Conformal Counterfactual Inference under Hidden Confounding (KDD'24)

Zonghao Chen[1]*, **Ruocheng Guo[2]***, Jean-Francois Ton[2], Yang Liu[2]
1 UCL
2 ByteDance Research
* Equal contribution

Introduction

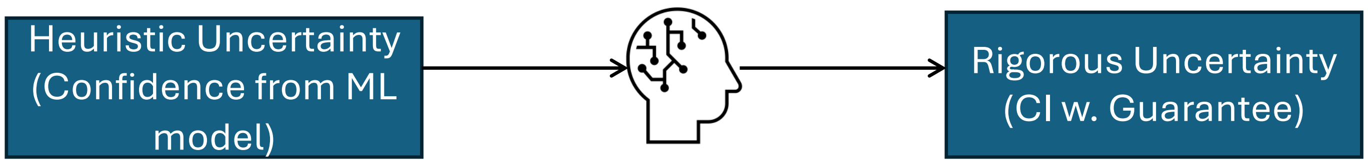
- Point estimate of outcome is often not enough for decision making in high-stake applications.

Example

A confidence interval, or at least a p-value, is required by the U.S. Food and Drug Administration to approve a drug, in order to guarantee sufficient evidence and confidence in favor of the drug [1].

[1] Lei, Lihua, and Emmanuel J. Candès. "Conformal inference of counterfactuals and individual treatment effects." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.5 (2021): 911-938.

Conformal Prediction (CP)



- CP Predicts confidence interval that has guaranteed probability to cover the ground truth
  - Weak assumption: only need exchangeability between calibration and test data, no assumption on error distribution
  - Low cost: only need inference on calibration set
  - Model agnostic: works with any ML models

How does CP work?

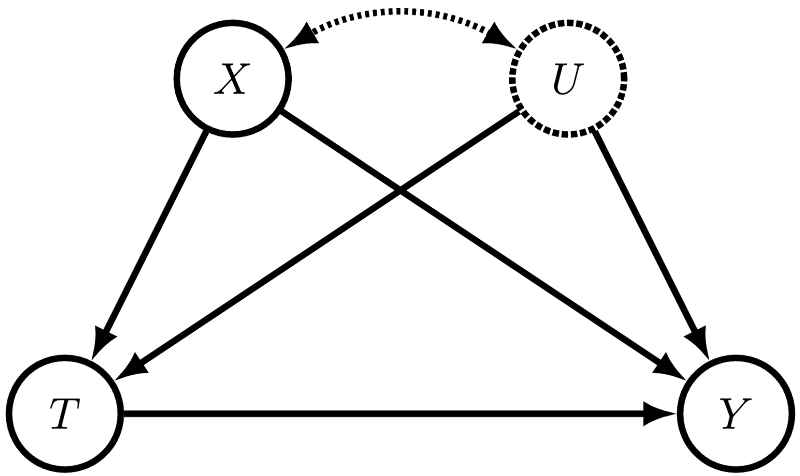
Example: Split CP for Regression

- Given a regression model, a calibration dataset, and a predefined coverage rate $1-\alpha$
    - Assumptions: calibration set is exchangeable with test set and has ground truth
- Make predictions on a calibration dataset
- Obtain distribution $\hat{F}$ of nonconformity scores $\mathrm{s_i} = |\hat{y}_i - y_i|$
- Compute the $1-\alpha$-th quantile of nonconformity scores $q_{\hat{F}}$
- Create confidence interval $C_{SCP}(x_i) = [\hat{y}_i - q_{\hat{F}}, \hat{y}_i + q_{\hat{F}}]$
- With exchangeability, test data shares same distribution F as calibration data, $C_{SCP}(x_i)$ has guaranteed marginal coverage rate $1-\alpha$

Motivating example

- What is the effect of the treatment T (pills or surgery) on the outcome Y (recovery rate), given both observed confounding X (severity of disease) and unobserved confounding U (patient adherence to treatment)?

