An introduction to conformal prediction and distribution-free inference

(Lecture 4)

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

http://rinafb.github.io/

Topics — Lecture 4

Extensions beyond predictive coverage

- Motivation
- Beyond the zero/one loss: conformal risk control
- Beyond a single timepoint—the sequential setting
- Testing for outliers
- Conformal prediction with selective coverage
- Other targets?

Summary

- Summary & preview

Extensions beyond predictive coverage

Motivation

Motivation

All methods thus far have focused on

predictive inference with marginal coverage:

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

- This lecture: beyond predictive coverage—what other goals might we consider?

Extensions beyond predictive coverage

Beyond the zero/one loss: conformal risk control

Generalizing CP to other definitions of risk

CP methods bound $\mathbb{P}\{Y_{n+1} \notin \mathcal{C}(X_{n+1})\} = \underbrace{\mathbb{E}\left[\mathbb{1}_{Y_{n+1} \notin \mathcal{C}(X_{n+1})}\right]}_{\text{zero/one loss}}$

This may not be a good target for some settings...

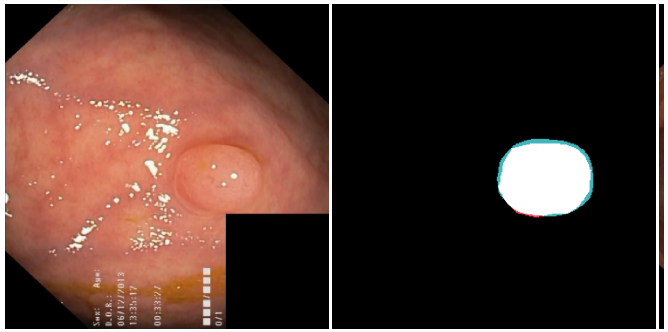
this may be too conservative

or may not capture a natural notion of risk

Generalizing CP to other definitions of risk: examples

Image segmentation:

Data: image $\in \mathbb{R}^{d_1 \times d_2}$. Goal: classify pixels as $+$ or $-$



(figure from Bates et al 2021)

- Coverage $Y \in \mathcal{C}(X) \leftrightarrow$ *every* pixel is correctly classified
- Better to use, e.g., false positive & false negative rates

Generalizing CP to other definitions of risk: examples

Image labeling:

Data: image $\in \mathbb{R}^{d_1 \times d_2}$. Goal: identify objects in the image

(figure from Bates et al 2021)

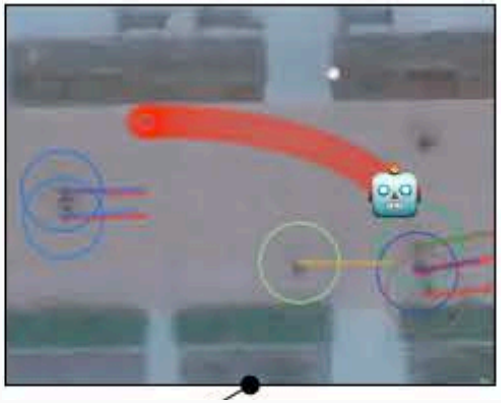
- Coverage $Y \in \mathcal{C}(X) \leftrightarrow$ the entire list is correctly identified
- Better to use, e.g., a measure of overlap of the lists

Generalizing CP to other definitions of risk: examples

Motion planning:

Data: a person's trajectory $\in \mathbb{R}^{d_1 \times d_2 \times T}$.

Goal: plan robot's trajectory to avoid collisions.



(figure from Lekeufack et al 2023)

- Coverage $Y \in \mathcal{C}(X) \leftrightarrow$ the entire trajectory is covered
- Better to directly control risk of collision

Generalizing CP to other definitions of risk

Idea: use CP-type approach to control any definition of risk[1,2,3,4]

Setting: we need to choose a parameter λ

e.g., λ is a threshold for selection / a margin of error / etc

(larger $\lambda \rightsquigarrow$ more conservative)

Notation:

- $\ell(Z; \lambda) =$ observed loss for data point Z if choose parameter λ
- $R(\lambda) = \mathbb{E}[\ell(Z; \lambda)] =$ expected risk
- Goal: use training data $Z_1, \ldots, Z_n$ to choose $\widehat{\lambda}$ s.t. $R(\widehat{\lambda}) \lessapprox \alpha$

