Quantifying predictive uncertainty with conformal inference

Rina Foygel Barber

http://rinafb.github.io/

The prediction problem

Setting:

- Training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, test point (X_{n+1}, Y_{n+1})

  observed want to predict

- If fitted model $\widehat{\mu}_n$ overfits to training data,

$$|Y_{n+1} - \widehat{\mu}_n(X_{n+1})| \gg \frac{1}{n}\sum_{i=1}^{n} |Y_i - \widehat{\mu}_n(X_i)|$$

  even if training & test data are from the same distribution

The prediction problem

Run algorithm $\mathcal{A}$ on the training data $\rightsquigarrow$ fitted model $\widehat{\mu}_n$

Prediction interval for Y_{n+1}:

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}_n(X_{n+1}) \pm \text{(margin of error)}$$

Use training residuals? ("naive")

Use a parametric model?

Use smoothness assumptions?

Use cross-validation?

The prediction problem

- Want to be distribution-free —

  $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$ w/o assumptions on data distrib.

- Want to be efficient — minimize width of interval $\widehat{C}_n(X_{n+1})$

The prediction problem

Outline:

1. Background: conformal prediction
2. The jackknife+ and jackknife+-after-bootstrap
3. Some extensions

Using a holdout set

- Using any algorithm, fit model

$$\widehat{\mu}_{n/2} = \mathcal{A}\Big((X_1, Y_1), \ldots, (X_{n/2}, Y_{n/2})\Big)$$

- Compute holdout residuals

$$R_i = |Y_i - \widehat{\mu}_{n/2}(X_i)|, \quad i = n/2 + 1\,, \ldots\,, n$$

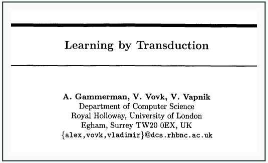
- Prediction interval:

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}_{n/2}(X_{n+1}) \pm \big(\text{the } (1-\alpha)\text{-quantile of } R_{n/2+1}, \ldots, R_n\big)$$

Conformal prediction

Background on the conformal prediction framework:

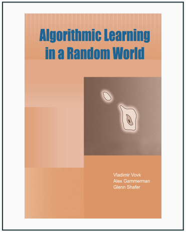
key idea = statistical inference via exchangeability of the data



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

Gammerman, Vovk, Vapnik
UAI 1998



Vovk, Gammerman, Shafer
2005 — see `alrw.net`

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 523, 1094–1111, Theory and Methods
https://doi.org/10.1080/01621459.2017.1307116

Taylor & Francis
Taylor & Francis Group

Check for updates

Distribution-Free Predictive Inference for Regression

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA

ABSTRACT

We develop a general framework for distribution-free predictive inference in regression, using conformal inference. The proposed methodology allows for the construction of a prediction band for the response variable using any estimator of the regression function. The resulting prediction band preserves the consistency properties of the original estimator under standard assumptions, while guaranteeing finite-sample marginal coverage even when these assumptions do not hold. We analyze and compare, both empirically and theoretically, the two major variants of our conformal framework: full conformal inference and split conformal inference, along with a related jackknife method. These methods offer different tradeoffs between statistical accuracy (length of resulting prediction intervals) and computational efficiency. As extensions, we develop a method for constructing valid in-sample prediction intervals called *rank-one-out* conformal inference, which has essentially the same computational efficiency as split conformal inference. We also describe an extension of our procedures for producing prediction bands with locally varying length, to adapt to heteroscedasticity in the data. Finally, we propose a model-free notion of variable importance, called *leave-one-covariate-out* or LOCO inference. Accompanying this article is an R package `conformalInference` that implements all of the proposals we have introduced. In the spirit of reproducibility, all of our empirical results can also be easily (re)generated using this package.

