Doubly Robust Calibration of Prediction Sets under Covariate Shift

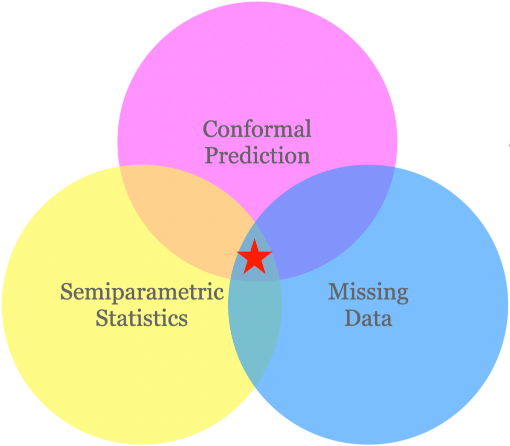
—the holy triad of
conformal prediction, semiparametrics, and missing data

Yachong Elsa Yang, University of Pennsylvania

Joint work with Prof. Arun Kumar Kuchibhotla, Carnegie Mellon University
Prof. Eric Tchetgen Tchetgen, University of Pennsylvania

http://arxiv.org/abs/2203.01761

Intersection of regimes



And many others:

- Covariate shift
- Transfer learning
- Causal inference
- ⋮

Table of contents

Prediction Problems

Usual Prediction Problem

Given i.i.d. pairs $(X_i, Y_i) \sim P, i = 1, \ldots, N$, for a distribution P on

$$\mathcal{X} \times \mathbb{R} \quad (\text{e.g. } \mathcal{X} = \mathbb{R}^d)$$

Goal. Build a prediction set $\widehat{C}_N$, such that for new i.i.d. pair (X_{N+1}, Y_{N+1}) :

$$\mathbb{P}\left(Y_{N+1} \in \widehat{C}_N(X_{N+1})\right) \geq 1 - \alpha,$$

(where the probability is over all $N + 1$ pairs).

Conformal prediction provides a simple and finite-sample valid solution to this problem without any assumptions on P.

Such prediction sets can be asymptotically efficient (i.e., the smallest) too.

Today's talk: Prediction under Covariate Shift

- Suppose we have

$$\underbrace{(X_i, Y_i) \overset{iid}{\sim} P_X \otimes P_{Y|X}}_{\text{labeled data}}, \; 1 \le i \le n \quad \text{and} \qquad \underbrace{X_i \overset{iid}{\sim} Q_X,}_{\text{unlabeled data}} \quad n+1 \le i \le N.$$

- The covariate distribution in the unlabeled data, Q_X, is allowed to be different from that in the labeled data: covariate shift.

Today's talk: Prediction under Covariate Shift

- Suppose we have

$$\underbrace{(X_i, Y_i) \overset{iid}{\sim} P_X \otimes P_{Y|X}}_{\text{labeled data}},\ 1 \le i \le n \quad \text{and} \qquad \underbrace{X_i \overset{iid}{\sim} Q_X,}_{\text{unlabeled data}} \quad n+1 \le i \le N.$$

- The covariate distribution in the unlabeled data, Q_X, is allowed to be different from that in the labeled data: covariate shift.
- When $P_X \neq Q_X$, this relates to transfer learning.

Today's talk: Prediction under Covariate Shift

- Suppose we have

$$\underbrace{(X_i, Y_i) \overset{iid}{\sim} P_X \otimes P_{Y|X}}_{\text{labeled data}},\ 1 \le i \le n \quad \text{and} \qquad \underbrace{X_i \overset{iid}{\sim} Q_X,}_{\text{unlabeled data}} \quad n+1 \le i \le N.$$

- The covariate distribution in the unlabeled data, Q_X, is allowed to be different from that in the labeled data: covariate shift.
- When $P_X \neq Q_X$, this relates to transfer learning.
- When $P_X = Q_X$, this relates to semi-supervised learning.

Today's talk: Prediction under Covariate Shift

- Suppose we have

$$\underbrace{(X_i, Y_i) \overset{iid}{\sim} P_X \otimes P_{Y|X}}_{\text{labeled data}},\ 1 \le i \le n \quad \text{and} \qquad \underbrace{X_i \overset{iid}{\sim} Q_X}_{\text{unlabeled data}}, \quad n+1 \le i \le N.$$

- The covariate distribution in the unlabeled data, Q_X, is allowed to be different from that in the labeled data: covariate shift.
- When $P_X \neq Q_X$, this relates to transfer learning.
- When $P_X = Q_X$, this relates to semi-supervised learning.

Goal. Build a prediction set $\widehat{C}_N$ such that

$$\mathbb{P}\big(Y_f \in \widehat{C}_N(X_f)\big) \ge 1 - \alpha, \tag{1}$$

whenever $(X_f, Y_f) \sim Q_X \otimes P_{Y|X}$.

Today's talk: Prediction under Covariate Shift

- Suppose we have

$$\underbrace{(X_i, Y_i) \overset{iid}{\sim} P_X \otimes P_{Y|X}}_{\text{labeled data}}, \; 1 \leq i \leq n \quad \text{and} \qquad \underbrace{X_i \overset{iid}{\sim} Q_X}_{\text{unlabeled data}}, \quad n+1 \leq i \leq N.$$

- The covariate distribution in the unlabeled data, Q_X, is allowed to be different from that in the labeled data: covariate shift.
- When $P_X \neq Q_X$, this relates to transfer learning.
- When $P_X = Q_X$, this relates to semi-supervised learning.

