Conformal prediction and its extensions

1 CP: Conformal prediction

Methods for distribution-free prediction.

Assumption: exchangeability

- The **data** $Z_i = (X_i, Y_i)$ are assumed to be exchangeable (for example, i.i.d.).
  - Definition (Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008). Suppose that for any collection of N values, the $N!$ different orderings are equally likely. Then we say that $Z_1, \ldots, Z_N$ are exchangeable. The exchangeability assumption is slightly weaker than the i.i.d. assumption.
- The **algorithm** which maps data to a fitted model $\hat{\mu} : \mathcal{X} \to \mathbb{R}$ is assumed to treat the data points symmetrically.
  - For example, OLS versus WLS.

1.1 Split conformal prediction

(inductive conformal prediction)

1. initial training data set: pre-trained model $\hat{\mu} : \mathcal{X} \to \mathbb{R}$.
2. holdout/calibration set: nonconformity scores $R_i = \left|Y_i - \hat{\mu}\left(X_i\right)\right|, \quad i = 1, \ldots, n$.
3. prediction set: $\widehat{C}_n\left(X_{n+1}\right) = \hat{\mu}\left(X_{n+1}\right) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \frac{1}{n+1} \cdot \delta_{R_i} + \frac{1}{n+1} \cdot \delta_{+\infty}\right)$.

   (the $\lceil (n+1)(1-\alpha) \rceil$th smallest of $R_1, \ldots, R_n$)

Drawback: the loss of accuracy due to sample splitting.

1.2 Full conformal prediction

(transductive conformal prediction)

1. training data & a hypothesized test point: $\widehat{\mu}^y = \mathcal{A}\left(\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right), \left(X_{n+1}, y\right)\right)$ for each $y \in \mathbb{R}$.
2. residuals: $R_i^y = \begin{cases} \left|Y_i - \widehat{\mu}^y\left(X_i\right)\right|, & i = 1, \ldots, n \\ \left|y - \widehat{\mu}^y\left(X_{n+1}\right)\right|, & i = n+1 \end{cases}$.
3. prediction set: $\widehat{C}_n\left(X_{n+1}\right) = \left\{y \in \mathbb{R} : R_{n+1}^y \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_i^y}\right)\right\}$.

Drawback: a steep computational cost.

THEOREM:

$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right\} \geq 1-\alpha$ holds true for both split conformal and full conformal.

1.3 Jackknife+

(close to cross-conformal prediction in Vovk (2013), offering a compromise between the computational and statistical costs)

1. training data with ith point removed: $\widehat{\mu}_{-i} = \mathcal{A}\left(\left(X_1, Y_1\right), \ldots, \left(X_{i-1}, Y_{i-1}\right), \left(X_{i+1}, Y_{i+1}\right), \ldots, \left(X_n, Y_n\right)\right)$.
2. residuals: $R_i^{\mathrm{LOO}} = \left|Y_i - \widehat{\mu}_{-i}\left(X_i\right)\right|$.
3. prediction set:

$$\left[\mathrm{Q}_\alpha\left(\sum_{i=1}^n \frac{1}{n+1} \cdot \delta_{\widehat{\mu}_{-i}\left(X_{n+1}\right)-R_i^{\mathrm{LOO}}} + \frac{1}{n+1} \cdot \delta_{-\infty}\right),\ \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^n \frac{1}{n+1} \cdot \delta_{\widehat{\mu}_{-i}\left(X_{n+1}\right)+R_i^{\mathrm{LOO}}} + \frac{1}{n+1} \cdot \delta_{+\infty}\right)\right]$$

Drawback: while in practice the it generally provides coverage close to the target level $1-\alpha$, its theoretical guarantee only ensures $1-2\alpha$ probability of coverage in the worst case.

THEOREM:

$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right\} \geq 1-2\alpha$ holds true for jackknife+.

2 Conformal time-series forecasting

Stankeviciute, M Alaa, and Schaar (2021): CF-RNNs

Multi-horizon time-series forecasting problem

Notation:

- the ith data point: $Z_i = (y_{1:t}^{(i)}, y_{t+1:t+H}^{(i)})$. Note that the label $y_{t+1:t+H}^{(i)}$ is now an H-dimensional value, in contrast with the scalar y value from before.

Assumption:

- exchangeable time-series observations

2.1 Methodology

(Split conformal prediction)

1. training set: train the underlying (auxiliary) model $\hat{\mu} : \mathbb{R}^t \to \mathbb{R}^H$, which produces multi-horizon forecasts **directly** (conditionally independent predictions).
2. calibration set: obtain the H-dimensional nonconformity scores

   $R_i = \left[\left|y^{(i)}_{t+1} - \hat{y}^{(i)}_{t+1}\right|, \dots, \left|y^{(i)}_{t+H} - \hat{y}^{(i)}_{t+H}\right|\right]^\top.$
3. prediction set: $\Gamma^\alpha_1\left(y^{(n+1)}_{(1:t)}\right), \dots, \Gamma^\alpha_H\left(y^{(n+1)}_{(1:t)}\right)$, where $\Gamma^\alpha_h\left(y^{(n+1)}_{(1:t)}\right) = \left[\hat{y}^{(n+1)}_{t+h} - \hat{\varepsilon}_h, \hat{y}^{(n+1)}_{t+h} + \hat{\varepsilon}_h\right]$, $\forall h \in \{1, \dots, H\}$ with the critical nonconformity scores $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_H$ become the $\lceil (n+1)(1-\alpha/H) \rceil$-th smallest residuals in the corresponding nonconformity score distributions. (Bonferroni correction)

THEOREM:

