Conformal Prediction in 2020

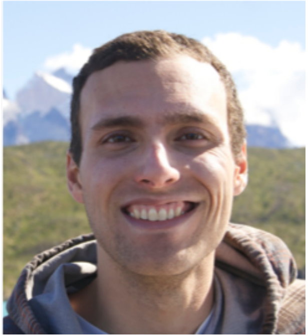
Emmanuel Candès

Tripods Distinguished Seminar

Thanks!

Rina Barber

Aaditya Ramdas



Ryan Tibshirani

Machine learning in sensitive applications

ML 15 years ago: predict movie ratings

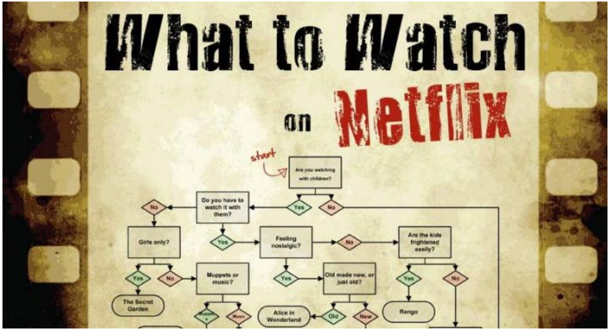


Image credit: Silveroak Casino

Machine learning in sensitive applications

ML 15 years ago: predict movie ratings

ML today:

CNN politics 45 Congress Supreme Court Facts First 2020 Election • LIVE TV Edition

Washington Post: ICE, FBI use state driver's license photos for facial-recognition scans

By Paul LeBlanc, CNN

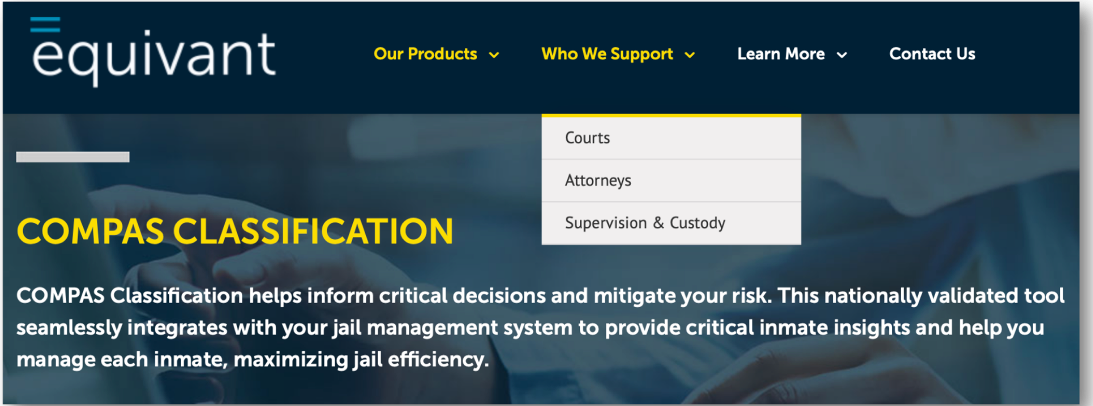
Updated 2:38 PM ET, Mon July 8, 2019

8 July 2019

Machine learning in sensitive applications

ML 15 years ago: predict movie ratings

ML today:



Machine learning in sensitive applications

ML 15 years ago: predict movie ratings

ML today:

THE WALL STREET JOURNAL.

Subscribe | Sign In

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

Search

CIO JOURNAL

HR Departments Turn to AI-Enabled Recruiting in Race for Talent

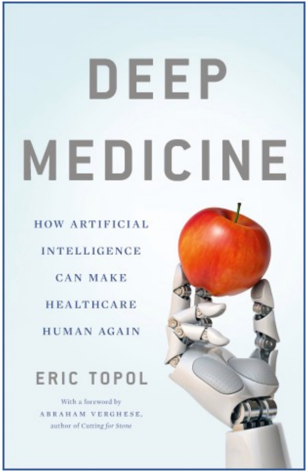
As the battle for talent becomes more competitive, companies are turning toward artificial intelligence to help with recruiting and other human resources tasks

14 March 2019

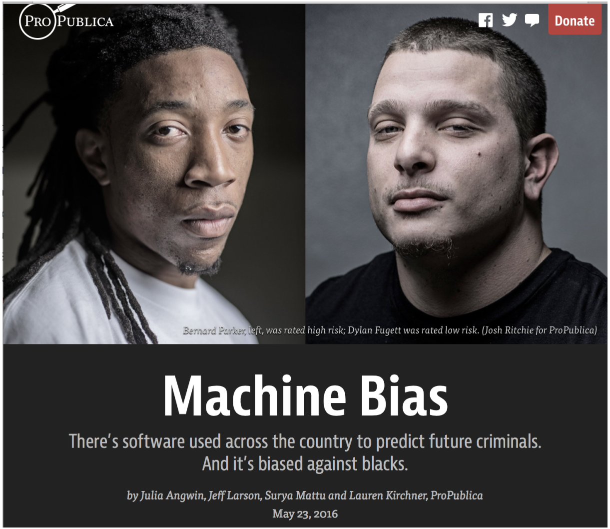
Machine learning in sensitive applications

ML 15 years ago: predict movie ratings

ML today:



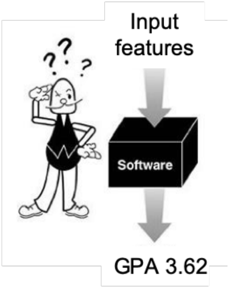
Growing pains



Data ethics 101: convey uncertainty and reliable outcomes

Imagine a *quantitative outcome* as GPA

Can we trust this?
$3.62 \pm\ ?$



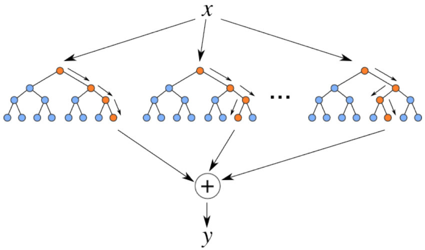
Desperately need reliable systems

Why don't we see prediction intervals more often?

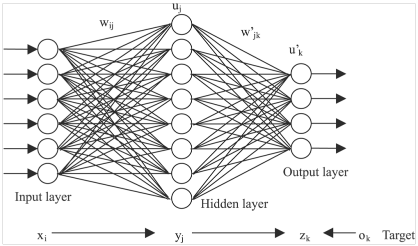
$$\mathbb{P}\{Y \in C(X)\} \approx 90\%$$

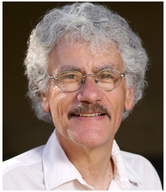
Today's predictive algorithms

random forests, gradient boosting



neural networks



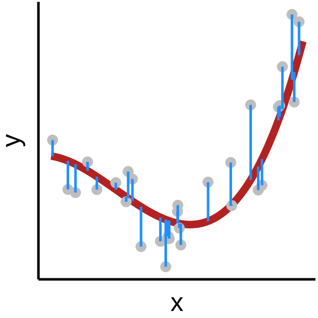


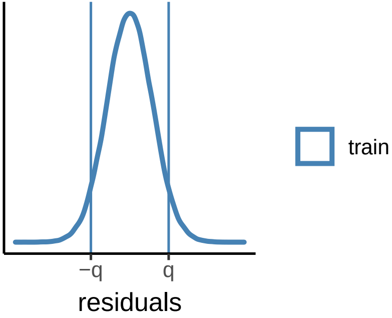
Breiman and Friedman

LeCun, Hinton and Bengio

Conformal prediction

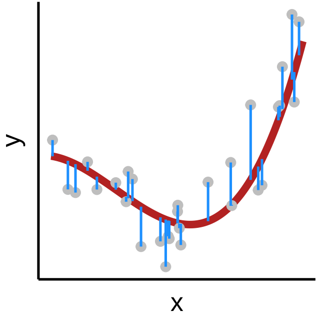
Predicting with confidence?

