Conformal prediction beyond exchangeability

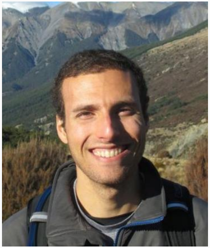
Rina Foygel Barber

http://rinafb.github.io/

Collaborators

Emmanuel Candès

Aaditya Ramdas



Ryan Tibshirani

- Thanks to American Institute of Math (AIM) for hosting & supporting our collaboration as an AIM SQuaRE

The prediction problem

Setting:

- Training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ $\rightsquigarrow$ fit model $\widehat{\mu}(X_i) \approx Y_i$

- Test point (X_{n+1}, Y_{n+1})

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

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

  even if training & test data are from the same distribution

The prediction problem

Goal: build prediction band C as a function of the training data,
such that $\widehat{C}_n(X_{n+1})$ is likely to contain Y_{n+1}

- Want to be <u>distribution-free</u> —
  coverage holds w/o assumptions on distrib. of (X, Y)
- Want to be <u>efficient</u> — minimize width of interval $\widehat{C}_n(X_{n+1})$

The holdout method

- 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:

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

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

The holdout method

Theorem:[1]

If $\underbrace{(X_{n/2+1}, Y_{n/2+1}), \ldots, (X_n, Y_n)}_{\text{holdout}}, \underbrace{(X_{n+1}, Y_{n+1})}_{\text{test}}$ are i.i.d. (or exch.),

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

[1] Vovk et al 2005, Papadopoulos 2008, Lei et al. 2018

The holdout method

Theorem:[1]

If $\underbrace{(X_{n/2+1}, Y_{n/2+1}), \ldots, (X_n, Y_n)}_{\text{holdout}}, \underbrace{(X_{n+1}, Y_{n+1})}_{\text{test}}$ are i.i.d. (or exch.),

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

Proof:

After conditioning on training data, holdout + test data is i.i.d. (or exch.)

$\Rightarrow$ residuals $R_{n/2+1}, \ldots, R_{n+1}$ are exchangeable

[1] Vovk et al 2005, Papadopoulos 2008, Lei et al. 2018

The holdout method

Theorem:[1]

If $\underbrace{(X_{n/2+1}, Y_{n/2+1}), \ldots, (X_n, Y_n)}_{\text{holdout}}, \underbrace{(X_{n+1}, Y_{n+1})}_{\text{test}}$ are i.i.d. (or exch.),

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

Proof:

After conditioning on training data, holdout + test data is i.i.d. (or exch.)

$\Rightarrow$ residuals $R_{n/2+1}, \ldots, R_{n+1}$ are exchangeable

$\Rightarrow \mathbb{P}\left\{ R_{n+1} \leq \big(\text{the } (1-\alpha)\text{-quantile of } R_{n/2+1}, \ldots, R_{n+1}\big) \right\} \geq 1 - \alpha$

$$\Updownarrow$$

$$R_{n+1} \leq \widehat{\mathrm{Q}}_{n/2,\alpha}(R_{n/2+1}, \ldots, R_n)$$

[1] Vovk et al 2005, Papadopoulos 2008, Lei et al. 2018

Full conformal prediction

Naive method vs Holdout method

$$\underbrace{\widehat{\mu}_n(X_{n+1})}_{\text{more accurate}} \pm \underbrace{\widehat{\mathrm{Q}}_{n,\alpha}(R_i)}_{\substack{\text{too small}\\\text{(overfitted)}}} \qquad \text{vs} \qquad \underbrace{\widehat{\mu}_{n/2}(X_{n+1})}_{\text{less accurate}} \pm \underbrace{\widehat{\mathrm{Q}}_{n/2,\alpha}(R_i)}_{\substack{\text{calibrated}\\\text{(but wider)}}}$$