- $\mathcal{D} = \left\{\left(y^{(i)}_{1:t}, y^{(i)}_{t+1:t+H}\right)\right\}_{i=1}^n$: **exchangeable** time-series observations.
- $\hat{\mu}$: model predicting H-step forecasts using **the direct strategy**.

$$\mathbb{P}\left(\forall h \in \{1, \dots, H\} \cdot y_{t+h} \in [\hat{y}_{t+h} - \hat{\varepsilon}_h, \hat{y}_{t+h} + \hat{\varepsilon}_h]\right) \geq 1 - \alpha.$$

3 NexCP: Conformal prediction beyond exchangeability

Barber et al. (2023)

1. Nonexchangeable conformal with a **symmetric** algorithm (weights)
2. Nonexchangeable conformal with **nonsymmetric** algorithms (weights & swap)

3.1 Notation

- the ith data point $Z_i = (X_i, Y_i)$
- the full data sequence $Z = (Z_1, \dots, Z_{n+1})$
- the sequence after swap $Z^i = (Z_1, \dots, Z_{i-1}, Z_{n+1}, Z_{i+1}, Z_n, Z_i)$

3.2 Methodology

1. Choose **fixed (non-data-dependent)** weights $w_1, \ldots, w_n \in [0,1]$ with the intuition that a higher weight should be assigned to a data point that is "trusted" more.
2. Normalize weights $\tilde{w}_i = \frac{w_i}{w_1+\cdots+w_n+1}, i = 1, \ldots, n$ and $\tilde{w}_{n+1} = \frac{1}{w_1+\cdots+w_n+1}$.
3. Generate "tagged" data points $(X_i, Y_i, t_i) \in \mathcal{X} \times \mathbb{R} \times \mathcal{T}$.
4. Swap data set, resulting in Z^K with $K \sim \sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_i$, i.e., two data points have swapped tags.
5. Apply algorithm $\mathcal{A}$ to Z^K in place of Z.

3.3 Split conformal prediction

- prediction set: $\widehat{C}_n\left(X_{n+1}\right) = \hat{\mu}\left(X_{n+1}\right) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{w}_i \cdot \delta_{R_i} + \tilde{w}_{n+1} \cdot \delta_{+\infty}\right)$

3.4 Full conformal prediction

1. training data & a hypothesized test point: $\hat{\mu}^{y,k} = \mathcal{A}\left(\left(X_{\pi_k(i)}, Y^y_{\pi_k(i)}, t_i\right) : i \in [n+1]\right)$ for any $y \in \mathbb{R}$ and $k \in [n+1]$, where π_k is the permutation on $[n+1]$ swapping indices k and $n+1$, and $Y_i^y = \begin{cases} Y_i, & i = 1, \ldots, n \\ y, & i = n+1 \end{cases}$.
2. residuals: $R_i^{y,k} = \begin{cases} \left|Y_i - \hat{\mu}^{y,k}\left(X_i\right)\right|, & i = 1, \ldots, n \\ \left|y - \hat{\mu}^{y,k}\left(X_{n+1}\right)\right|, & i = n+1 \end{cases}$.
3. prediction set: $\widehat{C}_n\left(X_{n+1}\right) = \left\{y : R_{n+1}^{y,K} \leq Q_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{R_i^{y,K}}\right)\right\}$.

THEOREM:

- Lower bounds on coverage.

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right\} \geq 1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i \cdot \mathrm{d}_{\mathrm{TV}}\left(R(Z), R\left(Z^i\right)\right)$$

- Upper bounds on coverage.

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right\} < 1 - \alpha + \tilde{w}_{n+1} + \sum_{i=1}^{n} \tilde{w}_i \cdot \mathrm{d}_{\mathrm{TV}}\left(R(Z), R\left(Z^i\right)\right),$$

if $R_1^{Y_{n+1},K}, \ldots, R_n^{Y_{n+1},K}, R_{n+1}^{Y_{n+1},K}$ are distinct with probability 1.

The results hold true for both nonexchangeable split conformal and full conformal.

So, if $\tilde{w}_{n+1} = \frac{1}{w_1+\cdots+w_n+1}$ is small (the effective sample size is large), then mild violations of exchangeability can only lead to mild undercoverage or to mild overcoverage.

3.5 Jackknife+

1. training data with ith point removed and data tag swapped:

   $\hat{\mu}_{-i}^{k} = \mathcal{A}\left(\left(X_{\pi_k(j)}, Y_{\pi_k(j)}, t_j\right) : j \in [n+1], \pi_k(j) \notin \{i, n+1\}\right).$

2. residuals: $R_i^{k,\mathrm{LOO}} = \left|Y_i - \hat{\mu}_{-i}^{k}\left(X_i\right)\right|.$

3. prediction set:

   $\left[\mathrm{Q}_\alpha\left(\sum_{i=1}^n \tilde{w}_i \cdot \delta_{\hat{\mu}_{-i}^{K}(X_{n+1}) - R_i^{K,\mathrm{LOO}}} + \tilde{w}_{n+1} \cdot \delta_{-\infty}\right),\ \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^n \tilde{w}_i \cdot \delta_{\hat{\mu}_{-i}^{K}(X_{n+1}) + R_i^{K,\mathrm{LOO}}} + \tilde{w}_{n+1} \cdot \delta_{+\infty}\right)\right]$

THEOREM:

$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right\} \geq 1 - 2\alpha - \sum_{i=1}^n \tilde{w}_i \cdot \mathrm{d}_{\mathrm{TV}}\left(R_{\text{jack }+}(Z), R_{\text{jack }+}\left(Z^i\right)\right)$

4 WCP: Conformal prediction under covariate shift

Tibshirani et al. (2019)

A weighted version of conformal prediction, using a quantile of a suitably weighted empirical distribution of nonconformity scores.

