Week 13: Conformal Prediction

MATH-517 Statistical Computation and Visualization

Linda Mhalla

2025-12-12

Motivation

- Prediction algorithms usually return point predictions
- In practice we also need a notion of uncertainty:
  - prediction intervals,
  - or more general prediction sets
- Classical approaches:
  - parametric models + asymptotic normality
  - bootstrap, jacknife, Bayesian intervals, etc
- Today: conformal prediction
  - wraps around any black-box predictor
  - gives finite-sample, distribution-free guarantees

Abmbitious aim

We observe i.i.d. training data $(X_i, Y_i) \sim P$, $i = 1, \dots, n$

Given a miscoverage level $\alpha \in [0, 1]$, we want a set-valued predictor

$$\hat{C}_\alpha : \mathcal{X} \to \{\text{subsets of } \mathbb{R}\}$$

such that, for a new i.i.d. pair (X_{n+1}, Y_{n+1}),

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\big) \geq 1 - \alpha$$

for any distribution P

- No asymptotics, i.e., finite sample guarantees
- No parametric assumptions (agnostic to P and the model/algorithm)
- Sets should still be reasonably short when prediction is easy

Exchangeability

- A sequence of random variables $(Z_1, Z_2, \dots, Z_{n+1})$ is **exchangeable** if its joint distribution does not change when we permute the indices, i.e., for any permutation π:

$$(Z_1, \dots, Z_{n+1}) \overset{d}{=} (Z_{\pi(1)}, \dots, Z_{\pi(n+1)})$$

- i.i.d. samples are always exchangeable, but exchangeability is slightly more general (think of $\mathcal{N}_d(\mu, \Sigma)$, with $\sigma_{ij} = \delta^2$, for $i \neq j$)
- Symmetry under permutations $\Rightarrow$ certain order-statistic arguments (e.g., about ranks or quantiles) hold **exactly in finite samples**: this is the key ingredient behind conformal prediction

Consequence of exchangeability I (Vovk et al., 2005)

Start with exchangeable $Y_1, \dots, Y_n$, and let $Y_{(1)} \le \dots \le Y_{(n)}$ denote their order statistics

For a target miscoverage level $\alpha \in (0, 1)$, define the empirical $(1 - \alpha)$-quantile as

$$\hat{q}_{1-\alpha} = \begin{cases} Y_{(\lceil (n+1)(1-\alpha) \rceil)}, & \text{if } \lceil (1-\alpha)(n+1) \rceil \le n, \\ \infty, & \text{otherwise} \end{cases}$$

For one more exchangeable sample Y_{n+1}, we obtain

$$\mathbb{P}\big(Y_{n+1} \le \hat{q}_{1-\alpha}\big) \;\ge\; 1 - \alpha$$

Additionally, under no ties, i.e., $Y_1, \dots, Y_n, Y_{n+1}$ are a.s. distinct,

$$\mathbb{P}(Y_{n+1} \le \hat{q}_{1-\alpha}) < (1 - \alpha) + \frac{1}{n+1}$$

Consequence of exchangeability II

- Exchangeability ⇒ the rank of Y_{n+1} among $(Y_1, \dots, Y_n, Y_{n+1})$ is uniformly distributed over $\{1, \dots, n+1\}$. Thus

$$\begin{aligned}\mathbb{P}(\mathrm{Rank}_{n+1} \leq \lceil (1-\alpha)(n+1) \rceil) &= \sum_{j=1}^{\lceil (1-\alpha)(n+1) \rceil} \mathbb{P}(\mathrm{Rank}_{n+1} = j) \\ &\geq \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \geq 1-\alpha \\ &\overset{\text{a.s. distinct}}{=} \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} < \frac{(1-\alpha)(n+1)+1}{n+1}\end{aligned}$$

- Note that $\hat{q}_{1-\alpha}$ is computed over $Y_1, \dots, Y_n$ thanks to the equivalence between

$$\mathbb{P}(Y_{n+1} \text{ is among the } \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } Y_1, \dots, Y_{n+1}) \geq 1-\alpha$$

$$\mathbb{P}(Y_{n+1} \text{ is among the } \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } Y_1, \dots, Y_n) \geq 1-\alpha$$

