Counterfactual Cross-Validation: Stable Model Selection Procedure for Causal Inference Models

Authors: Yuta Saito, Shota Yasui

September 11, 2020

Presented by: Serge Assaad

Overview

- **Goal:** Given candidate models $\hat{\tau}_1(\cdot), ..., \hat{\tau}_K(\cdot)$ for the individual treatment effect (ITE), how do we select the best one, given that we don't know the true ITE function $\tau(\cdot)$ (i.e. half of the potential outcomes are missing)?
- **Key idea:** Find a proxy $\tilde{\tau}$ (called "plug-in tau" in the paper) such that the *ranks* of the errors of the candidate models w.r.t. the proxy matches the *ranks* of the errors of the candidates w.r.t. the true ITE function $\tau(\cdot)$ – more on this later.

Note: this work makes the standard assumptions of ignorability, overlap, and consistency. $X \in \mathbb{R}^p$ *are covariates,* $T \in \{0, 1\}$ *are treatments, and* $Y \in \mathbb{R}$ *are outcomes.*

Evaluation of Individual Treatment Effect (ITE) models

Typically, the quality of an individual treatment effect (ITE) model is measured by a quantity known as the *precision in estimation of heterogeneous effect* (PEHE), defined as

$$\mathcal{R}_{true}(\hat{\tau}) \triangleq \mathbb{E}_X \left[\left(\tau(X) - \hat{\tau}(X) \right)^2 \right], \tag{1}$$

$\mathcal{R}_{true}(\hat{\tau})$ is also called the **true risk** of the model $\hat{\tau}(\cdot)$, and $\tau(\cdot)$ is the true ITE (which we do not have access to for real-world datasets).

Proxy for true risk

The key idea of this paper is that we don't need a good proxy for the true risk, we only need a proxy **which preserves the rank order of candidate models**.

More precisely, for a candidate model $\hat{\tau} \in \mathcal{M}$ (where $\mathcal{M}$ is a model class), we would like to design a proxy risk function $\hat{\mathcal{R}}(\hat{\tau})$ which satisfies:

$$\mathcal{R}_{true}\left(\hat{\tau}\right) \leq \mathcal{R}_{true}\left(\hat{\tau}'\right) \Rightarrow \widehat{\mathcal{R}}\left(\hat{\tau}\right) \leq \widehat{\mathcal{R}}\left(\hat{\tau}'\right), \ \forall\, \hat{\tau}, \hat{\tau}' \in \mathcal{M}. \qquad (2)$$

Let's refer to (2) as the "rank-preserving" property.

How do we design a good proxy $\widehat{\mathcal{R}}$?

Consider a risk proxy $\widehat{\mathcal{R}}$ of the form:

$$\widehat{\mathcal{R}}\left(\hat{\tau}\right) := \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\tau}\left(X_i, T_i, Y_i\right) - \hat{\tau}\left(X_i\right)\right)^2 \tag{3}$$

where $\tilde{\tau}$ is called a "plug-in". Hence, the problem boils down to finding a good plug-in, namely a plug-in $\tilde{\tau}$ such that $\widehat{\mathcal{R}}$ has the rank-preserving property.

Note: Using the above setup, if we were to obtain a good plug-in $\tilde{\tau}$, we could not use it as the estimate of τ itself (on test samples) since it uses X_i, T_i, Y_i as inputs – i.e. in this setup, $\tilde{\tau}$ can only be used for evaluation, not prediction.

Desirable properties of the plug-in $\tilde{\tau}$

Proposition

Suppose that a given plug-in $\tilde{\tau}$ is an unbiased estimator for the true ITE (i.e., $\mathbb{E}\left[\tilde{\tau}\left(X,T,Y\right) \mid X\right] = \tau(X)$), then, the expectation of the performance estimator $\widehat{\mathcal{R}}$ is decomposed into the true performance metric and the MSE of the given plug-in $\tilde{\tau}$:

$$\mathbb{E}\left[\widehat{\mathcal{R}}\left(\hat{\tau}\right)\right] = \mathcal{R}_{true}\left(\hat{\tau}\right) + \underbrace{\mathbb{E}\left[\left(\tau(X) - \tilde{\tau}(X,T,Y)\right)^2\right]}_{\text{independent of } \hat{\tau}} \tag{4}$$

Which yields, for 2 different candidate models $\hat{\tau}_1$ and $\hat{\tau}_2$:

$$\mathbb{E}\left[\widehat{\mathcal{R}}\left(\hat{\tau}_1\right)\right] - \mathbb{E}\left[\widehat{\mathcal{R}}\left(\hat{\tau}_2\right)\right] = \mathcal{R}_{true}\left(\hat{\tau}_1\right) - \mathcal{R}_{true}\left(\hat{\tau}_2\right) \tag{5}$$

This is close to the rank-preserving property, but not quite the same since we are taking the difference in the *expected values* of $\widehat{\mathcal{R}}$, not the $\widehat{\mathcal{R}}$ values themselves. This motivates us to consider the finite-sample uncertainty of $\widehat{\mathcal{R}}$.