ARTICLE HISTORY
Received April 2016
Revised February 2017

KEYWORDS
Distribution-free; Model misspecification; Prediction band; Regression; Variable importance

Lei, G'Sell, Rinaldo,
Tibshirani, Wasserman
JASA 2018

Split conformal prediction

Split conformal prediction interval (a.k.a. holdout):[1]

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}_{n/2}(X_{n+1}) \pm \widehat{Q}_{1-\alpha}\left\{R_{n/2+1}, \ldots, R_n\right\}$$

the $\lceil (1-\alpha)(n/2+1) \rceil$-th smallest value in the list

Theorem:

If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable (e.g., i.i.d.), then for any algorithm $\mathcal{A}$, the split conformal method satisfies

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

[1] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Split conformal prediction

Exchangeability:

Random variables $Z_1, \ldots, Z_m$ are *exchangeable* if, for any permutation σ,

$$(Z_1, \ldots, Z_m) \overset{\mathrm{d}}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(m)})$$

the distributions are equal

Split conformal prediction

Exchangeability:
Random variables $Z_1, \ldots, Z_m$ are *exchangeable* if, for any permutation σ,

$$(Z_1, \ldots, Z_m) \overset{\mathrm{d}}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(m)})$$

the distributions are equal

For prediction:
Assume exchangeability of the *pairs* $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$

Examples:

- (X_i, Y_i)'s are *drawn i.i.d.* from any distribution
- (X_i, Y_i)'s sampled *uniformly without replacement* from a fixed set

The nonconformity score

In the above construction,

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm [...] \ = \{ \text{ all } y \text{ values with } |y - \widehat{\mu}(X_{n+1})| \leq [...] \ \}$$

[2] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[3] Lei et al 2018, *Distribution-Free Predictive Inference for Regression*

[4] Romano et al 2019, *Conformalized quantile regression*

[5] Izbicki et al 2020, *Flexible distribution-free conditional predictive bands using density estimators*

[6] Romano et al 2020, *Classification with Valid and Adaptive Coverage*

The nonconformity score

In the above construction,

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm [...] \ = \{ \text{ all } y \text{ values with } |y - \widehat{\mu}(X_{n+1})| \leq [...] \ \}$$

Generalize to any score $\widehat{S}(x, y)$ measuring "nonconformity" of (x, y):[2]

$$\widehat{C}_n(X_{n+1}) = \{ \text{ all } y \text{ values with } \widehat{S}(X_{n+1}, y) \leq [...] \ \}$$

Can choose $\widehat{S}$ to...

- Adapt to nonconstant variance[3]
- Use quantile regression[4] or density estimation[5]
- Handle a categorical response[6]
- & many more

[2] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[3] Lei et al 2018, *Distribution-Free Predictive Inference for Regression*

[4] Romano et al 2019, *Conformalized quantile regression*

[5] Izbicki et al 2020, *Flexible distribution-free conditional predictive bands using density estimators*

[6] Romano et al 2020, *Classification with Valid and Adaptive Coverage*

Full conformal prediction

Full conformal prediction:[7] distrib.-free guarantee w/o sample splitting

[7] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Full conformal prediction

Full conformal prediction:[7] distrib.-free guarantee w/o sample splitting

- Fit model to training + test data

$$\widehat{\mu}_{n+1} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}))$$

- Compute residuals

$$R_i = |Y_i - \widehat{\mu}_{n+1}(X_i)| \text{ for } i \leq n; \;\; R_{n+1} = |\, Y_{n+1} - \widehat{\mu}_{n+1}(X_{n+1})|$$

- Check if $R_{n+1} \leq \big(\text{the } (1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\big)$

↖

If data points are exchangeable, and $\mathcal{A}$ treats data points symmetrically, then $R_1, \ldots, R_{n+1}$ are exchangeable

$\Rightarrow$ this event has $\geq 1 - \alpha$ probability

[7] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Full conformal prediction

Full conformal prediction:[7] distrib.-free guarantee w/o sample splitting

- Fit model to training + test data

$$\widehat{\mu}_{n+1} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, \cancelto{y}{Y_{n+1}}))$$

- Compute residuals

$$R_i = |Y_i - \widehat{\mu}_{n+1}(X_i)| \text{ for } i \leq n;\ \ R_{n+1} = |\cancelto{y}{Y_{n+1}} - \widehat{\mu}_{n+1}(X_{n+1})|$$

- Check if $R_{n+1} \leq \big(\text{the } (1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\big)$

If data points are exchangeable, and $\mathcal{A}$ treats data points symmetrically, then $R_1, \ldots, R_{n+1}$ are exchangeable

$\Rightarrow$ this event has $\geq 1 - \alpha$ probability if we plug in $y = Y_{n+1}$

[7] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Full conformal prediction

Full conformal prediction:[7] distrib.-free guarantee w/o sample splitting

- Fit model to training + test data

$$\widehat{\mu}_{n+1} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, \overset{y}{\cancel{Y_{n+1}}}))$$