Goal. Build a prediction set $\widehat{C}_N$ such that

$$\mathbb{P}\big(Y_f \in \widehat{C}_N(X_f)\big) \geq 1 - \alpha, \tag{1}$$

whenever $(X_f, Y_f) \sim Q_X \otimes P_{Y|X}$.

– This problem was first introduced by Tibshirani, Barber, Candès, Ramdas, *Conformal prediction under covariate shift, 2020*.

Connections to Missing data and Semiparametric theory

Missing Data Reformulation

- Choose and fix an arbitrary map $R(\cdot,\cdot)$ on $\mathcal{X} \times \mathbb{R}$. This is like a residual (conformal score), e.g. $|y - \widehat{\mu}(x)|$, with $\widehat{\mu}$ from an independent sample.
- For each (X_i, Y_i) with observed response, define $R_i = R(X_i, Y_i)$ and $T_i = 0$. If response is *unobserved*, then $T_i = 1$ and R_i also remains unobserved.
- The training data then are iid observations $(X_i, T_i, (1 - T_i)R_i)$ such that
  - $\mathbb{P}(X_i \in A | T_i = 0) =: P_X(A)$ and $\mathbb{P}(X_i \in A | T_i = 1) =: Q_X(A)$
  - $R_i \perp T_i \,|\, X_i$. This is the missing at random (MAR) assumption.

Missing Data Reformulation

- Choose and fix an arbitrary map $R(\cdot,\cdot)$ on $\mathcal{X} \times \mathbb{R}$. This is like a residual (conformal score), e.g. $|y - \widehat{\mu}(x)|$, with $\widehat{\mu}$ from an independent sample.
- For each (X_i, Y_i) with observed response, define $R_i = R(X_i, Y_i)$ and $T_i = 0$. If response is *unobserved*, then $T_i = 1$ and R_i also remains unobserved.
- The training data then are iid observations $(X_i, T_i, (1 - T_i)R_i)$ such that
  - $\mathbb{P}(X_i \in A | T_i = 0) =: P_X(A)$ and $\mathbb{P}(X_i \in A | T_i = 1) =: Q_X(A)$
  - $R_i \perp T_i \,|\, X_i$. This is the missing at random (MAR) assumption.
- Define r_α such that

$$\mathbb{P}(R_i \leq r_\alpha | T_i = 1) = 1 - \alpha. \tag{2}$$

- This implies that

$$\mathbb{P}_{(X,Y)\sim Q_X \otimes P_{Y|X}}(R(X, Y) \leq r_\alpha) = 1 - \alpha.$$

Missing Data Reformulation

- Choose and fix an arbitrary map $R(\cdot, \cdot)$ on $\mathcal{X} \times \mathbb{R}$. This is like a residual (conformal score), e.g. $|y - \widehat{\mu}(x)|$, with $\widehat{\mu}$ from an independent sample.
- For each (X_i, Y_i) with observed response, define $R_i = R(X_i, Y_i)$ and $T_i = 0$. If response is *unobserved*, then $T_i = 1$ and R_i also remains unobserved.
- The training data then are iid observations $(X_i, T_i, (1 - T_i)R_i)$ such that
  - $\mathbb{P}(X_i \in A | T_i = 0) =: P_X(A)$ and $\mathbb{P}(X_i \in A | T_i = 1) =: Q_X(A)$
  - $R_i \perp T_i \,|\, X_i$. This is the missing at random (MAR) assumption.
- Define r_α such that

$$\mathbb{P}(R_i \leq r_\alpha | T_i = 1) = 1 - \alpha. \tag{2}$$

- This implies that

$$\mathbb{P}_{(X,Y) \sim Q_X \otimes P_{Y|X}}(R(X, Y) \leq r_\alpha) = 1 - \alpha.$$

- Note that r_α is a semi-parametric functional here defined by (2).

Influence function and nuisance parameters

- Note that r_α satisfies

$$\mathbb{E}[\mathbb{P}(R_i \leq r_\alpha | T_i = 1, X_i)] = 1 - \alpha.$$

Influence function and nuisance parameters

- Note that r_α satisfies

$$\mathbb{E}[\mathbb{P}(R_i \leq r_\alpha | T_i = 1, X_i)] = 1 - \alpha.$$

- Hence r_α can be written in terms of two nuisance parameters:
  - one relating to conditional distribution of R given X

$$\begin{aligned} m^\star(\gamma, x) &:= \mathbb{P}(R \leq \gamma | T = 1, X = x) \\ &= \mathbb{P}(R \leq \gamma | X = x); \end{aligned}$$

  - one relating to conditional distribution of T given X

$$\pi^\star(x) := \mathbb{P}(T = 1 | X = x) / \mathbb{P}(T = 0 | X = x).$$

Influence function and nuisance parameters

- Note that r_α satisfies

$$\mathbb{E}[\mathbb{P}(R_i \leq r_\alpha | T_i = 1, X_i)] = 1 - \alpha.$$

- Hence r_α can be written in terms of two nuisance parameters:
  - one relating to conditional distribution of R given X

$$\begin{aligned} m^\star(\gamma, x) &:= \mathbb{P}(R \leq \gamma | T = 1, X = x) \\ &= \mathbb{P}(R \leq \gamma | X = x); \end{aligned}$$

  - one relating to conditional distribution of T given X

$$\pi^\star(x) := \mathbb{P}(T = 1 | X = x) / \mathbb{P}(T = 0 | X = x).$$

- The best way to estimate r_α is through the efficient influence function, which would give an estimator with second order (product) bias.