[2] Vovk, Gammerman, Shafer 2005

Full conformal prediction

$$\underset{\text{Naive method}}{\underbrace{\widehat{\mu}_n(X_{n+1})}_{\text{more accurate}} \pm \underbrace{\widehat{\mathrm{Q}}_{n,\alpha}(R_i)}_{\substack{\text{too small}\\\text{(overfitted)}}}} \quad \text{vs} \quad \underset{\text{Holdout method}}{\underbrace{\widehat{\mu}_{n/2}(X_{n+1})}_{\text{less accurate}} \pm \underbrace{\widehat{\mathrm{Q}}_{n/2,\alpha}(R_i)}_{\substack{\text{calibrated}\\\text{(but wider)}}}}$$

An alternative—the *full conformal* method:[2]

- Models fitted on all n training samples (no data splitting)
- Guaranteed distribution-free predictive coverage
- High computational cost

[2] Vovk, Gammerman, Shafer 2005

Full conformal prediction

- Suppose we observe training + test data:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$$

- Fit model to all $n+1$ data points,

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

  & compute residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)|, \ i = 1, \ldots, n, \quad R_{n+1} = |Y_{n+1} - \widehat{\mu}(X_{n+1})|$$

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

$\nwarrow$

If data points are i.i.d. (or exch.), and $\mathcal{A}$ treats data points symmetrically,
then $R_1, \ldots, R_{n+1}$ are exchangeable
$\Rightarrow$ this event has $\geq 1-\alpha$ probability

Full conformal prediction

- Suppose we observe training + test data:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, \quad y \quad)$$

- Fit model to all $n+1$ data points,

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

  & compute residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)|, \ i = 1, \ldots, n, \quad R_{n+1} = | \quad y \quad - \widehat{\mu}(X_{n+1})|$$

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

$\nwarrow$

If data points are i.i.d. (or exch.), 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 $y = Y_{n+1}$

Full conformal prediction

- Suppose we observe training + test data:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$

- Fit model to all $n+1$ data points,

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

  & compute residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)|,\ i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$

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

Full conformal prediction

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model to all $n+1$ data points,
$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)),$$
& compute residuals
$$R_i = |Y_i - \widehat{\mu}(X_i)|, \; i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$
- Check if $R_{n+1} \leq \big(\text{the } (1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\big)$

$$y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{\text{Yes}, \text{No}\}$$

Full conformal prediction

Test value $y \in \mathbb{R}$

- Suppose we observe training + test data:
$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$
- Fit model to all $n + 1$ data points,
$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)),$$
& compute residuals
$$R_i = |Y_i - \widehat{\mu}(X_i)|,\ i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$
- Check if $R_{n+1} \leq \big(\text{the } (1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\big)$

$$y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{\text{Yes}, \text{No}\}$$

if Yes: add $y \in \widehat{C}_n(X_{n+1})$

if No: discard y

Full conformal prediction

- In theory, need to run $\mathcal{A}$ on $(X_1, Y_1), \ldots, (X_{n+1}, y)$ for every $y \in \mathbb{R}$
- Can compute efficiently for some special cases (ridge, Lasso[3])
- In practice, run on a grid of y values (can be formalized[4])

[3] Lei 2017, *Fast Exact Conformalization of Lasso...*

[4] Chen et al 2017, *Discretized conformal prediction...*

Full conformal prediction

Theorem:[5]
If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are i.i.d. (or exchangeable), and the algorithm $\mathcal{A}$ that treats data points symmetrically,

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

Proof:

$$\mathbb{P}\left\{Y \in \widehat{C}_n(X_{n+1})\right\} =$$

$$\mathbb{P}\left\{\text{for test value } y = Y_{n+1}, \text{ answer is Yes}\right\} \geq 1 - \alpha$$

[5] Vovk, Gammerman, Shafer 2005

Full conformal prediction

Theorem:[5]
If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are i.i.d. (or exchangeable),
and the algorithm $\mathcal{A}$ that treats data points symmetrically,

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

Proof:

$$\mathbb{P}\left\{Y \in \widehat{C}_n(X_{n+1})\right\} = \mathbb{P}\left\{\text{for test value } y = Y_{n+1}\text{, answer is Yes}\right\} \geq 1 - \alpha$$

- The holdout method a.k.a. "split conformal" is a special case.