Section 1

Naive attempt for regression

Naive construction

- Fit any regression algorithm on the entire (training) data set:

$$\hat{f}_n : \mathcal{X} \to \mathbb{R}$$

- Training residuals:

$$R_i = |Y_i - \hat{f}_n(X_i)|, \quad i = 1, \dots, n$$

- Let $\hat{q}_{1-\alpha}$ be the empirical $(1-\alpha)$-quantile of $\{R_i\}$
- For a test point $X_{n+1} = x$, define prediction set

$$\hat{C}_\alpha(x) = \{y : |\, y - \hat{f}_n(x) \,| \leq \hat{q}_{1-\alpha}\} = [\, \hat{f}_n(x) - \hat{q}_{1-\alpha},\ \hat{f}_n(x) + \hat{q}_{1-\alpha}\,]$$

Looks reasonable, but:

- the test residual

$$R_{n+1} = |Y_{n+1} - \hat{f}_n(X_{n+1})|$$

  is not exchangeable with $R_1, \dots, R_n$, because $\hat{f}_n$ was trained on (X_i, Y_i) but not on (X_{n+1}, Y_{n+1})

- typically we get under-coverage (especially if $\hat{f}_n$ overfits the training data)

Recover exchangeability

To recover exchangeability we must:

- construct conformity scores (residuals) that treat calibration points and the test point symmetrically
- rely on data splitting or cross-validation

This leads to *split conformal prediction (SCP)*

Section 2

Split Conformal Prediction (Regression)

Algorithm

Split index set into two disjoint parts:

- proper training set D_1
- calibration set D_2

1. Fit a point predictor $\hat{f}_{n_1}$ using only $(X_i, Y_i)_{i \in D_1}$
2. For $i \in D_2$, define the conformity scores (calibration residuals)
$$R_i = |Y_i - \hat{f}_{n_1}(X_i)|$$
3. Let $\hat{q}_{1-\alpha}$ be the $\lceil (1-\alpha)(n_2+1) \rceil$-th smallest residual
4. For a test point x (exchangeable with D_2), define the prediction interval
$$\hat{C}_\alpha(x) = [\, \hat{f}_{n_1}(x) - \hat{q}_{1-\alpha},\ \hat{f}_{n_1}(x) + \hat{q}_{1-\alpha} \,]$$

Coverage guarantee (upper bound under assumption of no ties):

$$1 - \alpha \leq \mathbb{P}\big(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\big) \leq 1 - \alpha + \frac{1}{n_2 + 1}$$

Marginal versus conditional coverage

- Conformal prediction guarantees marginal coverage:

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\} \geq 1-\alpha,$$

averaged over the randomness in the calibration and test point (X_{n+1}, Y_{n+1})

- Conditioning on the calibration set, the distribution of the coverage is

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \mid \{(X_i, Y_i)\}_{i=1}^{n_2}\right) \sim \mathrm{Beta}(n_2 + 1 - l, l),$$

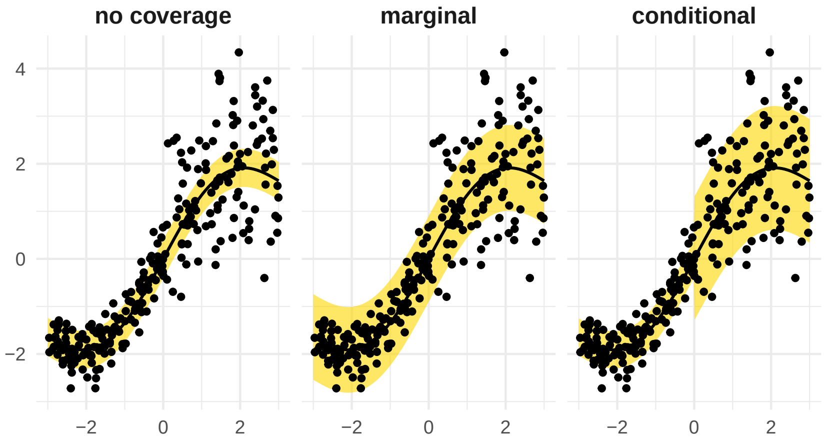
with $l = \lfloor (n_2+1)\alpha \rfloor \Rightarrow$ has mean$= \frac{\lceil (1-\alpha)(n_2+1) \rceil}{n_2+1}$ (sanity check)

- This is different from conditional coverage:

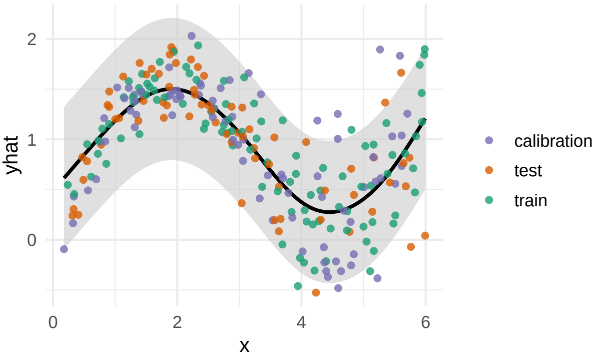
$$\Pr\{Y_{n+1} \in \widehat{C}_\alpha(x) \mid X_{n+1} = x\} \geq 1-\alpha \quad \text{for all } x,$$

which is generally impossible without assumptions on the distribution of the data; see Vovk (2012) and Lei and Wasserman (2013)

Marginal versus conditional coverage



Coverage is not adaptive



- better prediction algorithms (from proper training set) lead to smaller prediction sets (width measured on average over x), but
- split conformal bands with absolute residuals have width exactly constant in $x \Rightarrow$ no adaptivity to local hardness of the prediction

Section 3

Conformalised Quantile Regression (CQR)

Motivation: quantile models to improve local adaptivity

- Split conformal regression uses absolute residuals, producing constant-width bands
- But often the noise varies with x: heteroskedasticity, skewness, outliers ...
- Idea: instead of modelling mean+variance, model conditional quantiles:

$$f_\tau(x) \approx Q_{Y|X=x}(\tau)$$

- For a prediction point x, fit:
  - a lower model $\hat{f}_{\alpha/2}(x)$
  - an upper model $\hat{f}_{1-\alpha/2}(x)$

These already form a prediction band, but with no coverage guarantee!

CQR adds calibration to guarantee finite-sample coverage

Quantile regression

For a random variable Z with cdf F, the τ-quantile is

$$Q_Z(\tau) = \inf\{z : F(z) \geq \tau\}$$

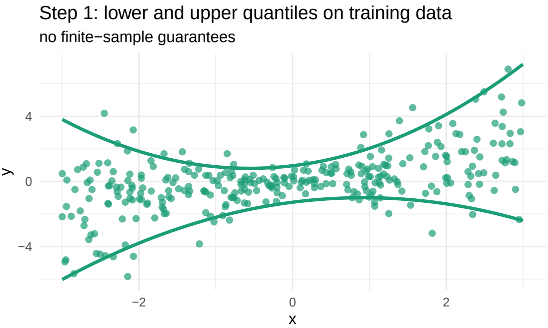
In regression, we model conditional quantiles:

$$f_\tau^\star(x) = Q_{Y|X=x}(\tau)$$

Quantile regression estimates $f_\tau^\star$ by minimising the pinball loss

$$\ell_\tau(y, u) = (\tau - \mathbf{1}\{y < u\})(y - u)$$

Fitting conditional quantiles on training data

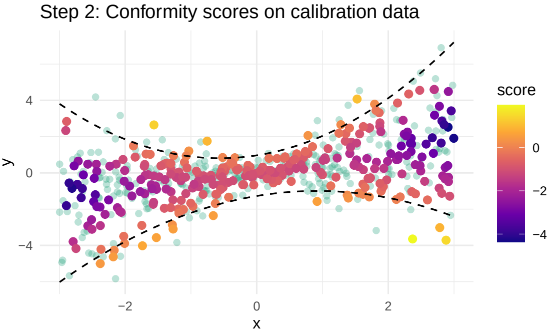


Calibration: checking how well the quantile band fits unseen data

We now compute how much calibration points fall outside the quantile band

Conformity score (negatively oriented) for each point $i \in D_2$:

$$R_i = \max\{\hat{f}_{\alpha/2}(X_i) - Y_i,\ Y_i - \hat{f}_{1-\alpha/2}(X_i)\}$$