4.1 Setup/assumption - Covariate shift

Focus on settings in which the data $\left(X_i, Y_i\right), i = 1, \ldots, n+1$ are no longer exchangeable. Specifically,

$$\begin{aligned} \left(X_i, Y_i\right) &\overset{\text{i.i.d.}}{\sim} P = P_X \times P_{Y|X}, \quad i = 1, \ldots, n, \\ \left(X_{n+1}, Y_{n+1}\right) &\sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}, \text{ independently.} \end{aligned} \tag{1}$$

the test and training covariate distributions differ, but the likelihood ratio between the two distributions, $\mathrm{d}\widetilde{P}_X / \mathrm{d}P_X$, must be known exactly or well approximated for correct coverage.

4.2 Methodology

Prediction set:

- CP: $\widehat{C}_n\left(X_{n+1}\right)=\hat{\mu}\left(X_{n+1}\right) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^n \frac{1}{n+1} \cdot \delta_{R_i}+\frac{1}{n+1} \cdot \delta_{+\infty}\right)$.
- NexCP: $\widehat{C}_n\left(X_{n+1}\right)=\hat{\mu}\left(X_{n+1}\right) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^n \tilde{w}_i \cdot \delta_{R_i}+\tilde{w}_{n+1} \cdot \delta_{+\infty}\right)$.
  - weights w are fixed
  - $\tilde{w}_i=\frac{w_i}{w_1+\cdots+w_n+1}, i=1, \ldots, n$ and $\tilde{w}_{n+1}=\frac{1}{w_1+\cdots+w_n+1}$
- WCP: $\widehat{C}_n\left(X_{n+1}\right)=\hat{\mu}\left(X_{n+1}\right) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^n p_i^w(x) \delta_{R_i}+p_{n+1}^w(x) \delta_{\infty}\right)$.
  - $w=\mathrm{d} \widetilde{P}_X / \mathrm{d} P_X$ or $w \propto \mathrm{d} \widetilde{P}_X / \mathrm{d} P_X$
  - $p_i^w(x)=\frac{w\left(X_i\right)}{\sum_{j=1}^n w\left(X_j\right)+w(x)}, i=1, \ldots, n$, and $p_{n+1}^w(x)=\frac{w(x)}{\sum_{j=1}^n w\left(X_j\right)+w(x)}$

The **weight function** $\hat{w}$ can be estimated using logistic regression, random forests, etc.

THEOREM:

Assume data from the model Equation 1. Assume $\widetilde{P}_X$ is absolutely continuous with respect to P_X, and denote $w=\mathrm{d} \widetilde{P}_X / \mathrm{d} P_X$. For any score function S, and any $\alpha \in(0,1)$, define for $x \in \mathbb{R}^d$. Then

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

4.3 Comparison to NexCP

1. Assumption
   - WCP: covariate shift
   - NexCP: do not make any assumption on the joint distribution of the $n+1$ points
2. Weights
   - WCP: a function of the data point (X_i, Y_i) to compensate for the **known** distribution shift.
   - NexCP: required to be fixed, can compensate for **unknown** violations of the exchangeability assumption, as long as the violations are **small** (to ensure a low coverage gap).
3. Nonsymmetric algorithm
   - WCP: No.
   - NexCP: Yes.
4. For exchangeable data

- WCP: does not have any coverage guarantee.
- NexCP: retains exact coverage.

5 ACP: Adaptive conformal inference under distribution shift

Gibbs and Candes (2021)

- No assumptions on the data-generating distribution.
- Modelling the distribution shift as a learning problem in a single parameter whose optimal value is varying over time and must be continuously re-estimated.
- Adjust significance level α based on rolling coverage of Y_t .

5.1 Methodology

Work with score function $S(\cdot)$ and quantile function $\hat{Q}(\cdot)$.

Some facts:

- If the distribution of the data is shifting over time, both functions should be regularly re-estimated to align with the most recent observations.
- The realized miscoverage rate $M_t(\alpha)$ also varies over time and may not be equal or close to α.

Assumptions:

Assume that there may be an alternative value $\alpha^* \in [0,1]$ such that $M_t(\alpha^*) \cong \alpha$.

Assume that with probability one, $\hat{Q}_t(\cdot)$ is continuous, non-decreasing and such that $\hat{Q}_t(0) = -\infty$ and $\hat{Q}_t(1) = \infty$.

Adaptive conformal inference:

- Under assumptions, $M_t(\cdot)$ will be non-decreasing on $[0,1]$ with $M_t(0) = 0$ and $M_t(1) = 1$.
- Define $\alpha_t^* := \sup\{\beta \in [0,1] : M_t(\beta) \leq \alpha\}$, then $M_t(\alpha_t^*) = \alpha$.
- Use a simple **update process** to perform the calibration.
  - Intuition: after examining the empirical miscoverage frequency of the previous prediction sets, decreasing (increasing) estimate of α_t^* if the prediction sets were historically under-covering (over-covering) Y_t.

- Let $\alpha_1 = \alpha$, consider the **update**

$$\alpha_{t+1} := \alpha_t + \gamma\left(\alpha - \mathrm{err}_t\right)$$

OR

$$\alpha_{t+1} = \alpha_t + \gamma\left(\alpha - \sum_{s=1}^{t} w_s\,\mathrm{err}_s\right),$$

where $\gamma > 0$ is a fixed step size parameter whose choice gives a tradeoff between adaptability and stability.

So the distribution is allowed to shift continuously over time.

THEOREM:

With probability one we have that for all $T \in \mathbb{N}$,

$$\left|\frac{1}{T}\sum_{t=1}^{T} err_t - \alpha\right| \leq \frac{\max\left\{\alpha_1, 1-\alpha_1\right\} + \gamma}{T\gamma}.$$

In particular, $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} err_t \overset{\text{a.s.}}{=} \alpha$.

