Distribution-free prediction: exchangeability and beyond

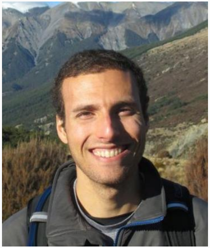
Rina Foygel Barber

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 $(X_1, Y_1), \ldots, (X_n, Y_n)$, 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 —

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

- 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+
3. Conformal prediction beyond exchangeability

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

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

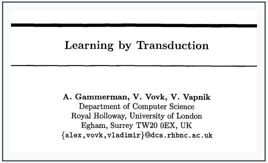
- Prediction interval:

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

Conformal prediction

Background on the conformal prediction framework:

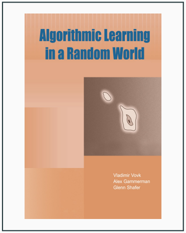
key idea = statistical inference via exchangeability of the data



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

Background: split conformal prediction

Split conformal prediction interval (a.k.a. holdout):

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

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

Theorem: [Vovk, Gammerman, Shafer 2005]

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

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

Proof:

After conditioning on $\widehat{\mu}_{n/2}$, holdout + test data is exchangeable

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

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

Proof:

After conditioning on $\widehat{\mu}_{n/2}$, holdout + test data is exchangeable

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

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

$$\Updownarrow$$

$$R_{n+1} \leq \widehat{\mathrm{Q}}_{1-\alpha}\{R_{n/2+1}, \dots, R_n\}$$

$$\Updownarrow$$

$$Y_{n+1} \in \widehat{C}_n(X_{n+1})$$

Background: full conformal prediction

Full conformal prediction:

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

$\nwarrow$

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

Background: full conformal prediction

Full conformal prediction:

- 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

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

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

$\nwarrow$

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 $\ge 1 - \alpha$ probability if we plug in $y = Y_{n+1}$

Background: full conformal prediction

Full conformal prediction:

- 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 $y = Y_{n+1}$

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

Background: full conformal prediction

Theorem: [Vovk, Gammerman, Shafer 2005]

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

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

Jackknife+

Outline:

1. Background: conformal prediction
2. The jackknife+
3. Conformal prediction beyond exchangeability

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:

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

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

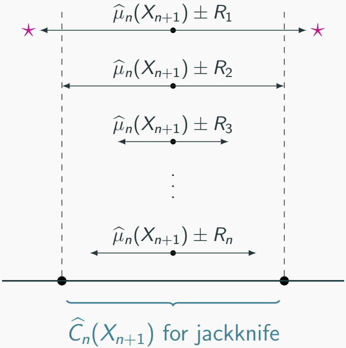
trained on data points $\{1, \ldots, n\} \setminus \{i\}$

- No distribution-free guarantees
- Predictive coverage holds assuming algorithmic stability:[1]

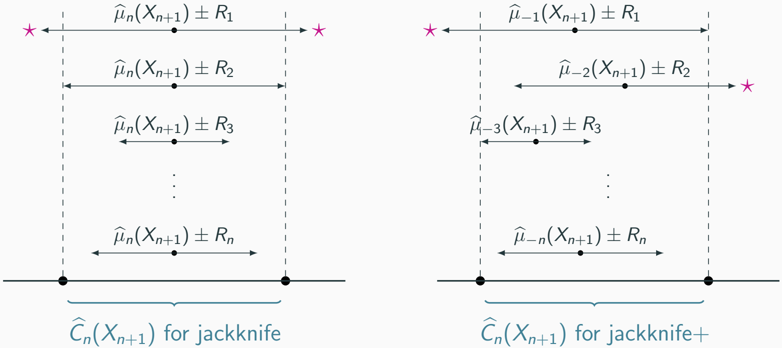
$$\widehat{\mu}_n(X_{n+1}) \approx \widehat{\mu}_{-i}(X_{n+1})$$

[1] 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+:

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

- CV+ = extension to K-fold cross-validation
- Closely related to the cross-conformal method[2]