Double Robustness Property

- The efficient influence function for estimating r_α when the nuisance functions are π and m is

$$\begin{aligned}\mathrm{IF}(\theta, x, r, t; \pi, m) \propto \mathbb{1}\{t=0\}\pi(x)\Big[\mathbb{1}\{r \le \theta\} - m(\theta, x)\Big] \\ + \mathbb{1}\{t=1\}\Big[m(\theta, x) - (1-\alpha)\Big].\end{aligned}$$

- This follows from the semiparametric theory and our missing data reformulation of the prediction problem under covariate shift.
- The connection to semiparametric theory also highlights the fact that our IF is doubly robust[1] for r_α in that

$$\mathbb{E}[\mathrm{IF}(r_\alpha, X, R, T; \pi, m)] = 0, \quad \text{if either } \pi \equiv \pi^\star \text{ or } m \equiv m^\star.$$

- We are now ready to state our methodology for prediction under covariate shift.

[1] James M. Robins and Heejung Bang (2005)

Methodology & Validity

Algorithm: Split Doubly Robust Prediction (Split-DRP)

Input: Data $(X_i, T_i, (1 - T_i)Y_i), 1 \leq i \leq N$, coverage probability $1 - \alpha$, a conformal score map $R(\cdot, \cdot)$, and estimators $\widehat{\pi}, \widehat{m}$, prediction point x.

1. Randomly split training data into two parts $\mathcal{D}_1$ and $\mathcal{D}_2$ each with $N/2$ observations.
2. Fit the estimators $\widehat{\pi}$ and $\widehat{m}$ on the first split of the data and compute the conformal scores R_i on the second split of the data.
3. Solve for $\theta = \widehat{r}_\alpha$ as the solution to $\mathbb{P}_{\mathcal{I}_2}[\mathrm{IF}(\hat{\theta}, X, R, T; \widehat{\pi}, \widehat{m})] = 0$, where

$$\begin{aligned}\mathbb{P}_{\mathcal{I}_2}[\mathrm{IF}(\hat{\theta}, X, R, T; \widehat{\pi}, \widehat{m})] \ :=\ & \frac{1}{N/2} \sum_{i \in \mathcal{D}_2} \mathbb{1}\{T_i = 0\} \widehat{\pi}(X_i) \big[\mathbb{1}\{R_i \leq \widehat{\theta}\} - \widehat{m}(\widehat{\theta}, X_i)\big] \\ & + \ \frac{1}{N/2} \sum_{i \in \mathcal{D}_2} \mathbb{1}\{T_i = 1\} [\widehat{m}(\widehat{\theta}, X_i) - (1 - \alpha)].\end{aligned}$$

4. **Output:** The prediction set $\widehat{C}_\alpha := \{y : R(x, y) \leq \widehat{r}_\alpha\}$.

Coverage Validity

Under i.i.d assumption, suppose estimators $\widehat{\pi}, \widehat{m}$ are bounded, then with probability at least $1-\delta$,

$$\mathbb{P}_{(X,Y)\sim Q_X \otimes P_{Y|X}}\left(Y \in \widehat{C}(\widehat{r}_\alpha; X) \mid \mathcal{D}^{\mathrm{tr}}\right) \geq 1-\alpha - \frac{\|\widehat{\pi}-\pi^\star\|_2 \sup_\theta \|\widehat{m}(\theta,\cdot)-m^\star(\theta,\cdot)\|_2}{\mathbb{P}(T=1)} - \frac{\mathfrak{C}}{\mathbb{P}(T=1)}\sqrt{\frac{\log(1/\delta)}{N}}.$$

- First term is the target coverage;
- Second term is for estimating $\pi^\star$ and $m^\star$;
- Third term is for replacing the population expectation of IF with the sample expectation.

Coverage Validity

Under i.i.d assumption, suppose estimators $\widehat{\pi}, \widehat{m}$ are bounded, then with probability at least $1-\delta$,

$$\mathbb{P}_{(X,Y)\sim Q_X \otimes P_{Y|X}}\left(Y \in \widehat{C}(\widehat{r}_\alpha; X) \,|\mathcal{D}^{\mathrm{tr}}\right) \geq 1-\alpha - \frac{\|\widehat{\pi}-\pi^\star\|_2 \sup_\theta \|\widehat{m}(\theta,\cdot)-m^\star(\theta,\cdot)\|_2}{\mathbb{P}(T=1)} - \frac{\mathfrak{C}}{\mathbb{P}(T=1)}\sqrt{\frac{\log(1/\delta)}{N}}.$$

- First term is the target coverage;
- Second term is for estimating $\pi^\star$ and $m^\star$;
- Third term is for replacing the population expectation of IF with the sample expectation.

The product bias term comes from the doubly robust IF.

Coverage Validity

PAC guarantee; With probability at least $1-\delta$,

$$\mathbb{P}_{(X,Y)\sim Q_X \otimes P_{Y|X}}\left(Y \in \widehat{C}(\widehat{r}_\alpha; X) \mid \mathcal{D}^{\mathrm{tr}}\right) \geq 1-\alpha - \frac{\|\widehat{\pi}-\pi^\star\|_2 \sup_\theta \|\widehat{m}(\theta,\cdot)-m^\star(\theta,\cdot)\|_2}{\mathbb{P}(T=1)} - \frac{\mathfrak{C}}{\mathbb{P}(T=1)}\sqrt{\frac{\log(1/\delta)}{N}}.$$