---

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

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

[3] Angelopolous et al 2022, *Conformal Risk Control*

[4] Lekeufack et al 2023, *Conformal Decision Theory: Safe Autonomous Decisions from Imperfect Predictions*

Conformal risk control

Assume a bounded loss: $\ell(Z;\lambda) \leq B$

The conformal risk control algorithm[5]

Define an empirical estimate of risk,

$$\widehat{R}(\lambda) = \frac{B + \sum_{i=1}^{n} \ell(Z_i;\lambda)}{n+1},$$

and then define

$$\widehat{\lambda} = \inf\left\{\lambda : \widehat{R}(\lambda) \leq \alpha\right\}.$$

[5] Angelopolous et al 2022, *Conformal Risk Control*

Conformal risk control

Relate back to predictive inference....

Split CP is a special case (with calibration set $= Z_1, \ldots, Z_n$)

- Parameter λ controls the size of the prediction set
  e.g., $\mathcal{C}_\lambda(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \lambda$ for pretrained $\widehat{\mu}$
- Zero/one loss $\ell((X, Y); \lambda) = \mathbb{1}_{Y \notin \mathcal{C}_\lambda(X)}$
- Risk $=$ marginal miscoverage, $R(\lambda) = \mathbb{P}\{Y \notin \mathcal{C}_\lambda(X)\}$

Conformal risk control

Theorem: distribution-free risk control

Assume $\lambda \mapsto \ell(Z; \lambda)$ is nonincreasing and right-continuous. Then for any P, if $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} P$ then

$$\mathbb{E}_P\left[R(\widehat{\lambda})\right] \leq \alpha.$$

Conformal risk control

Proof:

Let $Z_{n+1} \sim P$, then $\mathbb{E}_P\left[R(\widehat{\lambda})\right] = \mathbb{E}\left[\ell(Z_{n+1}; \widehat{\lambda})\right]$.

Define

$$\lambda_* = \inf\left\{\lambda : \frac{\sum_{i=1}^{n+1} \ell(Z_i; \lambda)}{n+1} \leq \alpha\right\}$$

Compare to

$$\widehat{\lambda} = \inf\left\{\lambda : \underbrace{\frac{B + \sum_{i=1}^{n} \ell(Z_i; \lambda)}{n+1}}_{=\widehat{R}(\lambda)} \leq \alpha\right\}$$

Loss bounded by $B \implies \lambda_* \leq \widehat{\lambda}$

Conformal risk control

$$\mathbb{E}\left[\ell(Z_{n+1}; \widehat{\lambda})\right] \leq \mathbb{E}\left[\ell(Z_{n+1}; \lambda_*)\right] = \mathbb{E}\left[\frac{1}{n+1}\sum_{i=1}^{n+1}\ell(Z_i; \lambda_*)\right] \leq \alpha.$$

