Sensitivity Analysis of Individual Treatment Effects: A Robust Conformal Inference Approach

Ying Jin [1] Zhimei Ren [2] Emmanuel J. Candès [1,3]

[1]Department of Statistics, Stanford University [2]Department of Statistics, University of Chicago [3]Department of Mathematics, Stanford University

Abstract

We propose a model-free framework for sensitivity analysis of individual treatment effects (ITEs), building upon ideas from conformal inference. For any unit, our procedure reports the Γ-*value*, a number which quantifies the minimum strength of confounding needed to explain away the evidence for ITE.

Our approach rests on the reliable predictive inference of counterfactuals and ITEs in situations where the training data is confounded. Under the marginal sensitivity model of [3], we characterize the shift between the distribution of the observations and that of the counterfactuals. We first develop a general method for predictive inference of test samples from a shifted distribution; we then leverage this to construct covariate-dependent prediction sets for counterfactuals. No matter the value of the shift, these prediction sets (resp. approximately) achieve marginal coverage if the propensity score is known exactly (resp. estimated). We describe a distinct procedure also attaining coverage, however, conditional on the training data. In the latter case, we prove a sharpness result showing that for certain classes of prediction problems, the prediction intervals cannot possibly be tightened. We verify the validity and performance of the new methods via simulation studies and apply them to analyze real datasets.

Paper preprint: `https://arxiv.org/abs/2111.12161`

Backgrounds

From average to individual treatment effects

The average treatment effect (ATE) or the conditional average treatment effect (CATE) might fail to provide reliable uncertainty quantification for individual responses: the knowledge that a drug might be effective for a whole population `on average' does not imply that it is effective on a particular patient. Instead, we focus on the *individual treatment effect* (ITE)

$$\Delta := Y(1) - Y(0)$$

for a new test sample, where $Y(0), Y(1)$ are the potential outcomes for being control and treated [1]. To quantify the uncertainty in ITE, our goal [2] is to construct prediction interval $\widehat{C}(X)$ for a test point with covariate X, such that

$$\mathbb{P}\big(\Delta \in \widehat{C}(X)\big) \geq 1-\alpha.$$

Potential outcomes and unmeasured confounding

We work under the potential outcome framework [1]

- Subjects $\big(X_i, Y_i(0), Y_i(1)\big) \overset{\text{i.i.d.}}{\sim} \mathcal{P}$
- Treatment assignment mechanism $T_i \sim \mathcal{T}$
- Population $(X_i, Y_i(1), Y_i(0), T_i) \sim \mathbb{P}$
- Partial observations: (X_i, T_i, Y_i), where $Y_i = Y(T_i)$ (SUTVA)

[2] studies this problem under the strong ignorability condition $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \,|\, X_i$. Absence of confounding is misleading but not testable in practice.

What happens if strong ignorability does not hold? How to reliably infer ITE in this situation?

Sensitivity models

Unobserved confounder U affects both T and $(Y(0), Y(1))$.

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i \mid (X_i, U_i)$$

Such confounders always exist: think about $U = (Y(1), Y(0))$. We impose bounds on selection bias by marginal Γ-selection condition [3, 4]:

$$\frac{1}{\Gamma} \leq \frac{\mathbb{P}(T=1 \mid X=x, U=u)}{\mathbb{P}(T=0 \mid X=x, U=u)} \cdot \frac{\mathbb{P}(T=0 \mid X=x)}{\mathbb{P}(T=1 \mid X=x)} \leq \Gamma$$

Our proposal

Robust predictive inference of ITEs: For any hypothesized confounding strength Γ, construct robust prediction intervals $\widehat{C}(X, \Gamma)$ for counterfactual and ITE.

Gauge the robustness of causal conclusion: For any individual, output a Γ-*value* that quantifies the strength of confounding that is need to explain out the evidence for ITE.

Distributional shift for counterfactuals

When inferring the counterfactual $Y(1)$ for a new control sample $T=0$, the training distribution is $\mathbb{P}_{X,Y(1)\,|\,T=1}$, while the counterfactual is from $\mathbb{P}_{X,Y(1)\,|\,T=0}$.