Coverage Validity

PAC guarantee; With probability at least $1-\delta$,

$$\mathbb{P}_{(X,Y)\sim Q_X\otimes P_{Y|X}}\Big(Y \in \widehat{C}(\widehat{r}_\alpha;X) \mid \mathcal{D}^{\mathrm{tr}}\Big) \geq 1-\alpha - \frac{\|\widehat{\pi}-\pi^\star\|_2 \sup_\theta \|\widehat{m}(\theta,\cdot)-m^\star(\theta,\cdot)\|_2}{\mathbb{P}(T=1)} - \frac{\mathfrak{C}}{\mathbb{P}(T=1)}\sqrt{\frac{\log(1/\delta)}{N}}.$$

Unconditional coverage:

$$\left|\mathbb{P}_{(X,Y)\sim Q_X\otimes P_{Y|X}}\Big(Y \in \widehat{C}(\widehat{r}_\alpha;X)\Big) - (1-\alpha)\right| \leq \frac{\|\widehat{\pi}-\pi^\star\|_2 \sup_\theta \|\widehat{m}(\theta,\cdot)-m^\star(\theta,\cdot)\|_2}{\mathbb{P}(T=1)} + \frac{\mathfrak{C}}{\mathbb{P}(T=1)\sqrt{N}}.$$

Algorithm: Full Doubly Robust Prediction (Full-DRP)

Input: Data $(X_i, T_i, (1-T_i)Y_i), 1 \le i \le N$, coverage probability $1-\alpha$, a conformal score map $R(\cdot,\cdot)$, and estimators $\widehat{\pi}, \widehat{m}$, prediction point x.

1. Fit the estimators $\widehat{\pi}$ and $\widehat{m}$ on the training data and compute the conformal scores R_i for each $i \in [N]$.
2. Solve for $\theta = \widehat{r}_\alpha$ as a solution to $\mathbb{P}_N[\mathrm{IF}(\hat{\theta}, X, R, T; \widehat{\pi}, \widehat{m})] = 0$, where

$$\begin{aligned}\mathbb{P}_N[\mathrm{IF}(\hat{\theta}, X, R, T; \widehat{\pi}, \widehat{m})] \;:=\; & \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{T_i = 0\}\widehat{\pi}(X_i)\big[\mathbb{1}\{R_i \le \widehat{\theta}\} - \widehat{m}(\widehat{\theta}, X_i)\big] \\ &+ \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{T_i = 1\}[\widehat{m}(\widehat{\theta}, X_i) - (1-\alpha)].\end{aligned}$$

3. **Output:** The prediction set $\widehat{C}_\alpha := \{y : R(x,y) \le \widehat{r}_\alpha\}$.

Simulations: Comparison between methods

Mean coverage and width from 500 monte carlo runs	Synthetic data		Real data	
	Coverage	Width	Coverage	Width
DRP w. full data	0.90	3.29	0.94	27.85
DRP w. splitting	0.90	3.30	0.90	25.79
WCP[2]	0.97	7.41	0.99	47.71

Table 1: Coverage and width of DRP and WCP on synthetic and real data.

[2] Weighted Conformal Prediction, Tibshirani et al. (2020)

Simulations: Comparison between methods

Mean coverage and width from 500 monte carlo runs	Synthetic data		Real data	
	Coverage	Width	Coverage	Width
DRP w. full data	0.90	3.29	0.94	27.85
DRP w. splitting	0.90	3.30	0.90	25.79
WCP[2]	0.97	7.41	0.99	47.71

Table 1: Coverage and width of DRP and WCP on synthetic and real data.

- WCP produces wider width and therefore, tends to over cover by a considerable amount (by more than 7% over the nominal coverage of 90%). See appendix for a description on WCP.

[2]Weighted Conformal Prediction, Tibshirani et al. (2020)

Simulations: Comparison between methods

Mean coverage and width from 500 monte carlo runs	Synthetic data Coverage	Synthetic data Width	Real data Coverage	Real data Width
DRP w. full data	0.90	3.29	0.94	27.85
DRP w. splitting	0.90	3.30	0.90	25.79
WCP[2]	0.97	7.41	0.99	47.71

Table 1: Coverage and width of DRP and WCP on synthetic and real data.

- WCP produces wider width and therefore, tends to over cover by a considerable amount (by more than 7% over the nominal coverage of 90%). See appendix for a description on WCP.
- Doubly robust prediction with full data and multiple splits have similar performance with valid coverage.

[2]Weighted Conformal Prediction, Tibshirani et al. (2020)