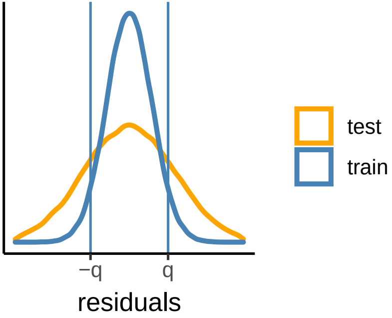




Naive approach: look at residuals and build predictive set $[\hat{\mu}(x) - q, \hat{\mu}(x) + q]$

Predicting with confidence?





Naive approach: look at residuals and build predictive set $[\hat{\mu}(x) - q, \hat{\mu}(x) + q]$

Doesn't work! residuals much smaller than on test points (extreme for neural nets)

(Jackknife is better, but still fails)

Enter conformal prediction

Learning by Transduction

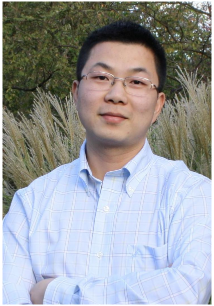
A. Gammerman, V. Vovk, V. Vapnik
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, UK
{alex,vovk,vladimir}@dcs.rhbnc.ac.uk

–UAI '98

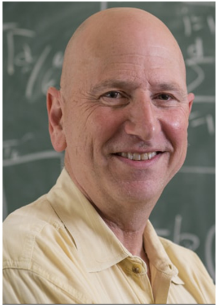
Predictive inference is possible under no assumptions!

Some pioneers

Vladmimir Vovk



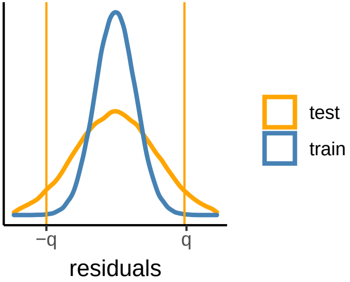
Jing Lei

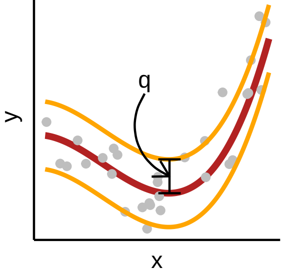


Larry Wasserman

Split conformal prediction

Main idea: look at holdout residuals

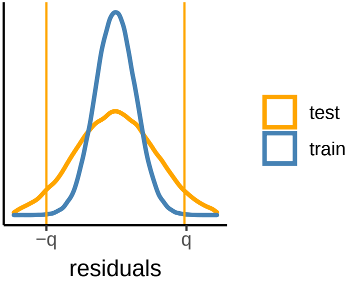


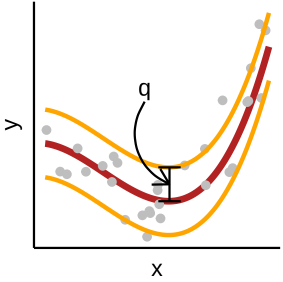


About 90% of future test points will fall within this band

Split conformal prediction

Main idea: look at holdout residuals





About 90% of future test points will fall within this band

Theorem (Papadopoulos, Proedrou, Vovk, Gammerman '02)

q is $\lceil (n+1)(1-\alpha) \rceil$ smallest value of $|y_i - \hat{\mu}(x_i)|$ on calibration set (not used for model fitting)

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

Beyond residuals

- Just used $s(x, y) = |y - \hat{\mu}(x)|$
- Why stop here? Can use *any conformity score* $s(x, y)$
- New predictive set: $C(x) = \{y : s(x, y) \leq q\}$

Beyond residuals

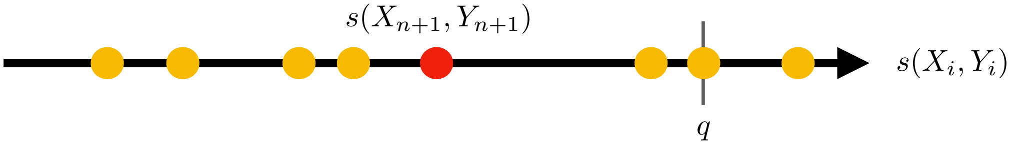
- Just used $s(x, y) = |y - \hat{\mu}(x)|$
- Why stop here? Can use *any conformity score* $s(x, y)$
- New predictive set: $C(x) = \{y : s(x, y) \leq q\}$

Theorem (Papadopoulos, Proedrou, Vovk, Gammerman '02)

q is $\lceil (n+1)(1-\alpha) \rceil$ smallest value of $s(X_i, Y_i)$ on calibration set. Then

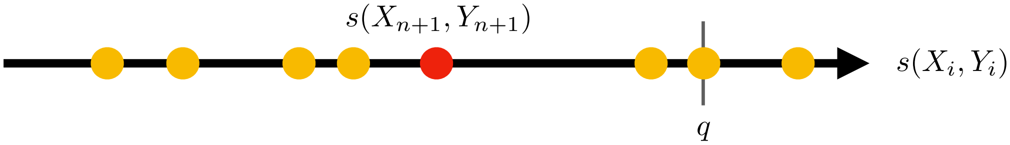
$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$$

Proof



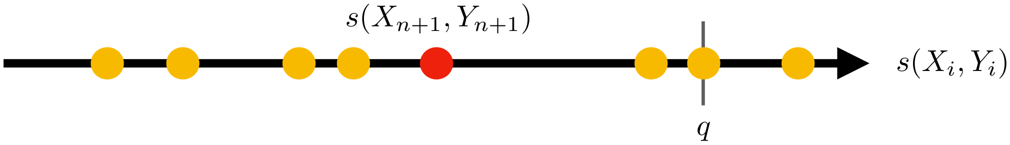
- Scores $s(X_i, Y_i)$ are exchangeable

Proof



- Scores $s(X_i, Y_i)$ are exchangeable
- $\rightsquigarrow$ rank of $s(X_{n+1}, Y_{n+1})$ is discrete uniform

Proof



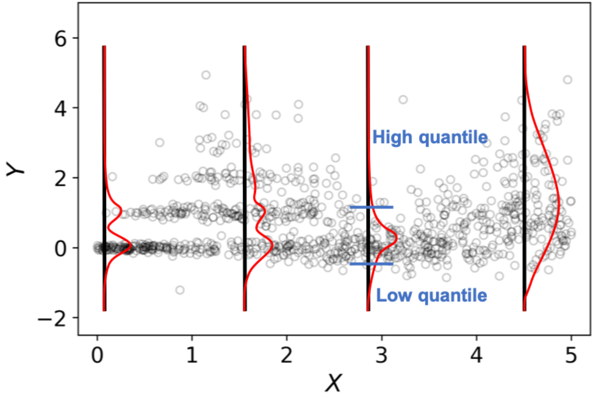
- Scores $s(X_i, Y_i)$ are exchangeable
- $\rightsquigarrow$ rank of $s(X_{n+1}, Y_{n+1})$ is discrete uniform
- $$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} = \mathbb{P}\{s(X_{n+1}, Y_{n+1}) \leq q\} \geq 1 - \alpha$$