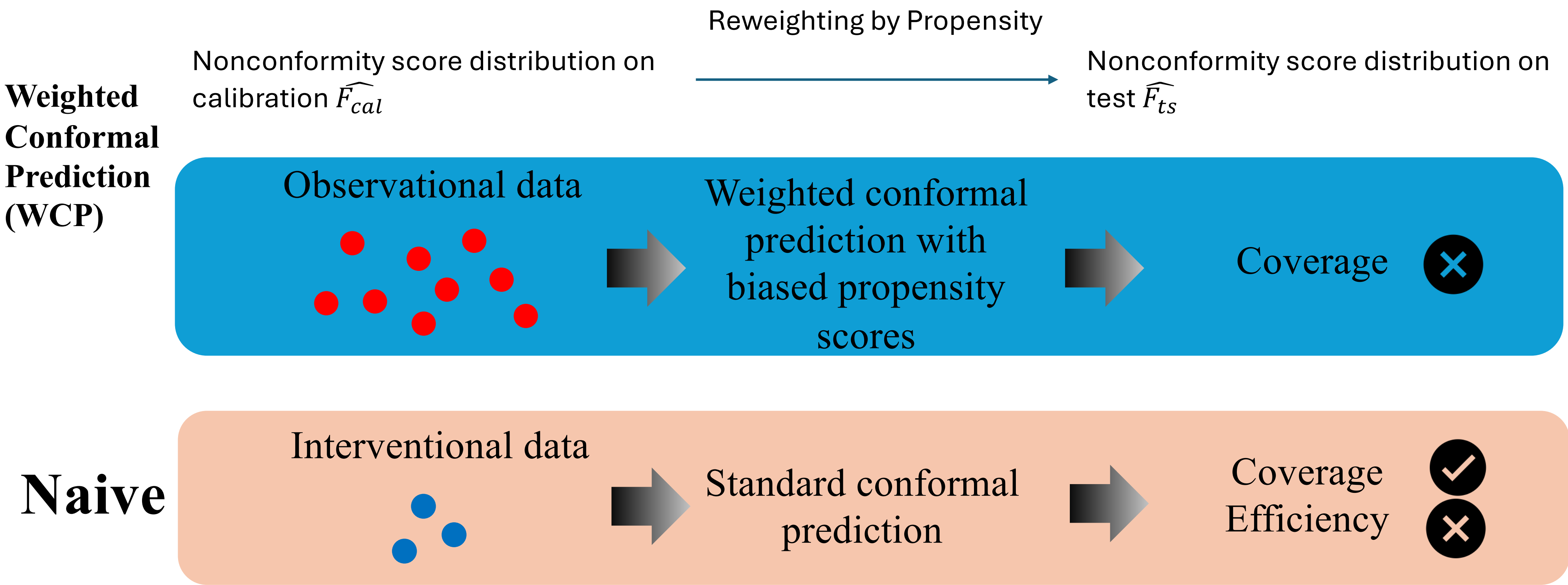
- For an individual, what is the estimated treatment effect? What is the confidence interval with guaranteed coverage of the estimate?

