Conformal Inference under Distributional Shift

Patrick Chao and Jeffrey Zhang

University of Pennsylvania

3/1/22 and 3/3/22

Table of Contents

Table of Contents

Conformal Prediction under Covariate Shift [Tibshirani et al., 2019]

Regression Setting: Observations $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i \in [n]$

Goal: Construct band $\hat{C}_n : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}\}$ based on the data $(X_i, Y_i)_{i=1}^n$ such that for a new (X_{n+1}, Y_{n+1})

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_n(X_{n+1})\right] \geq 1 - \alpha$$

Question: What can we do if the distribution of X_{n+1} is different?

Conformal Prediction

Notation:

- $Z_i = (X_i, Y_i)$, $Z_{1:n} = \{Z_1, \ldots, Z_n\}$, $Z_{-i} = Z_{1:n} \setminus \{Z_i\}$
- Score function S, small values of $S((x,y), Z)$ imply (x, y) conforms to Z
  - Example: $S((x,y), Z) = |y - \hat{\mu}(x)|$ for some regression function $\hat{\mu}$ trained on Z
- Nonconformity Scores:

$$V_i^{(x,y)} = S(Z_i, Z_{-i} \cup \{(x,y)\}), \quad V_{n+1}^{(x,y)} = S((x,y), Z_{1:n}),$$

- Let Quantile$(\beta; F)$ denote the β-th quantile of a distribution function F, i.e.

$$\text{Quantile}(\beta; F) = \inf\{z : F(z) \geq \beta\}$$

Use this to construct conformal confidence interval

$$\hat{C}_n(x) = \left\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq \text{Quantile}(1-\alpha; V_{1:n}^{(x,y)} \cup \{\infty\}) \right\}$$

Conformal Prediction Theorem

Theorem 1.

For exchangeable $(X_i, Y_i)_{i=1}^{n+1} \in \mathbb{R}^d \times \mathbb{R}$ *and any* $\alpha \in (0, 1)$, *then*

$$\hat{C}_n(x) = \left\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq \mathrm{Quantile}(1 - \alpha; V_{1:n}^{(x,y)} \cup \{\infty\}) \right\}$$

satisfies

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_n(X_{n+1}) \right] \geq 1 - \alpha.$$

If $V_1^{(X_{n+1}, Y_{n+1})}, \ldots, V_{n+1}^{(X_{n+1}, Y_{n+1})}$ *are distinct with probability* 1, *then the probability is upper bounded by* $1 - \alpha + 1/(n+1)$.

Covariate Shift

Consider the setting where the data are no longer exchangeable

$$
\begin{aligned}
(X_i, Y_i) &\overset{\text{i.i.d.}}{\sim} P = P_X \times P_{Y|X},\ i \in [n] \\
(X_{n+1}, Y_{n+1}) &\sim \tilde{P} = \tilde{P}_X \times P_{Y|X},\ \text{independently}
\end{aligned} \tag{1}
$$

Main Idea: If we preserve the conditional distribution $Y \mid X$ and we have knowledge of the covariate likelihood ratios $d\tilde{P}_X / dP_X$, can construct prediction band with marginal coverage

Covariate Shift

Idea: Use weights proportional to the likelihood ratio

- $w(X_i) = \frac{d\tilde{P}_X(X_i)}{dP_X(X_i)}$
- Conformal weights

$$p_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)}, \ i \in [n]$$

$$p_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}$$

Covariate Shift Conformal Prediction Theorem

Theorem 2.

Under model (1)*, for any score function* S *and* $\alpha \in (0,1)$*, then*

$$\hat{C}_n(x) = \left\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^w(x)\delta_{V_i^{(x,y)}} + p_{n+1}^w(x)\delta_\infty\right)\right\}$$

satisfies

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_n(X_{n+1})\right] \geq 1-\alpha.$$

Empirical Performance: Setup

Airfoil Dataset from UCI Machine Learning Repository

- $X \in \mathbb{R}^5$: log frequency, angle of attack, chord length, free-stream velocity, suction side log displacement thickness
- $Y \in \mathbb{R}$: scaled sound pressure of NASA airfoils
- $N = 1503$ observations
  - D_{pre}: 25% of the data to fit regression function μ_0
  - D_{train}: 25% of the data to compute residual quantiles for conformal prediction interval
  - D_{test}: 50% of the data for test set
  - D_{shift}: Sampled 25% from D_{test} with replacement with probabilities proportional to

    $$w(x) = \exp(x^\top \beta),\ \beta = (-1, 0, 0, 0, 1)$$

    Implies $d\tilde{P}_X \propto \exp(x^\top \beta) dP_X$[1]
- Simulations use 5000 random splits and $\alpha = 0.1$

[1] form of exponential tilting

Estimate Likehood Ratio

Consider $X_1, \ldots, X_n$ from D_{train} and $X_{n+1}, \ldots, X_{n+m}$ from D_{shift}

Procedure:

- Fit a classifier for (X_i, C_i) where $C_i = 0$ for $i = 1, \ldots, n$ and $C_i = 1$ for $i = n+1, \ldots, n+m$
- From Bayes Rule

$$\frac{\mathbb{P}[C = 1 \mid X = x]}{\mathbb{P}[C = 0 \mid X = x]} = \frac{\mathbb{P}[C = 1]}{\mathbb{P}[C = 0]} \frac{d\tilde{P}_X}{dP_X}(x)$$

- If $\hat{p}(x)$ is our estimate of $\mathbb{P}[C = 1 \mid X = x]$ from our classifier,

$$\hat{w}(x) = \frac{\hat{p}(x)}{1 - \hat{p}(x)}.$$

Estimate Likehood Ratio

Consider $X_1, \ldots, X_n$ from D_{train} and $X_{n+1}, \ldots, X_{n+m}$ from D_{shift}

Procedure:

- Fit a classifier for (X_i, C_i) where $C_i = 0$ for $i = 1, \ldots, n$ and $C_i = 1$ for $i = n+1, \ldots, n+m$
- From Bayes Rule