$\square$

Better conformity scores

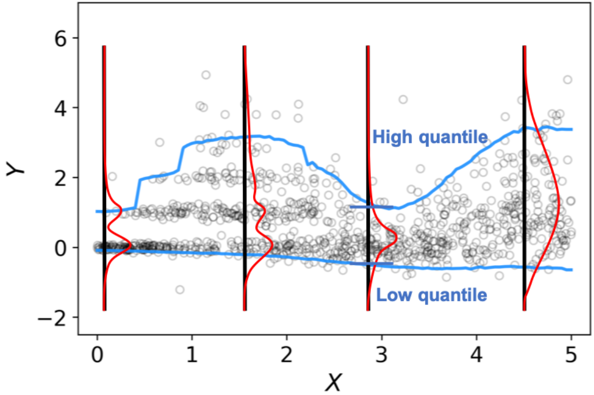
Setting with perfect knowledge

$P_{Y|X}$ known $\rightsquigarrow$ can fit upper and lower quantile functions



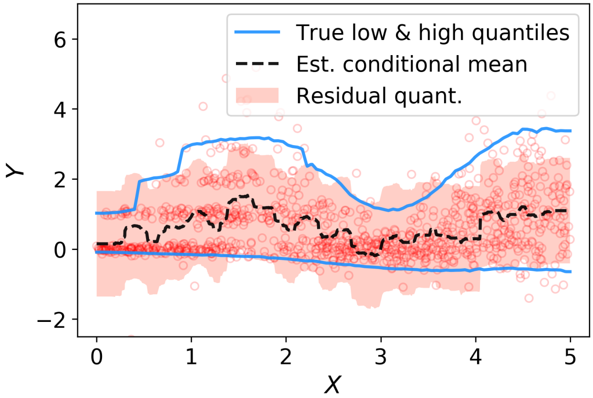
Setting with perfect knowledge

$P_{Y|X}$ known $\rightsquigarrow$ can fit upper and lower quantile functions



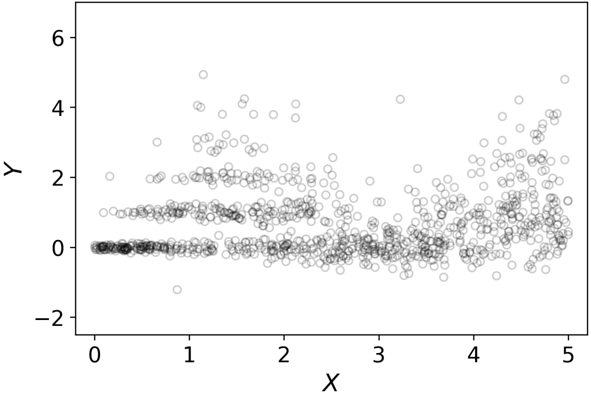
Length of interval can vary greatly

Fixed vs. adaptive intervals



Target coverage: 90%; Actual coverage (test data): 90.03%

No perfect knowledge, only a few samples from $P_{Y|X}$!



Econometrica, Vol. 46, No. 1 (January, 1978)

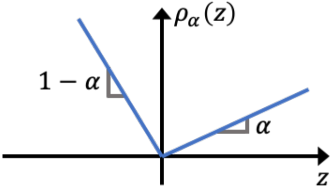
REGRESSION QUANTILES[1]

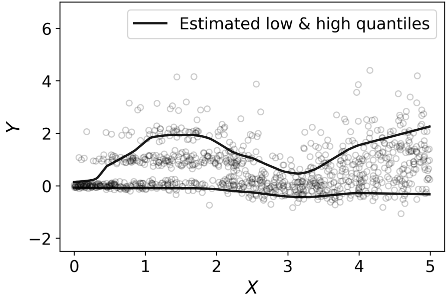
BY ROGER KOENKER AND GILBERT BASSETT, JR.

Formulate quantile estimation as a learning task

$$f(\cdot) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_i \rho_\alpha(Y_i - f(X_i)) + \mathcal{R}(f)$$

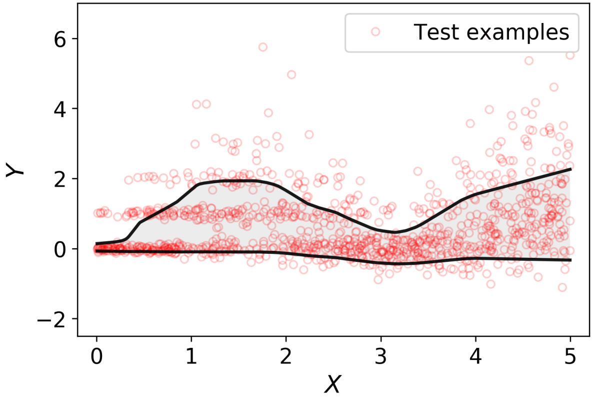
- $\mathcal{R}(f)$ is a possible regularizer
- ρ_α is pinball loss Koenker & Bassett '78





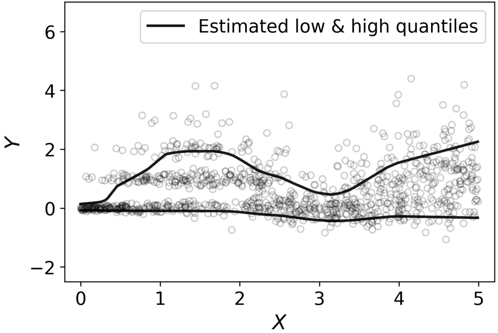
Validity for unseen data?

Valid? No (imagine training a neural net)

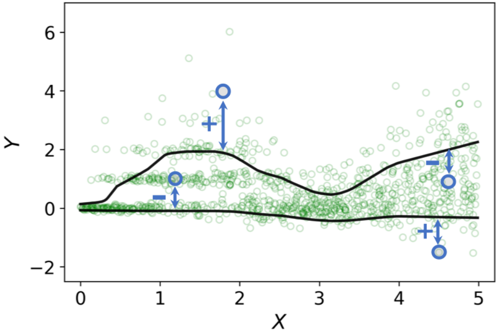


Target coverage level: 90%; Actual coverage: 72.31%

Calibration



Apply quantile regression



Calibrate

Calibrate: how?

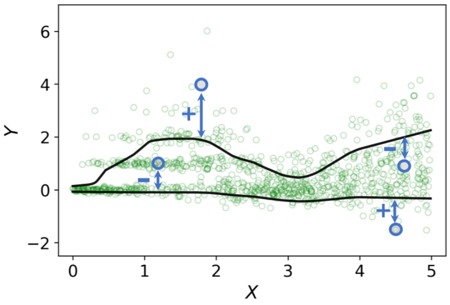
i. For ith point in calibration set

$$S_i = \max\{\text{lower}(X_i) - Y_i, Y_i - \text{upper}(X_i)\}$$

- S_i signed distance to boundary
- S_i negative if $\text{lower}(X_i) \leq Y_i \leq \text{upper}(X_i)$)
  positive otherwise

ii. Q is $(1-\alpha)$th quantile of S_i's

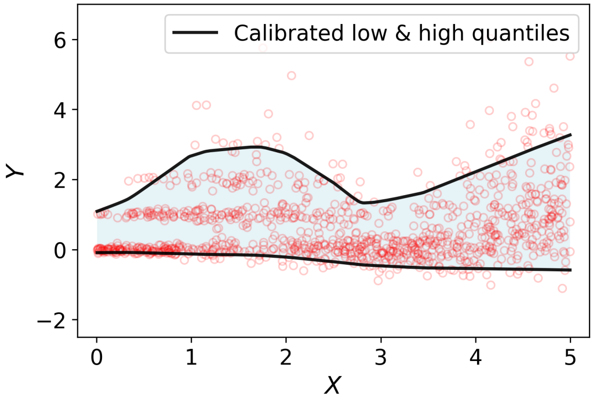
- Q is positive if "initial intervals are too small"



iii. Define the prediction interval as

$$C(x) = [\text{lower}(x) - Q,\ \text{upper}(x) + Q]$$

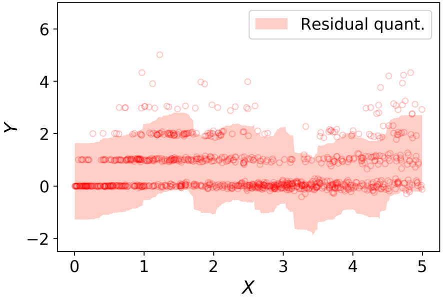
Validity on **new** data



Target coverage: 90%; Actual coverage: 90.01%

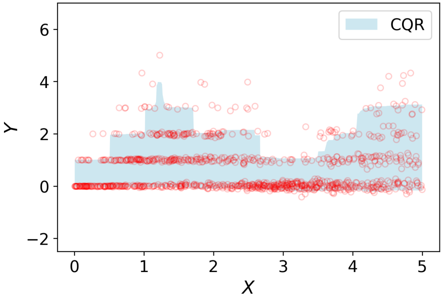
Comparison to split conformal: random forests regression