Simulations: Comparison between methods

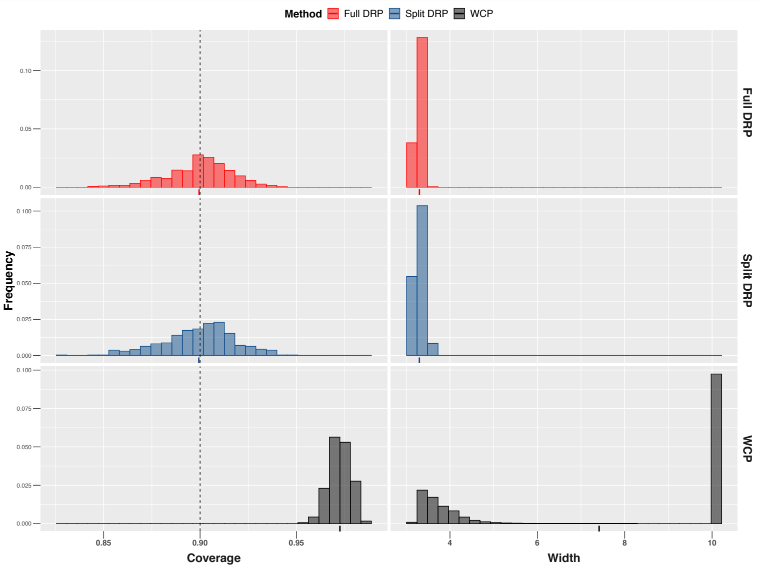


Figure 1: Coverage and width comparison on synthetic data

Simulations: Comparison between methods

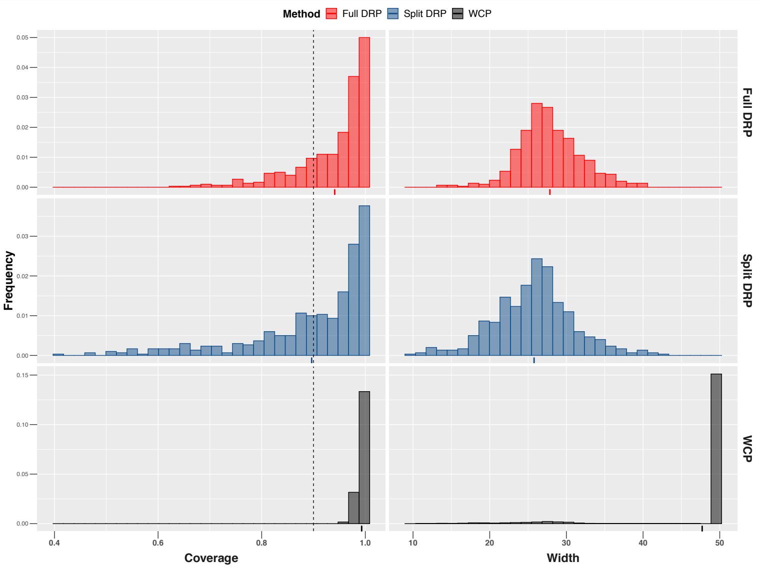


Figure 2: Coverage and width comparison on real data

Comparison with existing works & Extensions

Comparison with existing works

Our method

- Our method does not depend on the test point (new x) at which prediction is needed.
- Our method has double robustness for arbitrary conformal score and the coverage is guaranteed for any training method.
- The bias of our coverage is a product of two errors.

Weighted conformal prediction
Tibshirani et al. 2020

- This method *requires* the test point (new x) to be specified in advance.

Comparison with existing works

Our method

- Our method does not depend on the test point (new x) at which prediction is needed.
- Our method has double robustness for arbitrary conformal score and the coverage is guaranteed for any training method.
- The bias of our coverage is a product of two errors.

Weighted conformal prediction Tibshirani et al. 2020

- This method *requires* the test point (new x) to be specified in advance.

Lei and Candès, 2021

- Their result on double robustness holds only under a specific conformal score: conformal quantile regression (CQR).
- The bias of their coverage is the minimum (not product) of two errors.

Extensions

- This method can be combined with our previous work on efficiency first conformal prediction (EFCP)[3] to choose the prediction interval with the minimum width.

[3] Yang and Kuchibhotla (2021)

Extensions

- This method can be combined with our previous work on efficiency first conformal prediction (EFCP)[3] to choose the prediction interval with the minimum width.
- It can also be extended to provide prediction intervals for counterfactuals and individual treatment effects (ITE), following Lei and Candès (2021).

[3] Yang and Kuchibhotla (2021)

Extensions

- This method can be combined with our previous work on efficiency first conformal prediction (EFCP)[3] to choose the prediction interval with the minimum width.
- It can also be extended to provide prediction intervals for counterfactuals and individual treatment effects (ITE), following Lei and Candès (2021).
- We can relax the MAR assumption to MNAR (corresponding to *Unconfoundedness* condition in Causal) using sensitivity analysis.

[3] Yang and Kuchibhotla (2021)

Prediction intervals for counterfactuals and individual treatment effects (ITE)

Causal inference and Counterfactuals

Given N subjects, let $T_i \in \{0, 1\}$ be a binary treatment indicator, $(Y_i(1), Y_i(0))$ be the pair of potential outcomes, and X_i be the covariates.

- Assume $\big(Y_i(1), Y_i(0), T_i, X_i\big) \overset{\text{i.i.d.}}{\sim} (Y(1), Y(0), T, X)$

Causal inference and Counterfactuals

Given N subjects, let $T_i \in \{0, 1\}$ be a binary treatment indicator, $(Y_i(1), Y_i(0))$ be the pair of potential outcomes, and X_i be the covariates.

- Assume $\big(Y_i(1), Y_i(0), T_i, X_i\big) \overset{\text{i.i.d.}}{\sim} (Y(1), Y(0), T, X)$
- Predict individual treatment effect (ITE) $\tau_i := Y_i(1) - Y_i(0)$, unobserved

Causal inference and Counterfactuals

Given N subjects, let $T_i \in \{0, 1\}$ be a binary treatment indicator, $(Y_i(1), Y_i(0))$ be the pair of potential outcomes, and X_i be the covariates.

- Assume $\big(Y_i(1), Y_i(0), T_i, X_i\big) \overset{\text{i.i.d.}}{\sim} (Y(1), Y(0), T, X)$
- Predict individual treatment effect (ITE) $\tau_i := Y_i(1) - Y_i(0)$, unobserved
- For any treated unit i in the study, i.e. with $T_i = 1$, we construct a prediction interval $\widehat{C}_i^{\mathrm{ITE}}$ for τ_i such that $\widehat{C}_i^{\mathrm{ITE}} = Y_i^{\mathrm{obs}} - \widehat{C}_0(X_i)$, where $\widehat{C}_0(x)$ satisfies

$$\mathbb{P}\big(Y(0) \in \hat{C}_0(X) \mid T = 1\big) \geq 1 - \alpha; \qquad (3)$$

Causal inference and Counterfactuals

Given N subjects, let $T_i \in \{0, 1\}$ be a binary treatment indicator, $(Y_i(1), Y_i(0))$ be the pair of potential outcomes, and X_i be the covariates.