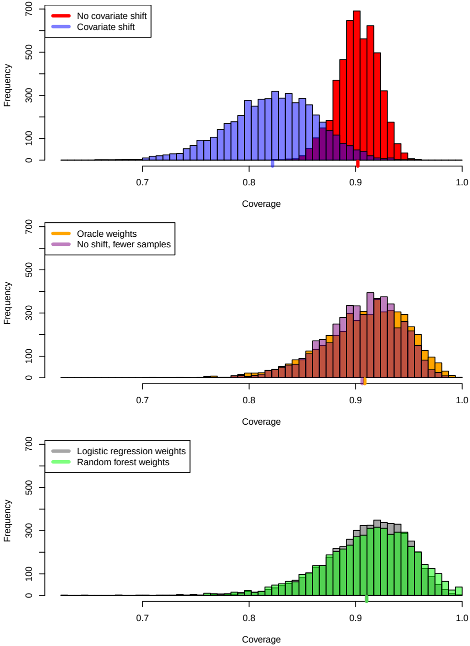
$$\frac{\mathbb{P}[C = 1 \mid X = x]}{\mathbb{P}[C = 0 \mid X = x]} = \frac{\mathbb{P}[C = 1]}{\mathbb{P}[C = 0]} \frac{d\tilde{P}_X}{dP_X}(x)$$

- If $\hat{p}(x)$ is our estimate of $\mathbb{P}[C = 1 \mid X = x]$ from our classifier,

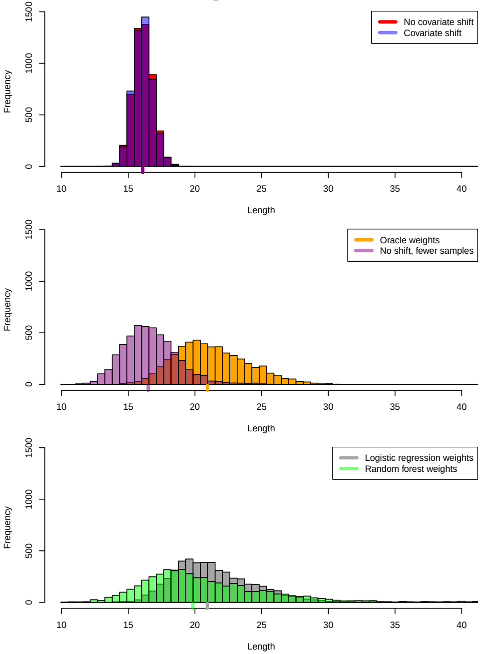
$$\hat{w}(x) = \frac{\hat{p}(x)}{1 - \hat{p}(x)}.$$

We do not need $\mathbb{P}[C = 1]$ since Theorem 2 holds when $w(x)$ is known up to a proportionality constant

Empirical Performance: Coverage



Empirical Performance: Length



Weighted Exchangeability

Definition 3.

Random variables $V_1, \ldots, V_n$ are weighted exchangeable with weight functions $w_1, \ldots, w_n$ if the density f of their joint distribution can be factorized as

$$f(v_1, \ldots, v_n) = \prod_{i=1}^{n} w_i(v_i) \cdot g(v_1, \ldots, v_n)$$

where g is any function that is independent on ordering, i.e.

$$g(v_{\sigma(1)}, \ldots, v_{\sigma(n)}) = g(v_1, \ldots, v_n)$$

for $\sigma \in S_n$.

Weighted Exchangeability

Definition 3.

Random variables $V_1, \ldots, V_n$ are weighted exchangeable with weight functions $w_1, \ldots, w_n$ if the density f of their joint distribution can be factorized as

$$f(v_1, \ldots, v_n) = \prod_{i=1}^{n} w_i(v_i) \cdot g(v_1, \ldots, v_n)$$

where g is any function that is independent on ordering, i.e.

$$g(v_{\sigma(1)}, \ldots, v_{\sigma(n)}) = g(v_1, \ldots, v_n)$$

for $\sigma \in S_n$.

Example 4.

If $Z_i \overset{\text{ind.}}{\sim} P_i$ for $i \in [n]$ and P_i are absolutely continuous w.r.t. P_1, then $Z_1, \ldots, Z_n$ are weighted exchangeable with $w_1 \equiv 1$ and $w_i = dP_i/dP_1$.

Weighted Conformal Prediction

Theorem 5.

Let Z_i for $i \in [n+1]$ be weighted exchangeable random variables with weights $w_1, \ldots, w_{n+1}$. Let $V_i = S(Z_i, Z_{-i})$ for any score function S. Define

$$p_i^w(z_1, \ldots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}$$

For $\alpha \in (0,1)$,

$$\mathbb{P}\left[V_{n+1} \leq \mathrm{Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^w(Z_{1:n+1})\delta_{V_i} + p_{n+1}^w(Z_{1:n+1})\delta_{\infty}\right)\right] \geq 1-\alpha.$$

Weighted Conformal Prediction: Proof Sketch

Condition on event E_z where $\{Z_1, \ldots, Z_{n+1}\} = \{z_1, \ldots, z_{n+1}\}$ and $v_i = S(z_i, z_{-i})$

p_i^w is constructed so that from weighted exchangeability,

$$\mathbb{P}[V_{n+1} = v_i \mid E_z] = p_i^w(z_1, \ldots, z_{n+1}) \implies$$

$$V_{n+1} \mid E_z \sim \sum_{i=1}^{n+1} p_i^w(z_1, \ldots, z_{n+1}) \delta_{V_i} \implies$$

$$\mathbb{P}\left[V_{n+1} \leq \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^w(Z_{1:n+1})\delta_{V_i} + p_{n+1}^w(Z_{1:n+1})\delta_\infty\right)\right] \geq 1-\alpha.$$

Table of Contents

Problem Statement

Consider the more general online setting

Observations: $\{(X_r, Y_r)\}_{1 \leq r \leq t-1}$ and new covariate X_t where $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$