A crucial fact: The distributional shift is bounded by identifiable functions of x.

$$\frac{1}{\Gamma} \cdot \frac{p}{1-p}\frac{1-e(x)}{e(x)} \leq \frac{\mathrm{d}\mathbb{P}_{X,Y(1)\,|\,T=0}}{\mathrm{d}\mathbb{P}_{X,Y(1)\,|\,T=1}} \leq \Gamma \cdot \frac{p}{1-p}\frac{1-e(x)}{e(x)}$$

under marginal Γ-selection, where $e(x) = \mathbb{P}(T=1\,|\,X=x)$, $p = \mathbb{P}(T=1)$.

A more general problem: robust inference

Train/calibration distribution: $\mathbb{P}_{XY} = \mathbb{P}_{Y|X} \cdot \mathbb{P}_X$

Unknown target distribution: $\widetilde{\mathbb{P}}_{XY} = \widetilde{\mathbb{P}}_{Y|X} \cdot \widetilde{\mathbb{P}}_X$

Assume for $\mathbb{P}$-almost all $x \in \mathcal{X}$,

$$\ell(x) \leq w(x,y) = \frac{\mathrm{d}\widetilde{\mathbb{P}}_{XY}}{\mathrm{d}\mathbb{P}_{XY}}(x,y) \leq u(x).$$

Goal: construct a predictive interval $\widehat{C}(X_{n+1})$ such that

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1})\big) \geq 1-\alpha.$$

Robust weighted conformal inference

Method 1: for a new unit from control group, we construct $\widehat{C}_0(X_{n+1})$ such that

$$\mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}_0(X_{n+1}) \,\big|\, T_{n+1}=0\big) \geq 1-\alpha,$$

where $\mathbb{P}$ is over all training data and the test point $(X_{n+1}, Y_{n+1}(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=0}$. This and the observed $Y_{n+1}(0)$ can be used to construct prediction interval for ITE:

$$\mathbb{P}\big(\Delta_{n+1} \in \widehat{C}_0(X_{n+1}) - Y_{n+1}(0) \,\big|\, T_{n+1}=0\big) \geq 1-\alpha.$$

Algorithm 1: Robust weighted conformal inference

- **Sample splitting**: Randomly split the $T_i = 1$ observations into two sets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{calib}}$
- **Machine learning**: Train any model of $Y(1) \mid X$ on $\mathcal{D}_{\text{train}}$, and form a residual $V(x,y)\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, e.g. $V(x,y) = |y - \widehat{f}(x)|$ or $V(x,y) = y - \widehat{f}(x)$
- **Train bounds**: If p and $e(x)$ are unknown, train $\widehat{p}$ and $\widehat{e}(x)$ on $\mathcal{D}_{\text{train}}$ to obtain

$$\widehat{\ell}(x) = \frac{1}{\Gamma} \cdot \frac{\widehat{p}}{1-\widehat{p}}\frac{1-\widehat{e}(x)}{\widehat{e}(x)}, \quad \widehat{u}(x) = \Gamma \cdot \frac{\widehat{p}}{1-\widehat{p}}\frac{1-\widehat{e}(x)}{\widehat{e}(x)}.$$

- **Calibrate ML models with robust weights**: Compute $V_i = V(X_i, Y_i)$, $\widehat{\ell}_i = \widehat{\ell}(X_i)$ and $\widehat{u}_i = \widehat{u}(X_i)$ for $i \in \mathcal{D}_{\text{calib}}$, let

$$\widehat{v} = \sup_{\widehat{\ell}_i \leq w(X_i,Y_i) \leq \widehat{u}_i} \text{Quantile}\left(1-\alpha; \sum_{i \in \mathcal{D}_{\text{calib}}} p_w^{(i)} \delta_{V_i} + p_w^{(n+1)} \delta_\infty\right),$$

where $p_w^{(i)} = w(X_i, Y_i) / \big(\sum_{j \in \mathcal{D}_{\text{calib}}} w(X_j, Y_j) + w(x,y)\big)$.

- **Conformal prediction**: Output $\widehat{C}_0(x) = \big\{y\colon V(x,y) \leq \widehat{v}\big\}$.

Then $\widehat{C}_0$ achieves valid coverage for the counterfactual under a given confounding level. In observational studies, the coverage is robust to estimation error of $\widehat{e}$.