Split conformal



Avg. Coverage 91.4%
Avg. Length 2.91

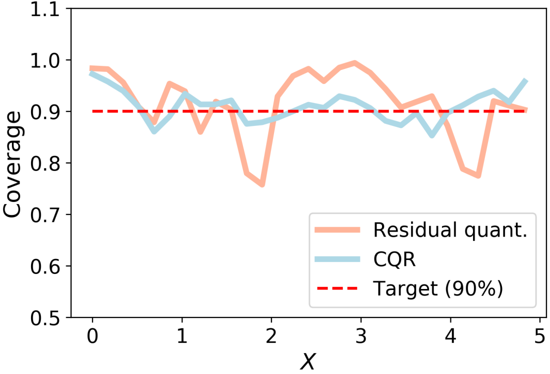
CQR

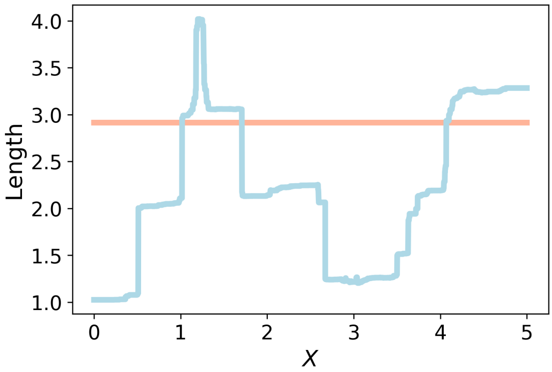


Avg. Coverage 91.0%
Avg. Length 2.18

CQR is adaptive while split conformal is not

Approx. conditional coverage and adaptive length





CQR is largely the right thing to do Sesia and C. ('19)

Predicting utilization of medical services

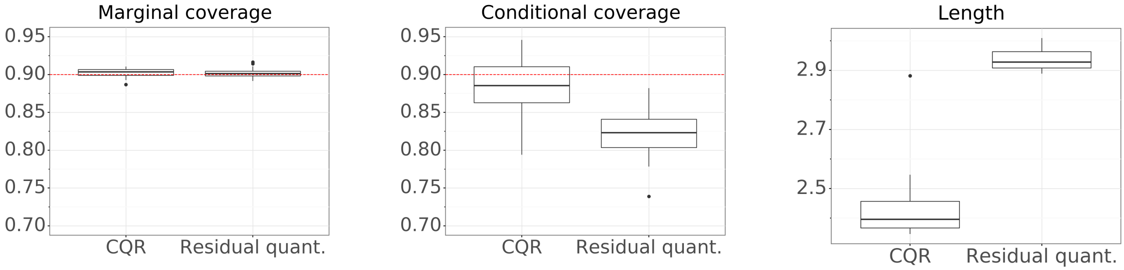
Medical Expenditure Panel Survey 2015

- X_i – age, marital status, race, poverty status, functional limitations, health status, health insurance type, ...
- Y_i – health care system utilization, reflecting # visits to doctor's office/hospital, ...
- $\approx 16,000$ subjects
- ≈ 140 features

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care

Results on MEPS data

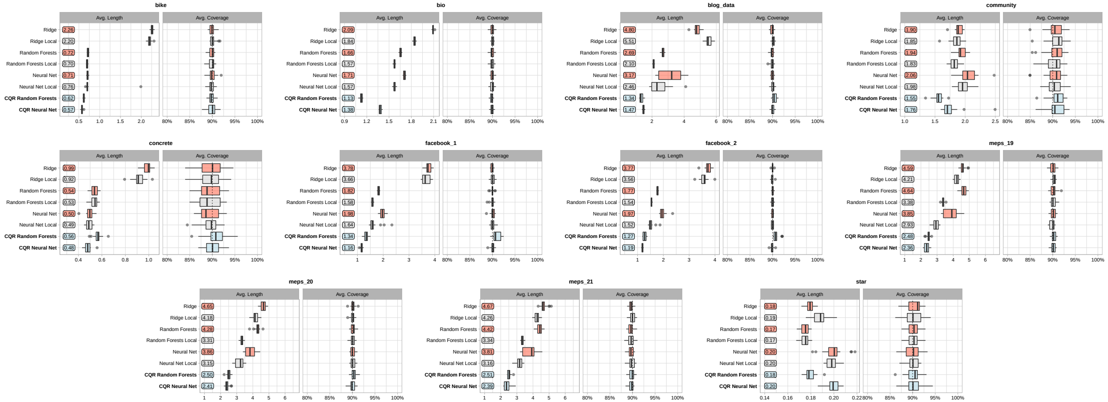
- NNet regression (MSE or pinball loss)
- Average across 20 random train-test (80%/20%) splits



Better conditional coverage* and shorter intervals

*measured over the worst slab Cauchois, Gupta, and Duchi ('20)

A more comprehensive study



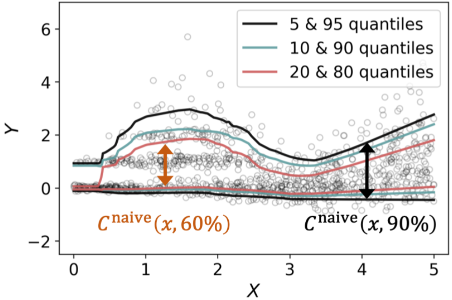
Prediction intervals using quantile regression outperform existing conformal methods in 10/11 regression datasets

Calibration via adaptive coverage

Kivaranovic, Johnson, Leeb ('19); Chernozhukov, Wüthrich, Zhu ('19); Gupta, Kuchibhotla, Ramdas ('19)
Romano, Sesia, & C. ('20); Bates, C., Romano, & Sesia ('20)

1. Uncalibrated guess for parameter τ

$$C^{\text{naive}}(x, 1-\tau) = [\hat{F}^{-1}_{Y|X}(\tau/2),\ \hat{F}^{-1}_{Y|X}(1-\tau/2)]$$



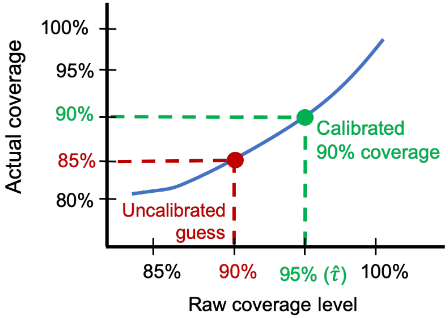
Calibration via adaptive coverage

Kivaranovic, Johnson, Leeb ('19); Chernozhukov, Wüthrich, Zhu ('19); Gupta, Kuchibhotla, Ramdas ('19)
Romano, Sesia, & C. ('20); Bates, C., Romano, & Sesia ('20)

1. Uncalibrated guess for parameter τ

$$C^{\text{naive}}(x, 1-\tau) = [\hat{F}^{-1}_{Y|X}(\tau/2),\ \hat{F}^{-1}_{Y|X}(1-\tau/2)]$$

2. Find $\hat{\tau}$ achieving 90% coverage on calibration set



3. Set

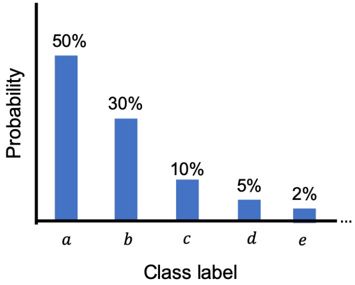
$$C(x) = C^{\text{naive}}(x, \hat{\tau})$$

"Choose 95% nominal to get 90% coverage on test data"

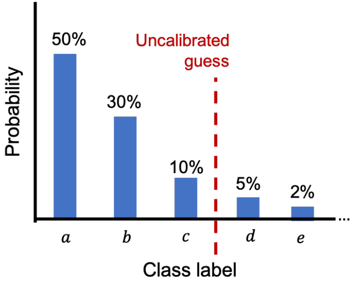
Discrete labels Romano, Sesia, & C. ('20)

- Estimate conditional probabilities $\hat{\pi}(y \mid x)$
  $\rightsquigarrow$ e.g., output of NNet's softmax layer