Goal: Construct prediction set $\hat{C}_t(X_t)$ for Y_t with marginal coverage

Challenge: No exchangeability or assumption on distribution of (X_i, Y_i), e.g. there may be distribution shifts for every observation rather than a train and shifted distribution

Prediction Set

Score Function: $S_t(X_t, y)$, e.g.

$$S_t(X_t, y) = |y - \hat{\mu}(X_t)|$$

Calibration Set: $\mathcal{D}_{\mathsf{cal}} \subseteq \{(X_r, Y_r)\}_{1 \leq r \leq t-1}$ distinct from data used to fit model

$$\hat{Q}_t(p) = \inf \left\{ s : \left(\frac{1}{|\mathcal{D}_{\mathsf{cal}}|} \sum_{(X_r, Y_r) \in \mathcal{D}_{\mathsf{cal}}} \mathbb{1}\{S(X_r, Y_r) \leq s\} \right) \geq p \right\}$$

Typical Prediction Set:

$$\hat{C}_t(X_t) = \{y : S_t(X_t, y) \leq \hat{Q}_t(1 - \alpha)\}$$

Miscoverage Rate

$$M_t(\alpha) = \mathbb{P}\left[S_t(X_t, Y_t) > \hat{Q}_t(1-\alpha)\right]$$

Ideally $M_t(\alpha) \approx \alpha$, but due to distributional shift this may be very different

Idea: There may be a different value $\alpha_t^\star \in [0,1]$ s.t. $M_t(\alpha_t^\star) = \alpha$

Question: Can we adaptively choose $\alpha_t^\star$ and use it as our cutoff?

Miscoverage Rate

$$M_t(\alpha) = \mathbb{P}\left[S_t(X_t, Y_t) > \hat{Q}_t(1-\alpha)\right]$$

Ideally $M_t(\alpha) \approx \alpha$, but due to distributional shift this may be very different

Idea: There may be a different value $\alpha_t^\star \in [0,1]$ s.t. $M_t(\alpha_t^\star) = \alpha$

Question: Can we adaptively choose $\alpha_t^\star$ and use it as our cutoff?

Answer: Yes! Field of Online Learning

Forecasters

Consider a Reasonably Young And Naive (RYAN) forecaster that uses the same α for every observation

- Large miscoverage rates $M_t(\alpha)$
- Cannot handle distribution shifts

Forecasters

Consider a Reasonably Young And Naive (RYAN) forecaster that uses the same α for every observation

- Large miscoverage rates $M_t(\alpha)$
- Cannot handle distribution shifts

Consider a BRilliant Adaptive Yet Nuanced (BRYAN) forecaster

- Can statistically guarantee coverage on average even under arbitrary distributional shift

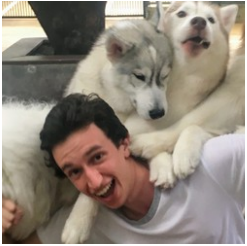


Figure: Ryan Brill, AMCS PhD Student

Forecasters

Consider a Reasonably Young And Naive (RYAN) forecaster that uses the same α for every observation

- Large miscoverage rates $M_t(\alpha)$
- Cannot handle distribution shifts

Consider a BRilliant Adaptive Yet Nuanced (BRYAN) forecaster

- Can statistically guarantee coverage on average even under arbitrary distributional shift

Figure: Bryan Frill, BRYAN forecaster

Adaptive Conformal Inference (ACI) [Gibbs and Candes, 2021]

The BRYAN forecaster monitors empirical miscoverage frequency and updates α_t

$$\hat{C}_t(\alpha_t) = \{y : S_t(X_t, y) \leq \hat{Q}_t(1 - \alpha_t)\}$$
$$\mathrm{err}_t = \mathbb{1}\{Y_t \notin \hat{C}_t(\alpha_t)\}$$

Adaptive Conformal Inference (ACI) [Gibbs and Candes, 2021]

The BRYAN forecaster monitors empirical miscoverage frequency and updates α_t

$$\hat{C}_t(\alpha_t) = \{y : S_t(X_t, y) \leq \hat{Q}_t(1 - \alpha_t)\}$$
$$\mathrm{err}_t = \mathbb{1}\{Y_t \notin \hat{C}_t(\alpha_t)\}$$

Online Updates with step size $\gamma > 0$

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t)$$

Adaptive Conformal Inference (ACI) [Gibbs and Candes, 2021]

The BRYAN forecaster monitors empirical miscoverage frequency and updates α_t

$$\hat{C}_t(\alpha_t) = \{y : S_t(X_t, y) \leq \hat{Q}_t(1 - \alpha_t)\}$$
$$\mathrm{err}_t = \mathbb{1}\{Y_t \notin \hat{C}_t(\alpha_t)\}$$

Online Updates with step size $\gamma > 0$

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t)$$

Smoothed Updates with $\sum_{s=1}^{t} w_s = 1$, e.g. $w_s \propto 0.95^{t-s}$

$$\alpha_{t+1} = \alpha_t + \gamma\left(\alpha - \sum_{s=1}^{t} w_s \mathrm{err}_s\right)$$

Coverage Guarantees I

Lemma 6.

With probability one, for all $t \geq 1$, $\alpha_t \in [-\gamma, 1+\gamma]$.

Coverage Guarantees I

Lemma 6.

With probability one, for all $t \geq 1$, $\alpha_t \in [-\gamma, 1+\gamma]$.

Proof.

For the sake of contradiction, assume $\alpha_{t+1} = \inf_i \alpha_i < -\gamma$. Then $\alpha_t < 0$ since

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t).$$

However if $\alpha_t < 0$, then

$$\alpha_t < 0 \implies \hat{Q}_t(1-\alpha_t) = \infty \implies$$
$$\mathrm{err}_t = 0 \implies \alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t) \geq \alpha_t$$

This is a contradiction, since $\alpha_{t+1} = \inf_i \alpha_i$ (similar argument for $\alpha_t \leq 1+\gamma$) □

Coverage Guarantees II

Theorem 7.