- Compute residuals

$$R_i = |Y_i - \widehat{\mu}_{n+1}(X_i)| \text{ for } i \leq n;\ \ R_{n+1} = |\overset{y}{\cancel{Y_{n+1}}} - \widehat{\mu}_{n+1}(X_{n+1})|$$

- Check if $R_{n+1} \leq \big(\text{the } (1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\big)$

↖

If data points are exchangeable, and $\mathcal{A}$ treats data points symmetrically,
then $R_1, \ldots, R_{n+1}$ are exchangeable
$\Rightarrow$ this event has $\geq 1-\alpha$ probability if we plug in $y = Y_{n+1}$

$\widehat{C}_n(X_{n+1}) = \{\text{all } y \in \mathbb{R} \text{ for which the event above holds}\}$

[7] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Full conformal prediction

Validity guarantee for full conformal:[8]

Theorem:

If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable (e.g., i.i.d.), and the algorithm $\mathcal{A}$ treats data points symmetrically, then full CP satisfies

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

[8] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Full conformal: computational challenges

Full conformal prediction requires that the algorithm $\mathcal{A}$ is re-run:

- For each test value X_{n+1} of interest
- For every possible value of Y_{n+1} (e.g, all $y \in \mathbb{R}$)

Approaches:

- In practice — restrict to a grid of y values (but no theory)
- Specialized methods for specific algorithms e.g. Lasso[9]
- Discretized CP — use a discretized version of $\mathcal{A}$
  to restore theoretical guarantees[10]

[9] Lei 2017, *Fast Exact Conformalization of Lasso using Piecewise Linear Homotopy*

[10] Chen, Chun, & B. 2017, *Discretized conformal prediction for efficient distribution-free inference*

Jackknife+

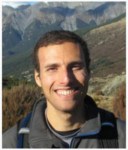
Outline:

1. Background: conformal prediction
2. The jackknife+ and jackknife+-after-bootstrap
3. Some extensions

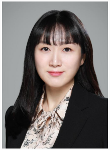
Collaborators:

Emmanuel Candès

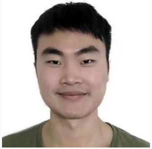
Aaditya Ramdas



Ryan Tibshirani



Byol Kim



Chen Xu

Jackknife+

Jackknife+-after-bootstrap

Jackknife+

Computational/statistical tradeoff:

	$\#$ calls to $\mathcal{A}$	Sample size for training
Split conformal (a.k.a. holdout)	1	$n/2$
Full conformal	∞	n

Can cross-validation type methods offer a compromise?

Jackknife+

Jackknife a.k.a. leave-one-out cross-validation:

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}_n(X_{n+1}) \pm \widehat{Q}_{1-\alpha}\{R_1, \dots, R_n\}$$

where $R_i = |Y_i - \widehat{\mu}_{-i}(X_i)|$ = leave-one-out residual

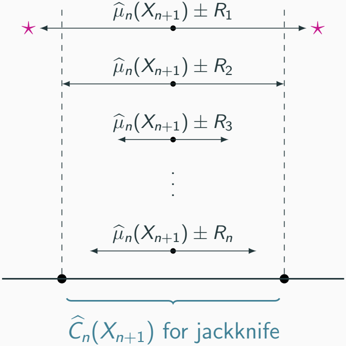
trained on data points $\{1, \dots, n\} \setminus \{i\}$

- No distribution-free guarantees
- Predictive coverage holds assuming algorithmic stability:[11]

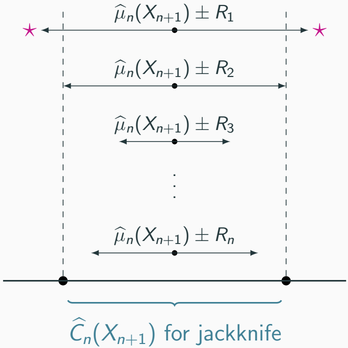
$$\widehat{\mu}_n(X_{n+1}) \approx \widehat{\mu}_{-i}(X_{n+1})$$

[11] Steinberger & Leeb 2018

Jackknife+


$\widehat{\mu}_n(X_{n+1}) \pm R_1$
$\widehat{\mu}_n(X_{n+1}) \pm R_2$
$\widehat{\mu}_n(X_{n+1}) \pm R_3$
$\widehat{\mu}_n(X_{n+1}) \pm R_n$
$\widehat{C}_n(X_{n+1})$ for jackknife