CQR interval: quantile band expanded just enough for finite-sample validity

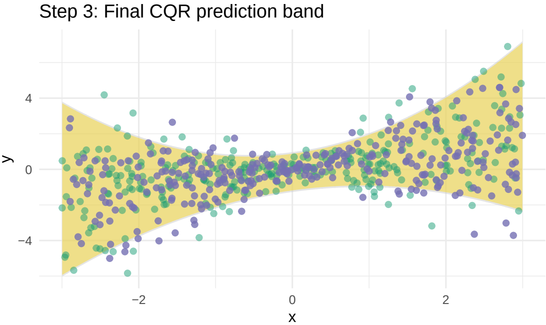
Compute the quantile of the conformity scores (on calibrated data)

$$\hat{q}_{1-\alpha} = R_{(\lceil (1-\alpha)(n_2+1) \rceil)}$$

Final interval

$$\hat{C}_\alpha(x) = [\hat{f}_{\alpha/2}(x) - \hat{q}_{1-\alpha}, \quad \hat{f}_{1-\alpha/2}(x) + \hat{q}_{1-\alpha}]$$

has valid finite-sample coverage, though coverage is only marginal!



CQR: Algorithm

1. On the train set D_1, fit two quantile models:
   - $\widehat{f}_{\alpha/2}(x)$ and $\widehat{f}_{1-\alpha/2}(x)$
2. On calibration set D_2, define scores

$$R_i = \max\{\widehat{f}_{\alpha/2}(X_i) - Y_i, Y_i - \widehat{f}_{1-\alpha/2}(X_i)\}$$

3. Let $\hat{q}_{1-\alpha}$ be the $\lceil(1-\alpha)(n_2+1)\rceil$-th smallest score
4. Final prediction band:

$$\widehat{C}_\alpha(x) = [\,\widehat{f}_{\alpha/2}(x) - \hat{q}_{1-\alpha},\ \widehat{f}_{1-\alpha/2}(x) + \hat{q}_{1-\alpha}\,],$$

   with the same marginal coverage as before, but better local adaptivity

The interval based on quantile regression is widened if $\hat{q}_{1-\alpha} > 0$ and tightened if $\hat{q}_{1-\alpha} < 0$

Section 4

Conformal Classification

From trees/forests to probabilities

From the previous lecture, a classification model, e.g., tree, forest, outputs estimated class probabilities

$$\hat{p}_k(x) \approx \mathbb{P}(Y = k \mid X = x), \qquad k = 1, \dots, K$$

- For a decision tree, these probabilities come from the empirical class proportions in the leaf that contains x
- For a random forest, we average the class proportions over many trees
- For more complex models (logistic regression, neural nets, etc.), the model directly outputs estimated probabilities

Idea for conformal prediction:

We only need the probability vector

$$\hat{f}(x) = (\hat{p}_1(x), \dots, \hat{p}_K(x)),$$

to construct adequate conformal scores and guarantee that the prediction set contains the true class of a new x with high probability

Split conformal classification with likelihood scores

A very common choice of conformity score for classification is

$$s(\hat{f}(x), y) = 1 - \hat{p}_y(x)$$

i.e. small score = high predicted probability for the candidate label y

1. Split data into train D_1 and calibration D_2 and fit $\hat{f}$ on D_1
2. On the calibration set, compute conformity scores

   $$S_i = s(\hat{f}(X_i), Y_i) = 1 - \hat{p}_{Y_i}(X_i), \quad i \in \mathcal{D}_2$$

   (1-likelihood assigned to correct class)
3. Let

   $$\hat{q}_{1-\alpha} = S_{(\lceil (1-\alpha)(n_2+1) \rceil)}$$

   be the $\lceil (1-\alpha)(n_2+1) \rceil$-th smallest score

Split conformal classification with likelihood scores

4. For a new x, define the conformal prediction set

$$\widehat{C}_\alpha(x) = \{y \in \mathcal{Y} : s(\hat{f}(x), y) \leq \hat{q}_{1-\alpha}\}$$

⇒ same marginal coverage guarantee as in regression with the smallest prediction sets on average

But, it has poor conditional coverage as the same threshold $\hat{q}_{1-\alpha}$ is applied to

- an "easy" point (probability mass on one class)
- a "hard" point (flat probability vector)

Example: moderate classifier

We follow Zaffran's toy example with miscoverage level $\alpha = 0.1$ and three labels $\mathcal{Y} = \{\text{dog}, \text{tiger}, \text{cat}\}$

