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

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: $R_i^{y,k} = \begin{cases} \left|Y_i - \hat{\mu}^{y,k}(X_i)\right|, & i = 1, \ldots, n \\ \left|y - \hat{\mu}^{y,k}(X_{n+1})\right|, & i = n+1 \end{cases}$.
3. prediction set: $\widehat{C}_n(X_{n+1}) = \left\{y : R_{n+1}^{y,K} \le Q_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{R_i^{y,K}}\right)\right\}$.

THEOREM:

- Lower bounds on coverage.

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

- Upper bounds on coverage.

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

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:

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

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

3. prediction set:

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

THEOREM:

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

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

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

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

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 $T \in \mathbb{N}$,

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

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

6 Others

6.1 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}).

6.2 EnbPI: predictive inference method around ensemble estimators

Xu and Xie (2021)

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

6.3 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.

6.4 Conformal PID Control for Time Series Prediction

Angelopoulos, Candes, and Tibshirani (2023)

PID: proportional-integral-derivative

7 References

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.

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.

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.