- Uncalibrated guess

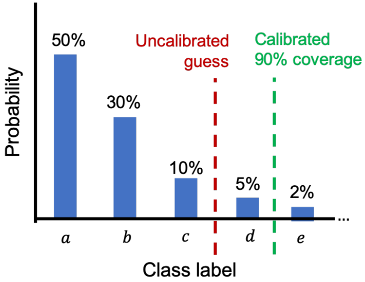


Sorted class probabilities

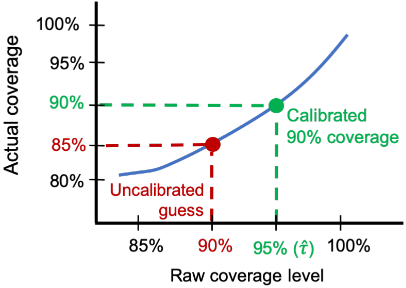


$C^{\text{naive}}(x, 90\%) = \{a, b, c\}$

Calibration via adaptive coverage



$$C^{\text{naive}}(x, {\color{green}95\%}) = \{a, b, c, d\}$$



Prediction set

$$C(x) = C^{\text{naive}}(x, {\color{green}\hat{\tau}})$$

"Choose 95% nominal to get 90% coverage on test data"

Correctness

Validity of CQR & adaptive CP holds regardless of choice/accuracy of quantile regression estimate

Theorem

If (X_i, Y_i), $i = 1, \ldots, n+1$ *are exchangeable, then*

$$1 - \alpha \leq \mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \leq 1 - \alpha + 1/(m+1)$$

- *m is size of calibration set*
- *Upper bound holds if conformity scores are a.s. distinct*

Early split conformal for classification

Lei, Robins, Wasserman '13; Vovk, Petej, Fedorova '14

- Use $\hat{\pi}(y \mid x)$ to construct a prediction set

$$C(x) = \{y \in \mathcal{Y} : \hat{\pi}(y \mid x) \geq Q\}$$

$Q := \alpha$th quantile of calibration scores $\hat{\pi}(Y_i \mid X_i)$

(1) Guess a label $y \in \mathcal{Y}$

(2) Is $\hat{\pi}(y \mid x)$ larger than most of the scores $\hat{\pi}(Y_i \mid X_i)$'s?

If yes $\rightsquigarrow$ include y in $C(x)$

Early split conformal for classification

Lei, Robins, Wasserman '13; Vovk, Petej, Fedorova '14

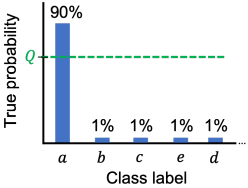
- Use $\hat{\pi}(y \mid x)$ to construct a prediction set

$$C(x) = \{y \in \mathcal{Y} : \hat{\pi}(y \mid x) \geq Q\}$$

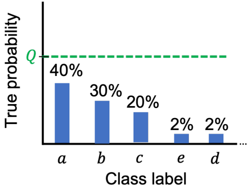
$Q := \alpha$th quantile of calibration scores $\hat{\pi}(Y_i \mid X_i)$

- **Main issue**: poor conditional coverage

Setting with *perfect knowledge* (90% target coverage)



Conformal set $= \{a\}$
Ideal set $= \{a\}$

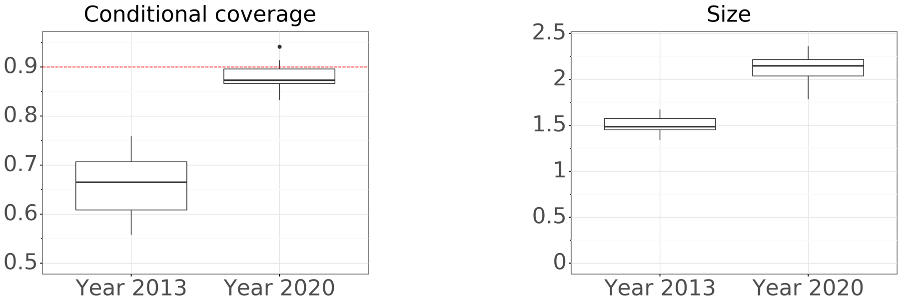


Conformal set $= \{\emptyset\}$
Ideal set $= \{a, b, c\}$

- Threshold Q is not adaptive to x

Adaptivity vs. not: simulation

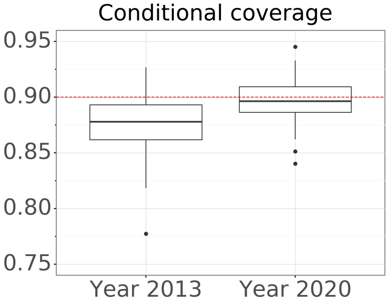
Ten-way classification via kernel SVM (simulated dataset)

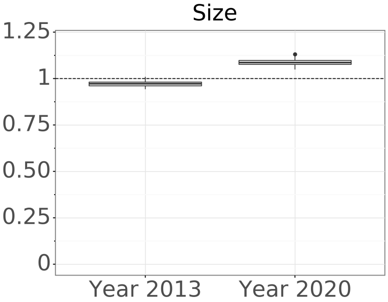


- Better conditional coverage
- May result in larger sets

Adaptivity vs. not: MNIST data

Classification of handwritten digits via NNets



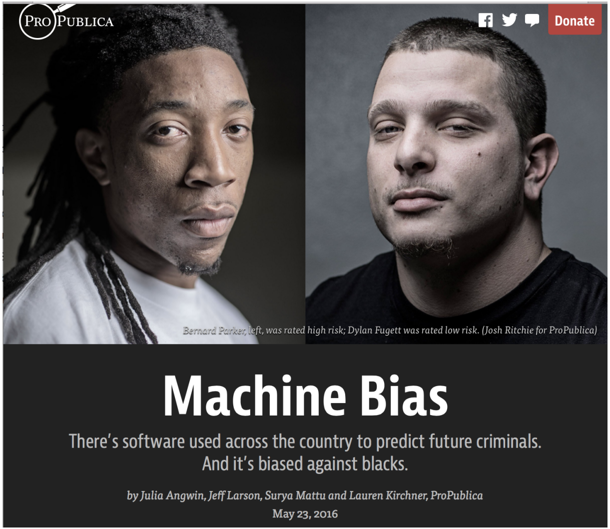


Equitable treatment via equalized coverage

With Malice Towards None:
Assessing Uncertainty via Equalized Coverage

Yaniv Romano[*] Rina Foygel Barber[†] Chiara Sabatti[*,‡] Emmanuel J. Candès[*,§]

Growing pains



Algorithms trained on biased data sets often recognize only the left-hand image as a bride.

Design AI so that it's fair

Identify sources of inequity, de-bias training data and develop algorithms that are robust to skews in data, urge **James Zou** and **Londa Schiebinger**.

On the use of ML to support important decisions

- How do we communicate uncertainty to decision makers?
- How do we not overstate what can be inferred from the black box?
- How do we treat everyone equitably?

Our take:

Decouple the statistical problem from the policy problem

Corbett-Davis and Goel, '19

Somewhat against current thinking in "algorithmic fairness in ML"

Predicting utilization of medical services

MEPS 2016 data set

- X_i – age, marital status, race, poverty status, functional limitations, health status, health insurance type, ...
- Y_i – health care system utilization, reflecting # visits to doctor's office/hospital, ...
- A_i – race (protected attribute)
- $\approx 9,600$ non-white individuals
- $\approx 6,000$ white individuals
- ≈ 140 features

AHRQ Agency for Healthcare Research and Quality
Advancing Excellence in Health Care

Some observations on 2016 MEPS data set

Fit a neural network regression function $\hat{\mu}(\cdot)$:

- NNet overestimates the response of the non-white group
- NNet underestimates the response of the white group

	Group	Avg. Coverage	Avg. Length
Marginal Conformal	Non-white	0.920	2.907
	White	0.871	2.907

Equalized coverage Romano, Barber, Sabatti, & C. '19

<u>Goal</u>: construct perfectly calibrated intervals across all groups

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1}) \mid A = ♂\} \geq 90\%$$
$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1}) \mid A = ♀\} \geq 90\%$$

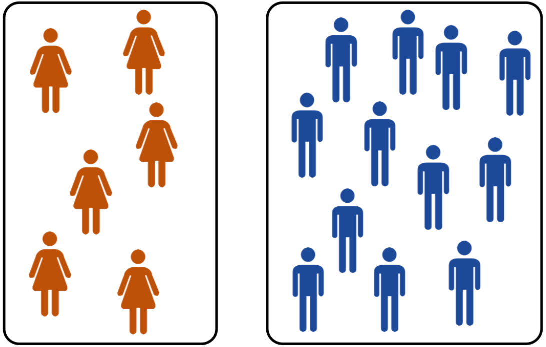
Summarizes what we have learned from ML s.t.

