Conformal inference for uncertainty quantification in official statistics

Nina Deliu & Brunero Liseo

Sapienza Università di Roma

nina.deliu@uniroma1.it; brunero.liseo@uniroma1.it

3rd Workshop on Methodologies for Official Statistics
December, 5th 2024

Overview

- What is a conformal prediction (CP)?
- Design-based CP
- Model-based CP
- CP as a possible agreement among statistical paradigms

What is CP?

Preliminary note on uncertainty quantification

Given a n-sample $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, and a generic *point* estimator for $Y \in \mathcal{Y}$, e.g., of an underlying regression model

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

Goal: to build a prediction interval for Y_{n+1}, say

$$\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1}) = \hat{f}_n(x_{n+1}) \pm \Delta_\alpha,$$

with $1 - \alpha$ coverage guarantees, that is, such that

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1})) \geq 1 - \alpha, \quad \alpha \in (0, 1).$$

What is CP?

Preliminary note on uncertainty quantification

Given a n-sample $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, and a generic *point* estimator for $Y \in \mathcal{Y}$, e.g., of an underlying regression model

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

Goal: to build a prediction interval for Y_{n+1}, say

$$\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1}) = \hat{f}_n(x_{n+1}) \pm \Delta_\alpha,$$

with $1 - \alpha$ coverage guarantees, that is, such that

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1})) \geq 1 - \alpha, \quad \alpha \in (0, 1).$$

$\hookrightarrow$ Alert! It is incorrect to use the *training* residuals $R_i = |y_i - \hat{f}_n(\boldsymbol{x}_i)|$, $i = 1, \ldots, n$ to estimate Δ_α: they may be to small (overfitting) when compared to that of the test point Y_{n+1}, with no coverage guarantees.

What is CP?

First idea: Split Conformal

- Fit $\hat{f}_{n/2}$ using half of your data: $\{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n/2\}$
- Then make a Bag of residuals with the other half

$$\{R_i = |y_i - \hat{f}_{n/2}(\boldsymbol{x}_i)|, \quad i = \frac{n}{2} + 1, \ldots, n\}.$$

- Construct the prediction interval as

$$\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1}) = \hat{f}_{n/2}(\boldsymbol{x}_{n+1}) \pm Q_{1-\alpha}(\text{Bag})$$

where $Q_{1-\alpha}$ is the $\lceil (1-\alpha)(\frac{n}{2}+1) \rceil$ smallest residual in the Bag.

What is CP?

First idea: Split Conformal

- Fit $\hat{f}_{n/2}$ using half of your data: $\{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n/2\}$
- Then make a Bag of residuals with the other half

$$\{R_i = |y_i - \hat{f}_{n/2}(\boldsymbol{x}_i)|, \quad i = \frac{n}{2} + 1, \ldots, n\}.$$

- Construct the prediction interval as

$$\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1}) = \hat{f}_{n/2}(\boldsymbol{x}_{n+1}) \pm Q_{1-\alpha}(\text{Bag})$$

where $Q_{1-\alpha}$ is the $\lceil (1-\alpha)(\frac{n}{2}+1) \rceil$ smallest residual in the Bag.

$\hookrightarrow$ Now: All the computed residuals are *exchangeable*, included that of the test point, avoiding overfitting and ensuring proper coverage.

What is CP?

Theoretical justification

Split Conformal Prediction enjoys finite sample guarantees, as proved by Vovk et al. [2005] and Lei and Wasserman [2014].

Theorem

Assume the pairs $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots n$, $n+1$, are exchangeable. Then

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1})) \geq 1 - \alpha$$

and the result holds for any finite sample size.

Proof: Easy, mainly based on quantiles, permutation, and exchangeability.

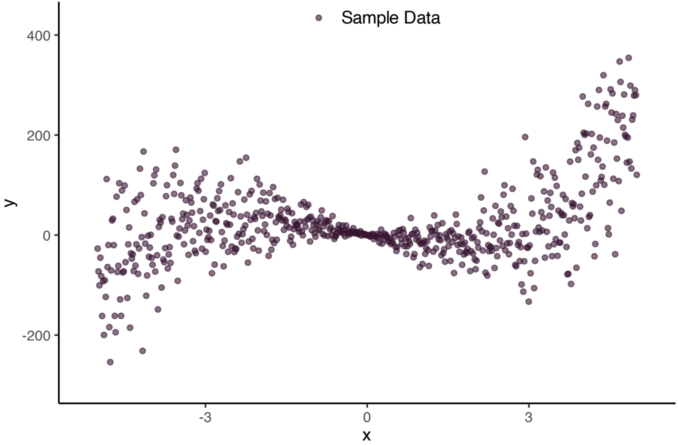
- Intuition: The set $\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1})$ consists of

$$\left\{ \text{all values of } Y \text{ such that } |Y - \hat{f}_n(\boldsymbol{x}_{n+1})| \leq k \right\}$$

  and k is a threshold constructed on the quantiles of the Bag.
- Here the residuals R_i play the role of conformity scores.

