An introduction to conformal prediction and distribution-free inference

(Lecture 3)

Rina Foygel Barber (University of Chicago)
Columbia Statistics short course May/June 2024

http://rinafb.github.io/

Topics — Lecture 3

Cross-validation based methods

- Challenges for distribution-free CV
- Jackknife & jackknife+
- Theory for jackknife+
- CV+ and cross-conformal
- Theory for cross-conformal (and CV+)
- Connection to algorithmic stability

Summary

- Summary & preview

Cross-validation based methods

Challenges for distribution-free CV

Returning to cross-validation

Summarizing different methods:

- Split CP fits $\widehat{\mu}$ to part of the data $\rightsquigarrow$ distrib.-free theory
- Full CP: use all the data for $\widehat{\mu}$ *and* achieves distrib.-free theory, but computationally very expensive
- Can cross-validation based methods offer a compromise?

Returning to cross-validation

Why does distribution-free theory hold for split CP but not for CV?

$$\mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q} \;\rightsquigarrow \text{coverage if } \underbrace{|Y_{n+1} - \widehat{\mu}(X_{n+1})|}_{=R_{n+1}} \leq \widehat{q}$$

- For split conformal, $\widehat{q}$ is quantile of calibration residuals

  $$R_i = |Y_i - \widehat{\mu}(X_i)|, \; i = n_0 + 1, \ldots, n$$

  and $\widehat{\mu}$ is pretrained $\Rightarrow R_{n_0+1}, \ldots, R_n, R_{n+1}$ are exchangeable
- For CV, $\widehat{q}$ is quantile of leave-one-out residuals

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

  $\Rightarrow R_1, \ldots, R_n, R_{n+1}$ are *not* exchangeable

Cross-validation based methods

Jackknife & jackknife+

Jackknife & jackknife+

Recall leave-one-out CV (also known as *jackknife*):

$$\mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Quantile}_{1-\alpha}\big(R_i\big)$$

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

- In practice, generally we see $\approx 1-\alpha$ coverage
- In theory, coverage may be zero!

  e.g., $\mathcal{A}$ runs least squares for n even, or neural net for n odd

Jackknife & jackknife+

Jackknife can equivalently be defined as:

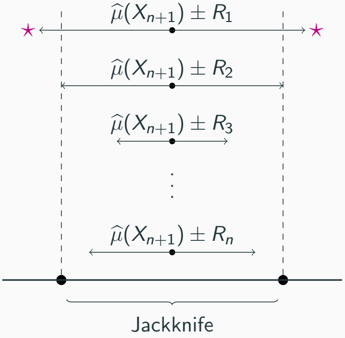
$$\mathcal{C}(X_{n+1}) = \Big[\cancel{\text{Quantile}_\alpha(\widehat{\mu}(X_{n+1}) + R_i)}, \text{Quantile}_{1-\alpha}(\widehat{\mu}(X_{n+1}) + R_i)\Big]$$
$$-\text{Quantile}_{1-\alpha}(-\widehat{\mu}(X_{n+1}) + R_i)$$

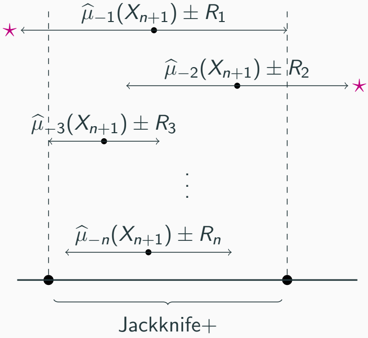
A modified version of the method: the jackknife+[1] .

$$\mathcal{C}(X_{n+1}) = \Big[-\text{Quantile}_{(1-\alpha)(1+1/n)}\big(-\widehat{\mu}_{-i}(X_{n+1}) + R_i\big), \text{Quantile}_{(1-\alpha)(1+1/n)}\big(\widehat{\mu}_{-i}(X_{n+1}) + R_i\big)\Big]$$

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

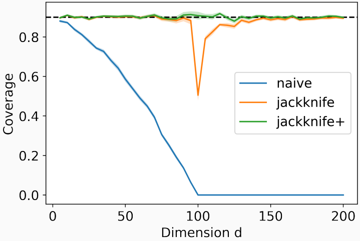
Jackknife & jackknife+

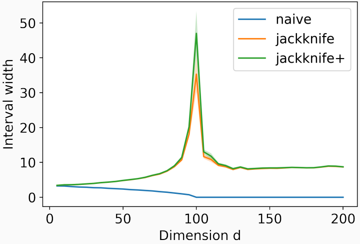




Jackknife & jackknife+

Empirical comparison (linear regression with $n = 100$):