- Assume $\big(Y_i(1), Y_i(0), T_i, X_i\big) \overset{\text{i.i.d.}}{\sim} (Y(1), Y(0), T, X)$
- Predict individual treatment effect (ITE) $\tau_i := Y_i(1) - Y_i(0)$, unobserved
- For any treated unit i in the study, i.e. with $T_i = 1$, we construct a prediction interval $\widehat{C}_i^{\mathrm{ITE}}$ for τ_i such that $\widehat{C}_i^{\mathrm{ITE}} = Y_i^{\mathrm{obs}} - \widehat{C}_0(X_i)$, where $\widehat{C}_0(x)$ satisfies

$$\mathbb{P}\big(Y(0) \in \hat{C}_0(X) \mid T = 1\big) \geq 1 - \alpha; \tag{3}$$

- Such construction has guaranteed coverage[4] for τ_i,

$$\mathbb{P}\big(Y_i(1) - Y_i(0) \in \widehat{C}_i^{\mathrm{ITE}}\big) \geq 1 - \alpha.$$

[4] See proof in appendix.

Relaxing the distribution shift (MAR) assumption

The original covariate shift problem

- Suppose we have

$$\underbrace{(X_i, Y_i) \overset{iid}{\sim} P_X \otimes P_{Y|X},}_{\text{labeled data}} \; 1 \le i \le n \quad \text{and} \qquad \underbrace{X_i \overset{iid}{\sim} Q_X,}_{\text{unlabeled data}} \quad n+1 \le i \le N.$$

- The covariate distribution in the unlabeled data, Q_X, is allowed to be different from that in the labeled data: covariate shift.

Goal. Build a prediction set $\widehat{C}_N$ such that

$$\mathbb{P}\big(Y_f \in \widehat{C}_N(X_f)\big) \ge 1 - \alpha, \tag{4}$$

whenever $(X_f, Y_f) \sim Q_X \otimes P_{Y|X}$.

Relaxing the covariate shift assumption

- Suppose we have

$$\underbrace{(X_i, Y_i) \overset{iid}{\sim} P_X \otimes P_{Y|X},}_{\text{labeled data}} \; 1 \le i \le n \quad \text{and} \quad \underbrace{X_i \overset{iid}{\sim} Q_X,}_{\text{unlabeled data}} \quad n+1 \le i \le N.$$

Goal. Build a prediction set $\widehat{C}_N$ such that

$$\mathbb{P}\big(Y_f \in \widehat{C}_N(X_f)\big) \ge 1 - \alpha, \tag{5}$$

whenever $(X_f, Y_f) \sim Q_X \otimes Q_{Y|X}$.

Relaxing the covariate shift assumption

- Suppose we have

$$\underbrace{(X_i, Y_i) \overset{iid}{\sim} P_X \otimes P_{Y|X},}_{\text{labeled data}} \; 1 \le i \le n \quad \text{and} \quad \underbrace{X_i \overset{iid}{\sim} Q_X,}_{\text{unlabeled data}} \quad n+1 \le i \le N.$$

Goal. Build a prediction set $\widehat{C}_N$ such that

$$\mathbb{P}\big(Y_f \in \widehat{C}_N(X_f)\big) \ge 1 - \alpha, \tag{5}$$

whenever $(X_f, Y_f) \sim Q_X \otimes Q_{Y|X}$.

Suppose we want to relax the $P_{Y|X} = Q_{Y|X}$ assumption,

$$\exp(\gamma(x, y)) = \frac{P_{Y=y|X=x}}{Q_{Y=y|X=x}} \frac{Q_{Y=y_0|X=x}}{P_{Y=y_0|X=x}},$$

where $\gamma(x, y)$ is a known sensitivity analysis function which satisfies $\gamma(x, y_0) = 0$ for any baseline y_0.

Sensitivity analysis

Sensitivity function:[5]

$$\exp(\gamma(x, y)) = \frac{P_{Y=y|X=x}}{Q_{Y=y|X=x}} \frac{Q_{Y=0|X=x}}{P_{Y=0|X=x}}.$$

As a special case when $\gamma(x, y) = 0$, this goes back to the original covariate shift problem.

[5] James M. Robins, Andrea Rotnitzky, and Daniel O. Sharfstein (1999)

Sensitivity analysis

Sensitivity function:[5]

$$\exp(\gamma(x, y)) = \frac{P_{Y=y|X=x}}{Q_{Y=y|X=x}} \frac{Q_{Y=0|X=x}}{P_{Y=0|X=x}}.$$

As a special case when $\gamma(x, y) = 0$, this goes back to the original covariate shift problem. An equivalent way is using the missing data notation:

$$\gamma(x, y) = \log \frac{\mathbb{P}(T = 0|X = x, Y = y)\mathbb{P}(T = 1|X = x, Y = 0)}{\mathbb{P}(T = 0|X = x, Y = 0)\mathbb{P}(T = 1|X = x, y)}.$$

[5] James M. Robins, Andrea Rotnitzky, and Daniel O. Sharfstein (1999)

Sensitivity analysis

Sensitivity function:[5]

$$\exp(\gamma(x,y)) = \frac{P_{Y=y|X=x}}{Q_{Y=y|X=x}} \frac{Q_{Y=0|X=x}}{P_{Y=0|X=x}}.$$