Jackknife+


$\widehat{\mu}_n(X_{n+1}) \pm R_1$
$\widehat{\mu}_n(X_{n+1}) \pm R_2$
$\widehat{\mu}_n(X_{n+1}) \pm R_3$
$\vdots$
$\widehat{\mu}_n(X_{n+1}) \pm R_n$
$\widehat{C}_n(X_{n+1})$ for jackknife
$\widehat{\mu}_{-1}(X_{n+1}) \pm R_1$
$\widehat{\mu}_{-2}(X_{n+1}) \pm R_2$
$\widehat{\mu}_{-3}(X_{n+1}) \pm R_3$
$\vdots$
$\widehat{\mu}_{-n}(X_{n+1}) \pm R_n$
$\widehat{C}_n(X_{n+1})$ for jackknife+

Jackknife+

Jackknife+:[12]

$$\widehat{C}_n(X_{n+1}) = \left[\widehat{Q}_\alpha\left\{\widehat{\mu}_{-i}(X_{n+1}) - R_i\right\},\ \widehat{Q}_{1-\alpha}\left\{\widehat{\mu}_{-i}(X_{n+1}) + R_i\right\}\right]$$

- CV+ = extension to K-fold cross-validation
- Closely related to the cross-conformal method[13]

[12] B., Candès, Ramdas, Tibshirani 2019, *Predictive inference with the jackknife+*

[13] Vovk 2015, Vovk et al 2018

Theory for jackknife+

Theorem:

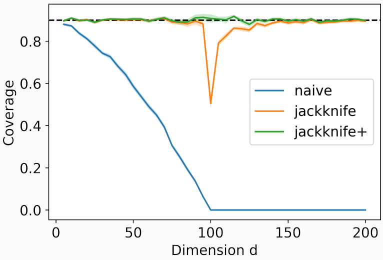
If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable (e.g., i.i.d.), and $\mathcal{A}$ treats data points symmetrically, then jackknife+ satisfies

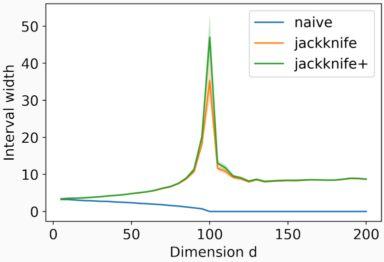
$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \right\} \geq 1 - 2\alpha.$$

- In practice, typically see $\approx 1 - \alpha$ coverage
- Can prove $\gtrapprox 1 - \alpha$ coverage if assume $\mathcal{A}$ is stable

Simulation

- $n = 100$, $d \in \{5, 10, \ldots, 200\}$
- $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\quad Y_i = X_i^\top \beta + \mathcal{N}(0,1)$
- $\mathcal{A}$ = "ridgeless" regression (least sq. with min ℓ_2 norm)
  Stable if $d \ll n$ or $d \gg n$, but if $d \approx n$ then unstable[14]





[14] Hastie et al 2019, *Ridgeless Least Squares Interpolation.*

Jackknife+-after-bootstrap

Suppose we would like to use an algorithm $\mathcal{A}_{\text{ens}}$
that is constructed with bootstrapping/subsampling:

- For $b = 1, \ldots, B$,
  - — Subsample/bootstrap new training set $S_b \subset \{1, \ldots n\}$ of size m
  - — Fit model $\widehat{\mu}^{(b)}$ on data set S_b using $\mathcal{A}_{\text{base}}$

- Then predict with aggregation:

$$\widehat{\mu}_\varphi(X_{n+1}) = \varphi\Big(\widehat{\mu}^{(1)}(X_{n+1}), \ldots, \widehat{\mu}^{(B)}(X_{n+1})\Big)$$

(E.g., φ is the mean / median / trimmed mean)

Jackknife+-after-bootstrap

The problem:

- Cost of $\mathcal{A}_{\text{ens}}$: B calls to $\mathcal{A}_{\text{base}}$
- Cost of jackknife+ for $\mathcal{A}_{\text{ens}}$: Bn calls to $\mathcal{A}_{\text{base}}$

Jackknife+-after-bootstrap

The J+aB algorithm:[15]