With probability one, for all $T \geq 1$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t - \alpha \right| \leq \frac{\max(\alpha_1, 1 - \alpha_1) + \gamma}{T\gamma}$$

and

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t \overset{\text{a.s.}}{=} \alpha.$$

Coverage Guarantees II

Theorem 7.

With probability one, for all $T \geq 1$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t - \alpha \right| \leq \frac{\max(\alpha_1, 1 - \alpha_1) + \gamma}{T\gamma}$$

and

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t \overset{\text{a.s.}}{=} \alpha.$$

Proof.

$$\alpha_{T+1} = \alpha_1 + \sum_{t=1}^{T} \gamma(\alpha - \mathrm{err}_t) \implies \frac{\alpha_1 - \alpha_{T+1}}{\gamma T} = \frac{\sum_{t=1}^{T} \mathrm{err}_t - \alpha}{T}$$

Apply Lemma 6 to obtain desired inequality □

Empirical Performance: Market Volatility Coverage

Setup: Daily open prices for stock $\{P_t\}_{1\leq t\leq T}$, Returns $R_t = (P_t - P_{t-1})/P_{t-1}$

Goal: Prediction set for volatility $V_t = R_t^2$ with $\alpha = 0.1$

Local Coverage Frequency:

$$\mathsf{localCov}_t = 1 - \frac{1}{500}\sum_{r=t-250+1}^{t+250} \mathrm{err}_r$$

Bernoulli Baseline: $\{I_t\}_{1\leq t\leq T}$ i.i.d. Bernoulli(0.1)

$$1 - \frac{1}{500}\sum_{r=t-250+1}^{t+250} I_r$$

Empirical Performance: Market Volatility Coverage

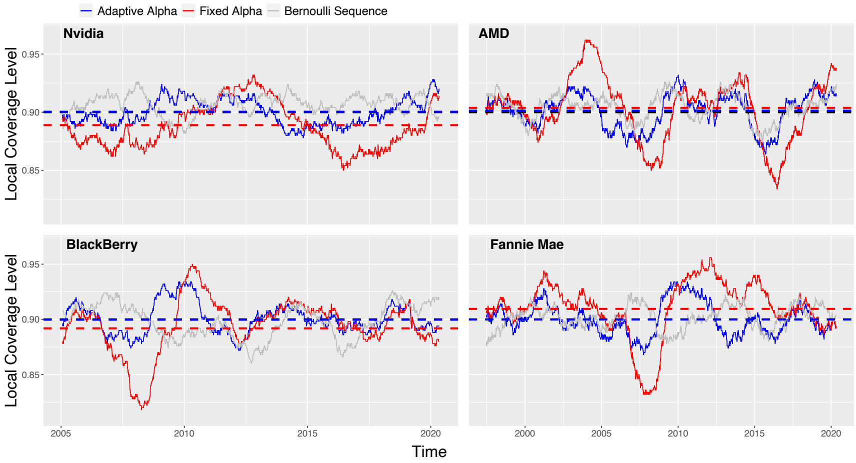


Table of Contents

A Standard Causal Inference Setting [Lei and Candes, 2021]

Setup/Notation:

- n subjects
- T_i the binary treatment indicator
- $(Y_i(1), Y_i(0))$ the potential outcomes under treatment and control
- X_i the vector of covariates
- Assume $(Y_i(1), Y_i(0), T_i, X_i) \overset{iid}{\sim} (Y(1), Y(0), T, X)$ (superpopulation)

Under the stable unit treatment value assumption, the observed dataset comprises triples $(Y_i^{\text{obs}}, T_i, X_i)$ where

$$Y_i^{\text{obs}} = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

The individual treatment effect τ_i is defined as $Y_i(1) - Y_i(0)$

Goal: Construct $\hat{C}_1 : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}\}$ such that

$$\mathbb{P}\left[Y(1) \in \hat{C}_1(X)\right] \geq 1 - \alpha$$

Connection to Weighted Conformal Prediction

Strong Ignorability: $(Y(1), Y(0)) \perp\!\!\!\perp T | X$

Observe that in the treated group, we have

$$\begin{aligned} P(X, Y^{\text{obs}} | T = 1) &= P(X | T = 1) P(Y^{\text{obs}} | X, T = 1) \\ &= P(X | T = 1) P(Y(1) | X, T = 1) \\ &= P(X | T = 1) P(Y(1) | X) \end{aligned}$$

From the treated group, we observe data from $P_{X|T=1} \times P_{Y(1)|X}$ but we want prediction regions for $P_X \times P_{Y(1)|X}$. This is exactly the setting from weighted conformal prediction with weights $dP_X / dP_{X|T=1} \propto 1/e(X)$ where $e(X) := \mathbb{P}[T = 1 | X]$ is the propensity score.

Connection to Weighted Conformal Prediction

If we were instead interested in prediction regions for $P_{X|T=1} \times P_{Y(1)|T=1,X}$, the weights would simply be 1. If we were instead interested in prediction regions for $P_{X|T=0} \times P_{Y(1)|T=0,X}$, the weights would be $dP_{X|T=0}/dP_{X|T=1} \propto (1 - e(X))/e(X)$. The below table summarizes weights for specific settings of interest:

Table 1. Summary of weight functions for different inferential targets

Inferential type	ATE	ATT	ATC	General
$w_1(x)$	$1/e(x)$	1	$(1-e(x))/e(x)$	$(dQ/dP)(x)/e(x)$
$w_0(x)$	$1/(1-e(x))$	$e(x)/(1-e(x))$	1	$(dQ/dP)(x)/(1-e(x))$

The weighted conformal framework requires the weights to be known; in practice they are estimated as the propensity score is unknown.

