# Not all distributional shifts are equal: Fine-grained robust conformal inference

Jiahao Ai[1] Zhimei Ren[2]

[1]Peking University [2]University of Pennsylvania

## Problem Setup

**Uncertainty Quantification under Distributional Shift**. Suppose we have

- Training sample $(X, Y) \sim P_{X,Y}$
- Test sample $(X, Y) \sim Q_{X,Y}$
- $P_{X,Y} \neq Q_{X,Y}$

Given a prediction model trained under $P$, how can we provide uncertain quantification for the model's performance under $Q$?

**Decomposing the Distributional Shift.** We decompose the difference between $Q_{X,Y}$ and $P_{X,Y}$ into two sources:

- the difference between $P_X$ and $Q_X \rightsquigarrow$ specified by $w(x) = \frac{dQ_X}{dP_X}(x)$ (estimable)
- the difference between $P_{Y\,|\,X}$ and $Q_{Y\,|\,X} \rightsquigarrow Q_{Y\,|\,X}$ lies in a "neighborhood ball" of $P_{Y\,|\,X}$:

$$\mathcal{P}(\rho; P_{X,Y}) := \{Q_{X,Y} \text{ s.t. } (X_{n+1}, Y_{n+1}) \sim Q_{X,Y} : D_f(Q_{Y\,|\,X} \,\|\, P_{Y\,|\,X}) \leq \rho \text{ a.s.}\}.$$

### Our Goal

Given $(X_i, Y_i)_{i=1}^n$ i.i.d. drawn from $P_{X,Y}$, suppose $Q_{X,Y}$ satisfies the above distribution shift model with parameter $\rho$, we aim to construct a prediction interval $\hat{C}_{f,\rho}(X_{n+1})$

$$\mathbb{P}_{(X_{n+1},Y_{n+1})\sim Q_{X,Y}}(Y_{n+1} \in \hat{C}_{f,\rho}(X_{n+1})) \geq 1 - \alpha,$$

for $\alpha \in (0, 1)$, with the probability over the randomness of the training and test data.

## Contributions

(1) We propose the **Weighted Robust Conformal Prediction (WRCP)** method $\rightsquigarrow$ approximate marginal coverage & miscoverage rate $\propto$ est. err. of $\mathrm{d}Q_X/\mathrm{d}P_X$

(2) We provide a **debiased variant of WRCP (D-WRCP)** $\rightsquigarrow$ double-robustness property & miscoverage rate $\propto$ est. err. of $\mathrm{d}Q_X/\mathrm{d}P_X \times$ est. err. of $Y\,|\,X$

(3) As a special example, our methodology can be adapted for sensitivity analysis of individual treatment effects (ITEs) under the $f$-sensitivity model

(4) We empirically validate the methods in simulations and real data, showing their improved efficiency over existing benchmarks

## Algorithm

**Intuition.**

- **Weight the samples** to account for the shift in $X$
- Consider an **inflated confidence level** to account for the worst-case $Y\,|\,X$-shift

**Preparations.**

- Randomly split the training set into two disjoint subsets and use one subset to determine the nonconformity score function $s(\cdot, \cdot)$
- Compute the nonconformity scores $S_i = s(X_i, Y_i)$ for $i$ in the other subset (calibration set)

**Constructing the Prediction Interval.** Recall that $w(x) = \frac{dQ_X}{dP_X}(x)$. We construct our prediction set as

$$\hat{C}_{f,\rho}(x) = \left\{y \in \mathbb{R} : s(x, y) \leq \text{Quantile}\left(g_{f,\rho}^{-1}(1-\alpha), \textstyle\sum_i p_i(x)\delta_{S_i} + p_{n+1}(x)\delta_\infty\right)\right\}, \quad (1)$$

$$\text{where } p_i(x) = \frac{w(X_i)}{\sum_j w(X_j) + w(x)}, \text{ and } p_{n+1}(x) = \frac{w(x)}{\sum_j w(X_j) + w(x)}. \quad (2)$$

