related but there are two distinctions. First, the distributional perturbations are different. In the causal minimax framework, the perturbation is often on the distribution of covariates X and we keep the relation between X and Y. On the other hands, in robust regression, we are generally more interested in perturbing the noise distribution (to heavier tails or a mixture of noises). The second distinction comes from the focus of research. In the causal minimax framework, we are often more interested in learning the model that has a good predictive performance across all scenarios. In robust statistics, often we are thinking of finding the model that has good predictive performance under a fixed scenario (without outliers/perturbations) from a sample that is contaminated with outliers. See Section 4 of [M2018] for some discussions on this.

Distributionally robust learning. The minimax framework shows a direction connection to the distributionally robust learning (DRL). The usual setup of a distributionally robust learning is the following empirical risk minimization problem:

$$\widetilde{\beta} = \mathsf{argmin}_{\beta} \sup_{P_{X,Y} \in \mathcal{P}_{X,Y}} \mathbb{E}(L(X,Y;\beta)),$$

where L is something like a loss function. In DRL, the distributional perturbation can be on $p(y|x)$ as well, which is the major distinction to the causal minimax framework. For recent research along this line, see the following two papers:

1. Gao, R., Chen, X., & Kleywegt, A. J. (2017). Wasserstein distributional robustness and regularization in statistical learning. arXiv preprint arXiv:1712.06050.
2. Sinha, A., Namkoong, H., & Duchi, J. (2017). Certifying some distributional robustness with principled adversarial training. arXiv preprint arXiv:1710.10571.