As a special case when $\gamma(x,y) = 0$, this goes back to the original covariate shift problem. An equivalent way is using the missing data notation:

$$\gamma(x,y) = \log \frac{\mathbb{P}(T=0|X=x, Y=y)\mathbb{P}(T=1|X=x, Y=0)}{\mathbb{P}(T=0|X=x, Y=0)\mathbb{P}(T=1|X=x, y)}.$$

The efficient influence function is given by

$$\begin{aligned}
&\mathrm{IF}(r_\alpha, x, y, r, t; \eta^\star, m^\star, \gamma^\star) \\
&\propto \mathbb{1}\{t=0\} \frac{\mathbb{P}(T=1|X=x, Y=y)}{\mathbb{P}(T=0|X=x, Y=y)} \Big[\mathbb{1}\{r \le r_\alpha\} - \mathbb{P}(R \le r_\alpha | X=x, T=1)\Big] \\
&\quad + \mathbb{1}\{t=1\}\Big[\mathbb{P}(R \le r_\alpha | X=x, T=1) - (1-\alpha)\Big].
\end{aligned} \tag{6}$$

[5] James M. Robins, Andrea Rotnitzky, and Daniel O. Sharfstein (1999)

Double robustness and nuisance functions

Formally, let $\gamma^\star(x, y)$ be the sensitivity function defined by

$$\gamma^\star(x, y) = \log \frac{\mathbb{P}(T = 0|X = x, Y = y)\mathbb{P}(T = 1|X = x, Y = 0)}{\mathbb{P}(T = 0|X = x, Y = 0)\mathbb{P}(T = 1|X = x, y)},$$

and denote two nuisance functions by

$$\begin{cases} \eta^\star(x) := \log \frac{\mathbb{P}(T=0|X=x,Y=0)}{\mathbb{P}(T=1|X=x,Y=0)}; \\ m^\star(\theta, x) := \mathbb{P}(R \leq \theta|X = x, T = 1). \end{cases} \tag{7}$$

$\mathrm{IF}(r_\alpha, x, y, r, t; \eta, m, \gamma^\star)$ satisfies the double robustness property that

$$\mathbb{E}\big[\mathrm{IF}(r_\alpha, x, y, r, t; \eta, m, \gamma^\star)\big] = 0, \tag{8}$$

if either $\eta = \eta^\star$ or $m = m^\star$.

Estimating nuisance functions

Note that the two nuisance parameters can not be directly estimated from data:

$$\begin{cases} \eta^\star(x) = \log \frac{\mathbb{P}(T=0|X=x,Y=0)}{\mathbb{P}(T=1|X=x,Y=0)}; \\ m^\star(\theta, x) = \mathbb{P}(R \le \theta | X = x, T = 1). \end{cases} \tag{9}$$

Estimating nuisance functions

Note that the two nuisance parameters can not be directly estimated from data:

$$\begin{cases} \eta^\star(x) = \log \frac{\mathbb{P}(T=0|X=x,Y=0)}{\mathbb{P}(T=1|X=x,Y=0)}; \\ m^\star(\theta, x) = \mathbb{P}(R \leq \theta | X = x, T = 1). \end{cases} \quad (9)$$

Ghassami et al. (2021)[6] provides a general framework to estimate nuisance parameters, that can establish the regularity and asymptotic normality of the doubly robust estimator of r_α, see appendix.

[6] AmirEmad Ghassami, Andrew Ying, Ilya Shpitser, and Eric Tchetgen Tchetgen (2021)

Comparison with existing works on sensitivity analysis

Our method

- Our method does not depend on the test point (new x) at which prediction is needed.
- Our method has double robustness for arbitrary conformal score and the coverage is guaranteed for any training method.
- The bias of our coverage is a product of two errors.

Jin, Ren and Candès, 2021

- A different sensitivity framework.

Comparison with existing works on sensitivity analysis

Our method

- Our method does not depend on the test point (new x) at which prediction is needed.
- Our method has double robustness for arbitrary conformal score and the coverage is guaranteed for any training method.
- The bias of our coverage is a product of two errors.

Jin, Ren and Candès, 2021

- A different sensitivity framework.

Robust conformal prediction

- This method *requires* the test point (new x) to be specified in advance.
- The bias of their coverage is a first order bias.

Comparison with existing works on sensitivity analysis

Our method

- Our method does not depend on the test point (new x) at which prediction is needed.
- Our method has double robustness for arbitrary conformal score and the coverage is guaranteed for any training method.
- The bias of our coverage is a product of two errors.

Jin, Ren and Candès, 2021

- A different sensitivity framework.

Robust conformal prediction

- This method *requires* the test point (new x) to be specified in advance.
- The bias of their coverage is a first order bias.

Robust conformal prediction: the PAC procedure

- Depends on an addition parameter δ.

Take home message

- We have provided the methodology and theoretical guarantees with an arbitrary conformal score.

Take home message

- We have provided the methodology and theoretical guarantees with an arbitrary conformal score.
- Our method utilizes a doubly robust influence function, and brings together conformal prediction, semiparametric statistics and missing data; As a result, the bias of our coverage is of second order.

Take home message

- We have provided the methodology and theoretical guarantees with an arbitrary conformal score.
- Our method utilizes a doubly robust influence function, and brings together conformal prediction, semiparametric statistics and missing data; As a result, the bias of our coverage is of second order.
- Our result can be easily extended to choosing the prediction interval with the minimum width and to provide prediction intervals for counterfactuals and individual treatment effects.