Decomposition of $\widehat{\mathcal{R}}(\hat{\tau})$ for finite samples

As a reminder:

$$\widehat{\mathcal{R}}(\hat{\tau}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\tau}\left(X_i, T_i, Y_i\right) - \hat{\tau}\left(X_i\right)\right)^2, \tag{6}$$

which can be decomposed as:

$$\begin{aligned}\widehat{\mathcal{R}}\left(\hat{\tau}\right) = & \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \hat{\tau}(X_i))^2}_{\text{converges to } \mathcal{R}_{true}(\hat{\tau})} \\ & \underbrace{- \frac{2}{n}\sum_{i=1}^{n}\left(\hat{\tau}\left(X_i\right) - \tau\left(X_i\right)\right)\left(\tilde{\tau}\left(X_i, T_i, Y_i\right) - \tau\left(X_i\right)\right)}_{\mathcal{W}\text{:source of uncertainty}} \\ & \underbrace{+ \frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \tilde{\tau}(X_i, T_i, Y_i))^2}_{\text{independent of } \hat{\tau}}.\end{aligned} \tag{7}$$

Variance minimization

Considering the previous expression, we can write the difference in $\widehat{\mathcal{R}}(\hat{\tau}_1)$ and $\widehat{\mathcal{R}}(\hat{\tau}_2)$ as:

$$\widehat{\mathcal{R}}(\hat{\tau}_1) - \widehat{\mathcal{R}}(\hat{\tau}_2) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \hat{\tau}_1(X_i))^2}_{\text{converges to } \mathcal{R}_{true}(\hat{\tau}_1)} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \hat{\tau}_2(X_i))^2}_{\text{converges to } \mathcal{R}_{true}(\hat{\tau}_2)}$$
$$- \frac{2}{n}\sum_{i=1}^{n}\left(\hat{\tau}_1\left(X_i\right) - \tau\left(X_i\right)\right)\left(\tilde{\tau}\left(X_i, T_i, Y_i\right) - \tau\left(X_i\right)\right)$$
$$+ \frac{2}{n}\sum_{i=1}^{n}\left(\hat{\tau}_2\left(X_i\right) - \tau\left(X_i\right)\right)\left(\tilde{\tau}\left(X_i, T_i, Y_i\right) - \tau\left(X_i\right)\right) \tag{8}$$

Hence, for $\widehat{\mathcal{R}}$ to satisfy the rank preserving property, we must optimize $\tilde{\tau}$ so that the last 2 terms cancel out. This motivates the minimization of $\mathbb{V}(\mathcal{W})$ ($\mathcal{W}$ defined on the previous slide).

Variance upper-bound

Theorem

Suppose that the plug-in $\tilde{\tau}$ is unbiased for the ITE and the output of the plug-in $\tilde{\tau}$ for an instance is independent of that of other instances. Then, we have the upper bound of the variance of $\mathcal{W}$ as follows:

$$\mathbb{V}(\mathcal{W}) \leq 4C_{\max} n^{-1} \, \mathbb{E}_X \left[\mathbb{V}\left(\tilde{\tau}(X,T,Y) \mid X \right) \right], \tag{9}$$

where $C_{\max} = \sup_{x \in \mathcal{X}} (\tau(x) - \hat{\tau}(x))^2$.

Hence, we can write the variance upper-bound minimization as:

$$\begin{aligned} &\min_{\tilde{\tau} \in \Theta} \mathbb{E}_X \left[\mathbb{V}\left(\tilde{\tau}(X,T,Y) \mid X \right) \right], \\ &\text{s.t. } \mathbb{E}[\tilde{\tau}(X,T,Y) \mid X] = \tau(X). \end{aligned} \tag{10}$$

where Θ is a pre-defined class of plug-ins $\tilde{\tau}$.