6 PID: Conformal PID control for time series prediction

Angelopoulos, Candes, and Tibshirani (2023)

- PID: proportional-integral-derivative (control theory).
- reactive + forward-looking.

6.1 Methodology

The proposed conformal PID controller is given by

$$q_{t+1} = \underbrace{\eta \left(\text{err}_t - \alpha\right)}_{\text{P}} + \underbrace{r_t \left(\sum_{i=1}^{t} \left(\text{err}_i - \alpha\right)\right)}_{\text{I}} + \underbrace{g'_t}_{\text{D}},$$

where $g_t = err_t - \alpha$.

The goal is to achieve **long-run coverage** in time, i.e., $\frac{1}{T}\sum_{t=1}^{T} \text{err}_t = \alpha + o(1)$, where $o(1)$ denotes a quantity that tends to zero as $T \to \infty$.

- P: Quantile tracking. ACI can be regarded as a special case but it can sometimes output infinite or null prediction sets, while quantile tracking on the scale of the original score sequence does not have this behavior.
- I: Error integration. r_t should be a saturation function to achieve long-run coverage.
- D: Scorecasting (forward-looking). It trains a second model to predict the quantile of the next score.

The applications only consider one-step-ahead forecast.

7 Others

7.1 EnbPI: predictive inference method around ensemble estimators

Xu and Xie (2021)

- For Dynamic time series
- Ensemble point forecasts + update residuals

7.2 SPCI: Sequential Predictive Conformal Inference

Xu and Xie (2023)

- Adaptively re-estimate the conditional quantile of non-conformity scores, upon leveraging the temporal dependency among residuals. Random Forest for quantile regression is used.

7.3 Kath & Ziel (2021, IJF)

Kath and Ziel (2021)

Method 1.

The steps are:

1. Re-arrange time series data as form $Z = \{(x_1, y_1), \dots, (x_L, y_L)\}$.
2. Randomly split observations into subsets π (training set) and $1 - \pi$ (calibration set).
3. Train a random forecast model using the training set and generate out-of-sample forecasts.
4. Calculate non-conformity score.
5. Obtain final prediction interval estimation.

Method 2.

The differences compared to Method 1 are:

1. Here an adjusted normalized non-conformity score $\lambda_{i,h} = \frac{|y_{i,h} - \hat{y}_{i,h}|}{|\hat{\varepsilon}_{i,h}|}$ is used.
2. A second explicit error estimation model is trained to estimate the error associated with the prediction $\hat{y}_{t,h}$, i.e., $\varepsilon_{t,h}$.
3. The new interval forecast is now given by $y_{\alpha,T+1,h} = \hat{y}_{T+1,h} \pm \lambda^{\alpha}_{L+1,h} \left|\hat{\varepsilon}_{L+1,h}\right|$.

The applications only consider one-step-ahead forecast.

7.4 TimeGPT

Garza and Mergenthaler-Canseco (2023)

For uncertainty quantification, they perform rolling forecasts on the latest available data to estimate the model's errors in forecasting the particular target time series.

The h-step-ahead forecast is generated based on h-step-ahead non-conformity errors on the calibration set separately.

7.5 Weighted quantile estimators

Akinshin (2023)

- The goal is to estimate the distribution at the tail of a time series, which reflects the latest state of the underlying system.

Desired properties of the weighted quantile estimators:

- **Requirement R1:** consistency with existing quantile estimators.

$$\mathrm{Q}^*(\mathbf{x}, \mathbf{w}, p) = \mathrm{Q}(\mathbf{x}, p), \text{ where } \mathbf{w} = \{1, 1, \dots, 1\}.$$

- **Requirement R2:** zero weight support.

$$\mathrm{Q}^*\left(\{x_1, x_2, \dots, x_{n-1}, x_n\}, \{w_1, w_2, \dots, w_{n-1}, 0\}, p\right) = \mathrm{Q}^*\left(\{x_1, x_2, \dots, x_{n-1}\}, \{w_1, w_2, \dots, w_{n-1}\}, p\right).$$

- **Requirement R3:** stability (continuity of the quantile estimations with respect to the weight coefficient).

$$\lim_{\varepsilon_i \to 0} Q^*\left(\{x_1, x_2, \dots, x_n\}, \{w_1 + \varepsilon_1, w_2 + \varepsilon_2, \dots, w_n + \varepsilon_n\}, p\right) \to \mathrm{Q}^*\left(\{x_1, x_2, \dots, x_n\}, \{w_1, w_2, \dots, w_n\}, p\right).$$

Some issues:

1. Some popular R implementations of weighted quantiles show **violations of Requirement R3**.

```r
x <- c(0, 1, 100)
wA <- c(1, 0.00000, 1)
wB <- c(1, 0.00001, 1)
message(modi::weighted.quantile(x, wA, 0.5), " | ",
        modi::weighted.quantile(x, wB, 0.5))
## 100 | 1
message(laeken::weightedQuantile(x, wA, 0.5), " | ",
        laeken::weightedQuantile(x, wB, 0.5))
## 100 | 1
message(MetricsWeighted::weighted_quantile(x, wA, 0.5), " | ",
        MetricsWeighted::weighted_quantile(x, wB, 0.5))
## 100 | 1
message(spatstat.geom::weighted.quantile(x, wA, 0.5), " | ",
        spatstat.geom::weighted.quantile(x, wB, 0.5))
## 1 | 0.499999999994449
message(matrixStats::weightedMedian(x, wA), " | ",
        matrixStats::weightedMedian(x, wB))
## 50 | 1.00000000000003
```

2. In order to satisfy Requirement R2, the sample size needs adjustments. Consider using Kish's **effective sample size** given by

$$n^*(\mathbf{w}) = \frac{\left(\sum_{i=1}^n w_i\right)^2}{\sum_{i=1}^n w_i^2} = \frac{1}{\sum_{i=1}^n \bar{w}_i^2},$$

where $\mathbf{w}$ is the vector of non-negative weight coefficients, $\bar{\mathbf{w}}$ is the vector of normalized weights where $\bar{w}_i = \frac{w_i}{\sum_{i=1}^n w_i}$. So $n^*(\{1,1,1\}) = 3$ and $n^*(\{1,1,1,0,0\}) = 3$.