- "Ridgeless" regression — minimum-ℓ_2-norm solution, if $d > n$
- Note: ridgeless regression is stable except the $d \approx n$ regime[2]

[2] Hastie et al 2022, *Surprises in High-Dimensional Ridgeless Least Squares Interpolation*

Cross-validation based methods

Theory for jackknife+

Jackknife+ coverage guarantee

Theorem: coverage for jackknife+

If $Z_1, \ldots, Z_{n+1}$ are exchangeable, and $\mathcal{A}$ is symmetric, then jackknife+ satisfies

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - 2\alpha$$

(In contrast, jackknife may have zero coverage, in the worst case)

Jackknife+ coverage guarantee — proof

Step 1: a matrix of residuals

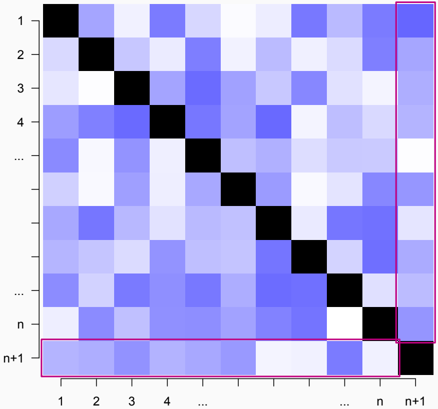
For each $i \neq j \in \{1, \ldots, n+1\}$ define

$$\widetilde{\mu}_{-ij} = \mathcal{A}\big(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_{n+1}\big)$$

and define residuals, $\widetilde{R}_{ij} = |Y_i - \widetilde{\mu}_{-ij}(X_i)|$

Jackknife+ coverage guarantee — proof

Matrix of residuals $\widetilde{R} \in \mathbb{R}^{(n+1)\times(n+1)}$



$\widetilde{R}_{i,n+1} = R_i$

$\widetilde{R}_{n+1,i} = |Y_{n+1} - \widehat{\mu}_{-i}(X_{n+1})|$

Jackknife+ coverage guarantee — proof

Exchangeability of $\widetilde{R} \in \mathbb{R}^{(n+1)\times(n+1)}$:

- The $n(n+1)$ nondiagonal entries of $\widetilde{R}$ are *not* exchangeable:
  e.g., expect higher correlation for $\widetilde{R}_{1,2}$&$\widetilde{R}_{1,3}$ than $\widetilde{R}_{1,2}$&$\widetilde{R}_{3,4}$
- However, $\widetilde{R}$ satisfies $\widetilde{R} \stackrel{d}{=} \Pi\widetilde{R}\Pi^\top$ for any permutation matrix Π

Jackknife+ coverage guarantee — proof

Proof that $\widetilde{R} \overset{d}{=} \Pi\widetilde{R}\Pi^\top$

Let σ be the permutation corresponding to Π.
Let $f : (\mathcal{X} \times \mathcal{Y})^{n+1} \to \mathbb{R}^{(n+1)\times(n+1)}$ map data to $\widetilde{R}$:

$$\big(f(Z_1, \ldots, Z_{n+1})\big)_{ij} = \big|Y_i - \big[\mathcal{A}\big((Z_k)_{k\neq i,j}\big)\big](X_i)\big|$$

Then

$$\widetilde{R} = f(Z_1, \ldots, Z_{n+1}) \overset{d}{=} f(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) = \Pi\widetilde{R}\Pi^\top.$$

exchangeability of the data

symmetry of $\mathcal{A}$

Jackknife+ coverage guarantee — proof

Step 2: relate $\widetilde{R}$ to the jackknife+

$Y_{n+1} \notin \mathcal{C}(X_{n+1})$, i.e., jackknife+ fails to cover $\Longleftrightarrow$

either $\underbrace{Y_{n+1} > \text{Quantile}_{(1-\alpha)(1+1/n)}\big(\widehat{\mu}_{-i}(X_{n+1}) + R_i\big)}$

or $\underbrace{Y_{n+1} < -\text{Quantile}_{(1-\alpha)(1+1/n)}\big(-\widehat{\mu}_{-i}(X_{n+1}) + R_i\big)}$