How can we get plug-in $\tilde{\tau}$ in practice?

First, we define a class of *plug-in* $\tilde{\tau}$ building on the doubly robust estimator.

Definition

The doubly robust plug-in $\tilde{\tau}_{DR}$ for a given triplet (X, T, Y) is defined as follows:

$$\begin{aligned} &\tilde{\tau}_{DR}(X, T, Y; f_1, f_0) \\ &:= \frac{T - e(X)}{e(X)(1 - e(X))}(Y - f_T(X)) + f_1(X) - f_0(X), \end{aligned} \tag{11}$$

where f_1, f_0 are regressors, and $e(X) := P(T = 1|X)$.

Given true propensity scores and a regression function, the proposed plug-in $\tilde{\tau}$ is unbiased against the true ITE, i.e.,

$$\mathbb{E}\left[\tilde{\tau}_{DR}(X, T, Y; f_1, f_0) \mid X\right] = \tau(X). \tag{12}$$

Training $\tilde{\tau}_{DR}$

The previously described minimization problem is:

$$\begin{aligned} &\min_{\tilde{\tau} \in \Theta} \mathbb{E}_X \left[\mathbb{V} \left(\tilde{\tau}(X, T, Y) \mid X \right) \right], \\ &\text{s.t. } \mathbb{E}[\tilde{\tau}(X, T, Y) \mid X] = \tau(X). \end{aligned} \tag{13}$$

Given the true propensity score $e(X)$, $\tilde{\tau}_{DR}$ has the advantage of satisfying the unbiasedness constraint. But how do we minimize the variance of $\tilde{\tau}_{DR}$?

Training $\tilde{\tau}_{DR}$

Proposition

Given true propensity scores and a regression function, the expected conditional variance of the proposed plug-in $\tilde{\tau}_{DR}$ can be represented as:

$$\begin{aligned}
&\mathbb{E}_X \left[\mathbb{V} \left(\tilde{\tau}_{DR}(X, T, Y; f_1, f_0) \mid X \right) \right] \\
&= \zeta + \mathbb{E}_X \left[\{ \sum_{t \in \{0,1\}} \sqrt{w_t(X)} (f_t(X) - m_t(X)) \}^2 \right], \qquad (14)
\end{aligned}$$

where

$$w_t(X) := \frac{t(1 - 2e(X)) + e(X)^2}{e(X)(1 - e(X))}, m_t(x) := \mathbb{E}_{Y(t)}[Y(t) \mid X = x],$$

and ζ is a constant w.r.t. the models f_1, f_0.

Training $\tilde{\tau}_{DR}$

Based on the previous expression for the variance, we can formulate the minimization problem as:

$$\min_{f \in \mathcal{F}} \mathbb{E}_X \left[\{ \sum_{t \in \{0,1\}} \sqrt{w_t(X)} (f_t(X) - m_t(X)) \}^2 \right], \qquad (15)$$

but unfortunately, this is impossible to minimize directly, since we do not have access to counterfactual outcomes (i.e., we would be missing a value of $m_t(X)$ for every X).
The solution? Find a tractable upper bound!

Another upper bound

Theorem

Let G be a family of functions $g : \mathcal{R} \to \mathcal{Y}$ and suppose that, for any given $t \in \{0, 1\}$ and $w : \mathcal{X} \times \{0, 1\} \to \mathcal{R}_{\geq 0}$, there exists a positive constant B_Φ such that the per-unit expected loss functions obey $\frac{1}{B_\Phi} \cdot \ell^w_{h,\Phi}(\Psi(r), t) \in G$ where Ψ is the inverse image of Φ. Then, the following inequality holds:

$$\begin{aligned}
&\mathbb{E}_X[\{\sum_{t\in\{0,1\}} \sqrt{w_t(X)}(f_t(X) - m_t(X))\}^2] \\
&\leq 2\big(\epsilon^{w_1}_{F_1}(h, \Phi) + \epsilon^{w_0}_{F_0}(h, \Phi) \\
&+ B_\Phi \mathrm{IPM}_G\left(p^\Phi_0, p^\Phi_1\right) - 2\sigma^2\big), \qquad (16)
\end{aligned}$$

where $\epsilon^{w_t}_{F_t}(h, \Phi) := \int_{\mathcal{X}} \ell^w_{h,\Phi}(x, T = t)p_t(x)dx$ for $t \in \{0, 1\}$ and σ is a constant w.r.t. f_t.