Method 1: Weighted Harrell–Davis quantile estimator

- a linear combination of all order statistics.
- **pros:** provide higher statistical efficiency in the cases of light-tailed distributions.
- **cons:** not robust (its breakdown point is zero).

The weighted Harrell–Davis quantile estimator is defined as

$$\mathrm{Q}^*_{\mathrm{HD}}(\mathbf{x}, \mathbf{w}, p) = \sum_{i=1}^n W^*_{\mathrm{HD},i} \cdot x_{(i)}, \quad W^*_{\mathrm{HD},i} = I_{t^*_i}(\alpha^*, \beta^*) - I_{t^*_{i-1}}(\alpha^*, \beta^*),$$

where $\alpha^* = (n^*+1)\,p$, $\beta^* = (n^*+1)\,(1-p)$, $t^*_i = s_i(\bar{\mathbf{w}})$ is the partial sum of normalized weight coefficients, $I_{t^*_i}(\alpha^*, \beta^*)$ is the CDF of the beta distribution $\mathrm{Beta}(\alpha, \beta)$. $W^*_{\mathrm{HD},i}$ can be considered as probabilities of observing the target quantile at the given position. See details from Examples 7 and 8.

For all $p \in (0;1)$, $\mathrm{Q}^*_{\mathrm{HD}}(\mathbf{x}, \mathbf{w}, p)$ satisfies R1, R2, and R3.

Method 2: Weighted trimmed Harrell–Davis quantile estimator

- **a trimmed modification** of the HD method.
- idea: since most of the linear coefficients $W^*_{\mathrm{HD},i}$ are pretty small, they do not have a noticeable impact on efficiency, but they significantly reduce the breakdown point.
- pros: allow customizing trade-off between robustness and efficiency.
- cons: use the rule of thumb to decide $[L^*; R^*]$.

For $p \in (0;1)$, the weighted trimmed Harrell–Davis quantile estimator based on the beta distribution highest density interval $[L^*; R^*]$ of the given size D^* (the rule of thumb: $D^* = 1/\sqrt{n^*}$) is defined as

$$\mathrm{Q}^*_{\mathrm{THD}}(\mathbf{x},\mathbf{w},p)=\sum_{i=1}^{n} W^*_{\mathrm{THD},i}\cdot x_{(i)},\quad W^*_{\mathrm{THD},i}=F^*_{\mathrm{THD}}\left(t^*_i\right)-F^*_{\mathrm{THD}}\left(t^*_{i-1}\right),\quad t^*_i=s_i(\overline{\mathbf{w}}),$$

where

$$F^*_{\mathrm{THD}}(t)=\begin{cases}0 & \text{for } t<L^*,\\ \left(I_t\left(\alpha^*,\beta^*\right)-I_{L^*}\left(\alpha^*,\beta^*\right)\right)/\left(I_{R^*}\left(\alpha^*,\beta^*\right)-I_{L^*}\left(\alpha^*,\beta^*\right)\right) & \text{for } L^*\le t\le R^*,\\ 1 & \text{for } R^*<t.\end{cases}$$

Method 3: Weighted traditional quantile estimators

- Hyndman and Fan (1996) summarized types 1-9.
  - Types 1–3 have discontinuities, so the corresponding estimators fail to satisfy R3.
  - Here only Types 4-9 are considered (using a linear combination of **two order statistics**).
- Pros: extremely robust, not so efficient.

Table: The Hyndman & Fan (1996) taxonomy of quantile estimators.

Type	h	Equation
1	np	$x_{(\lceil h\rceil)}$
2	$np+1/2$	$\left(x_{(\lceil h-1/2\rceil)}+x_{(\lceil h+1/2\rceil)}\right)/2$
3	np	$x_{(\lfloor h\rceil)}$
4	np	$x_{(\lfloor h\rfloor)}+(h-\lfloor h\rfloor)\left(x_{(\lceil h\rceil)}-x_{(\lfloor h\rfloor)}\right)$
5	$np+1/2$	$x_{(\lfloor h\rfloor)}+(h-\lfloor h\rfloor)\left(x_{(\lceil h\rceil)}-x_{(\lfloor h\rfloor)}\right)$
6	$(n+1)p$	$x_{(\lfloor h\rfloor)}+(h-\lfloor h\rfloor)\left(x_{(\lceil h\rceil)}-x_{(\lfloor h\rfloor)}\right)$
7	$(n-1)p+1$	$x_{(\lfloor h\rfloor)}+(h-\lfloor h\rfloor)\left(x_{(\lceil h\rceil)}-x_{(\lfloor h\rfloor)}\right)$
8	$(n+1/3)p+1/3$	$x_{(\lfloor h\rfloor)}+(h-\lfloor h\rfloor)\left(x_{(\lceil h\rceil)}-x_{(\lfloor h\rfloor)}\right)$
9	$(n+1/4)p+3/8$	$x_{(\lfloor h\rfloor)}+(h-\lfloor h\rfloor)\left(x_{(\lceil h\rceil)}-x_{(\lfloor h\rfloor)}\right)$