For each calibration point (X_i, Y_i), we have predicted probabilities $(\hat{p}_{\text{dog}}(X_i), \hat{p}_{\text{tiger}}(X_i), \hat{p}_{\text{cat}}(X_i))$, and define the score $S_i = 1 - \hat{p}_{Y_i}(X_i)$

i	Y_i	$\hat{p}_{\text{dog}}$	$\hat{p}_{\text{tiger}}$	$\hat{p}_{\text{cat}}$	S_i
1	dog	0.95	0.02	0.03	0.05
2	dog	0.90	0.05	0.05	0.10
3	dog	0.85	0.10	0.05	0.15
4	tiger	0.15	0.60	0.25	0.40
5	tiger	0.15	0.55	0.30	0.45
6	tiger	0.20	0.50	0.30	0.50
7	tiger	0.15	0.45	0.40	0.55
8	cat	0.25	0.40	0.35	0.65
9	cat	0.20	0.45	0.35	0.65
10	cat	0.20	0.35	0.45	0.55

- The split-conformal threshold is $\hat{q}_{1-\alpha} = S_{([(1-\alpha)(n_2+1)])} = S_{(10)} = 0.65$

For a test point with $(p_{\text{dog}}, p_{\text{tiger}}, p_{\text{cat}}) = (0.05, 0.60, 0.35)$, the scores are

$$S(\text{dog}) = 0.95, \quad S(\text{tiger}) = 0.40, \quad S(\text{cat}) = 0.65$$

Thus, $\widehat{C}_\alpha\,(x_{\text{test}}\,) = \{y : S(y) \le \hat{q}_{1-\alpha}\} = \{\text{tiger}, \text{cat}\}$

Same example: sharper classifier, smaller sets

i	Y_i	$\hat{p}_{\text{dog}}$	$\hat{p}_{\text{tiger}}$	$\hat{p}_{\text{cat}}$	S_i
1	dog	0.95	0.02	0.03	0.05
2	dog	0.90	0.05	0.05	0.10
3	dog	0.85	0.10	0.05	0.15
4	tiger	0.15	0.60	0.25	0.40
5	tiger	0.15	0.55	0.30	0.45
6	tiger	0.20	0.50	0.30	0.50
7	tiger	0.15	0.45	0.40	0.55
8	cat	0.25	0.40	0.35	0.65
9	cat	0.20	0.45	0.35	0.65
10	cat	0.20	0.35	0.45	0.55

The new threshold is $\hat{q}_{1-\alpha} = S_{([(1-\alpha)(n_2+1)])} = S_{(10)} = 0.45$

For the same test point,

$$S(\text{dog}) = 0.95, \quad S(\text{tiger}) = 0.40, \quad S(\text{cat}) = 0.65$$

Now the prediction set is

$$\widehat{C}_\alpha\left(x_{\text{test}}\right) = \{y : S(y) \leq 0.45\} = \{\text{tiger}\}$$

Better classifier on calibration set yields smaller prediction sets

Adaptive Prediction Sets (Angelopoulos et al., 2021)

Idea: We trade the previous method yielding smallest average size with a new method that adapts to the hardness of the problem and uses predicted probabilities of all classes (not only the true class) to build conformity scores

Let $\hat{f}(x) = (\hat{p}_1(x), \dots, \hat{p}_K(x))$ be predicted class probabilities

For each point x:

1. Sort classes by decreasing probability:

$$\hat{p}_{\pi_x(1)}(x) \geq \hat{p}_{\pi_x(2)}(x) \geq \cdots \geq \hat{p}_{\pi_x(K)}(x)$$

2. For each calibration point (X_i, Y_i), $i \in D_2$, compute a conformity score

$$R_i = \sum_{j=1}^{k_i} \hat{p}_{\pi_i(j)}(X_i), \quad \text{where } \pi_i(k_i) = Y_i$$

This is the cumulative probability of classes at least as likely as the true one

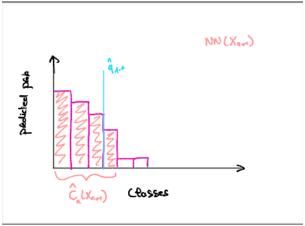
Adaptive Prediction Sets (Angelopoulos et al., 2021)