$$\Leftrightarrow \sum_{i=1}^n \underbrace{\mathbb{1}\left\{Y_{n+1} > \widehat{\mu}_{-i}(X_{n+1}) + R_i\right\}} \geq n \cdot (1-\alpha)(1+1/n)$$

$$= \mathbb{1}\left\{Y_{n+1} - \widehat{\mu}_{-i}(X_{n+1}) > \widetilde{R}_{i,n+1}\right\}$$

$$\leq \mathbb{1}\left\{\widetilde{R}_{n+1,i} > \widetilde{R}_{i,n+1}\right\}$$

same calculation

Jackknife+ coverage guarantee — proof

Summary of step 2:

$$Y_{n+1} \notin \mathcal{C}(X_{n+1}) \implies \sum_{i=1}^{n} \mathbb{1}\left\{\widetilde{R}_{n+1,i} > \widetilde{R}_{i,n+1}\right\} \geq (1-\alpha)(n+1)$$

Combining step 1 & step 2:

we need to show that exchangeability of $\widetilde{R}$ implies

$$\mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{1}\left\{\widetilde{R}_{n+1,i} > \widetilde{R}_{i,n+1}\right\} \geq (1-\alpha)(n+1)\right\} \leq 2\alpha$$

Jackknife+ coverage guarantee — proof

Step 3: the tournament matrix

Consider a tournament played between $n+1$ teams:

$$A_{ij} = \mathbb{1}\left\{\widetilde{R}_{ij} > \widetilde{R}_{ji}\right\} \text{ for } i \neq j$$

(and $A_{ii} = 0$)

- $A \stackrel{d}{=} \Pi A \Pi^\top$ for perm. matrix Π (inherits this property from $\widetilde{R}$)
- $Y_{n+1} \notin \mathcal{C}(X_{n+1}) \implies \underbrace{\sum_i A_{n+1,i}}_{\substack{\#\text{ games won}\\ \text{by team } n+1}} \geq (1-\alpha)(n+1)$

Jackknife+ coverage guarantee — proof

$$\mathbb{P}\left\{Y_{n+1} \notin \mathcal{C}(X_{n+1})\right\} \leq \mathbb{P}\left\{\sum_{i=1}^{n+1} A_{n+1,i} \geq (1-\alpha)(n+1)\right\}$$

by exchangeability

$$= \frac{1}{n+1}\sum_{j=1}^{n+1} \mathbb{P}\left\{\sum_{i=1}^{n+1} A_{ji} \geq (1-\alpha)(n+1)\right\}$$

$$= \mathbb{E}\left[\frac{1}{n+1}\underbrace{\sum_{j=1}^{n+1} \mathbb{1}\left\{\sum_{i=1}^{n+1} A_{ji} \geq (1-\alpha)(n+1)\right\}}_{= \text{ \# of teams that win } \geq (1-\alpha)(n+1) \text{ games}}\right]$$

Jackknife+ coverage guarantee — proof

Step 4: a deterministic bound

How many teams can win $\geq (1-\alpha)(n+1)$ many games?

Let $S = \{j : \sum_i A_{ji} \geq (1-\alpha)(n+1)\}$

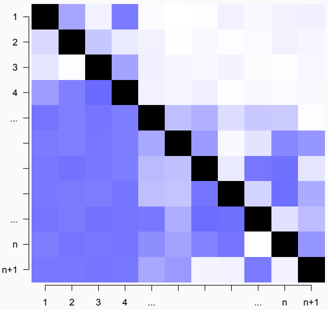
$$
\begin{aligned}
|S| \cdot (1-\alpha)(n+1) &\leq \sum_{j \in S} \sum_{i=1}^{n+1} A_{ji} \\
&= \sum_{j \in S} \sum_{i \in S} A_{ji} + \sum_{j \in S} \sum_{i \in S^c} A_{ji} \\
&= \underbrace{\frac{1}{2} \cdot \sum_{j \in S} \sum_{i \in S} \underbrace{(A_{ji} + A_{ij})}_{\leq 1}}_{\leq \frac{1}{2}|S|^2} + \underbrace{\sum_{j \in S} \sum_{i \in S^c} A_{ji}}_{\leq |S|(n+1-|S|)}
\end{aligned}
$$

$\implies |S| \leq 2\alpha(n+1)$

The worst case scenario, to achieve $|S| \approx 2\alpha(n+1)$:

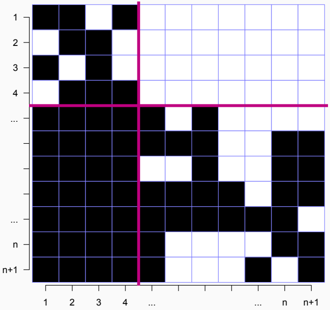
- $\approx 2\alpha(n+1)$ "good" teams, $\approx (1-2\alpha)(n+1)$ "bad" teams
- A "good" team always beats a "bad" team
- All other games are random

Residual matrix $\widetilde{R}$



$\rightsquigarrow$

Tournament matrix A



Jackknife+: intuition for the factor of 2

This worst case scenario is connected to *instability*—
$\widehat{\mu} = \mathcal{A}(Z_1, \ldots, Z_n)$ is very sensitive to changing a single Z_i

Among "good" teams $i \neq j \neq k$:

- May have $\underbrace{\widetilde{R}_{ij} > \widetilde{R}_{ji}}_{\text{"}Y_i \text{ is harder to predict than } Y_j\text{"}}$ and $\underbrace{\widetilde{R}_{jk} > \widetilde{R}_{kj}}_{\text{"}Y_j \text{ is harder to predict than } Y_k\text{"}}$ and $\underbrace{\widetilde{R}_{ki} > \widetilde{R}_{ik}}_{\text{"}Y_k \text{ is harder to predict than } Y_i\text{"}}$

We will return to this later.

Cross-validation based methods

CV+ and cross-conformal

From leave-one-out to K-fold

To avoid computational cost of leave-one-out CV —
K-fold CV (e.g., $K = 5$ or $K = 10$)

- Partition $\{1, \dots, n\}$ into K folds $A_1 \cup \dots \cup A_K$
- Fit model $\widehat{\mu}_{-A_k} = \mathcal{A}\Big(\{(X_i, Y_i) : i \in \{1, \dots, n\} \backslash A_k\}\Big)$
- For $i \in A_k$ define $R_i = |Y_i - \widehat{\mu}_{-A_k}(X_i)|$

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

From leave-one-out to K-fold

Generalize jackknife+ to the K-fold setting $\rightsquigarrow$ CV+

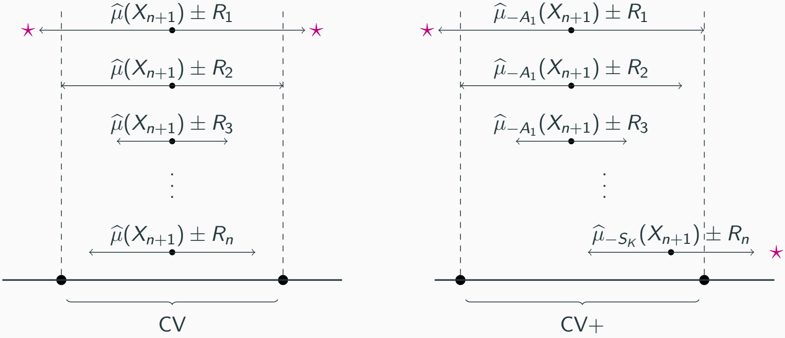
K-fold CV+[3]

- Partition $\{1, \ldots, n\}$ into K folds $A_1 \cup \cdots \cup A_K$
- Fit model $\widehat{\mu}_{-A_k} = \mathcal{A}\Big(\{(X_i, Y_i) : i \in \{1, \ldots, n\} \backslash A_k\}\Big)$
- For $i \in A_k$ define $R_i = |Y_i - \widehat{\mu}_{-A_k}(X_i)|$
- Prediction set

$$\mathcal{C}(X_{n+1}) = \Big[-\text{Quantile}_{(1-\alpha)(1+1/n)}\big(\{-\widehat{\mu}_{-A_k}(X_{n+1}) + R_i\}\big), \\ \text{Quantile}_{(1-\alpha)(1+1/n)}\big(\{\widehat{\mu}_{-A_k}(X_{n+1}) + R_i\}\big)\Big]$$

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

From leave-one-out to K-fold



Cross-conformal prediction

CV+ is related to a more general method:

Cross-conformal prediction[4,5]

- Partition $\{1, \dots, n\}$ into K folds $A_1 \cup \dots \cup A_K$
- Fit score function $s^{(k)} = \mathcal{A}\Big(\{(X_i, Y_i) : i \in \{1, \dots, n\} \backslash A_k\}\Big)$
- For $i \in A_k$ define $S_i = s^{(k)}(X_i, Y_i)$
- Prediction set