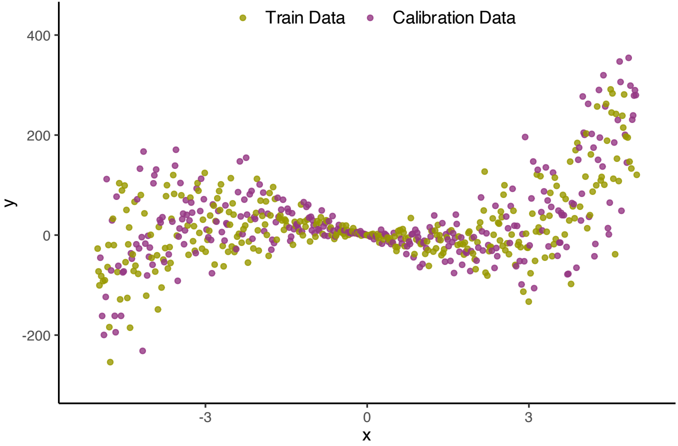
An illustrative example

Sample data $\text{Data}_n^{\texttt{Sample}}$



An illustrative example

Sample data $\text{Data}_n^{\texttt{Sample}}$

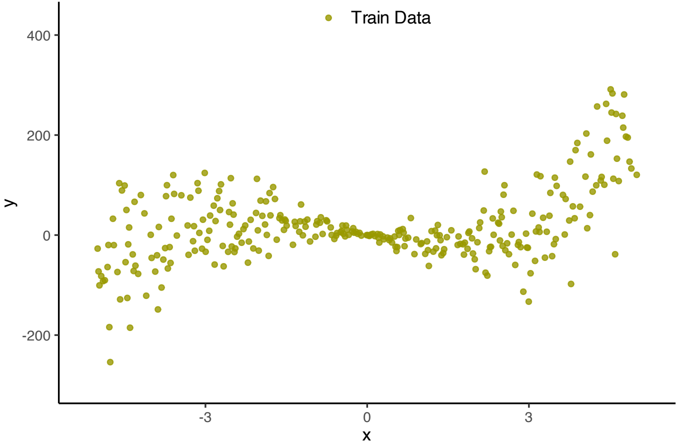


$\hookrightarrow$ *Split* the sample data:

$$\text{Data}^{\texttt{Sample}} = \text{Data}^{\texttt{Train}} \sqcup \text{Data}^{\texttt{Cal}} \quad \text{with} \quad \text{Data}^{\texttt{Train}} \cap \text{Data}^{\texttt{Cal}} = \emptyset$$

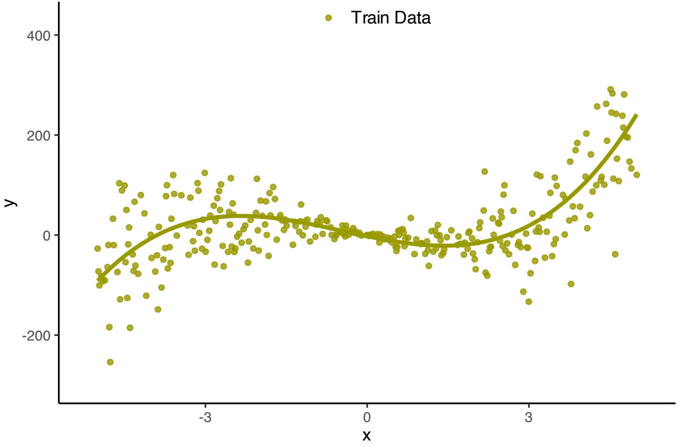
An illustrative example

Train data: $\text{Data}_{n_T}^{\texttt{Train}}$



An illustrative example

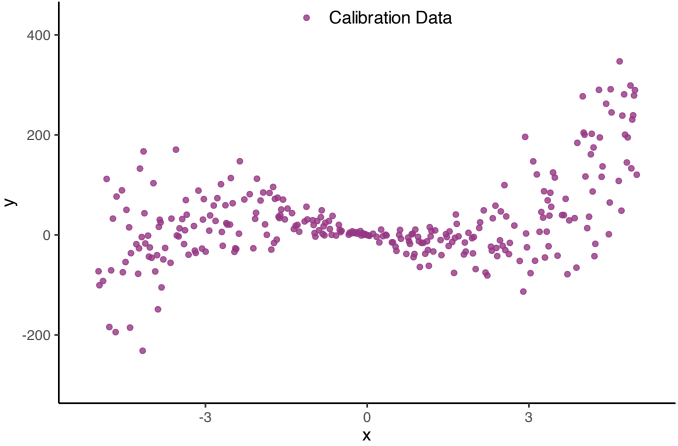
Train data: $\text{Data}_{n_T}^{\texttt{Train}}$