since loss is monotone

since $\ell(Z_1; \lambda_*), \ldots, \ell(Z_{n+1}; \lambda_*)$ are exchangeable

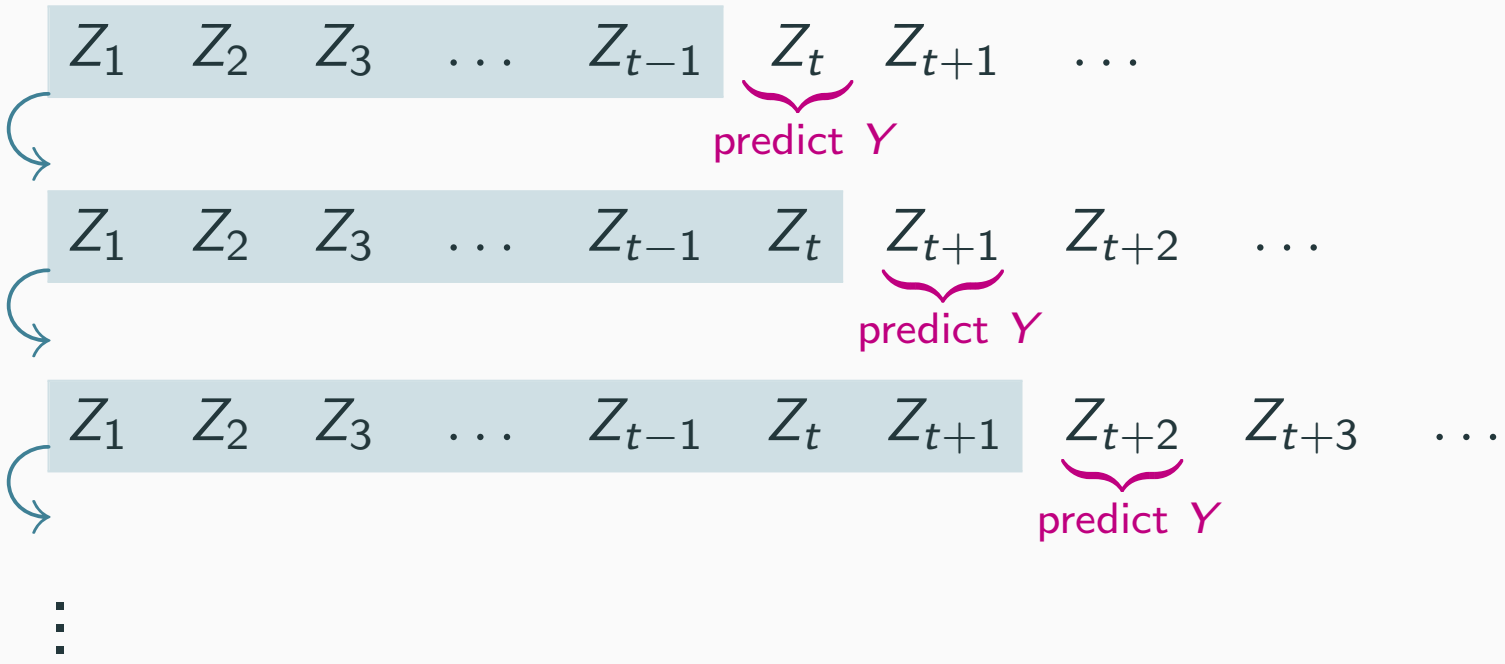
by def. of λ_*

Extensions beyond predictive coverage

Beyond a single timepoint—the sequential setting

Prediction in a streaming setting

In practice, we want to predict "in real time":



Prediction in a streaming setting

$\mathcal{C}_t =$ the prediction interval constructed for prediction at time t

Conformal method guarantee for each t:

$$\mathbb{P}\{Y_t \in \mathcal{C}_t(X_t)\} \geq 1 - \alpha$$

Is this sufficient for practical purposes?

A high-dependence scenario...
The distribution of the data stream is such that

- with probability $1 - \alpha$, $\quad Y_t \in \mathcal{C}_t(X_t)$ for every t
- with probability α, $\quad Y_t \notin \mathcal{C}_t(X_t)$ for every t

Returning to the conformal p-value

Recall:

Full conformal prediction (p-value version)[6]

For each $y \in \mathcal{Y}$ define a conformal p-value:

$$p^y = \frac{1 + \sum_{i=1}^{n} \mathbb{1}\left\{s^y(X_i, Y_i) \geq s^y(X_{n+1}, y)\right\}}{n+1}$$

where

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

$Y_{n+1} \notin \mathcal{C}(X_{n+1}) \;\Leftrightarrow\; p \leq \alpha$, where

$$p = p^{Y_{n+1}} = \frac{1+\sum_{i=1}^{n} \mathbb{1}\{s(X_i,Y_i) \geq s(X_{n+1},Y_{n+1})\}}{n+1} \text{ for } s = \mathcal{A}(Z_1, \ldots, Z_{n+1})$$

[6] Vovk et al 2005, *Algorithmic Learning in a Random World*

Returning to the conformal p-value

Change notation to accommodate streaming data:

- $s_t = \mathcal{A}(Z_1, \ldots, Z_t)$
- $Y_t \notin \mathcal{C}_t(X_t) \iff p_t \leq \alpha$, where

$$p_t = \frac{1 + \sum_{i=1}^{t-1} \mathbb{1}\left\{s_t(X_i, Y_i) \geq s_t(X_t, Y_t)\right\}}{t}$$

A high-dependence scenario...
The distribution of the data stream is such that

- with probability $1 - \alpha$, $p_t > \alpha$ for every t
- with probability α, $p_t \leq \alpha$ for every t

Conformal p-values are independent

Theorem: conformal p-values in streaming time[7]

Assume scores $\{s_t(Z_i)\}_{i=1,\ldots,t}$ are distinct at each t (no ties).