[5] Vovk, Gammerman, Shafer 2005

Related methods

Computational/statistical tradeoff:

	Holdout method (a.k.a. split conformal)	vs	Full conformal
# calls to $\mathcal{A}$	1		∞
Sample size used by $\mathcal{A}$	$n/2$		n

[6] Vovk 2015, Vovk et al 2018
[7] Barber et al 2019

Related methods

Computational/statistical tradeoff:

	Holdout method (a.k.a. split conformal)	vs	Full conformal
# calls to $\mathcal{A}$	1		∞
Sample size used by $\mathcal{A}$	$n/2$		n

A compromise—cross-validation type methods:

- Cross-conformal[6]
- Jackknife+ and CV+[7]

[6] Vovk 2015, Vovk et al 2018
[7] Barber et al 2019

The role of exchangeability

Theory for split/full conformal (and jackknife+ etc) 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

Why? Need exchangeability of residuals when we plug in $y = Y_{n+1}$:

- Fit model to all $n + 1$ data points,
$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)),$$
& compute residuals
$$R_i = |Y_i - \widehat{\mu}(X_i)|, \ i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$

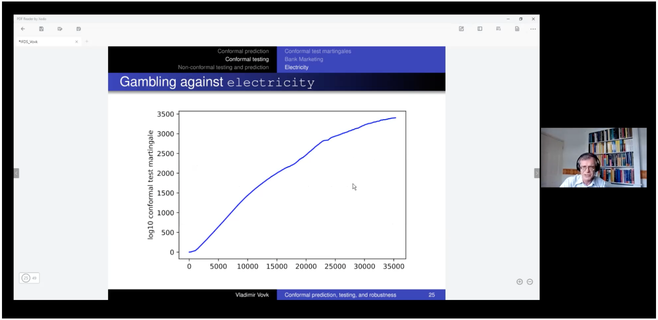
The role of exchangeability

Challenges in practice:

1. $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ 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, ...)

The role of exchangeability

Inspired by Vladimir Vovk's talk at IFDS MADlab workshop 2021:



- ELEC2 data set — tracking electricity demand in Australia
- Run conformal at each time point
- Under exchangeability, the process in the figure should be mean-zero

The role of exchangeability

Exchangeability $\Longleftrightarrow$ (X_{n+1}, Y_{n+1}) is a random draw from empirical distribution $\frac{1}{n+1}\sum_{i=1}^{n+1} \delta_{(X_i,Y_i)}$

δ_z = point mass at z

The role of exchangeability

Exchangeability $\Longleftrightarrow$ (X_{n+1}, Y_{n+1}) is a random draw from empirical distribution $\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{(X_i,Y_i)}$

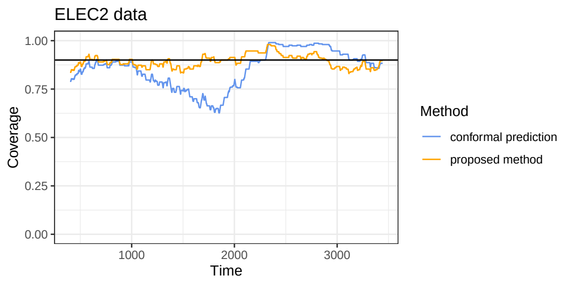
δ_z = point mass at z

Intuitively, with distribution drift, can we approximate (X_{n+1}, Y_{n+1}) as a random draw from $\sum_{i=1}^{n+1} w_i \cdot \delta_{(X_i,Y_i)}$?

weights $w_{n+1} \geq w_n \geq \cdots \geq w_1 \geq 0$ with $\sum_i w_i = 1$

The role of exchangeability

In practice, using weights appears to correct for nonexchangeability:



The role of exchangeability

In theory, using weights seems to violate the exchangeability argument.