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \sum_{k=1}^{K} \sum_{i \in A_k} \mathbb{1}\left\{ S_i \geq s^{(k)}(X_{n+1}, y) \right\} \geq \alpha(n+1) \right\}$$

[4] Vovk 2015, *Cross-conformal predictors*

[5] Vovk et al 2018, *Cross-conformal predictive distributions*

Cross-conformal prediction

For the residual score function $s(x, y) = |y - \widehat{\mu}(x)|$,

$$\mathcal{C}_{\text{cross-conf.}}(X_{n+1}) \subseteq \mathcal{C}_{\text{CV+}}(X_{n+1})$$

Proof (for the case $K = n$, i.e., jackknife+)

Recall from the proof of the jackknife+ theorem:

$$Y_{n+1} \notin \mathcal{C}_{\text{j+}}(X_{n+1}) \;\Rightarrow\; \sum_{i=1}^{n} \mathbb{1}\left\{\widetilde{R}_{n+1,i} > \widetilde{R}_{i,n+1}\right\} \geq (1-\alpha)(n+1)$$

Rewrite as:

$$Y_{n+1} \in \mathcal{C}_{\text{j+}}(X_{n+1}) \;\Leftarrow\; \underbrace{\sum_{i=1}^{n} \mathbb{1}\left\{\widetilde{R}_{n+1,i} > \widetilde{R}_{i,n+1}\right\}}_{\substack{= \sum_{i=1}^{n} \mathbb{1}\left\{s^{(i)}(X_{n+1}, Y_{n+1}) > S_i\right\} \\ = n - \underbrace{\sum_{i=1}^{n} \mathbb{1}\left\{S_i \geq s^{(i)}(X_{n+1}, Y_{n+1})\right\}}_{\geq \alpha(n+1) \;\Leftrightarrow\; Y_{n+1} \in \mathcal{C}_{\text{cross-conf.}}(X_{n+1})}}} \leq (1-\alpha)(n+1) - 1$$

Cross-validation based methods

Theory for cross-conformal (and CV+)

Theory for cross-conformal

Theorem: coverage for CV+ and cross-conformal

If $Z_1, \ldots, Z_{n+1}$ are i.i.d., and $\mathcal{A}$ is symmetric, then K-fold cross-conformal satisfies

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq \begin{cases} 1 - 2\alpha - 2/K & {}^{6} \\ 1 - 2\alpha - 2K/n & {}^{7} \end{cases}$$

As a special case, the same is true for K-fold CV+.

For any K, then,

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - 2\alpha - \frac{2}{\sqrt{n}}.$$

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

[7] Vovk et al 2018, *Cross-conformal predictive distributions*

Theory for cross-conformal

Theorem part 1—proved via tournament matrix

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

Imagine that we sample additional test data:

$$\underbrace{(X_1, Y_1), \ldots, (X_n, Y_n)}_{K \text{ training folds } A_1, \ldots, A_K}, \underbrace{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+n/K}, Y_{n+n/K})}_{1 \text{ test fold } A_{K+1}} \stackrel{\text{iid}}{\sim} P$$

For each $k \neq k' \in \{1, \ldots, K+1\}$ define

$$\widetilde{s}_{-kk'} = \mathcal{A}\left((Z_i)_{i \in [n+n/K] \setminus (A_k \cup A_{k'})}\right)$$

and define scores, $\widetilde{S}_{ij} = \widetilde{s}_{-kk'}(X_i, Y_i)$ for $i \in A_k$, $j \in A_{k'}$

Theory for cross-conformal

For a permutation σ on $[n + n/K]$, we say that σ preserves folds if

$$i, j \text{ in same fold } \iff \sigma(i), \sigma(j) \text{ in same fold}$$

If σ preserves folds, and Π is the corresponding perm. matrix,

$$\widetilde{S} \stackrel{\mathrm{d}}{=} \Pi \widetilde{S} \Pi^\top$$

Define the tournament matrix $A \in \{0, 1\}^{(n+n/K)\times(n+n/K)}$:

$$A_{ij} = \mathbb{1}\left\{\widetilde{S}_{ij} > \widetilde{S}_{ji}\right\} \text{ (or } A_{ij} = 0 \text{ if } i, j \text{ in same fold)}$$