Then:

- For each t, $p_t \sim \text{Uniform}\{\frac{1}{t}, \frac{2}{t}, \ldots, 1\}$
- And, $p_1, p_2, \ldots$ are mutually independent

Implication: $\sum_{i=1}^{t} \mathbb{1}\{p_i \leq \alpha\} \leq \text{Binomial}(t, \alpha)$
(& this also holds if ties allowed)

$\rightsquigarrow$ the high-dependence scenario cannot occur

[7] Vovk et al 2003, *Testing Exchangeability On-Line*

Conformal p-values are independent

Proof that $p_t \sim \text{Unif}\{\frac{1}{t}, \ldots, 1\}$:

Follows from exchangeability of $s_t(Z_1), \ldots, s_t(Z_t)$

Proof that $p_1, p_2, \ldots$ are mutually independent:

Key idea: work in reverse time.

- Sufficient to show $p_1, \ldots, p_T$ mutually indep. for all $T \geq 1$
- $\rightsquigarrow$ Sufficient to show $p_t \perp\!\!\!\perp (p_{t+1}, \ldots, p_T)$ for all $T \geq t \geq 1$

Permute data $\rightsquigarrow Z_{\sigma(1)}, \ldots, Z_{\sigma(t)}, Z_{t+1}, \ldots, Z_T$

- p-values $p_{t+1}, \ldots, p_T$ are *unchanged* for permuted data
- And, p_t is *uniform* on $\{\frac{1}{t}, \ldots, 1\}$ when $\sigma \sim \text{Unif}(\mathcal{S}_t)$

Conformal p-values are independent

More formally....

- Let $p_i(Z) =$ p-value at time i for data $Z = (Z_1, \ldots, Z_T)$
- Let $Z_\sigma = (Z_{\sigma(1)}, \ldots, Z_{\sigma(t)}, Z_{t+1}, \ldots, Z_T)$
- Then $p_i(Z_\sigma) = p_i(Z)$ for $t < i \leq T$

$$\big(p_t(Z), p_{t+1}(Z), \ldots, p_T(Z)\big) \overset{\mathrm{d}}{=} \big(p_t(Z_\sigma), \underbrace{p_{t+1}(Z), \ldots, p_T(Z)}_{=p_{t+1}(Z_\sigma),\ldots,p_T(Z_\sigma)}\big)$$

- Take $\sigma \sim \mathrm{Unif}(\mathcal{S}_t)$
- $\implies p_t(Z_\sigma) \mid \big(p_{t+1}(Z), \ldots, p_T(Z)\big) \sim \mathrm{Unif}\{\frac{1}{t}, \ldots, 1\}$
- $\implies p_t(Z) \mid \big(p_{t+1}(Z), \ldots, p_T(Z)\big) \sim \mathrm{Unif}\{\frac{1}{t}, \ldots, 1\}$

Extensions beyond predictive coverage

Testing for outliers

Conformal outlier detection

Conformal prediction is closely related to testing for outliers:

$$\text{Is } Z_{n+1} \text{ exchangeable with } Z_1, \ldots, Z_n?$$

$$\Updownarrow$$

$$\text{Is } Z_{n+1} \text{ an outlier relative to } Z_1, \ldots, Z_n?$$

Conformal outlier detection

Comparing the problems (via the conformal p-value)