- Under exchangeability, the key step in the proof:

$$\mathbb{P}\left\{R_{n+1} \leq \widehat{\mathrm{Q}}_{1-\alpha}\left(\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{R_i}\right)\right\} = \mathbb{P}\left\{R_k \leq \widehat{\mathrm{Q}}_{1-\alpha}\left(\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{R_i}\right)\right\}$$

The role of exchangeability

In theory, using weights seems to violate the exchangeability argument.

- Under exchangeability, the key step in the proof:

$$\mathbb{P}\left\{R_{n+1} \le \widehat{\mathrm{Q}}_{1-\alpha}\left(\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{R_i}\right)\right\} = \mathbb{P}\left\{R_k \le \widehat{\mathrm{Q}}_{1-\alpha}\left(\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{R_i}\right)\right\}$$

- But with weights...

$$\mathbb{P}\left\{R_{n+1} \le \widehat{\mathrm{Q}}_{1-\alpha}\left(\sum_{i=1}^{n+1} w_i \cdot \delta_{R_i}\right)\right\} \stackrel{???}{=} \mathbb{P}\left\{R_k \le \widehat{\mathrm{Q}}_{1-\alpha}\left(\sum_{i=1}^{n+1} w_i \cdot \delta_{R_i}\right)\right\}$$

Even if the data points (X_i, Y_i) are exchangeable,

- $w_i \not\equiv \frac{1}{n+1} \rightsquigarrow \widehat{\mathrm{Q}}_{1-\alpha}(\dots)$ is not symmetric function of the R_i's
- If also $\mathcal{A}$ treats data nonsymmetrically $\rightsquigarrow R_i$'s not exch.

Aims

Our aim is to construct a procedure that...

- Uses weighted quantiles to be more robust to distribution drift etc
- Allows for nonsymmetric algorithms, for a more accurate model
- Guarantees exact coverage (if data is in fact exchangeable)
- Guarantees bounded loss of coverage (if data has bounded violation of exchangeability)

Nonexchangeable conformal prediction (nexCP)

nexCP method (symmetric algorithm case):

- Suppose we observe training + test data:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$

- Fit model to all $n+1$ data points,

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

  & compute residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)|, \; i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$

- Check if $R_{n+1} \leq \big(\text{the } (1-\alpha) \text{ quantile of } \sum_{i=1}^{n+1} w_i \cdot \delta_{R_i}\big)$
- $\widehat{C}_n(X_{n+1}) = \{\text{all } y \in \mathbb{R} \text{ for which the above holds}\}$

Nonexchangeable conformal prediction (nexCP)

nexCP method (general algorithm case):

- Suppose we observe training + test data:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$

- Fit model to all $n+1$ data points,

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

for a random index $K \sim \sum_{i=1}^{n+1} w_i \cdot \delta_i$,

& compute residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)|,\ i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$

- Check if $R_{n+1} \leq \big(\text{the } (1-\alpha) \text{ quantile of } \sum_{i=1}^{n+1} w_i \cdot \delta_{R_i}\big)$
- $\widehat{C}_n(X_{n+1}) = \{\text{all } y \in \mathbb{R} \text{ for which the above holds}\}$

Nonexchangeable conformal prediction (nexCP)

Extensions — can define analogous nonexchangeable versions of:

- Split conformal (note: symmetry of $\mathcal{A}$ doesn't matter for this case)
- Jackknife+ and CV+

Nonexchangeable conformal prediction (nexCP)

Note—our methods require the weights $w_1, \ldots, w_{n+1}$ to be fixed.