Algorithm

The authors propose to use a weighted version of conformalized quantile regression. The algorithm is below:

Algorithm 1 Weighted split-CQR

Input: level α, data $\mathcal{Z} = (X_i, Y_i)_{i \in \mathcal{I}}$, testing point x, function $\hat{q}_\beta(x; \mathcal{D})$ to fit β-th conditional quantile and function $\hat{w}(x; \mathcal{D})$ to fit the weight function at x using $\mathcal{D}$ as data

Procedure:

1: Split $\mathcal{Z}$ into a training fold $\mathcal{Z}_{\text{tr}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{tr}}}$ and a calibration fold $\mathcal{Z}_{\text{ca}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{ca}}}$

2: For each $i \in \mathcal{I}_{\text{ca}}$, compute the score $V_i = \max\{\hat{q}_{\alpha_{\text{lo}}}(X_i; \mathcal{Z}_{\text{tr}}) - Y_i, Y_i - \hat{q}_{\alpha_{\text{hi}}}(X_i; \mathcal{Z}_{\text{tr}})\}$

3: For each $i \in \mathcal{I}_{\text{ca}}$, compute the weight $W_i = \hat{w}(X_i; \mathcal{Z}_{\text{tr}}) \in [0, \infty)$

4: Compute the normalized weights $\hat{p}_i(x) = \frac{W_i}{\sum_{i \in \mathcal{I}_{\text{ca}}} W_i + \hat{w}(x; \mathcal{Z}_{\text{tr}})}$ and $\hat{p}_\infty(x) = \frac{\hat{w}(x; \mathcal{Z}_{\text{tr}})}{\sum_{i \in \mathcal{I}_{\text{ca}}} W_i + \hat{w}(x; \mathcal{Z}_{\text{tr}})}$

5: Compute $\eta(x)$ as the $(1-\alpha)$-th quantile of the distribution $\sum_{i \in \mathcal{I}_{\text{ca}}} \hat{p}_i(x) \delta_{V_i} + \hat{p}_\infty(x) \delta_\infty$

Output: $\hat{C}(x) = [\hat{q}_{\alpha_{\text{lo}}}(x; \mathcal{Z}_{\text{tr}}) - \eta(x), \hat{q}_{\alpha_{\text{hi}}}(x; \mathcal{Z}_{\text{tr}}) + \eta(x)]$

Validity Theorem

Theorem 8.

Let $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} (X, Y) \sim P_X \times P_{Y|X}$ *and* Q_X *be another distribution. Set* $N = |\mathcal{Z}_{\mathrm{tr}}|$ *and* $n = |\mathcal{Z}_{\mathrm{ca}}|$. *Further, let* $\hat{q}_{\beta,N}(x)$ *be an estimate of the* β*-th conditional quantile* $q_\beta(x)$ *of* $Y \mid X = x, \hat{w}_N(x)$ *be an estimate of* $w(x) = (dQ_X/dP_X)(x)$, *and* $\hat{C}_{N,n}(x)$ *be the conformal interval resulting from Algorithm 1. Then*

$$\mathbb{P}_{(X,Y)\sim Q_X \times P_{Y|X}}\left(Y \in \hat{C}_{N,n}(X)\right) \geq 1 - \alpha - \frac{1}{2}\mathbb{E}_{X\sim P_X}\left|\hat{w}_N(X) - w(X)\right|.$$

Proof Intuition

For $\mathcal{E}(V)$ the set of unordered V and for $\mathcal{E}\left(v^*\right)$ a particular set of v^*, $\left(V_{n+1} \mid \mathcal{E}(V)=\mathcal{E}\left(v^*\right), \mathcal{Z}_{\mathrm{tr}}\right) \sim \sum_{i=1}^{n+1} p_i\left(X_{n+1}\right) \delta_{v_i^*}$

Next, define a measure $\tilde{Q}_X$ such that $\tilde{Q}_X=\hat{w}(x) d P_X$
$\left(\tilde{V}_{n+1} \mid \mathcal{E}(\tilde{V})=\mathcal{E}\left(v^*\right), \mathcal{Z}_{\mathrm{tr}}\right) \sim \sum_{i=1}^{n+1} \hat{p}_i\left(\tilde{X}_{n+1}\right) \delta_{v_i^*}$

There is a lemma that states
$d_{\mathrm{TV}}\left(Q_X \times P_{Y \mid X}, \tilde{Q}_X \times P_{Y \mid X}\right)=d_{\mathrm{TV}}\left(Q_X, \tilde{Q}_X\right)$

$$
\begin{aligned}
\left|\mathbb{P}\left(Y_{n+1} \in \hat{C}\left(X_{n+1}\right) \mid \mathcal{Z}_{\mathrm{tr}}, \mathcal{Z}_{\mathrm{ca}}\right)-\mathbb{P}\left(\tilde{Y}_{n+1} \in \hat{C}\left(\tilde{X}_{n+1}\right) \mid \mathcal{Z}_{\mathrm{tr}}, \mathcal{Z}_{\mathrm{ca}}\right)\right| & \\
\leq d_{\mathrm{TV}}\left(Q_X, \tilde{Q}_X\right) \ldots & \\
\mathbb{P}\left(Y_{n+1} \in \hat{C}\left(X_{n+1}\right) \mid \mathcal{Z}_{\mathrm{tr}}\right) \geq 1-\alpha-d_{\mathrm{TV}}\left(Q_X, \tilde{Q}_X\right) \ldots & \\
\mathbb{P}\left(Y_{n+1} \in \hat{C}\left(X_{n+1}\right)\right) \geq 1-\alpha-\frac{1}{2} \mathbb{E}_{X \sim P_X}|\hat{w}(X)-w(X)| &
\end{aligned}
$$