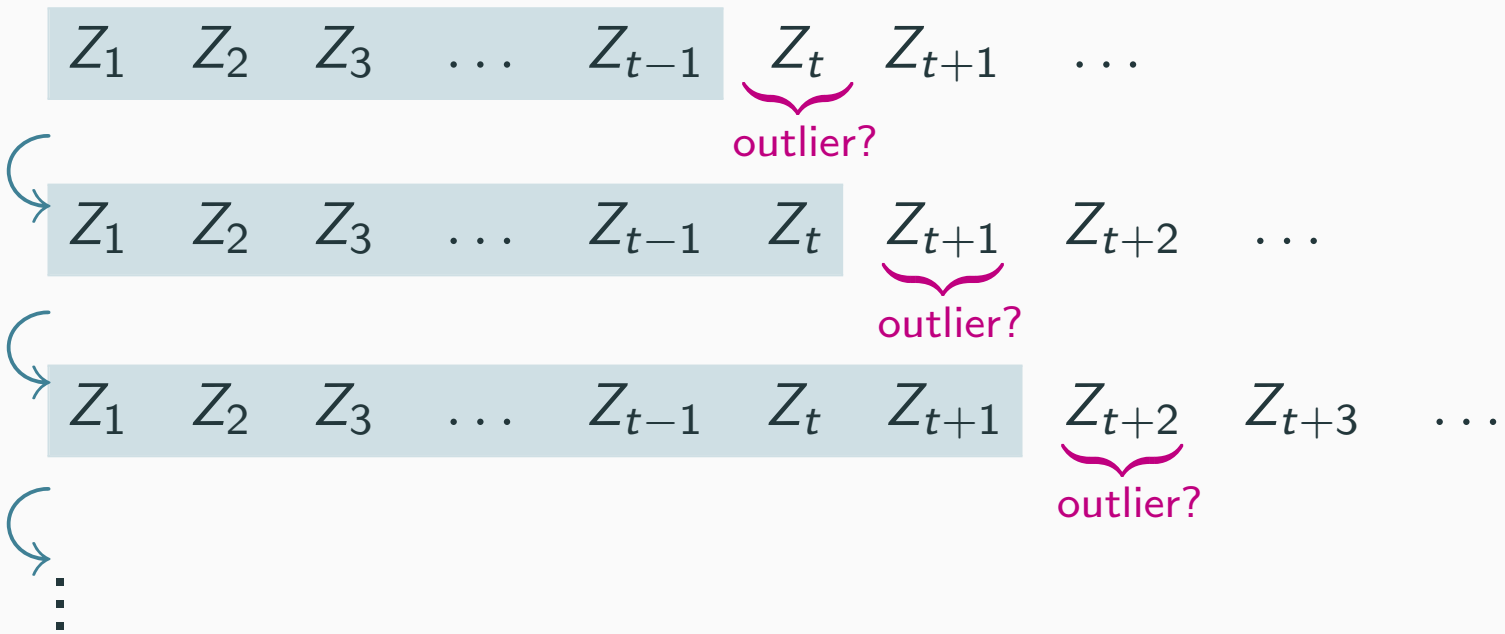
- To run full CP, Y_{n+1} is unobserved $\rightsquigarrow$ calculate p^y for every y
- For outlier detection, Y_{n+1} observed $\rightsquigarrow$ only need $p = p^{Y_{n+1}}$

What is the conformal p-value actually testing?

- H_0: data points $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exch.
- So in theory, might reject due to *any* violation of exch.
- But, p is designed to look for an unusual point (X_{n+1}, Y_{n+1})

Conformal outlier detection

In practice, we test a stream of data for outliers:



Let p_t be the conformal p-value at time t:

$$p_t = \frac{1 + \sum_{i=1}^{t-1} \mathbb{1}_{s_t(Z_i) \geq s_t(Z_t)}}{t} \text{ where } s_t = \mathcal{A}(Z_1, \dots, Z_t)$$

Conformal outlier detection

The challenge with streaming outlier detection:

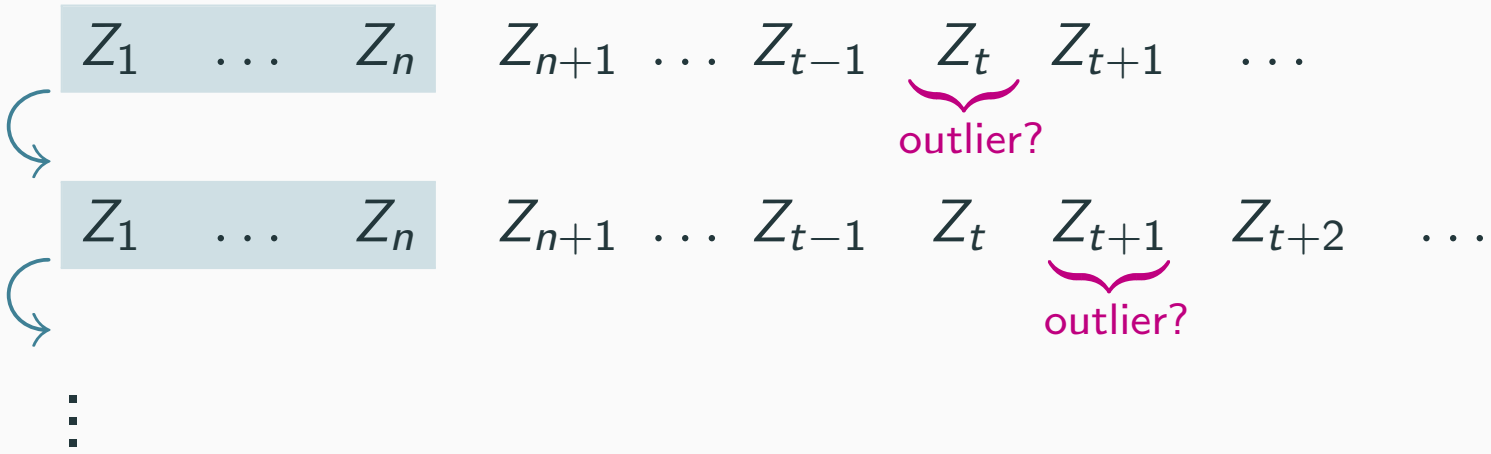
p_t is a p-value for testing if Z_t is an outlier relative to $Z_1, \ldots, Z_{t-1}$
Is this data coming from the "correct" distribution?