Putting it all together: model selection in practice

Algorithm (Counterfactual Cross-Validation (CF-CV))

Require: *A set of candidate ITE predictors* $\mathcal{M} = \{\hat{\tau}_1, ..., \hat{\tau}_{|\mathcal{M}|}\}$*; an observational validation dataset* $\mathcal{V} = \{X_i, T_i, Y\}_{i=1}^{n}$*; and a trade-off hyperparameter* α*.*

1: *Estimate the propensity score.*
2: *Train* $f(X, T)$ *by minimizing the factual errors + IPM loss using samples in* $\mathcal{V}$*.*
3: *Calculate the plug-in* $\tilde{\tau}_{DR}$ *for samples in* $\mathcal{V}$*.*
4: *Estimate performance of candidate predictors in* $\mathcal{M}$ *based on the performance estimator* $\widehat{\mathcal{R}}$ *and* $\tilde{\tau}_{DR}$*.*

Ensure: *A selected predictor:* $\hat{\tau}^* = \arg\min_{\hat{\tau} \in \mathcal{M}} \widehat{\mathcal{R}}(\hat{\tau})$.

Experimental Procedure

- Conducted experiments on IHDP100 dataset [1] (a dataset with real covariates, but with simulated potential outcomes & treatment assignment – consists of 100 realizations, each with 35/35/30 train/val/test split).
- For each realization, trained 25 candidate models (e.g. decision tree, random forest, ridge regressor), ranked the candidate models using the proxy $\widehat{\mathcal{R}}$ on the validation set, and examined the ranked candidates' relative performances on the test set (measured by $\mathcal{R}_{true}$).
- Measured the quality of the proxy $\widehat{\mathcal{R}}$ (relative to $\mathcal{R}_{true}$) via Regret and Spearman Rank Correlation.
- Regret is defined as $\textit{Regret} = \frac{\mathcal{R}_{true}(\hat{\tau}_{selected}) - \mathcal{R}_{true}(\hat{\tau}_{best})}{\mathcal{R}_{true}(\hat{\tau}_{best})}$, where $\hat{\tau}_{selected} = \arg\min_{\hat{\tau} \in \mathcal{M}} \widehat{\mathcal{R}}(\hat{\tau})$ is the model selected by $\widehat{\mathcal{R}}$ and $\hat{\tau}_{best} = \arg\min_{\hat{\tau} \in \mathcal{M}} \mathcal{R}_{true}(\hat{\tau})$ is the best model in $\mathcal{M}$.

Baselines

The first two baselines used are proxies of the form

$$\widehat{\mathcal{R}} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}(X_i, T_i, Y_i) - \hat{\tau}(X_i))^2, \tag{17}$$

for different choices of $\tilde{\tau}$:

- **IPW validation:** $\tilde{\tau}_{IPW}(X_i, T_i, Y_i) = \frac{T_i}{e(X_i)}Y_i - \frac{1-T_i}{1-e(X_i)}Y_i$
- **Plug-in validation:** $\tilde{\tau}_{plug-in}(X_i) = \tilde{\tau}_i^{(1)} - \tilde{\tau}_i^{(0)}$, where $\tilde{\tau}_i^{(1)}$ and $\tilde{\tau}_i^{(0)}$ are predictions for the potential outcomes. CFR [2] is used as the predictor, for a fair comparison with CF-CV
- τ**-risk:**

$$\hat{\mathcal{R}}_\tau(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n}\left((Y_i - m(X_i)) - (T_i - e(X_i))\hat{\tau}(X_i)\right)^2$$

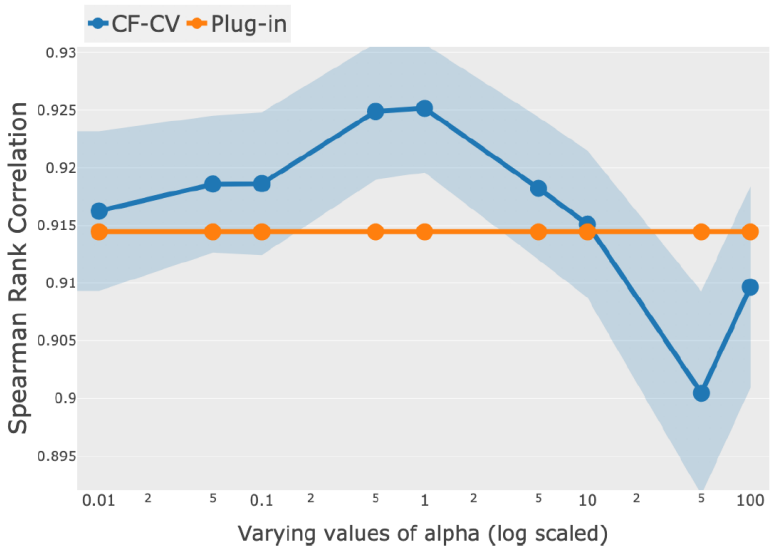
where $m(\cdot)$ is the expectation of observed outcome $\mathbb{E}[Y|X]$.