Simulation Study

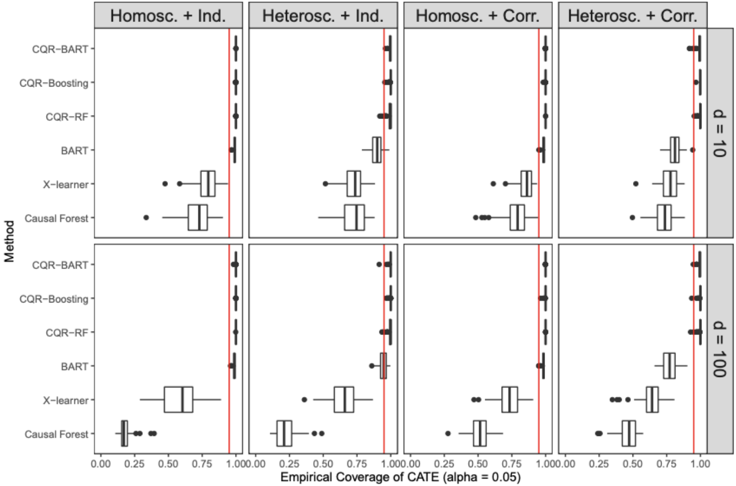
Setup: Generate 100 datasets of 1000 subjects and 10000 test subjects. Compare weighted CQR with BART, Causal Forest, and X-Learner on marginal coverage of the CATE, defined as

$$(1/n_{\text{test}}) \sum_{i=1}^{n_{\text{test}}} \mathbf{1}(\tau(X_i) \in \hat{C}_{ITE}(X_i))$$

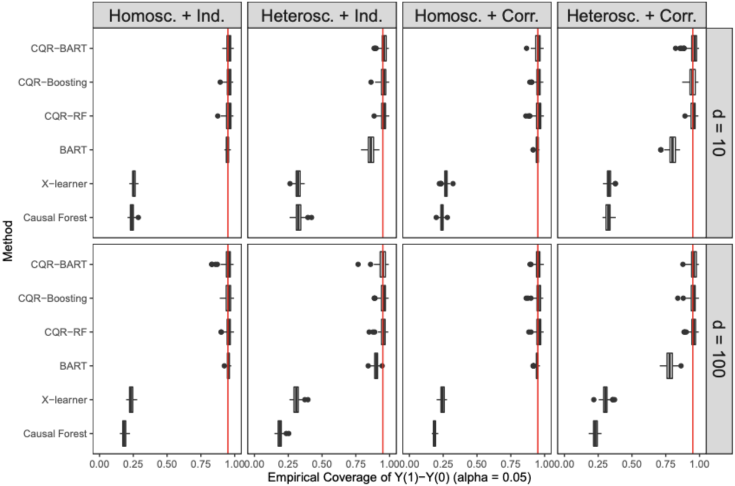
and marginal coverage of the ITE, defined as

$$(1/n_{\text{test}}) \sum_{i=1}^{n_{\text{test}}} \mathbf{1}(Y_i(1) - Y_i(0) \in \hat{C}_{ITE}(X_i))$$

Simulation Results



Simulation Results



Violation of Strong Ignorability [Jin et al., 2021]

Strong ignorability is typically an untenable assumption. Assume instead that for some unmeasured confounder U, we have

$$(Y(1), Y(0)) \perp\!\!\!\perp T | X, U$$

Without further assumptions, cannot proceed. Instead, bound the "strength" of the unmeasured confounder.

Definition 9.

A distribution $\mathbb{P}$ over $(X, U, T, Y(0), Y(1))$ satisfies the marginal Γ-selection condition if for $\mathbb{P}$-almost all x and u

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

Goal: Construct $\hat{C}_1 : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}\}$ such that

$$\mathbb{P}\left[Y(1) \in \hat{C}_1(X, \Gamma)\right] \geq 1 - \alpha$$

for all $\mathbb{P}^{\text{sup}}$ that satisfy the Γ-marginal selection.

A Review of Weighted Conformal Inference

Suppose we have (X_i, Y_i) from distribution $\mathbb{P}$ and a test point (X_{n+1}, Y_{n+1}) from a different distribution $\tilde{\mathbb{P}}$, where $\tilde{\mathbb{P}}$ is within a bounded distance of $\mathbb{P}$ in that it belongs to the set

$$\{\tilde{\mathbb{P}} : l(x) \leq \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}(x, y) \leq u(x)\}$$

If we assume $w(x, y) = \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}(x, y)$ is known exactly, we can directly apply the results of Tibshirani et al. [2019].

- Non-conformity score V
- $p_i^w(x, y) := \frac{w(X_i, Y_i)}{\sum_{j=1}^n w(X_j, Y_j) + w(x, y)}, i = 1, \ldots, n,$
  $p_{n+1}^w(x, y) := \frac{w(x, y)}{\sum_{j=1}^n w(X_j, Y_j) + w(x, y)}$
- $\widehat{C}\left(X_{n+1}\right) = \left\{y : V\left(X_{n+1}, y\right) \leq \widehat{V}_{1-\alpha}(y)\right\}$ where $\widehat{V}_{1-\alpha}(y) =$ Quantile $\left(1 - \alpha, \sum^n p_i^w\left(X_{n+1}, y\right) \cdot \delta_{V_i} + p_{n+1}^w\left(X_{n+1}, y\right) \cdot \delta_{\infty}\right)$

Robust Weighted Procedure

- Let $V_{[1]} \leq \ldots \leq V_{[n]}$ be a reordering of the conformity scores on the calibration set
- Let $\hat{l}(x)$ and $\hat{u}(x)$ be estimated upper and lower bounds on $w(x, y)$
- Define $l_i = \hat{l}(X_i)$ and $u_i = \hat{u}(X_i)$

Take the prediction interval $\widehat{C}\left(X_{n+1}\right) = \left\{y : V\left(X_{n+1}, y\right) \leq \widehat{V}_{k^*}\right\}$ where

$k^* = \min\{k : \widehat{F}(k) \geq 1 - \alpha\}, \quad \widehat{F}(k) = \frac{\sum_{i=1}^{k} \ell_{[i]}}{\sum_{i=1}^{k} \ell_{[i]} + \sum_{i=k+1}^{n} u_{[i]} + u_{n+1}}$

The $\widehat{F}(k)$ can be shown to be the solution to

$$\operatorname{minimize}_W \frac{\sum_{i=1}^{k} W_{[i]}}{\sum_{i=1}^{n} W_i + W_{n+1}}$$

subject to $\widehat{\ell}\left(X_i\right) \leq W_i \leq \widehat{u}\left(X_i\right), \quad \forall i \in \mathcal{D}_{\text{calib}} \cup \{n+1\}$

Robust Weighted Procedure

The procedure is summarized below:

Algorithm 1 Robust conformal prediction: the marginal procedure

Input: Calibration data $\mathcal{D}_{\text{calib}}$, bounds $\widehat{\ell}(\cdot)$, $\widehat{u}(\cdot)$, non-conformity score function $V\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, test covariate x, target level $\alpha \in (0,1)$.