- If we use all past data to test $Z_t \rightsquigarrow$ distrib. is contaminated
- If remove detected outliers $\rightsquigarrow$ distrib. is effectively truncated

Conformal outlier detection: "clean data" version

A different setting:

- We have "clean" data $Z_1, \ldots, Z_n$ (the reference set)
- Want to test $Z_{n+1}, Z_{n+2}, \ldots$ against the reference set



$$p_t = \frac{1 + \sum_{i=n_0+1}^{n} \mathbb{1}_{s(Z_i) \geq s(Z_t)}}{n_1 + 1} \text{ where } s = \mathcal{A}(Z_1, \ldots, Z_{n_0})$$

Conformal outlier detection: "clean data" version

Should we label Z_t as an outlier whenever $p_t \leq \alpha$?

- Problem: if outliers are rare $\rightsquigarrow$high false discovery proportion (FDP)

$$\text{FDP} = \frac{\#\text{ data points } \textit{falsely} \text{ labeled as outliers}}{\#\text{ data points labeled as outliers}}$$

- Approach: apply the Benjamini–Hochberg procedure[8]

BH procedure for the outlier detection problem:[9]

- Compute p-value p_t for each $t = n+1, \ldots, n+m$
- Define $\hat{k}$ as the max value such that $\hat{k}$ many p_t's are $\leq \alpha\hat{k}/m$
- Label any t with $p_t \leq \alpha\hat{k}/m$ as an outlier

[8] Benjamini & Hochberg 1995, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*

[9] Bates et al 2021, *Testing for outliers with conformal p-values*

Conformal outlier detection: "clean data" version

BH procedure guarantee:[10]

if the p-values are independent, then $\mathbb{E}[\text{FDP}] \leq \alpha$

For the outlier detection setting:

- Each p_t is a valid p-value (as in split CP)
- However, p_t's are *not* independent

  (dependence due to using the same calibration set)

[10] Benjamini & Hochberg 1995, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*

The PRDS condition

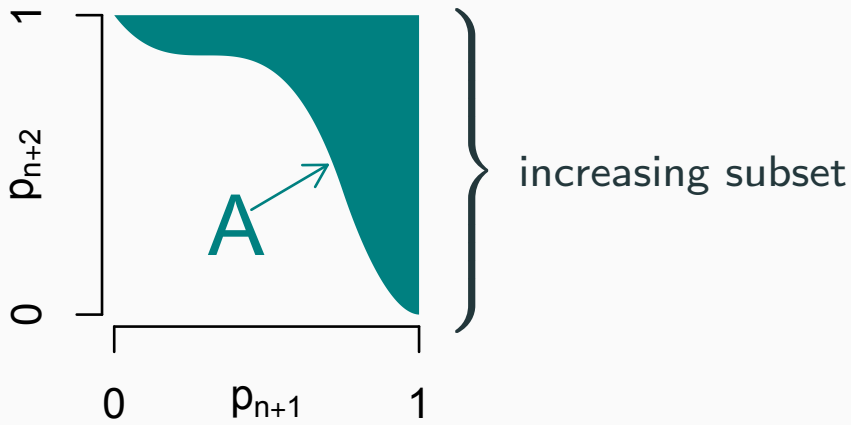
If p-values are not independent, BH also guarantees $\mathbb{E}[\text{FDP}] \leq \alpha$ if the p-values satisfy the PRDS condition:[11]

Positive regression dependence on a subset

For any increasing subset $A \subseteq [0,1]^m$,

$$t \mapsto \mathbb{P}\left\{(p_{n+1}, \ldots, p_{n+m}) \in A \mid p_{n+j} \geq t\right\}$$

is nondecreasing in t.