This is very different from *weighted conformal prediction*, where data-dependent weights $w_i = w(X_i)$ are used as an *exact* correction for covariate shift[8]

[8] Tibshirani et al 2019; Candès, Lei, Ren 2021; Lei & Candès 2021

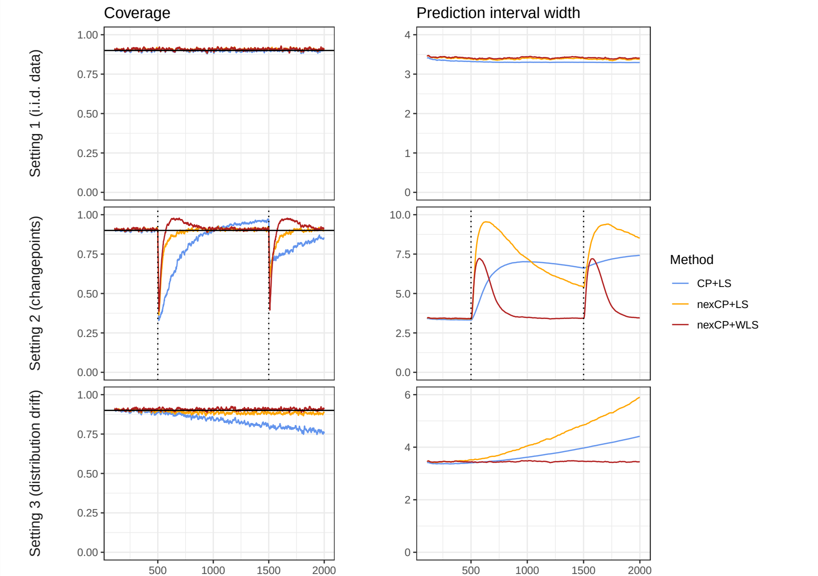
Empirical results

Compare 3 methods:

1. **CP+LS**: conformal prediction with $\mathcal{A}$ = least squares
2. **nexCP+LS**: nonexch. conformal prediction with $w_i \propto 0.99^{-i}$, with $\mathcal{A}$ = least squares
3. **nexCP+WLS**: nonexch. conformal prediction with $w_i \propto 0.99^{-i}$, with $\mathcal{A}$ = weighted least squares with weights $\propto 0.99^{-i}$

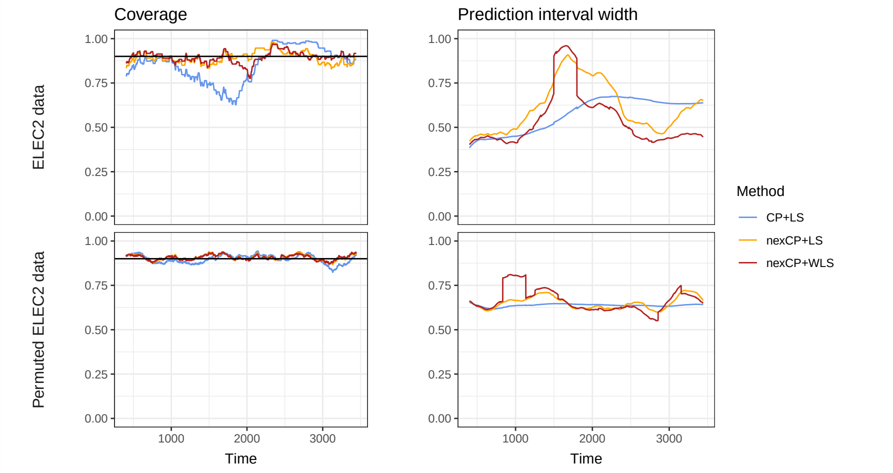
Empirical results

Simulated data



Empirical results

Real data — the ELEC2 dataset[9]



[9] Harries 1999; `https://www.kaggle.com/yashsharan/the-elec2-dataset`

Theoretical guarantee

Theorem: Let $w_1, \ldots, w_{n+1} \geq 0$ be fixed, with

$$\sum_i w_i = 1, \quad w_{n+1} = \max_i w_i.$$

Then nonexchangeable conformal prediction satisfies

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha - \sum_{k=1}^{n} w_k \cdot \mathrm{d}_{\mathrm{TV}}\big(R(\mathrm{data}), R(\mathrm{data}_{\mathrm{swap}(k)})\big)$$