1: For each $i \in \mathcal{D}_{\text{calib}}$, compute $V_i = V(X_i, Y_i)$
2: For each $i \in \mathcal{D}_{\text{calib}}$, compute $\ell_i = \widehat{\ell}(X_i)$ and $u_i = \widehat{u}(X_i)$.
3: Compute $u_{n+1} = \widehat{u}(x)$.
4: For each $1 \le k \le n$, compute $\widehat{F}(k)$ as in (8).
5: Compute $k^* = \min\{k\colon \widehat{F}(k) \ge 1 - \alpha\}$.

Output: Prediction set $\widehat{C}(x) = \{y\colon V(x,y) \le V_{[k^*]}\}$.

with the guarantee ...

Robust Weighted Procedure

Theorem 10.

Assume $(X_i, Y_i) i.i.d. \sim \mathbb{P}$*, and the independent test point* $(X_{n+1}, Y_{n+1}) \sim \tilde{\mathbb{P}}$ *has likelihood ratio* $w(x,y) = \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}(x,y)$*. Then for any target level* $\alpha \in (0,1)$*, the output of Algorithm 1 satisfies*

$$\tilde{P}(Y_{n+1} \in \widehat{C}(X_{n+1}) \geq 1 - \alpha - \hat{\Delta}$$

where $\widehat{\Delta} =$
$\|1/\widehat{\ell}(X)\|_q \cdot \left(\left\| (\widehat{\ell}(X) - w(X,Y))_+ \right\|_p + \|(\widehat{u}(X) - w(X,Y))_-\|_p \right.$
$\left. + \frac{1}{n} \left\| w(X,Y)^{1/p} \cdot (\widehat{u}(X) - w(X,Y))_- \right\|_p \right)$

Bounding the Likelihood Ratio

Under the Γ-selection condition, it is possible to show the following:

Lemma 11.

Suppose a distribution $\mathbb{P}$ *over* $(X, U, T, Y(0), Y(1))$ *satisfies the marginal* Γ*-selection condition. Then for any* $t \in \{0, 1\}$ *it holds for* $\mathbb{P}$*-almost all* x, y *that*

$$1/\Gamma \leq \frac{dP_{Y(t)|X,T=t}}{dP_{Y(t)|X,T=1-t}}(x, y) \leq \Gamma$$

Note that the observed data always factors as $\mathbb{P}_{X,Y(t)|T=t}$. We may be interested in $\mathbb{P}_{X,Y(t)}$. Bayes rule shows

$$\frac{d\mathbb{P}_{X,Y(1)}}{d\mathbb{P}_{X,Y(1)|T=1}} = \mathbb{P}(T = 1) \times \left(1 + \frac{dP_{Y(1)|X,T=1}}{dP_{Y(1)|X,T=0}} \times \frac{e(X)}{1 - e(X)}\right)$$

Bounding the Likelihood Ratio

Applying the Lemma, we get that

$$\mathbb{P}(T=1)\times\left(1+1/\Gamma\times\frac{e(X)}{1-e(X)}\right)\leq\frac{d\mathbb{P}_{X,Y(1)}}{d\mathbb{P}_{X,Y(1)|T=1}}$$
$$\leq\mathbb{P}(T=1)\times\left(1+\Gamma\times\frac{e(X)}{1-e(X)}\right)$$

So the likelihood ratio of what we are interested in and what we have observed has been bounded by a function of X and Γ.

Counterfactual	Bound	ATE-type	ATT-type	ATC-type	General
$Y(1)$	$\ell(x)$	$p_1\cdot\left(1+\frac{1}{\Gamma\cdot r(x)}\right)$	1	$\frac{p_1}{p_0}\cdot\left(\frac{1}{\Gamma\cdot r(x)}\right)$	$p_1\cdot\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}(x)\cdot\left(1+\frac{1}{\Gamma\cdot r(x)}\right)$
	$u(x)$	$p_1\cdot\left(1+\frac{\Gamma}{r(x)}\right)$	1	$\frac{p_1}{p_0}\cdot\frac{\Gamma}{r(x)}$	$p_1\cdot\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}(x)\cdot\left(1+\frac{\Gamma}{r(x)}\right)$
$Y(0)$	$\ell(x)$	$p_0\cdot\left(1+\frac{r(x)}{\Gamma}\right)$	$\frac{p_0}{p_1}\cdot\frac{r(x)}{\Gamma}$	1	$p_0\cdot\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}(x)\cdot\left(1+\frac{r(x)}{\Gamma}\right)$
	$u(x)$	$p_0\cdot\left(1+\Gamma\cdot r(x)\right)$	$\frac{p_0}{p_1}\cdot\Gamma\cdot r(x)$	1	$p_0\cdot\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}(x)\cdot\left(1+\Gamma\cdot r(x)\right)$

Table 1: Summary of the upper and lower bounds of the likelihood ratio for different inferential targets. For $t\in\{0,1\}$, $p_t=\mathbb{P}(T=t)$ and $r(x)=e(x)/(1-e(x))$ is the odds ratio of the propensity score. The training distribution for $Y(t)$ is always $\mathbb{P}_{X,Y(t)\,|\,T=t}$. For target distributions, ATE-type refers to $\mathbb{P}_{X,Y(t)}$; ATT-type refers to $\mathbb{P}_{X,Y(t)\,|\,T=1}$; ATC-type refers to $\mathbb{P}_{X,Y(t)\,|\,T=0}$; General refers to $\mathbb{Q}_X\times\mathbb{P}_{Y(t)\,|\,X}$.