[2] Vovk 2015, Vovk et al 2018

Jackknife+

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

Theory for jackknife+

Theorem: [B., Candès, Ramdas, Tibshirani]

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

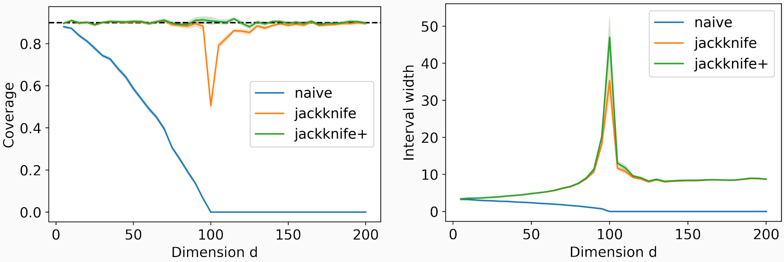
Theory for jackknife+

Challenge: jackknife+ construction doesn't appear exchangeable:
fitted models $\widehat{\mu}_{-i}$ for $i \in \{1, \ldots, n\}$

Proof idea: embed jackknife+ into a larger exchangeable problem:
models $\tilde{\mu}_{-ij}$ for each $i, j \in \{1, \ldots, n+1\}$

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[3]



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

Beyond exchangeability

Outline:

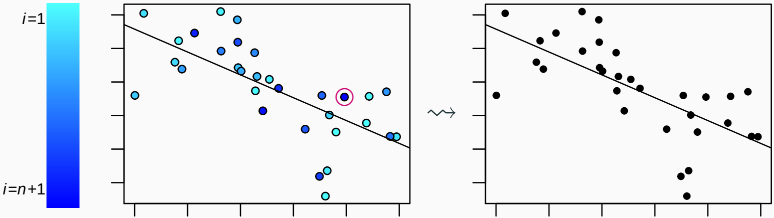
1. Background: conformal prediction
2. The jackknife+
3. Conformal prediction beyond exchangeability

Beyond exchangeability

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

$\Rightarrow$ when $\widehat{\mu}_{n+1}$ is fitted to training + test data,
(X_{n+1}, Y_{n+1}) is equally likely to be any of the $n+1$ data points:



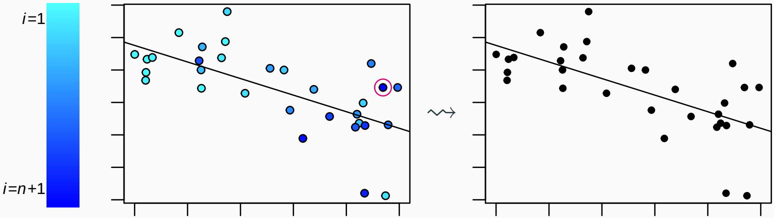
Beyond exchangeability

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, ...)

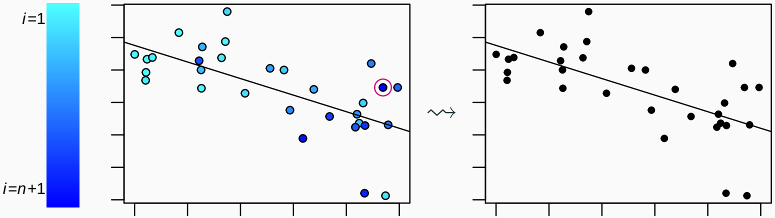
Beyond exchangeability

Example: distribution drift

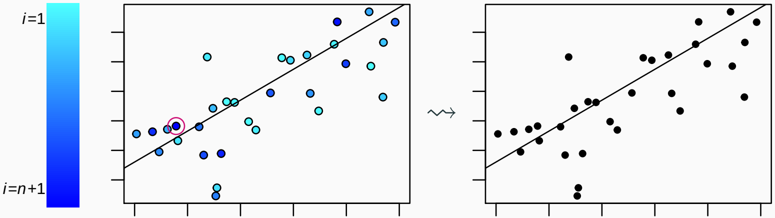


Beyond exchangeability

Example: distribution drift



Example: $\mathcal{A}$ is weighted least sq.