- Rigorously quantifies uncertainty

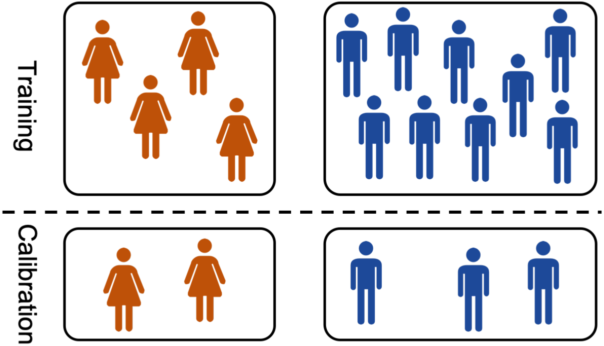
  Honest reporting: interval is long? $\rightsquigarrow$ model can say little

- Treats individuals equitably

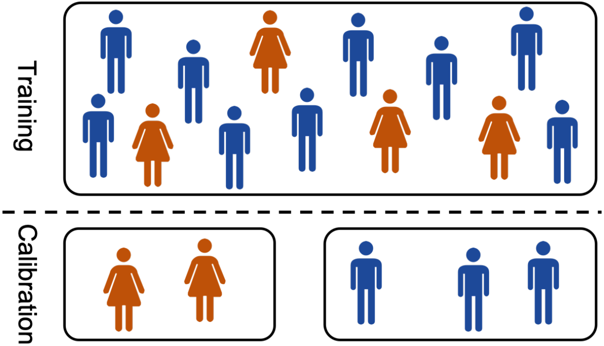
Minority and majority groups



Separate training + separate calibration



Joint training + separate calibration



Performance

- Average across 40 random train-test (80%/20%) splits

Method	Group	Avg. Coverage	Avg. Length
Residual quant. (separate train.)	Non-white	0.903	2.764
	White	0.901	3.182
Residual quant. (joint train.)	Non-white	0.904	2.738
	White	0.902	3.150
CQR (separate train.)	Non-white	0.904	2.567
	White	0.900	3.203
CQR (joint train.)	Non-white	0.902	**2.527**
	White	0.901	**3.102**

- CQR produces shorter intervals
- Joint training is more powerful

Bits of a data ethics framework...

- **Recognize that data analysis is non-neutral**

$\Longrightarrow$ Make sure the way we summarize information does not lead to discriminatory/unfair practices

- **Do not conflate data analysis with a decision rule**

$\Longrightarrow$ Our job is to empower the user, not to play God

- **First, do no harm**

$\Longrightarrow$ Be a professional, not a “hacker”: stakes are high

Counterfactual inference

Counterfactual inference

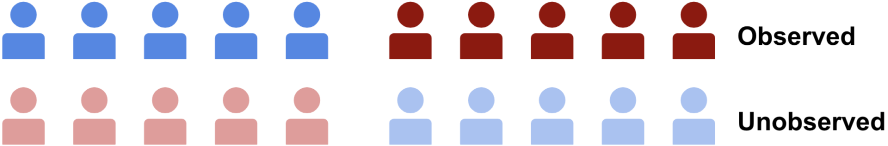
Assign treatment by a coin toss for each subject based on the **propensity score** $e(x)$

$$\mathbb{P}(\text{treated} \mid X = x) = e(x)$$

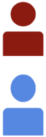
$$\mathbb{P}(\text{control} \mid X = x) = 1 - e(x)$$

Counterfactual inference

Each subject has potential outcomes $(Y(1), Y(0))$ and the observed outcome Y^{obs}



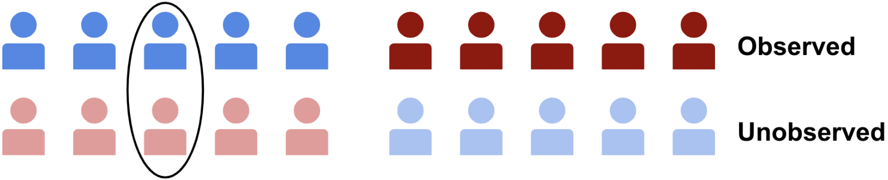
SUTVA



$$Y^{\text{obs}} = Y(1)$$

$$Y^{\text{obs}} = Y(0)$$

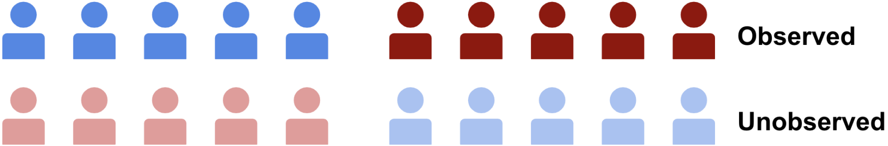
Counterfactual inference



How to infer $Y(1)$ of ?

Counterfactual inference

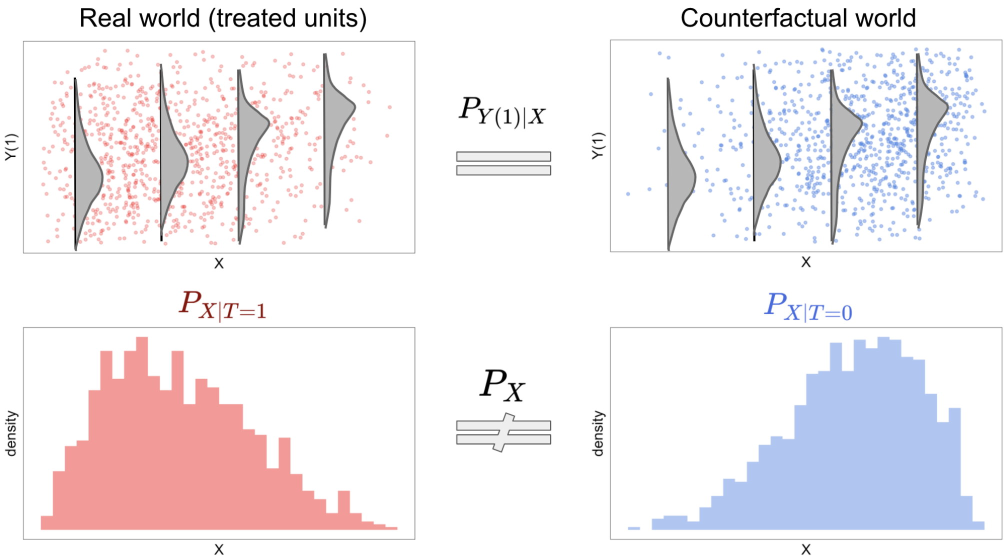
$P_{X|T=0} \times P_{Y(0)|X}$ $\quad$ $P_{X|T=1} \times P_{Y(1)|X}$



$P_{X|T=0} \times P_{Y(1)|X}$ $\quad$ $P_{X|T=1} \times P_{Y(0)|X}$

Distribution mismatch! Covariate shift

The counterfactual inference problem and covariate shift



Adapting conformal inference to covariate shift

Goal: Use i.i.d. samples $(X_i, Y_i) \sim P_X \times P_{Y|X}$ to construct $\hat{C}(x)$ with

$$\mathbb{P}(Y \in \hat{C}(X)) \geq 1 - \alpha \quad \text{with } (X, Y) \sim Q_X \times P_{Y|X}$$

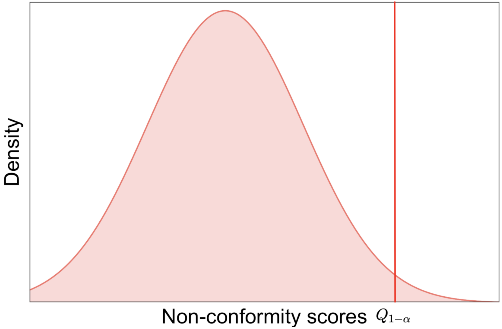
Covariate shift $\quad w(x) \triangleq \frac{dQ_X}{dP_X}(x)$