- For $b = 1, \ldots, B$,
  - Subsample/bootstrap new training set $S_b \subset \{1, \ldots n\}$ of size m
  - Fit model $\widehat{\mu}^{(b)}$ on data set S_b using $\mathcal{A}_{\text{base}}$
- Compute leave-one-out models & out-of-bag residuals:

$$\widehat{\mu}_{\varphi\setminus i}(x) = \varphi\Big(\widehat{\mu}^{(b)}(X_{n+1}) : i \notin S_b\Big), \quad R_i = |Y_i - \widehat{\mu}_{\varphi\setminus i}(X_i)|$$

- Prediction interval:

$$\widehat{C}_n(X_{n+1}) = \Big[\widehat{\mathrm{Q}}_\alpha\Big(\mu_{\varphi\setminus i}(X_{n+1}) - R_i\Big),\ \widehat{\mathrm{Q}}_{1-\alpha}\Big(\mu_{\varphi\setminus i}(X_{n+1}) + R_i\Big)\Big]$$

[15] Kim, Xu, B. 2020, *Predictive Inference Is Free with the Jackknife+-after-Bootstrap*
(Related ideas in the literature: Zhang et al, Devetyarov & Nouretdinov, Löfström et al, Boström et al, Linusson et al)

Jackknife+-after-bootstrap

- Cost of $\mathcal{A}_{\text{ens}}$: B calls to $\mathcal{A}_{\text{base}}$
- Cost of jackknife+ for $\mathcal{A}_{\text{ens}}$: Bn calls to $\mathcal{A}_{\text{base}}$
- Cost of jackknife+-after-bootstrap on $\mathcal{A}_{\text{base}}$: B calls to $\mathcal{A}_{\text{base}}$

(Assuming model fitting is the dominant cost — not aggregation / evaluation)

Jackknife+-after-bootstrap

Theorem:
For any distrib. P & any $\mathcal{A}_{\text{base}}$, jack+-after-bootstrap satisfies

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - 2\alpha$$

if the # of ensembled models is a *random* $B \sim \text{Binomial}(\tilde{B}, \theta)$ with

$$\theta = \begin{cases} \left(1 - \frac{1}{n+1}\right)^m, & \text{for bootstrapped samples of size } m, \\ 1 - \frac{m}{n+1}\ , & \text{for subsamples of size } m \end{cases}$$

Extensions

Outline:

1. Background: conformal prediction
2. The jackknife+ and jackknife+-after-bootstrap
3. Some extensions

Nonexchangeable conformal prediction (nexCP)

Theory for full conformal relies on:

1. $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable (e.g., i.i.d.)
2. Regression algorithm $\mathcal{A}$ treats input data points symmetrically

Nonexchangeable conformal prediction (nexCP)

Theory for full conformal relies on:

1. $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable (e.g., i.i.d.)
2. Regression algorithm $\mathcal{A}$ treats input data points symmetrically

Challenges in practice:

1. $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ may be nonexchangeable
   (e.g., distribution drift, dependence over time, ...)
2. May want to choose $\mathcal{A}$ that treats data nonsymmetrically
   (e.g., weighted regression, autoregressive model, ...)

Nonexchangeable conformal prediction (nexCP)

The method:[16] draw a random index K with $\mathbb{P}\{K = i\} = w_i$, then:

- Fit model to training + test data
$$\widehat{\mu}_{n+1} = \mathcal{A}((X_1, Y_1), \dots, \underbrace{(X_{n+1}, y)}_{\text{in position } K}, \dots, (X_n, Y_n), (X_K, Y_K))$$
- Compute residuals
$$R_i = |Y_i - \widehat{\mu}_{n+1}(X_i)| \text{ for } i \leq n; \;\; R_{n+1} = |y - \widehat{\mu}_{n+1}(X_{n+1})|$$
- Check if $R_{n+1} \;\leq\; \big(\text{the } (1-\alpha) \text{ quantile of } \{R_i \text{ with weight } w_i\}\big)$

fixed weights $w_i \geq 0$ with $\sum_i w_i = 1$

$$\widehat{C}_n(X_{n+1}) = \{\text{all } y \in \mathbb{R} \text{ for which the above holds}\}$$

[16] B., Candès, Ramdas, Tibshirani 2022, *Conformal prediction beyond exchangeability*

Nonexchangeable conformal prediction (nexCP)