Proof Sketch of Lemma

By rearranging the Γ-selection condition and using Bayes rule, we can derive $\frac{1}{\Gamma} \leq \frac{\mathrm{d}\mathbb{P}_{U|X,T=1}}{\mathrm{d}\mathbb{P}_{U|X,T=0}}(u,x) \leq \Gamma$ for almost every u, x. Also, for any measurable B,

$$\begin{aligned}\mathbb{P}(Y(1) \in B, T = t \mid X) &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{Y(1)\in B\}}\mathbf{1}_{\{T=t\}} \mid X, U\right] \mid X\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{Y(1)\in B\}} \mid X, U\right] \cdot \mathbb{E}\left[\mathbf{1}_{\{T=t\}} \mid X, U\right] \mid X\right]\end{aligned}$$

Rearranging the Γ-selection condition, we have

$$\begin{aligned}\frac{1}{\Gamma} \cdot \mathbb{E}\left[\mathbf{1}_{\{T=0\}} \mid X, U\right] \cdot \frac{\mathbb{E}\left[\mathbf{1}_{\{T=1\}} \mid X\right]}{\mathbb{E}\left[\mathbf{1}_{\{T=0\}} \mid X\right]} &\leq \mathbb{E}\left[\mathbf{1}_{\{T=1\}} \mid X, U\right] \\ &\leq \Gamma \cdot \mathbb{E}\left[\mathbf{1}_{\{T=0\}} \mid X, U\right] \cdot \frac{\mathbb{E}\left[\mathbf{1}_{\{T=1\}} \mid X\right]}{\mathbb{E}\left[\mathbf{1}_{\{T=0\}} \mid X\right]}\end{aligned}$$

Proof Sketch Continued

Multiplying the previous by $\mathbb{E}[\mathbf{1}\{Y(1) \in B\}|X, U]$ and taking expectation conditional on X,

$$\begin{aligned}
&\frac{1}{\Gamma} \cdot \mathbb{P}(Y(1) \in B, T = 0 \mid X) \cdot \frac{\mathbb{E}\left[\mathbf{1}_{\{T=1\}} \mid X\right]}{\mathbb{E}\left[\mathbf{1}_{\{T=0\}} \mid X\right]} \\
&\quad \leq \mathbb{P}(Y(1) \in B, T = 1 \mid X) \\
&\quad \leq \Gamma \cdot \mathbb{P}(Y(1) \in B, T = 0 \mid X) \cdot \frac{\mathbb{E}\left[\mathbf{1}_{\{T=1\}} \mid X\right]}{\mathbb{E}\left[\mathbf{1}_{\{T=0\}} \mid X\right]}
\end{aligned}$$

$$\frac{1}{\Gamma} \cdot \frac{1 - e(x)}{e(x)} \leq \frac{\mathbb{P}(Y(1) \in B, T = 0 \mid X = x)}{\mathbb{P}(Y(1) \in B, T = 1 \mid X = x)} \leq \Gamma \cdot \frac{1 - e(x)}{e(x)}$$

Note

$$\frac{\mathbb{P}(Y(1) \in B \mid X = x, T = 1)}{\mathbb{P}(Y(1) \in B \mid X = x, T = 0)} = \frac{\mathbb{P}(Y(1) \in B, T = 1 \mid X = x)}{\mathbb{P}(Y(1) \in B, T = 0 \mid X = x)} \cdot \frac{1 - e(x)}{e(x)}$$

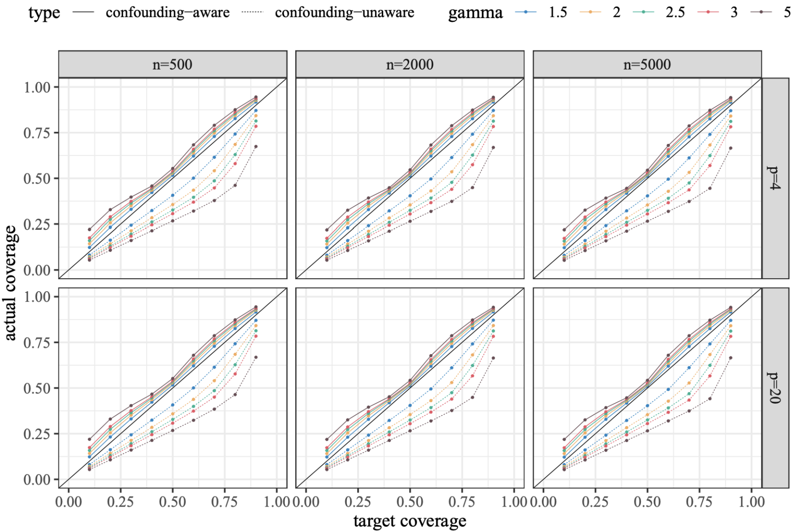
Conclude $\frac{1}{\Gamma} \leq \frac{d\mathbb{P}_{Y(1)|X,T=1}}{d\mathbb{P}_{Y(1)|X,T=0}}(x, y) \leq \Gamma$.

Simulation Study

Setup: Over sample sizes 500,2000,5000 and dimension 4 and 20, generate treatment with different confounding levels.

$$
\begin{aligned}
&X \sim \mathsf{Unif}[0,1]^p \\
&U|X \sim N(0, 1 + 0.5 \times (2.5X_1)^2) \\
&Y(1) = \beta^T X + U \\
&e(x) = \mathsf{logit}(\beta^T x) \\
&e(x,u) = a(x)\mathbf{1}\{|u| > t(x)\} + b(x)\mathbf{1}\{|u| \le t(x)\} \\
&a(x) = \frac{e(x)}{e(x) + \Gamma(1 - e(x))}, b(x) = \frac{e(x)}{e(x) + 1/\Gamma(1 - e(x))}
\end{aligned}
$$

Simulation Results



References I

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, 2021. URL https://openreview.net/forum?id=6vaActvpcp3.

Ying Jin, Zhimei Ren, and Emmanuel J. Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach, 2021.

Lihua Lei and Emmanuel Candes. Conformal inference of counterfactuals and individual treatment effects. Journal of the Royal Statistical Society: Series B, 2021.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.