- $R(\mathrm{data}) = (|Y_1 - \widehat{\mu}(X_1)|, \ldots, |Y_{n+1} - \widehat{\mu}(X_{n+1})|)$ for $\widehat{\mu} = \mathcal{A}\big((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})\big)$
- $\mathrm{data}_{\mathrm{swap}(k)}$ = data with points k and $n+1$ swapped

Theoretical guarantee

Implications:

- If (X_i, Y_i) are i.i.d. (or exch.), $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$
- If (X_i, Y_i) are independent, then

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha - 2\sum_i w_i \cdot \mathrm{d}_{\mathsf{TV}}((X_i, Y_i), (X_{n+1}, Y_{n+1}))$$

Proof sketch — exchangeable case

Suppose we generate *exchangeable* $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$,
then swap data points $n+1$ and K for $K \sim \sum_i w_i \cdot \delta_i$

After observing the *swapped* data set, which is the original test point?

- Since the data is exchangeable,
  $\mathbb{P}\{K = i \mid \text{(swapped data)}\} = \mathbb{P}\{K = i\} = w_i$

Proof sketch — exchangeable case

Suppose we generate *exchangeable* $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$,
then swap data points $n+1$ and K for $K \sim \sum_i w_i \cdot \delta_i$

After observing the *swapped* data set, which is the original test point?

- Since the data is exchangeable,
  $\mathbb{P}\{K = i \mid (\text{swapped data})\} = \mathbb{P}\{K = i\} = w_i$
- Equivalently, can view R_{n+1} as a draw from $\sum_i w_i \delta_{R(\text{data}_{\text{swap}(K)})_i}$

$$\Rightarrow \mathbb{P}\left\{R_{n+1} \leq \Big(\text{the } (1-\alpha)\text{-quantile of } \sum_i w_i \delta_{R(\text{data}_{\text{swap}(K)})_i}\Big)\right\} \geq 1 - \alpha$$

Proof sketch — general case

Suppose we generate *nonexchangeable* $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$, then swap data points $n+1$ and K for $K \sim \sum_i w_i \cdot \delta_i$

After observing the *swapped* data set, which is the original test point?

- For exchangeable data, $\mathbb{P}\{K = i \mid (\text{swapped data})\} = w_i$
- In general, distrib. of $K|(\text{swapped data})$ is $\approx \sum_i w_i \cdot \delta_i$
- $\mathrm{d}_{\mathrm{TV}}\Big((\text{data}, K), (\text{data}_{\text{swap}(K)}, K)\Big) \leq \sum_i w_i \mathrm{d}_{\mathrm{TV}}(\text{data}, \text{data}_{\text{swap}(k)})$

An aside — permutation tests beyond subgroups

Permutation tests for testing

$$H_0 : X = (X_1, \ldots, X_n) \text{ is exchangeable}$$

with a test statistic $T(X) = T(X_1, \ldots, X_n)$

A valid p-value, for any subgroup $G \subseteq \mathcal{S}_n$:[10]

$$P = \frac{\sum_{\sigma \in G} \mathbb{1}\{T(X_\sigma) \geq T(X)\}}{|G|}.$$

[10] Hemerik & Goeman 2018

An aside — permutation tests beyond subgroups

Permutation tests for testing

$$H_0 : X = (X_1, \ldots, X_n) \text{ is exchangeable}$$

with a test statistic $T(X) = T(X_1, \ldots, X_n)$

A valid p-value, for any subgroup $G \subseteq \mathcal{S}_n$:[10]

$$P = \frac{\sum_{\sigma \in G} \mathbb{1}\{T(X_\sigma) \geq T(X)\}}{|G|}.$$

For a subset $S \subset \mathcal{S}_n$ that is not a subgroup, this p-value is *not* valid:

$$P = \frac{\sum_{\sigma \in S} \mathbb{1}\{T(X_\sigma) \geq T(X)\}}{|S|}.$$

[10] Hemerik & Goeman 2018

An aside — permutation tests beyond subgroups

Example:

- $n = 4$
- S contains 3 permutations:

$$\text{Identity}, \quad 1 \leftrightarrow 3 \ \& \ 2 \leftrightarrow 4, \quad 1 \leftrightarrow 4 \ \& \ 2 \leftrightarrow 3$$

- Test statistic $T(X) = T(X_1, X_2, X_3, X_4) = X_1 + X_2$

An aside — permutation tests beyond subgroups

Example:

- $n = 4$
- S contains 3 permutations:

$$\text{Identity}, \quad 1 \leftrightarrow 3 \;\&\; 2 \leftrightarrow 4, \quad 1 \leftrightarrow 4 \;\&\; 2 \leftrightarrow 3$$

- Test statistic $T(X) = T(X_1, X_2, X_3, X_4) = X_1 + X_2$
- Calculate p-value:

$$P = \frac{1 + \mathbb{1}\{T(X_3, X_4, X_1, X_2) \geq T(X)\} + \mathbb{1}\{T(X_4, X_3, X_2, X_1) \geq T(X)\}}{3}$$

$$= \begin{cases} \frac{1}{3}, & X_1 + X_2 > X_3 + X_4, \\ 1, & \text{otherwise} \end{cases}$$

$\Rightarrow$ can have $\mathbb{P}\left\{P \leq \frac{1}{3}\right\} = \frac{1}{2}$

An aside — permutation tests beyond subgroups

Theorem: Let $S \subseteq \mathcal{S}_n$ be an arbitrary subset, and let $\sigma_* \in S$ be drawn uniformly at random. Then

$$P = \frac{\sum_{\sigma \in S} \mathbb{1}\{T(X_{\sigma \circ \sigma_*^{-1}}) \geq T(X)\}}{|S|}$$

is a valid p-value.

A related result—Besag & Clifford's "parallel" construction for exchangeable samples from an MCMC[11]

[11] Besag & Clifford 1989

An aside — permutation tests beyond subgroups

Theorem: Let $S \subseteq \mathcal{S}_n$ be an arbitrary subset, and let $\sigma_* \in S$ be drawn uniformly at random. Then

$$P = \frac{\sum_{\sigma \in S} \mathbb{1}\{T(X_{\sigma \circ \sigma_*^{-1}}) \geq T(X)\}}{|S|}$$

is a valid p-value.

A related result—Besag & Clifford's "parallel" construction for exchangeable samples from an MCMC[11]

Connection to nonexchangeable conformal prediction:

- swap random K with $n+1$ before running $\mathcal{A} \longleftrightarrow$ applying σ_*^{-1}
- compare R_{n+1} vs $\sum_i w_i \cdot \delta_{R_i} \longleftrightarrow$ compare $T(X)$ vs $T(X_{\sigma \circ \sigma_*^{-1}})$'s

[11] Besag & Clifford 1989

Summary

- If we use *fixed* w_i's, no loss of coverage under exchangeability
  (via a new exchangeability proof technique),
  and robustness to violations of exchangeability
- If we also want to use a nonsymmetric algorithm,
  get the *same coverage guarantee* with a randomization step
  (swap (X_K, Y_K) with (X_{n+1}, Y_{n+1}) before running $\mathcal{A}$)

$\rightsquigarrow$ Conformal prediction methods can now be applied to nonstationary data; models with a drift term or an autoregressive term; etc

Open questions

- Can we check robustness for a particular data set / distrib.?
  (i.e., if $w_i \propto \rho^{-i}$, how to tune ρ?)
- Can we use data-dependent w_i's?
  (e.g., w_i depends on distance(X_i, X_{n+1}))
- How to perform inference on multiple models / streaming data / other settings?