These estimators can be rewritten in a form that matches the definition of $\mathrm{Q}^*_{\mathrm{THD}}$ given by

$$Q^*_k(\mathbf{x},\mathbf{w},p)=\sum_{i=1}^{n} W^*_{F^*_k,i}\cdot x_{(i)},\quad W^*_{F^*_k,i}=F^*_k\left(t^*_i\right)-F^*_k\left(t^*_{i-1}\right),\quad t^*_i=s_i(\bar{\mathbf{w}}),$$

where

$$F_k^*(t) = \begin{cases} 0 & \text{for} \quad t < (h^*-1)/n^*, \\ tn^* - h^* + 1 & \text{for} \quad (h^*-1)/n^* \le t \le h^*/n^*, \\ 1 & \text{for} \quad t > h^*/n^*. \end{cases}$$

7.6 LCP: Localized conformal prediction

Guan (2022)

The **weight** on data point i is determined as **a function of the distance** $\|X_i - X_{n+1}\|_2$, to enable predictive coverage that holds **locally** (in neighborhoods of X space, that is, an approximation of prediction that holds conditional on the value of X_{n+1}).

8 Simulation

8.1 Setup

Simulate a time series y with length $T = 5000$ from an AR(2) model with $\phi_1 = 0.8$, $\phi_2 = -0.5$, and $\sigma^2 = 1$.

Only consider one-step-ahead forecasting, i.e., $h = 1$.

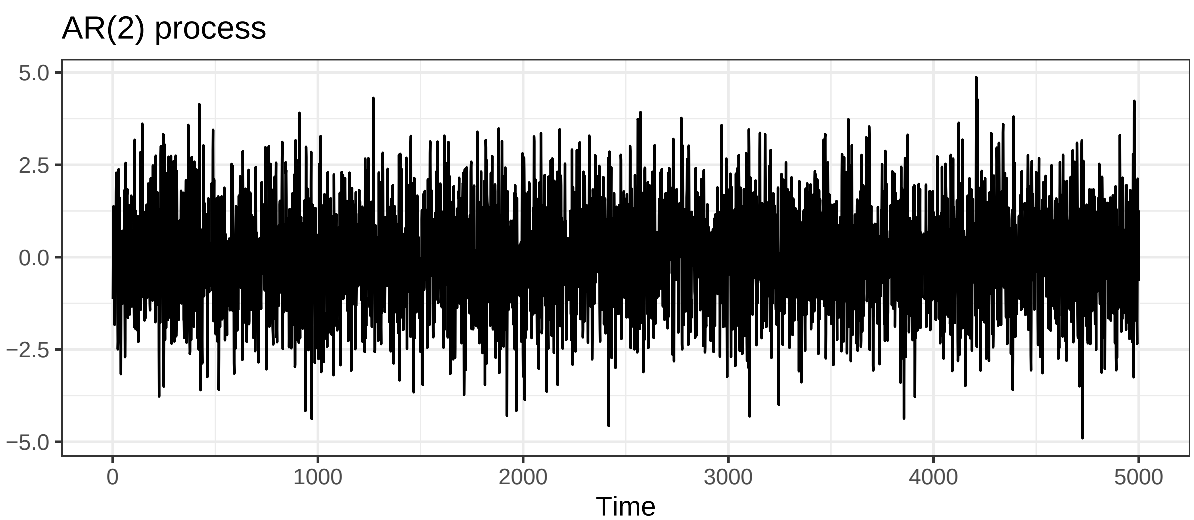

Figure 1: Simulated time series from an AR(2) model.

8.2 Split conformal prediction with fixed training set

8.2.1 Details

Let n be the number of data used to fit an AR(2) model, and m be the number of data in the calibration set.

- Step 1. Rearrange the time series as $Z_i = (X_i, Y_i)$, where $X_i = y_{i:(i+1)}$ and $Y_i = y_{i+2}$, $i = 1, 2, \dots, T-2$.
- Step 2. Train an AR(2) model $\hat{\mu}$ based on training set with length n, $Z_{tr} = (Z_1, Z_2, \dots, Z_n)$.
- Step 3. Calculate nonconformity scores (absolute residuals) based on calibration set with length m, i.e., $R_i = |Y_i - \hat{\mu}(X_i)|$, $i = n+1, n+2, \dots, n+m$.
- Step 4. Generate PI on test set. $\hat{C_m}(X_i) = \hat{\mu}(X_i) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{j=n+1}^{n+m} w_j \cdot \delta_{R_j} + w_i \cdot \delta_{+\infty}\right)$ for $i = n+m+i, \dots, T-2$.

8.2.2 Results

Let $n = m = 500$ and fit the AR(2) model using the `lm` function.

Consider methods: CP, WCP.LR, WCP.RF, NexCP with $\alpha = 0.1$.

Issues:

- PIs have constant width over the test set. We can generate PIs with varying local width by using a function to perform training on the absolute residuals i.e., to produce an estimator of $\mathrm{E}(R|X)$.
- When generating weights via GLM or RF in WCP method, we need to use all data from the test set, which is not reasonable.
- Regression-based model.

8.3 Split conformal prediction with rolling training set

8.3.1 Details

Let n be the number of observations used to fit an AR(2) model, and m be the number of observations in the calibration set.