[11] Benjamini & Yekutieli 2001, *The control of the false discovery rate in multiple testing under dependency*

Conformal outlier detection: "clean data" version

Conformal p-values are PRDS[12]

For conformal outlier detection, if there are no ties among scores
& p-values computed w.r.t. a reference set,
the p_t's satisfy the PRDS condition $\implies \mathbb{E}\,[\text{FDP}] \leq \alpha$ for BH

Intuition—

If $s(Z_{n_0+1}), \ldots, s(Z_n)$ are a bit low (or high) by random chance,
then $p_{n+1}, p_{n+2}, \ldots$ all biased to be lower (higher) than uniform
$\rightsquigarrow$ positive dependence among $p_{n+1}, p_{n+2}, \ldots$

[12] Bates et al 2021, *Testing for outliers with conformal p-values*

Conformal p-values are independent

Proof sketch of PRDS property:

Simple case: $m = 2$ test points

$$\underbrace{s(Z_{n_0+1}), \ldots, s(Z_n)}_{\text{reference set scores}}, \underbrace{s(Z_{n+1}), s(Z_{n+2})}_{\text{test scores}} \quad \rightsquigarrow \quad \underbrace{r_{n_0+1}, \ldots, r_n, r_{n+1}, r_{n+2}}_{\substack{\text{ranks = uniform perm. of } [n_1+2] \\ \text{(largest score } \leftrightarrow \text{ rank 1)}}}$$

$$(p_{n+1}, p_{n+2}) = \left(\frac{r_{n+1} - \mathbb{1}_{r_{n+1} > r_{n+2}}}{n_1 + 1}, \frac{r_{n+2} - \mathbb{1}_{r_{n+1} < r_{n+2}}}{n_1 + 1} \right)$$