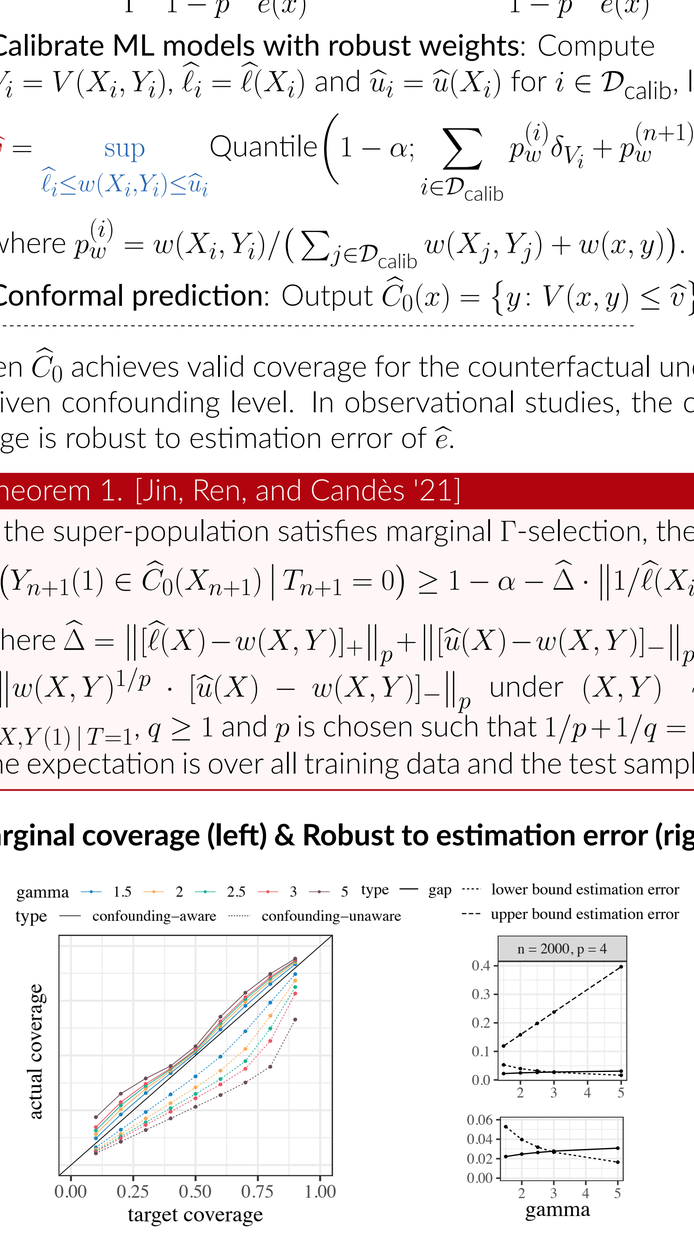
Theorem 1. [Jin, Ren, and Candès '21]

If the super-population satisfies marginal Γ-selection, then

$$\widetilde{\mathbb{P}}\big(Y_{n+1}(1) \in \widehat{C}_0(X_{n+1}) \,\big|\, T_{n+1}=0\big) \geq 1-\alpha-\widehat{\Delta} \cdot \big\|1/\widehat{\ell}(X_i)\big\|_q,$$

where $\widehat{\Delta} = \big\|[\widehat{\ell}(X) - w(X,Y)]_+\big\|_p + \big\|[\widehat{u}(X) - w(X,Y)]_-\big\|_p + \frac{1}{n}\big\|w(X,Y)^{1/p} \cdot [\widehat{u}(X) - w(X,Y)]_-\big\|_p$ under $(X,Y) \sim \mathbb{P}_{X,Y(1)\,|\,T=1}$, $q \geq 1$ and p is chosen such that $1/p + 1/q = 1$; the expectation is over all training data and the test sample.

Marginal coverage (left) & Robust to estimation error (right)



Calibration-conditional robust conformal inference

Method 2: for a new unit from control group, we construct $\widehat{C}_0(X_{n+1})$ such that with probability at least $1-\delta$ over the calibration data,

$$\mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}_0(X_{n+1}) \,\big|\, T_{n+1}=0, \mathcal{D}_{\text{calib}}\big) \geq 1-\alpha,$$

where $\mathbb{P}$ is over one test point $(X_{n+1}, Y_{n+1}(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=0}$. This and the observed $Y_{n+1}(0)$ can be used to construct prediction interval for ITE.