For $i = n+1, n+2, ..., T$:

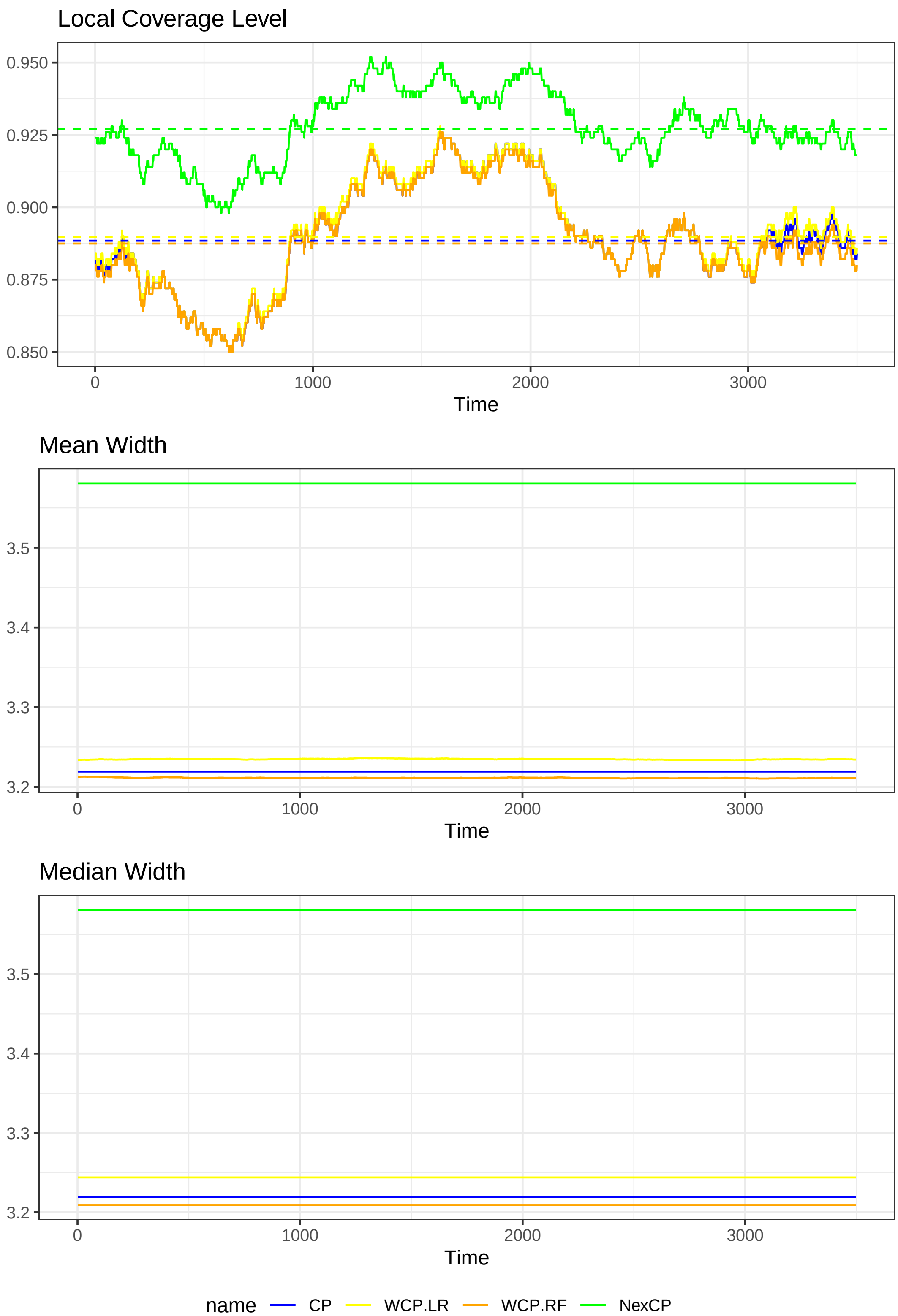

Figure 2: Local coverage frequencies and width for split conformal prediction methods with fixed training set (k=500).

- Step 1. Fit an AR(2) model $\hat{\mu}_{i-1}$ based on observations $y_{(i-n):(i-1)}$, generate one-step-ahead forecast $\hat{y}_i$.
- When $i > n + m$:
  - Step 2. Calculate weights for the updated calibration set with length m and the updated test set with length 1 using different methods. Here WCP can not be applied because we only have a test set with length equal to one.
  - Step 3. Generate PI on test set. $\hat{y}_i \pm \mathrm{Q}_{1-\alpha}\left(\sum_{j=i-m}^{i-1} w_i \cdot \delta_{R_j} + w_i \cdot \delta_{+\infty}\right)$.
  - Setp 4 for ACP. Update α based on the recent empirical miscoverage frequency.
- Step 5. Calculate nonconformity scores (absolute residuals) $R_i = |y_i - \hat{y}_i|$.

8.3.2 Results

Let $n = m = 500$ and fit AR(2) models using the `Arima` function with setting `order = c(2,0,0), include.mean = TRUE, method = "CSS"` to make it comparable with the previous result.

Consider methods: AR, CP, NexCP, ACP with $\alpha = 0.1$.

Features:

- The fitted model is updating over time.
- PIs have changing width over the test set.
- Other non-regression-based models can be considered.
- Time-consuming because of the rolling window approach.

9 To do list

- Multi-step ahead forecasting
- Other data-dependent weight.
- Papers from other journals such as IJF.
- Theoretical proof.
- Hierarchical time series.