$\Rightarrow A \stackrel{\mathrm{d}}{=} \Pi A \Pi^\top$ if Π preserves folds

Theory for cross-conformal

Verify via construction of $\mathcal{C}(X_{n+1})$:

$$Y_{n+1} \notin \mathcal{C}(X_{n+1}) \iff \sum_{i=1}^{n+n/K} A_{n+1,i} \geq (1-\alpha)(n+1)$$

$$\mathbb{P}\{Y_{n+1} \notin \mathcal{C}(X_{n+1})\} = \mathbb{P}\left\{\begin{array}{c}\text{in a tournament with } n+n/K \\ \text{teams, team } n+1 \text{ wins} \\ \geq \underbrace{(1-\alpha)(n+1)}_{=(1-[\alpha+(1-\alpha)\frac{n/K-1}{n+n/K}])(n+n/K)} \text{ games}\end{array}\right\}$$

$$\leq 2\left[\alpha + (1-\alpha)\frac{n/K-1}{n+n/K}\right] \leq 2\alpha + 2/K$$

via counting argument
(as for jackknife+ proof)

Theory for cross-conformal

Theorem part 2—proved via averaging p-values

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

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : 1 + \sum_{k=1}^{K} \sum_{i \in A_k} \mathbb{1}\left\{ S_i \geq s^{(k)}(X_{n+1}, y) \right\} \overset{>}{\cancel{\geq}} \alpha(n+1) \right\}$$

$$= \left\{ y \in \mathcal{Y} : \frac{1 + \sum_{k=1}^{K} \sum_{i \in A_k} \mathbb{1}\left\{ S_i \geq s^{(k)}(X_{n+1}, y) \right\}}{n+1} > \alpha \right\}$$

Theory for cross-conformal

For each fold $k = 1, \ldots, K$ define a p-value

$$p_k^y = \frac{1 + \sum_{i \in A_k} \mathbb{1}\left\{S_i \geq s^{(k)}(X_{n+1}, y)\right\}}{\frac{n}{K} + 1}$$

- As for split CP, $p_k^{Y_{n+1}}$ is a valid p-value (holdout & test scores are exch.)
- Define $\bar{p}^y = \frac{1}{K} \sum_{k=1}^K p_k^y$
  Then $\bar{p}^{Y_{n+1}}$ is a valid p-value up to a factor of 2: [8]

$$\mathbb{P}\left\{\bar{p}^{Y_{n+1}} \leq a\right\} \leq 2a, \text{ for all } a \in [0, 1].$$

[8] Rüschendorf 1982, *Random variables with maximum sums*; Vovk & Wang 2020, *Combining p-values via averaging*

Theory for cross-conformal

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \frac{1 + \sum_{k=1}^{K} \overbrace{\sum_{i \in A_k} \mathbb{1}\left\{S_i \geq s^{(k)}(X_{n+1}, y)\right\}}^{=p_k^y(\frac{n}{K}+1)-1}}{n+1} > \alpha \right\}$$

$$= \left\{ y \in \mathcal{Y} : \underbrace{\frac{1}{K} \sum_{k=1}^{K} p_k^y}_{=\bar{p}^y} > \alpha + \frac{K-1}{n+K}(1-\alpha) \right\}$$

$$\implies \mathbb{P}\left\{Y_{n+1} \notin \mathcal{C}(X_{n+1})\right\} = \mathbb{P}\left\{\bar{p}^{Y_{n+1}} \leq \alpha + \frac{K-1}{n+K}(1-\alpha)\right\}$$

$$\leq 2\left(\alpha + \frac{K-1}{n+K}(1-\alpha)\right) \leq 2\alpha + 2K/n$$

Cross-validation based methods

Connection to algorithmic stability

Jackknife & stability

Why might jackknife/CV fail to cover?
Why might jackknife+/CV+ have coverage $1 - 2\alpha$ not $1 - \alpha$?