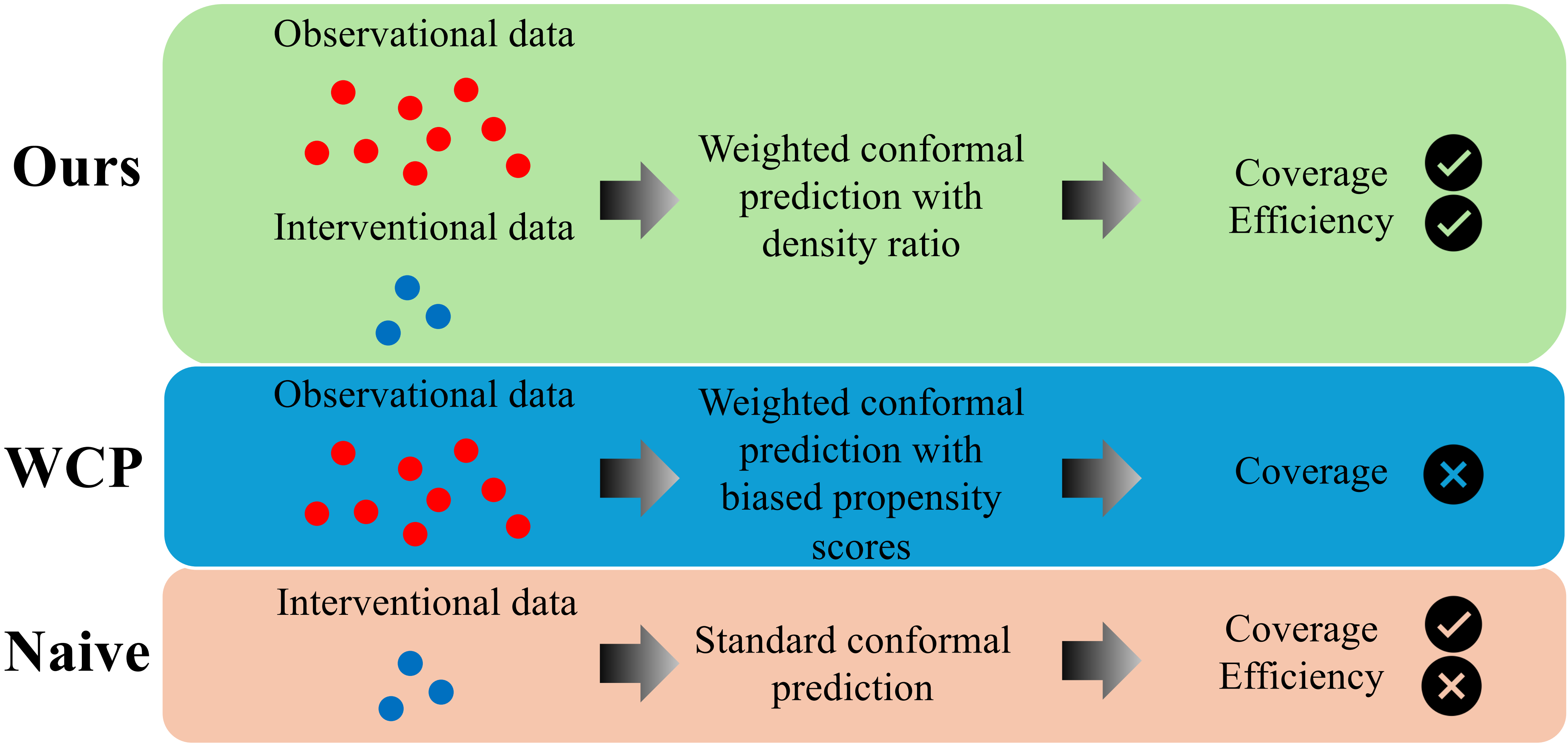


Problem

- A merged dataset with n observational and m interventional data ($n >> m$)
  - $\mathcal{D} = \{(x_i, y_i)_{i=1}^{n} \sim P_{\{X,Y\}}\} \cup \{(x_i, y_i)_{i=n+1}^{n+m} \sim P'_{\{X,Y\}}\}$.
- Goal: construct confidence interval $C(x_i)$ with guaranteed coverage rate $1-\alpha$ for potential outcomes given an unseen test sample $x_i, i > n+m$.

Existing methods




Ours
Observational data
Interventional data
Weighted conformal prediction with density ratio
Coverage
Efficiency
WCP
Observational data
Weighted conformal prediction with biased propensity scores
Coverage
Naive
Interventional data
Standard conformal prediction
Coverage
Efficiency

Theoretical Results

- Our method has guaranteed coverage.
- Under additive Gaussian noise model, our method is highly likely to have narrower confidence intervals than the naïve method.

Our method (wSCP-DR-Inexact)

- A merged dataset with n observational and m interventional data
  - $\mathcal{D} = \{(x_i, y_i)_{i=1}^{n} \sim P_{\{X,Y\}}\} \cup \{(x_i, y_i)_{i=n+1}^{n+m} \sim P'_{\{X,Y\}}\}$.
- First, we compute weights, which will be used to handle distribution shift between observational and interventional data.
- Weight functions
  - $w(x, y) = 1$ if $(x, y) \sim P_{\{X,Y\}}$ and $w(x, y) = \frac{d\, P_{\{X,Y\}}}{d\, P'_{\{X,Y\}}}(x, y)$ if $(x, y) \sim P'_{\{X,Y\}}$
- Let ${p_i}_{i=1}^{|\mathcal{D}|}$ denote the "normalized" weight functions.