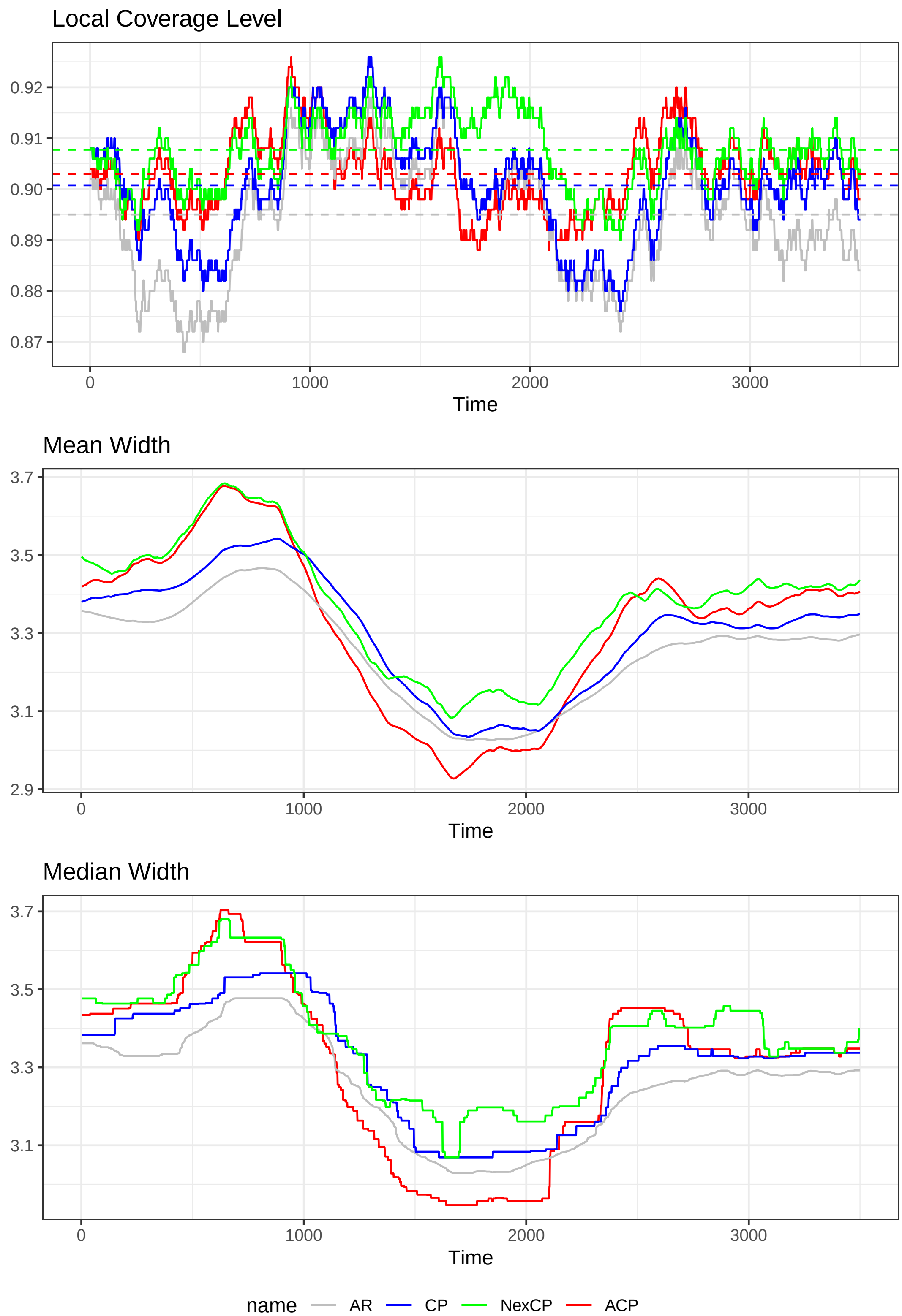

Figure 3: Local coverage frequencies and width for split conformal prediction methods with rolling training set and calibration set (k=500).

10 References

Akinshin, Andrey. 2023. "Weighted Quantile Estimators." *arXiv Preprint arXiv:2304.07265*.

Angelopoulos, Anastasios N, Emmanuel J Candes, and Ryan J Tibshirani. 2023. "Conformal PID Control for Time Series Prediction." *arXiv Preprint arXiv:2307.16895*.

Barber, Rina Foygel, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2023. "Conformal Prediction Beyond Exchangeability." *The Annals of Statistics* 51 (2). https://doi.org/10.1214/23-aos2276.

Garza, Azul, and Max Mergenthaler-Canseco. 2023. "TimeGPT-1." *arXiv Preprint arXiv:2310.03589*.

Gibbs, Isaac, and Emmanuel Candes. 2021. "Adaptive Conformal Inference Under Distribution Shift." *Advances in Neural Information Processing Systems* 34: 1660–72.

Guan, Leying. 2022. "Localized Conformal Prediction: A Generalized Inference Framework for Conformal Prediction." *Biometrika* 110 (1): 33–50. https://doi.org/10.1093/biomet/asac040.

Hyndman, Rob J, and Yanan Fan. 1996. "Sample Quantiles in Statistical Packages." *The American Statistician* 50 (4): 361–65.

Kath, Christopher, and Florian Ziel. 2021. "Conformal Prediction Interval Estimation and Applications to Day-Ahead and Intraday Power Markets." *International Journal of Forecasting* 37 (2): 777–99.

Shafer, Glenn, and Vladimir Vovk. 2008. "A Tutorial on Conformal Prediction." *Journal of Machine Learning Research* 9 (3).

Stankeviciute, Kamile, Ahmed M Alaa, and Mihaela van der Schaar. 2021. "Conformal Time-Series Forecasting." *Advances in Neural Information Processing Systems* 34: 6216–28.

Tibshirani, Ryan J, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. 2019. "Conformal Prediction Under Covariate Shift." *Advances in Neural Information Processing Systems* 32.

Vovk, Vladimir. 2013. "Cross-Conformal Predictors." *Annals of Mathematics and Artificial Intelligence* 74 (1-2): 9–28. https://doi.org/10.1007/s10472-013-9368-4.

Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag. https://doi.org/10.1007/b106715.

Xu, Chen, and Yao Xie. 2021. "Conformal Prediction Interval for Dynamic Time-Series." In *International Conference on Machine Learning*, 11559–69. PMLR.

———. 2023. "Sequential Predictive Conformal Inference for Time Series." In *International Conference on Machine Learning*, 38707–27. PMLR.