The method:[16] draw a random index K with $\mathbb{P}\{K=i\}=w_i$, then:

- Fit model to training + test data

$$\widehat{\mu}_{n+1} = \mathcal{A}((X_1,Y_1),\dots,\underbrace{(X_{n+1},y)}_{\text{in position } K},\dots,(X_n,Y_n),(X_K,Y_K))$$

- Compute residuals

$$R_i = |Y_i - \widehat{\mu}_{n+1}(X_i)| \text{ for } i\le n;\ \ R_{n+1} = |y - \widehat{\mu}_{n+1}(X_{n+1})|$$

- Check if $R_{n+1} \ \le\ \big(\text{the } (1-\alpha) \text{ quantile of } \{R_i \text{ with weight } w_i\}\big)$

fixed weights $w_i \ge 0$ with $\sum_i w_i = 1$

$$\widehat{C}_n(X_{n+1}) = \{\text{all } y\in\mathbb{R} \text{ for which the above holds}\}$$

- If data is i.i.d. or exchangeable, coverage $\ge 1-\alpha$
- If exchangeability is violated, control loss of coverage
  by choosing w_i to be low for "risky" data points (e.g., old data)

[16] B., Candès, Ramdas, Tibshirani 2022, *Conformal prediction beyond exchangeability*

Other definitions of risk

Conformal prediction methods bound

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1})\right\} = \mathbb{E}\Big[\underbrace{\mathbf{1}\{Y_{n+1} \notin \widehat{C}_n(X_{n+1})\}}_{\text{zero-one loss}}\Big]$$

Can use a conformal approach to control other definitions of risk.[17,18]

Examples:

- FDR for flagging out-of-distribution data points
- False pos./neg. rates if $Y =$ a set of labels
- Accuracy rate for selecting pixels within an image

[17] Angelopolous et al 2021, *Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control*

[18] Bates et al 2021, *Distribution-Free, Risk-Controlling Prediction Sets*

The streaming setting

Conformal prediction can also be applied to an online setting:

- If data points are iid,
  conformal p-values are valid (and $\perp\!\!\!\perp$) at each time t
  $\Rightarrow$ can use conformal to predict / to test for changepoints[19]
- Can bound cumulative error under arbitrary distribution drift[20,21]
- If data points form a time series,
  CP achieves asymptotic coverage under some assumptions[22]

[19] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[20] Gibbs & Candès 2021, *Adaptive conformal inference under distribution shift*

[21] Feldman et al 2022, *Conformalized Online Learning: Online Calibration Without a Holdout Set*

[22] Xu & Xie 2021, *Conformal prediction interval for dynamic time-series*

Weighted conformal prediction

The covariate shift setting:

- Marginal distribution of X is different in training vs. test data
  (e.g., some groups are under-represented in training data)
- But, distribution of $Y|X$ is the same
- If the shift is $\approx$ known, can apply *weighted conformal prediction*[23]

[23] Tibshirani, B., Ramdas, & Candès 2019, *Conformal prediction under covariate shift*

[24] Candès, Lei, Ren 2021, *Conformalized survival analysis*

[25] Lei & Candès 2020, *Conformal inference of counterfactuals and individual treatment effects*

[26] Fannjiang et al 2022, *Conformal prediction for the design problem*

[27] Podkopaev & Ramdas 2021, *Distribution-free uncertainty quantification for classification under label shift*

Weighted conformal prediction

The covariate shift setting:

- Marginal distribution of X is different in training vs. test data (e.g., some groups are under-represented in training data)
- But, distribution of $Y|X$ is the same
- If the shift is $\approx$ known, can apply *weighted conformal prediction*[23]

Applications:

- Survival analysis & censored data[24]
- Estimating individual treatment effects[25]
- Prediction in the design problem (active learning)[26]
- A related problem — label shift (for categorical Y / classification)[27]

[23] Tibshirani, B., Ramdas, & Candès 2019, *Conformal prediction under covariate shift*

[24] Candès, Lei, Ren 2021, *Conformalized survival analysis*

[25] Lei & Candès 2020, *Conformal inference of counterfactuals and individual treatment effects*

[26] Fannjiang et al 2022, *Conformal prediction for the design problem*

[27] Podkopaev & Ramdas 2021, *Distribution-free uncertainty quantification for classification under label shift*

Thank you!

Website:

http://rinafb.github.io/