Counterfactual inference $\quad w(x) \triangleq \frac{dP_{X|T=0}}{dP_{X|T=1}}(x) \propto \frac{1 - e(x)}{e(x)}$

Conformal inference of counterfactuals

Conformal inference without covariate shift: non-conformity score $S(x, y)$

$$y \in \hat{C}(x) \iff S(x, y) \leq Q_{1-\alpha}\left(\sum_{i=1}^{n} \frac{1}{n+1}\delta_{S(X_i, Y_i)} + \frac{1}{n+1}\delta_{S(x,y)}\right)$$

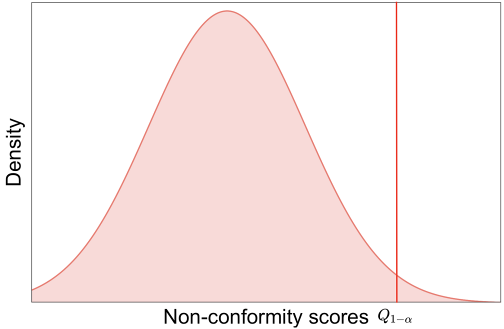


Unweighted histogram

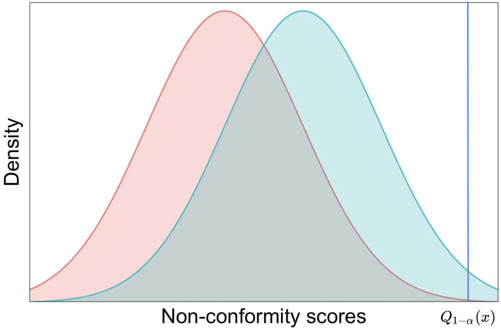
Conformal inference of counterfactuals

Weighted Conformal Inference (Tibshirani, Barber, C., Ramdas '19)

$$y \in \hat{C}(x) \Longleftrightarrow S(x,y) \le Q_{1-\alpha}\left(\sum_{i=1}^{n} p(X_i)\delta_{S(X_i,Y_i)} + p(x)\delta_{S(x,y)}\right), \quad p(X_i) \propto w(X_i)$$



Unweighted histogram



Weighted histogram

Near-exact counterfactual inference in finite samples

Theorem (Lei and C., 2020)

Set $w(x) = (1 - e(x))/e(x)$ ($e(x)$ known) in weighted conformal inference. Then

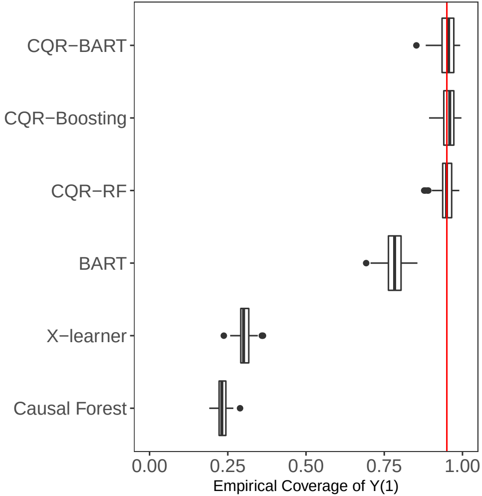
$$1 - \alpha \leq \mathbb{P}(Y_{n+1}(1) \in \hat{C}(X_{n+1})) \leq 1 - \alpha + \frac{C}{n}$$

- *Lower bound holds without extra assumption*
- *Upper bound holds if scores are a.s. distinct & an overlap condition holds*

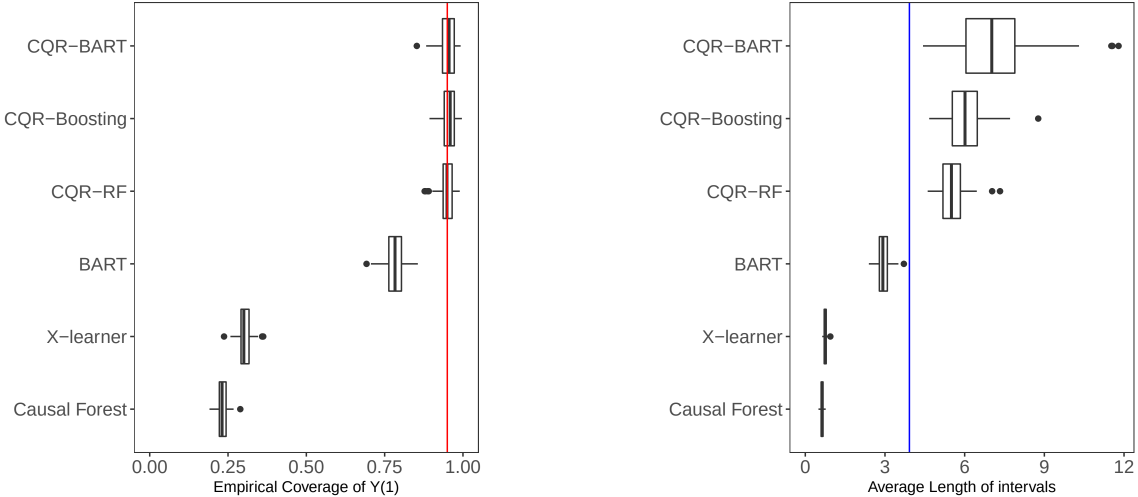
- Applicable to randomized experiments with perfect compliance
- Holds approximately if *either* $e(x)$ *or* $q(Y(1) \mid X)$ are estimated well (**double robustness**)

Simulation: marginal coverage

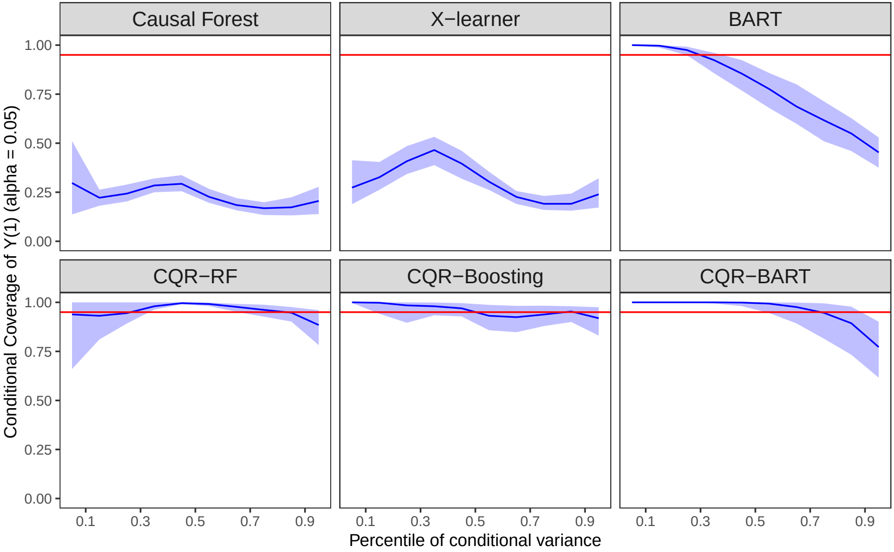
- 100 covariates
- Smooth mean
- Heteroscedastic errors
- Smooth propensity score



Simulation: average interval length



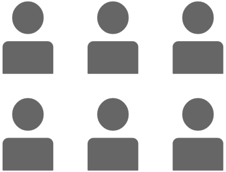
Simulation: conditional coverage



Conformal inference of individual treatment effects

$$(X_i, Y_i^{\text{obs}}) \overset{i.i.d.}{\sim} P_{X|T=0} \times P_{Y(0)|X}$$

$$(X_i, Y_i^{\text{obs}}) \overset{i.i.d.}{\sim} P_{X|T=1} \times P_{Y(1)|X}$$



$$X \sim Q_X$$

Lei and C. '20

Prediction interval for **individual treatment effect** $Y(1) - Y(0)$ **of unseen individual**

$$\mathbb{P}_{X \sim Q_X}\left(Y(1) - Y(0) \in \hat{C}_{\text{ITE}}(X)\right) \geq 1 - \alpha$$