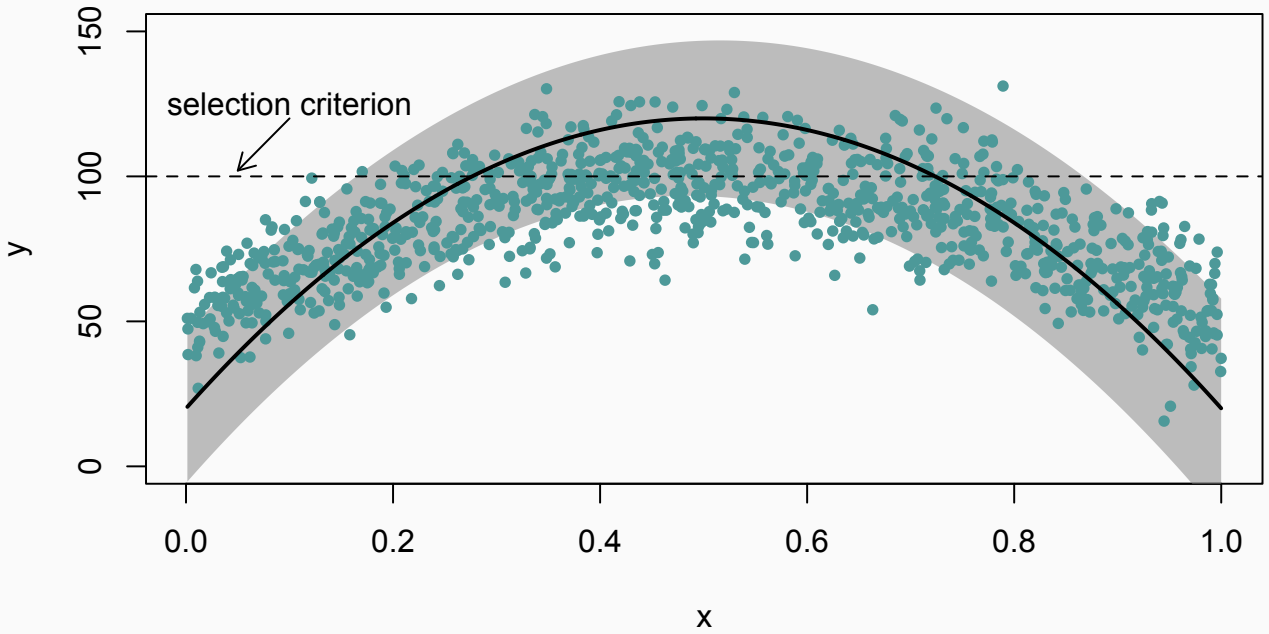
Extensions beyond predictive coverage

Conformal prediction with selective coverage

Prediction after selection

A hypothetical scenario:

- Test points $X_{n+1}, \ldots, X_{n+m}$
- We build a prediction interval for each $Y_{n+1}, \ldots, Y_{n+m}$
- Interested in test points X_{n+j} for which predictions are highest



Prediction after selection

Goal—valid predictive inference conditional on selection:[13,14,15,16]

$$\mathbb{P}\left\{ Y_{n+1} \in \mathcal{C}(X_{n+1}) \;\middle|\; \begin{array}{c}\text{data point } n+1 \text{ is} \\ \text{selected for further study}\end{array} \right\} \geq 1-\alpha$$

e.g., select data point $n+1$ if $\widehat{\mu}(X_{n+1}) \geq c$

---

[13] Jin & Candès 2023, *Selection by Prediction with Conformal p-values*

[14] Jin & Candès 2023, *Model-free selective inference under covariate shift via weighted conformal p-values*

[15] Bao et al 2024, *Selective Conformal Inference with False Coverage-statement Rate Control*

[16] Jin & Ren 2024, *Confidence on the Focal: Conformal Prediction with Selection-Conditional Coverage*

A selective version of exchangeability

Define a selection rule:

$$\mathcal{I} : (\mathcal{X} \times \mathcal{Y})^{n+1} \to \{\text{subsets of } [n+1]\}$$

Assume symmetry—

$$\mathcal{I}(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) = \sigma(\mathcal{I}(Z_1, \ldots, Z_{n+1}))$$

Then the following subsequence of data points is exchangeable:

$$(Z_i)_{i \in \mathcal{I}(Z_1, \ldots, Z_{n+1})}$$

Conformal prediction after selection

Full CP after selection[17,18]

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : n+1 \in \mathcal{I}^y, \text{ and } \right.$$
$$\left. S^y_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/|\mathcal{I}^y_n|)}\left(S^y_i : i \in \mathcal{I}^y_n\right) \right\}$$

where

$$\mathcal{I}^y = \mathcal{I}\left((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\right), \quad \mathcal{I}^y_n = \mathcal{I}^y \cap [n]$$

and scores S^y_i are defined as before

[17] Jin & Candès 2023, *Selection by Prediction with Conformal p-values*

[18] Jin & Ren 2024, *Confidence on the Focal: Conformal Prediction with Selection-Conditional Coverage*

Extensions beyond predictive coverage

Other targets?

Other targets of inference

Target of inference from lectures 1–4:

- $Y_{n+1} \in \mathcal{C}(X_{n+1})$
- $Y_{n+1} \in \mathcal{C}(X_{n+1})$, conditional on selection
- $Y_t \in \mathcal{C}(X_t)$ for $t = n+1, n+2, \ldots$
- Generalized risk bounds

} observable—can verify accuracy on a holdout set

Other potential targets of interest:

- Regression—what is $\mathbb{E}[Y \mid X]$?
- Hypothesis testing—$Y \perp\!\!\!\perp X_j \mid X_{-j}$?
- etc.

} unobservable?

Summary

Summary & preview

Summary: lecture 4

Lecture 4 topics: conformal style approaches to....

- Controlling any measure of risk
- Predictive inference / outlier detection for streaming data
- Prediction after selection

In general, the conformal framework can be applied to any problem where we can use the data to calibrate our accuracy

Preview: lectures 5 & 6

For predictive inference, all methods thus far have focused on marginal coverage as the guarantee:

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

the guarantee is always averaged over the distrib. of training + test data

- Lectures 5 & 6: can we provide a guarantee that is stronger than marginal coverage?