Results

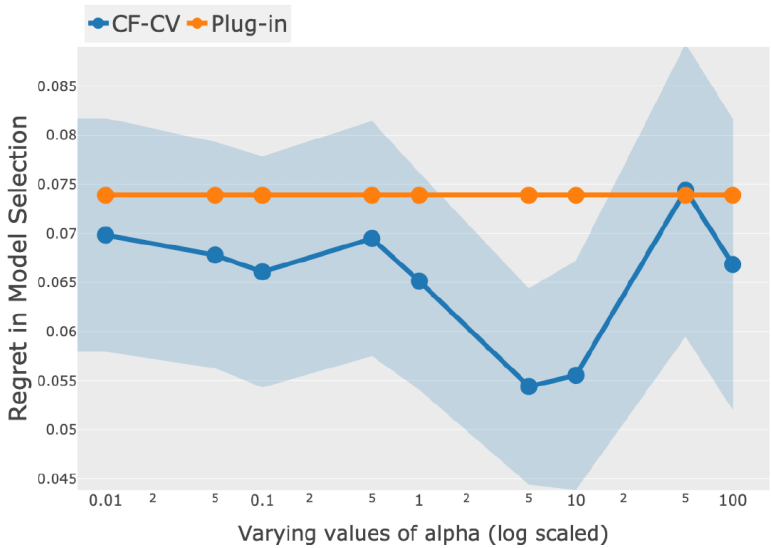
Table 1. Comparison of Model Selection and Hyperparameter Tuning Performance of Alternative Evaluation Metrics.

Methods	Rank Correlation		Regret		NRMSE	
	Mean ±StdErr	Worst-Case	Mean ±StdErr	Worst-Case	Mean ±StdErr	Worst-Case
IPW	0.195 ±0.039	-0.749	1.032 ±0.100	6.779	0.336 ±0.013	0.737
τ-risk	0.312 ±0.030	-0.553	1.392 ±0.130	7.884	0.324 ±0.013	0.700
Plug-in	0.914 ±0.006	0.591	0.073 ±0.012	0.780	0.257 ±0.010	0.490
CF-CV (ours)	**0.921 ±0.005**	**0.666**	**0.066 ±0.012**	**0.562**	**0.256 ±0.009**	**0.483**

Notes: Mean with standard errors (StdErr), and worst-case performance of the compared evaluation metrics over 100 realizations are reported. The **red fonts** represent the best performance in each performance measure.



(a) *Rank correlation* of CF-CV with different values of α



(b) *Regret* of CF-CV with different values of α

Figure 1. Comparing CF-CV with varying α and the plug-in validation. CF-CV (the blue lines) outperforms the plug-in validation (the orange lines) in most cases and demonstrates its robustness to the choice of α.

Conclusion

- This paper presented a way to select the "best" model from a set $\mathcal{M}$ of candidate models via a proxy $\widehat{\mathcal{R}}$, which is shown to approximately satisfy the rank-preserving property.
- The authors employ a plug-in model $\tilde{\tau}$ which enjoys an unbiasedness property (w.r.t. the true effect τ) given the true propensity $e(X)$ (though this has to be estimated in practice), and show that we can approach the rank-preserving property by minimizing the conditional variance of $\tilde{\tau}$ – an upper bound of which yields a CFR-type objective.

References

Jennifer L. Hill. "Bayesian Nonparametric Modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 217–240. DOI: 10.1198/jcgs.2010.08162. eprint: https://doi.org/10.1198/jcgs.2010.08162. URL: https://doi.org/10.1198/jcgs.2010.08162.

Uri Shalit, Fredrik D. Johansson, and David Sontag. *Estimating individual treatment effect: generalization bounds and algorithms*. 2016. arXiv: 1606.03976 [stat.ML].