Data re-use (when data is scarce)

Standard approach in CP (full conformal) is computationally prohibitive

Data re-use (when data is scarce)

Standard approach in CP (full conformal) is computationally prohibitive

- Jackknife/CV can fail (coverage can be zero)
- Modification: Jackknife+/CV+ has guaranteed coverage
  Barber, C., Ramdas and Tibshirani '19
- Related to cross-conformal prediction
  Vovk, '15
- Can be adapted to any conformity score, continuous/discrete labels, ...
  Gupta, Kuchibhotla, Ramdas '19; Romano, Sesia, & C. '20

Jackknife+/CV+

Barber, C., Ramdas and Tibshirani '19

K folds and leave-out residuals

$$R_i^{\mathrm{LOO}} = |Y_i - \hat{\mu}_{-K(i)}(X_i)|$$

- **Jackknife/CV**

$$\hat{\mu}(X_{n+1}) \pm R_i^{\mathrm{LOO}} \quad \Longleftrightarrow \quad \left[\text{10th perc. } \{\hat{\mu}(X_{n+1}) - R_i^{\mathrm{LOO}}\},\ \text{90th perc. } \{\hat{\mu}(X_{n+1}) + R_i^{\mathrm{LOO}}\}\right]$$

Jackknife+/CV+

Barber, C., Ramdas and Tibshirani '19

K folds and leave-out residuals

$$R_i^{\text{LOO}} = |Y_i - \hat{\mu}_{-K(i)}(X_i)|$$

- **Jackknife/CV**

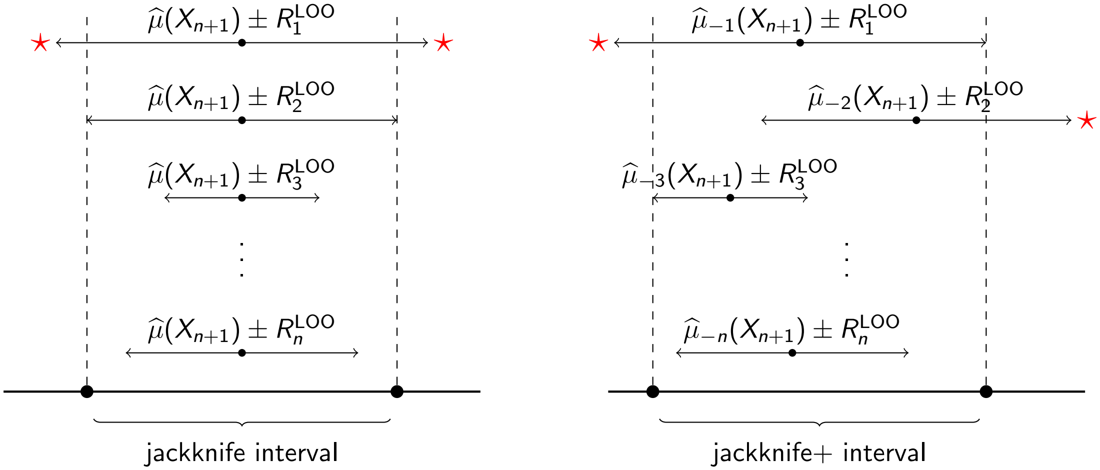
$$\hat{\mu}(X_{n+1}) \pm R_i^{\text{LOO}} \quad \Longleftrightarrow \quad \left[\text{10th perc. } \{\hat{\mu}(X_{n+1}) - R_i^{\text{LOO}}\},\ \text{90th perc. } \{\hat{\mu}(X_{n+1}) + R_i^{\text{LOO}}\}\right]$$

- **Jackknife+/CV+**

$$\left[\text{10th perc. } \{\hat{\mu}_{-K(i)}(X_{n+1}) - R_i^{\text{LOO}}\},\ \text{90th perc. } \{\hat{\mu}_{-K(i)}(X_{n+1}) + R_i^{\text{LOO}}\}\right]$$

- Related to cross-conformal prediction (Vovk, '15)
- Improved performance over split conformal when n is not large

Jackknife vs. Jackknife+



On either side interval boundary is exceeded by a sufficiently small prop. of two sided arrows (marked with $\star$)

Distribution-free guarantee

Theorem (Barber, C., Ramdas and Tibshirani 2019)

If (X_i, Y_i)*,* $i = 1, \ldots, n+1$ *are exchangeable, then*

$$\mathbb{P}\{Y_{n+1} \in C^{\text{jackknife+/CV+}}(X_{n+1})\} \geq 1 - 2\alpha$$

- Jackknife – coverage can be zero; i.e. can have $\mathbb{P}\{Y_{n+1} \in C^{\text{jackknife}}(X_{n+1})\} = 0$
- Coverage is usually (but not always) $1 - \alpha$

Example

- 100 samples
- 100 features
- $Y|X$ follows a linear model
- Regression method – least squares (minimal ℓ_2-norm solution)
- Average over 50 trials

Method	**Coverage**
Jackknife	0.475
Jackknife+	0.913

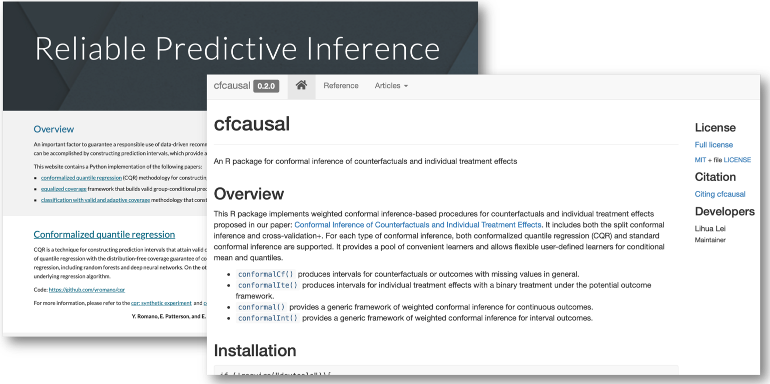
Extensions Gupta, Kuchibhotla, Ramdas '19; Romano, Sesia, & C. '20

- Arbitrary scores
- Discrete/categorical labels

$$\hat{\mathcal{C}}^{\mathrm{CV+}}_{n,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \quad \sum_{i=1}^{n} \mathbf{1}\left[s\left(X_i, Y_i, \hat{\pi}^{-k(i)}\right) < s\left(X_{n+1}, y, \hat{\pi}^{-k(i)}\right)\right] < (1-\alpha)(n+1) \right\}$$

$\hat{\pi}^{-k(i)}$ is model fitted on folds not containing the ith sample

Websites & code



- Effective conformity scores: `https://sites.google.com/view/cqr/`
- Counterfactual and individual treatment effects: `https://lihualei71.github.io/cfcausal/index.html`

Summary

- Personal tour of conformal prediction
- Importance of uncertainty quantification
- Ideas from conformal prediction applicable to meet the highest professional standards

Synthetic data experiment: classification

- Labels $Y \in \{1, 2, \ldots, 10\}$
- Features $X \in \mathbb{R}^{10}$ (two usnbalanced groups)

$$X_1 = \begin{cases} 1 & \text{w.p. } 1/5 \\ -8 & \text{otherwise} \end{cases} \qquad X_2, \ldots, X_{10} \sim \mathcal{N}(0, 1)$$

- $Y \mid X$ follows a linear multiclass logistic model with coefficients $\sim \mathcal{N}(0, 1)$
- Kernel SVM classifier
- 1000 training points
- 5000 test points

MNIST data experiment

- 10 class labels, 28×28 images
- NNet classifier fitted on PCA-reduced features ($p = 50$)
- 5000 training points
- 5000 test points

Synthetic data experiment for counterfactual inference

- Total sample size $n = 1000$
- $X \in \mathbb{R}^{100}$ correlated Gaussian
- $Y(1) \mid X \sim N(\mu(X), \sigma(X)^2)$:

  $\mu(X)$ depends on X_1, X_2 smoothly

  $\sigma(X) = -\log(1 - \Phi(X_1))$ (heteroscedastic)
- $e(X) \in [0.25, 0.5]$ depends on X_1 smoothly