Beyond exchangeability

Our aims:

- Allow for nonsymmetric algorithms (for a more accurate model)
- Guarantee exact coverage if data is exchangeable,
  & bounded loss of coverage under bounded violation of exch.

Nonexchangeable conformal prediction (nexCP)

nexCP method (symmetric algorithm case)

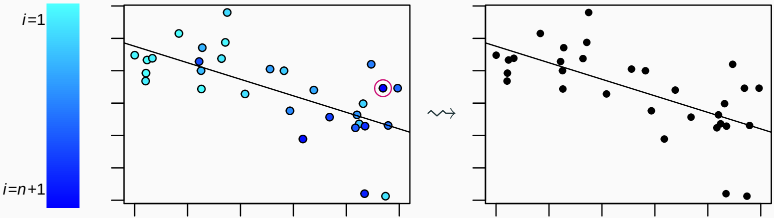
- Fit model to training + test data
$$\widehat{\mu}_{n+1} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$R_i = |Y_i - \widehat{\mu}_{n+1}(X_i)| \text{ for } i \leq n; \;\; R_{n+1} = |y - \widehat{\mu}_{n+1}(X_{n+1})|$$
- Check if $R_{n+1} \leq \big(\text{the } (1-\alpha)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

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

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

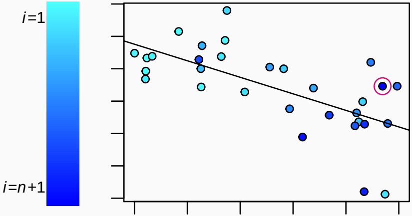
Nonexchangeable conformal prediction (nexCP)

Example: distribution drift

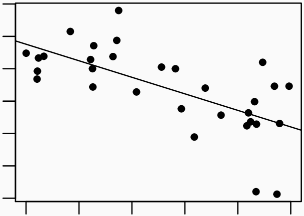


Nonexchangeable conformal prediction (nexCP)

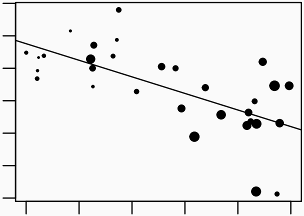
Example: distribution drift



$\rightsquigarrow$



Use weights $w_1 \leq \ldots \leq w_n \leq w_{n+1}$ $\rightsquigarrow$



Nonexchangeable conformal prediction (nexCP)

nexCP method (nonsymmetric algorithm case)

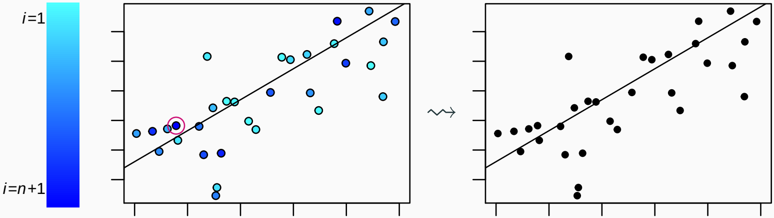
Draw a random index K with $\mathbb{P}\{K = i\} = w_i$, then:

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

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

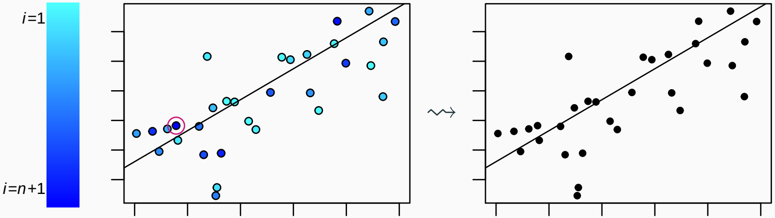
Nonexchangeable conformal prediction (nexCP)

Example: $\mathcal{A}$ is weighted least sq.