Examples we have seen:

- $\mathcal{A} =$ least squares for even n, neural net for odd n
- Least squares with $n \approx d$

These examples are *unstable*—
$\widehat{\mu} = \mathcal{A}(Z_1, \ldots, Z_n)$ is very sensitive to changing a single Z_i

Jackknife & stability

Algorithmic stability[9]

Algorithm $\mathcal{A}$ is (ϵ, ν)-stable if

$$\mathbb{P}\left\{|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})| \leq \epsilon\right\} \geq 1 - \nu$$

Note that this property depends on both $\mathcal{A}$ and distrib. of data

- Example: ridge regression is stable (due to strong convexity)
- Example: K-nearest-neighbors is stable (if $K \ll n$)
- Next week: stability of bootstrap / ensembled methods

[9] Bousquet & Elisseeff 2002, *Stability and Generalization*

Jackknife & stability

Define the inflated jackknife interval:

$$\mathcal{C}^{\epsilon}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \big(\mathrm{Quantile}_{1-\alpha}(R_i) + \epsilon\big)$$

Theorem: jackknife under stability[10]

If $\mathcal{A}$ satisfies (ϵ, ν)-stability, then

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}^{\epsilon}(X_{n+1})\right\} \geq 1 - \alpha - 2\sqrt{\nu}$$

(similar results hold for jackknife+)

[10] B., Candès, Ramdas, Tibshirani 2019, *Predictive inference with the jackknife+*; see also Steinberger & Leeb, *Conditional predictive inference for stable algorithms*

Jackknife & stability

A related method:

The conformal jackknife

Consider a leave-one-out version of full conformal:

$$\widetilde{\mathcal{A}}: \text{ data } (Z_1, \ldots, Z_{n+1}) \mapsto \text{ scores } (S_1, \ldots, S_{n+1})$$

where

$$S_i = |Y_i - [\mathcal{A}(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_{n+1})](X_i)|$$

If $\mathcal{A}$ overfits, this performs better than full CP with residual score

Validity:

- Symmetry of $\mathcal{A} \Rightarrow$ symmetry of $\widetilde{\mathcal{A}}$
- $\rightsquigarrow$ if data is exch., then $\mathbb{P}\{Y_{n+1} \in \mathcal{C}_{\text{conf. jack.}}(X_{n+1})\} \geq 1 - \alpha$

Proof for jackknife under stability:

Key idea: relate $\mathcal{C}^{\epsilon}_{\text{jack}}(X_{n+1})$, to conformal jackknife run at $\alpha' > \alpha$

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}^{\epsilon}_{\text{jack}}(X_{n+1})\right\} \geq \overbrace{\mathbb{P}\left\{Y_{n+1} \notin \mathcal{C}^{\alpha'}_{\text{jack-CP}}(X_{n+1})\right\}}^{\geq 1-\alpha'} - \mathbb{P}\left\{\mathcal{C}^{\epsilon}_{\text{jack}}(X_{n+1}) \not\supseteq \mathcal{C}^{\alpha'}_{\text{jack-CP}}(X_{n+1})\right\}$$

Jackknife & stability

Define models & residuals:

$$\widehat{\mu}_{-i} = \mathcal{A}(Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n) \quad \rightsquigarrow R_i = |Y_i - \widehat{\mu}_{-i}(X_i)|$$
$$\widetilde{\mu}_{-i} = \mathcal{A}(Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n, Z_{n+1}) \quad \rightsquigarrow S_i = |Y_i - \widetilde{\mu}_{-i}(X_i)|$$

The intervals:

$$\mathcal{C}^{\epsilon}_{\text{jack}}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \text{Quantile}_{1-\alpha}(R_i + \epsilon)$$
$$\mathcal{C}^{\alpha'}_{\text{jack-CP}}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \text{Quantile}_{(1-\alpha')(1+1/n)}(S_i)$$

$$\Rightarrow \mathcal{C}^{\epsilon}_{\text{jack}}(X_{n+1}) \supseteq \mathcal{C}^{\alpha'}_{\text{jack-CP}}(X_{n+1}) \text{ if } \underbrace{R_i + \epsilon \geq S_i \text{ for sufficiently many } i}_{\text{holds w.h.p. under stability assumption}}$$

Summary

Summary & preview

Summary: lectures 1, 2, 3

Lecture 1–3 topics:

- The exchangeability framework
- The distribution-free prediction problem
- Versions of the methodology:
  split CP, full CP, and CV-type conformal methods

Conformal methods allow for distribution-free predictive inference
using any base model, & assuming only exchangeability

Preview: lectures 4, 5, 6

All methods thus far have focused on
predictive inference with marginal coverage:

$$\underbrace{\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1})\right\}}_{\text{the goal is always coverage}} \quad \geq \quad \underbrace{1-\alpha}_{\substack{\text{the guarantee is always}\\ \text{averaged over the distrib.}\\ \text{of training + test data}}}$$

- Lecture 4: beyond predictive coverage (& more next week)
- Lectures 5 & 6: beyond the marginal coverage guarantee