Take home message

- We have provided the methodology and theoretical guarantees with an arbitrary conformal score.
- Our method utilizes a doubly robust influence function, and brings together conformal prediction, semiparametric statistics and missing data; As a result, the bias of our coverage is of second order.
- Our result can be easily extended to choosing the prediction interval with the minimum width and to provide prediction intervals for counterfactuals and individual treatment effects.
- We have adapted our method to the sensitivity analysis framework, thus relaxing the MAR/unconfoundedness assumption.

Take home message

- We have provided the methodology and theoretical guarantees with an arbitrary conformal score.
- Our method utilizes a doubly robust influence function, and brings together conformal prediction, semiparametric statistics and missing data; As a result, the bias of our coverage is of second order.
- Our result can be easily extended to choosing the prediction interval with the minimum width and to provide prediction intervals for counterfactuals and individual treatment effects.
- We have adapted our method to the sensitivity analysis framework, thus relaxing the MAR/unconfoundedness assumption.

Thank You!

Take home message

- We have provided the methodology and theoretical guarantees with an arbitrary conformal score.
- Our method utilizes a doubly robust influence function, and brings together conformal prediction, semiparametric statistics and missing data; As a result, the bias of our coverage is of second order.
- Our result can be easily extended to choosing the prediction interval with the minimum width and to provide prediction intervals for counterfactuals and individual treatment effects.
- We have adapted our method to the sensitivity analysis framework, thus relaxing the MAR/unconfoundedness assumption.

Thank You!

http://arxiv.org/abs/2203.01761

References

Lei, Robins, and Wasseerman (2013) Distribution-free prediction sets. *JASA.*

Tibshirani, Barber, Candès, and Ramdas (2020) Conformal prediction under covariate shift. *NeurIPS.*

Lei and Candès (2021) Conformal inference of counterfactuals and individual treatment effects. *JRSSB.*

Yang and Kuchibhotla (2021) Finite-sample efficient conformal prediction. *Arxiv.*

Scharfstein, Rotnitzky, and Robins (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association.*

Bang and Robins (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics.*

References

Robins (2000) Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association.*

Ghassami, Ying, Shpitser, and Tchetgen Tchetgen (2021) Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. *Arxiv.*

Jin, Ren, and Candès (2021) Sensitivity Analysis of Individual Treatment Effects: A Robust Conformal Inference Approach. *Arxiv.*

Appendix

Weighted Conformal Prediction

In the weighted conformal prediction method, the prediction interval is given by

$$\widehat{C}_n(x) = \mu_0(x) \pm \text{ Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^w(x)\delta_{|Y_i-\mu_0(X_i)|} + p_{n+1}^w(x)\delta_{\infty}\right), \tag{10}$$

where $p^w(x)$ depends on the likelihood ratio between P_X and Q_X, or $\pi^\star(x)$.

- When the distribution shift is too "large", i.e. $p^{\omega}_{n+1}(x)$ is larger than α, the width becomes ∞.

Back to main

Some proofs

Coverage for ITE:

$$
\begin{aligned}
&\mathbb{P}\big(Y_i(1) - Y_i(0) \in \widehat{C}_i^{\mathrm{ITE}}\big) \\
&= \mathbb{P}(T_i = 1)\mathbb{P}(Y_i(0) \in \widehat{C}_0(X_i) | T_i = 1) + \mathbb{P}(T_i = 0)\mathbb{P}(Y_i(1) \in \widehat{C}_1(X_i) | T_i = 0) \\
&\geq (1-\alpha)\big(\mathbb{P}(T_i = 1) + \mathbb{P}(T_i = 0)\big) \\
&= 1 - \alpha.
\end{aligned}
$$

Back to main

Estimating nuisance parameters

Leverage a doubly robust influence function such as $\mathrm{IF}(\cdots)$ to generate an objective function for each nuisance parameter.

Specifically, in the case where one specifies $\mu = \exp(-\eta)$ and m as elements of Reproducing Kernel Hilbert Spaces $\mathcal{R}$ and $\mathcal{M}$ equipped with the RKHS norms $\|\cdot\|_{\mathcal{R}}$ and $\|\cdot\|_{\mathcal{M}}$ respectively,

$$\widehat{\mu} = \underset{\mu\in\mathcal{R}}{\arg\min}\ \sup_{m\in\mathcal{M}} \mathbb{P}_N\Big\{m(\theta,X)\big[-\mu(x)\exp(-\gamma^\star(X,Y))\mathbb{1}\{T=0\}+\mathbb{1}\{T=1\}\big]-m^2(\theta,X)\Big\}-\lambda^q_{\mathcal{M}}\|m\|^2_{\mathcal{M}}+\lambda^q_{\mathcal{R}}\|\mu\|^2_{\mathcal{R}};$$

$$\widehat{m} = \underset{m\in\mathcal{M}}{\arg\min}\ \sup_{\mu\in\mathcal{R}} \mathbb{P}_N\Big\{-\mu(X)\exp(-\gamma^\star(X,Y))\mathbb{1}\{T=0\}\big[m(\theta,X)-\mathbb{1}\{R\le\theta\}\big]-\mu^2(X)\Big\}-\lambda^m_{\mathcal{R}}\|\mu\|^2_{\mathcal{R}}+\lambda^m_{\mathcal{M}}\|m\|^2_{\mathcal{M}},$$

where hyper parameters $\lambda^q_{\mathcal{M}}, \lambda^q_{\mathcal{R}}, \lambda^m_{\mathcal{R}}$, and $\lambda^m_{\mathcal{M}}$, as well as the kernel bandwidth are chosen by cross validation.

Back to main