$\hookrightarrow$ Use $\text{Data}_{n_T}^{\texttt{Train}}$ to fit a point predictor $\hat{f}_{n_T} : \mathcal{X} \rightarrow \mathcal{Y}$

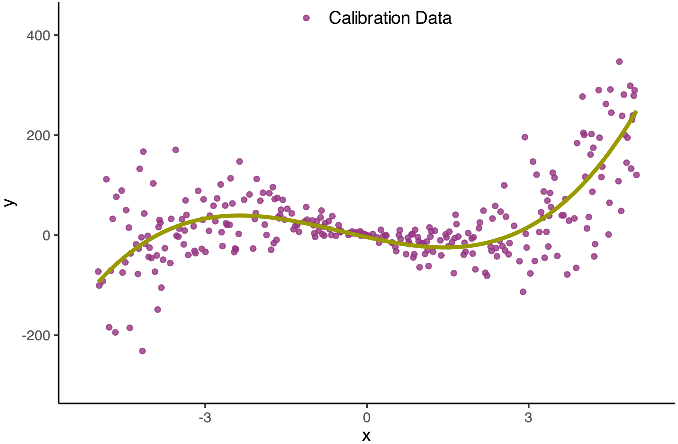
An illustrative example

Calibration data: $\text{Data}_{n_C}^{\texttt{Cal}}$



An illustrative example

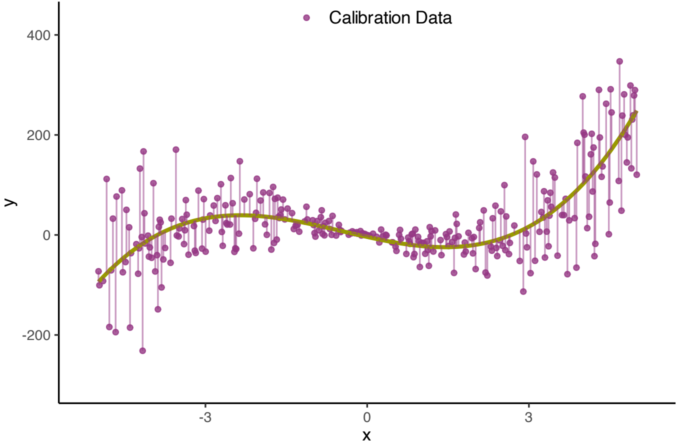
Calibration data: $\text{Data}_{n_C}^{\texttt{Cal}}$



✓ Get n_C predictions on $\text{Data}_{n_C}^{\texttt{Cal}}$: $\hat{f}_{n_T}(X_j)$, $j \in \text{Data}_{n_C}^{\texttt{Cal}}$

An illustrative example

Calibration data: $\text{Data}_{n_C}^{\texttt{Cal}}$



✓ Get calibration/*conformity* scores: $R_j = |Y_j - \hat{f}_{n_T}(X_j)|$, $j \in \text{Data}_{n_C}^{\texttt{Cal}}$