3. Compute $\hat{q}_{1-\alpha}$ the $\lceil (1-\alpha)(n_2+1) \rceil$-th smallest score
4. For a test point x, form

$$\widehat{C}_\alpha(x) = \{\pi_x(1), \dots, \pi_x(k_x)\},$$

where k_x is the smallest index such that

$$\sum_{j=1}^{k_x} \hat{p}_{\pi_x(j)}(x) \geq \hat{q}_{1-\alpha}$$



Adaptive Prediction Sets: example

- Scores on the calibration set

i	Y_i	$\hat{p}_{\text{dog}}$	$\hat{p}_{\text{tiger}}$	$\hat{p}_{\text{cat}}$	S_i
1	dog	0.95	0.02	0.03	0.95
2	dog	0.90	0.05	0.05	0.90
3	dog	0.85	0.10	0.05	0.85
4	tiger	0.05	0.85	0.10	0.85
5	tiger	0.05	0.80	0.15	0.80
6	tiger	0.05	0.75	0.20	0.75
7	tiger	0.10	0.75	0.15	0.75
8	cat	0.25	0.40	0.35	0.75
9	cat	0.10	0.30	0.60	0.60
10	cat	0.15	0.30	0.55	0.55

- The split-conformal threshold is $\hat{q}_{1-\alpha} = 0.95$
- For a test point with $(p_{\text{dog}}, p_{\text{tiger}}, p_{\text{cat}}) = (0.05, 0.45, 0.5)$, $k_x = 2$
- For a test point with $(p_{\text{dog}}, p_{\text{tiger}}, p_{\text{cat}}) = (0.03, 0.95, 0.02)$, $k_x = 1$

Section 5

Beyond Splitting: Full CP and Jackknife+

Full conformal prediction

Idea: The most probable labels Y_{n+1} live in $\mathcal{Y}$, and have a low enough conformity score. By looping over all possible $y \in \mathcal{Y}$, the ones leading to the smallest conformity scores will be found

For each test point x_{n+1} and candidate label y:

1. Augment the dataset with (x_{n+1}, y)
2. Fit the algorithm on all $n + 1$ points
3. Compute scores for each observation (including the test one)
4. Keep y in the prediction set if its score is not too extreme

- Uses all data both for training and calibration, but
- Extremely expensive: re-fit the model for many candidate y values

Full conformal prediction

Finite sample guarantees are obtained under exchangeability (train points and test point) and symmetry of the algorithm

Theorem (Vovk, 2005)

Suppose that

- $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, and
- the algorithm $\mathcal{A}$ is symmetric (its output depends only on the set of training points, not on their order)

Then, full CP applied on $(X_i, Y_i)_{i=1}^{n} \cup \{X_{n+1}\}$ outputs $\widehat{C}_\alpha(\cdot)$ such that

$$1 - \alpha \ \leq\ \mathbb{P}\Big(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\Big) \ \leq\ 1 - \alpha + \frac{1}{n+1},$$

where the upper bound holds if the scores are a.s. distinct

Jackknife and CV variants

- Compute conformity scores based on LOO predictors $\widehat{f}^{-i}$, trained without sample i
- Jackknife+ and CV+ use these LOO (or K-fold) models to produce conformal-like intervals that:
  - re-use data more efficiently than a single split
  - still enjoy marginal coverage guarantees under mild conditions

We will not go into the full formulas here; idea is to connect to the CV / bagging ideas you have already seen.

Section 6

Summary

Take-home messages

- Conformal prediction wraps around any black-box predictor
- Under exchangeability, we get finite-sample marginal coverage
- Split CP is simple and robust:
  - use proper training set to get point predictor
  - use calibration set to get residual (or score) quantile
- Choice of score matters:
  - residuals, studentised residuals, CQR scores, likelihood scores for classification
- Classification version connects nicely with probability trees / random forests

References

Introductions and tutorials

- Zaffran M. (2023)
- Angelopoulos A. & Bates S. (2023)
- Tibshirani R. (2024)
- Shafer G. & Vovk V. (2008)

Foundational work

- Vovk V., Gammerman A., Shafer G. (2005)

Methodological developments

- Romano Y., Patterson E., Candès E. (2019)
- Lei J., G'Sell M., Rinaldo A., Tibshirani R., Wasserman L. (2018)
- Barber R., Candès E., Ramdas A. (2021)