Our method (wSCP-DR-Inexact)

- Obtain distribution $\widehat{F'}$ of nonconformity scores $\mathrm{s_i} = |\widehat{y_i} - y_i|$
- Reweight $\widehat{F'}$ to estimate distribution of nonconformity scores on interventional data $\hat{F} = \sum_{i \in \mathcal{D}_{cal}} p_i \delta_{s_i}$
- Compute the $1 - \alpha$-th quantile of nonconformity scores $q_{\hat{F}}$
- Confidence interval for calibration $C_{SCP}(x_i) = [\hat{y}_i - q_{\hat{F}}, \hat{y}_i + q_{\hat{F}}]$
- Fit ML models to predict lower/upper bounds using datasets $\{(x_i, \hat{y}_i - q_{\hat{F}})\}_{i \in \mathcal{D}_{cal}}$ and $\{(x_i, \hat{y}_i + q_{\hat{F}})\}_{i \in \mathcal{D}_{cal}}$
- Use these models to predict lower/upper bounds for any test sample

Experiments

- Datasets
  - Synthetic data with controllable hidden confounding
  - Real-world recommendation datasets (rating prediction)
    - Yahoo!R3
    - Coat
- Evaluation metrics
  - Coverage rate
    - probability of true potential outcome / ITE in the predicted interval
  - Interval width

Results

- Synthetic data with hidden confounding

Table 2: Results for counterfactual outcomes and ITEs on the synthetic data. We compare our methods wSCP-DR (Inexact), wSCP-DR (Inexact), and wTCP-DR with baselines. Results are shown for coverage and confidence interval width on the synthetic data with $n = 10,000$ and $m = 250$. Boldface and underlining are used to highlight the top and second-best interval width among the methods with coverage close to 0.9.

Method	Coverage $Y(0)$ ↑	Interval Width $Y(0)$ ↓	Coverage $Y(1)$ ↑	Interval Width $Y(1)$ ↓	Coverage ITE ↑	Interval Width ITE ↓
wSCP-DR(Inexact)	0.891 ± 0.026	<u>0.414</u> ± 0.008	0.889 ± 0.019	**0.421** ± 0.013	0.942 ± 0.017	**0.835** ± 0.016
wSCP-DR(Exact)	0.934 ± 0.026	0.496 ± 0.010	0.935 ± 0.023	<u>0.503</u> ± 0.010	0.957 ± 0.018	0.998 ± 0.015
wTCP-DR	0.899 ± 0.028	**0.386** ± 0.013	0.923 ± 0.015	0.576 ± 0.066	0.953 ± 0.015	<u>0.962</u> ± 0.074
WCP	0.572 ± 0.039	0.222 ± 0.007	0.608 ± 0.042	0.227 ± 0.009	0.710 ± 0.027	0.449 ± 0.012
Naive	0.932 ± 0.018	0.508 ± 0.042	0.930 ± 0.023	0.560 ± 0.049	0.952 ± 0.018	1.068 ± 0.098

Empirical Results

- Recommendation system data
  - Rating prediction with distribution shift

Table 3: Coverage and interval width results on Yahoo and Coat. Boldface and underlining are used to highlight the top and second-best interval width among the methods with coverage close to 0.9.

	Yahoo		Coat	
Method	Coverage ↑	Interval Width ↓	Coverage ↑	Interval Width ↓
wSCP-DR(Inexact)	0.892 ± 0.019	**4.353** ± 0.019	0.919 ± 0.008	**3.787** ± 0.045
wSCP-DR(Exact)	0.952 ± 0.001	5.140 ± 0.001	0.959 ± 0.001	4.565 ± 0.228
wSCP-DR*(Inexact)	0.892 ± 0.020	**4.353** ± 0.020	0.919 ± 0.008	<u>3.789</u> ± 0.046
wSCP-DR*(Exact)	0.952 ± 0.001	5.140 ± 0.001	0.960 ± 0.001	4.571 ± 0.233
WCP-NB	0.825 ± 0.002	4.036 ± 0.002	0.912 ± 0.005	3.635 ± 0.040
Naive	0.899 ± 0.001	6.047 ± 0.001	0.896 ± 0.003	7.725 ± 0.018

Take away

- We propose a simple yet effective method to handle hidden confounding for conformal counterfactual inference.
- Our method reweights nonconformity scores with density ratio of joint distributions instead of propensity scores (WCP).
- Theoretically, we prove the proposed method guarantees coverage as well as is more efficient than the naïve method.
- Empirically, experimental results support our claims.

Our paper and code can be found at

https://arxiv.org/abs/2405.12387

https://github.com/rguo12/KDD24-Conformal

Thanks for attending my talk!

Question time