The quantity $g_{f,\rho}^{-1}(1-\alpha)$ corresponds to the "inflated level" define through:

$$g_{f,\rho}(\beta) := \inf\left\{z \in [0,1] : \beta f\left(\frac{z}{\beta}\right) + (1-\beta) f\left(\frac{1-z}{1-\beta}\right) \leq \rho\right\},$$

and its inverse

$$g_{f,\rho}^{-1}(\tau) := \sup\{\beta \in [0,1] : g_{f,\rho}(\beta) \leq \tau\}.$$

**What If $w$ is Unknown?** Suppose multiple test units (only $X$ observed but missing $Y$).

- Using additional test units to estimate $w$
- Construct $\hat{C}_{f,\rho}(X)$ with $w(x)$ replaced by $\hat{w}^{(k)}(x)$

**The debiased variant.**

- Use the set-aside data to obtain an estimate $\hat{m}(x; t)$ of $\mathbb{E}[\mathbb{1}\{s(X, Y) \leq t\}\,|\,X = x]$
- Define the debiased CDF estimator

$$\hat{p}(t) = \frac{\sum_{i \text{ calib}} \hat{w}(X_i) \cdot (\mathbb{1}\{S_i \leq t\} - \hat{m}(X_i; t))}{\sum_{i \text{ calib}} \hat{w}^{(k)}(X_i)} + \frac{1}{|\mathcal{I}_{\text{test},j}|} \sum_{i \in \mathcal{I}_{\text{test},j}} \hat{m}(X_i; t),$$

where $\mathcal{I}_{\text{test},j} = \mathcal{I}_{\text{test}} \backslash \{j\}$

- We construct our prediction set as

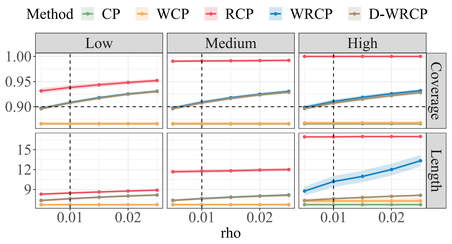
$$\hat{C}_{f,\rho,n+j}^{\text{DR}}(X_{n+j}) = \{y : s(X_{n+j}, y) \leq \hat{q}\}, \text{ where } \hat{q} = \inf\left\{t \in \mathbb{R} : \inf_{t' \geq t} \hat{p}(t') \geq g_{f,\rho}^{-1}(1-\alpha)\right\}.$$

## Experiments

**Benchmarks.** (1) `CP`: standard conformal prediction, (2)`WCP`: weighted conformal prediction, and (3) `RCP`: robust conformal prediction.

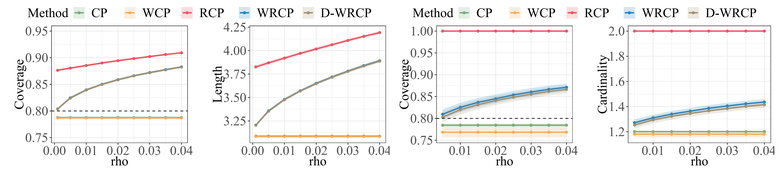
The horizontal dashed line $\rightsquigarrow$ target coverage rate; the vertical dashed line $\rightsquigarrow$ true robust parameter.

**Simulations**. Low, medium and high levels of covariate shift.




**Real data.**

- Left: National Study of Learning Mindsets. Predicting the potential outcome of instilling a growth mindset in the control group.
- Right: ACS Income Dataset. Predicting whether an individual's annual income is above 50,000 dollars, where we choose the data from New York as the training set, and that from South Dakota as the target.




## Future work

1. Investigate methods for identifying (an upper bound) of the robust parameter $\rho$ when we have a small amount of supervised data from the target population.
2. Extend this fine-grained approach to other distributional shift models (e.g., the multi-group model) and improves the efficiency of the corresponding methodologies.
3. Find optimal decomposition in different settings and the corresponding treatments.

**CODE**: `https://github.com/zhimeir/finegrained-conformal-paper`.