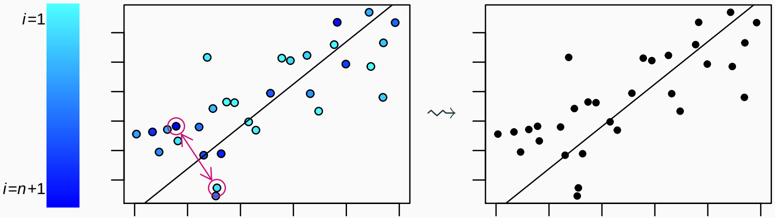


Nonexchangeable conformal prediction (nexCP)

Example: $\mathcal{A}$ is weighted least sq.



Apply $\mathcal{A}$ after swapping data points K & $n+1$



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+

Theoretical guarantee

Theorem: [B., Candès, Ramdas, Tibshirani]

Let $w_i \geq 0$ be fixed, with

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

Then nonexchangeable conformal prediction satisfies

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \right\} \geq$$
$$1 - \alpha - \textstyle\sum_i w_i \cdot \mathrm{d}_{\mathsf{TV}}\big(R(\text{data}), R(\text{data}_{\text{swap}(i)})\big)$$

$R(\text{data}) = (|Y_1 - \widehat{\mu}_{n+1}(X_1)|, \ldots, |Y_{n+1} - \widehat{\mu}_{n+1}(X_{n+1})|)$ $\quad$ $\text{data}_{\text{swap}(i)}$ = swap points i & $n+1$

$\Rightarrow$ If data is i.i.d. or exchangeable, coverage $\geq 1 - \alpha$

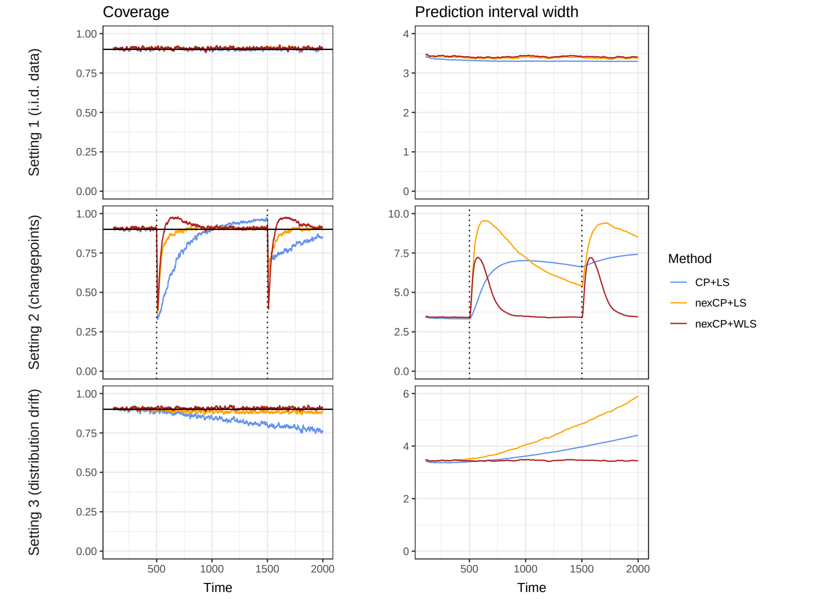
Empirical results

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



Summary

Under exchangeability:

- Jackknife+ allows for a compromise between split and full conformal
  for tradeoff of computation & accuracy
- Coverage guarantee is $\geq 1 - 2\alpha$ (but $\gtrapprox 1 - \alpha$ with stability)

Summary

Under exchangeability:

- Jackknife+ allows for a compromise between split and full conformal for tradeoff of computation & accuracy
- Coverage guarantee is $\geq 1 - 2\alpha$ (but $\gtrapprox 1 - \alpha$ with stability)

Beyond exchangeability:

- Robust to violations of exchangeability
- Swap trick allows for a nonsymmetric algorithm

$\Rightarrow$ can apply CP to nonstationary data / models with drift / etc

Summary

Thank you!

Papers:

B., Candès, Ramdas, Tibshirani, *Predictive inference with the jackknife+*

B., Candès, Ramdas, Tibshirani, *Conformal prediction beyond exchangeability*

Website:

`http://rinafb.github.io/`