Algorithm 2: PAC robust conformal inference

- **Sample splitting** + **Machine learning** + **Train bounds**: same as Alg. 1
- **Robust quantile functions**: Construct envelope function

$$G(t) = \max\big\{\mathbb{E}[\mathbb{1}_{\{V(X,Y)\leq t\}}\,\widehat{\ell}(X)], 1 - \mathbb{E}[\mathbb{1}_{\{V(X,Y)>t\}}\,\widehat{u}(X)]\big\}$$

and a pointwise lower bound $\widehat{G}(t)$ from $\mathcal{D}_{\text{calib}}$, such that for any fixed t,

$$\mathbb{P}_{\mathcal{D}_{\text{calib}}}\big(\widehat{G}(t) \leq G(t)\big) \geq 1-\delta.$$

- **Calibrate ML models with robust quantile**: Let

$$\widehat{v} = \inf\big\{t : \widehat{G}_n(t) \geq 1-\alpha\big\}.$$

- **Conformal prediction**: Output $\widehat{C}_0(x) = \big\{y\colon V(x,y) \leq \widehat{v}\big\}$.

Theorem 2. [Jin, Ren, and Candès '21]

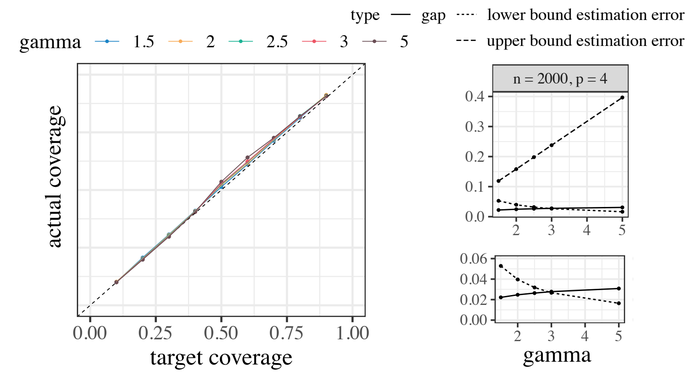
If the super-population satisfies marginal Γ-selection, then

$$\widetilde{\mathbb{P}}\big(Y_{n+1}(1) \in \widehat{C}_0(X_{n+1}) \,\big|\, \mathcal{D}, T_{n+1}=0\big) \geq 1-\alpha-\widehat{\Delta}$$

with probability at least $1-\delta$ with respect to $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{calib}}$, where

$$\widehat{\Delta} = \max\Big\{\mathbb{E}\big[(\widehat{\ell}(X_i) - w(X_i,Y_i)_+\big],\ \mathbb{E}\big[(\widehat{u}(X_i) - w(X_i,Y_i))_-\big]\Big\}.$$

0.05-quantile of coverage & Robust to estimation error



Sensitivity analysis of ITEs

For any hypothesized confounding strength Γ, our methods allow to construct robust prediction intervals $\widehat{C}(X, \Gamma)$ for ITE.

Gauge the robustness of causal conclusions

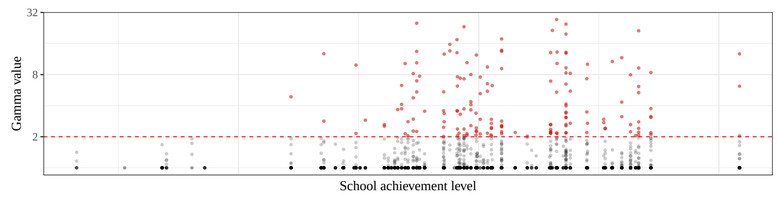
Suppose we are interested in whether $\Delta \in C$ for some set C, e.g., $C = (-\infty, 0]$. We define the Γ-value of a test individual with covariate X as

$$\widehat{\Gamma} = \sup\{\Gamma\colon C \cap \widehat{C}(X, \Gamma) = \varnothing\}.$$

This individual has positive ITE unless the confounding strength is above $\widehat{\Gamma}$.

Application to ACIC 2018 challenge dataset

Γ-values versus individual school achievement levels



References

[1] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[2] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*, 2020.

[3] Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[4] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *arXiv preprint arXiv:1711.11286*, 2017.