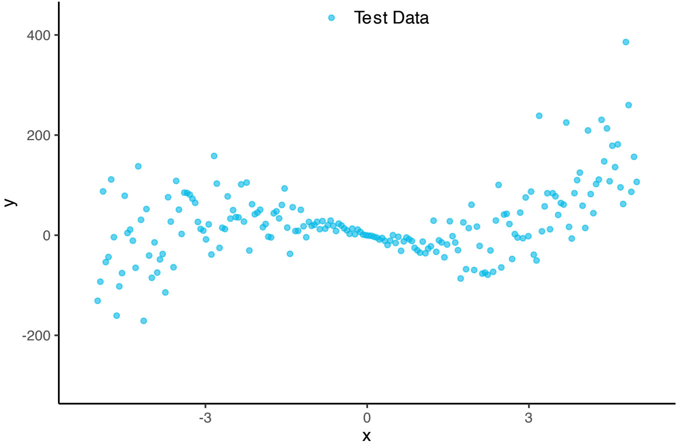
An illustrative example

Calibration data: $\text{Data}_{n_C}^{\texttt{Cal}}$

Use $\{R_i : i \in \text{Data}^{\texttt{Cal}}\}$ to get

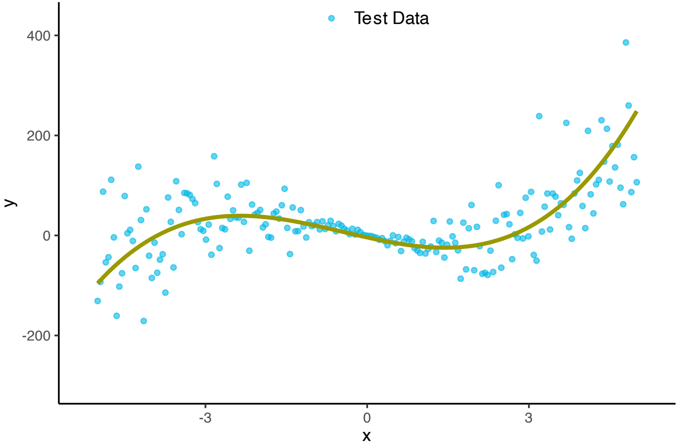
An illustrative example

Test data: $\text{Data}_{n^*}^{\texttt{Test}}$



An illustrative example

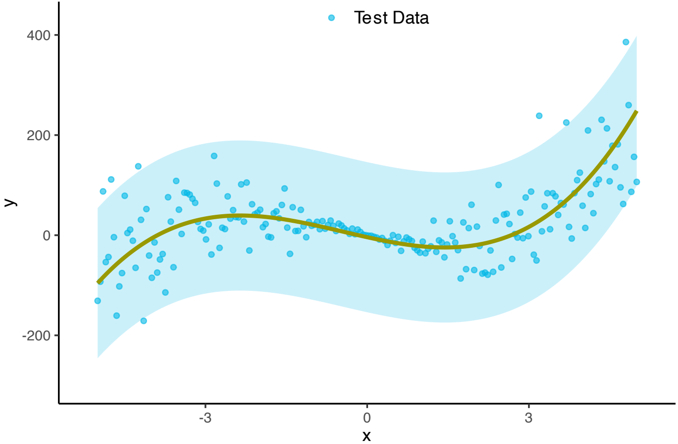
Test data: $\text{Data}_{n^*}^{\texttt{Test}}$



✓ Get n^* predictions on $\text{Data}_{n^*}^{\texttt{Test}}$: $\hat{f}_{n_T}(X_{j^*})$, $j^* \in \text{Data}_{n^*}^{\texttt{Test}}$

An illustrative example

Test data: $\text{Data}_{n^*}^{\texttt{Test}}$



$\hookrightarrow$ Split-CP: $\mathcal{C}_{n,1-\alpha}^{\texttt{split}}(X_{j^*}) = [\hat{f}_{n_T}(X_{j^*}) \pm q_{n,1-\alpha}], j^* \in \text{Data}_{n^*}^{\texttt{Test}}$

Conformal Prediction in Official Statistics

CP in Official Statistics

Consider the following set-up

Unit	Sample Membership I	Covariate X_1	$\dots$	Covariate X_p	Outcome Y_1
1	$i_1 = 1$	x_{11}	$\dots$	x_{1p}	y_1
2	$i_2 = 0$	x_{21}	$\dots$	x_{2p}	$\hat{y}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	$i_j = 1$	x_{j1}	$\dots$	x_{jp}	y_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$i_N = 0$	x_{N1}	$\dots$	x_{Np}	$\hat{y}_N$

Inferences are made based on the (sample) data:

$$\text{Data}_n^{\texttt{Sample}} = \{(\boldsymbol{X}_j, Y_j) : j \in \mathcal{S}_n\}, \quad \mathcal{S}_n = \{j : I_j = 1\}.$$

CP in Official Statistics

Consider the following set-up

Unit	Sample Membership I	Covariate X_1	$\dots$	Covariate X_p	Outcome Y_1
1	$i_1 = 1$	x_{11}	$\dots$	x_{1p}	y_1
2	$i_2 = 0$	x_{21}	$\dots$	x_{2p}	$\hat{y}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	$i_j = 1$	x_{j1}	$\dots$	x_{jp}	y_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$i_N = 0$	x_{N1}	$\dots$	x_{Np}	$\hat{y}_N$

Inferences are made based on the (sample) data:

$$\text{Data}_n^{\texttt{Sample}} = \{(\boldsymbol{X}_j, Y_j) : j \in \mathcal{S}_n\}, \quad \mathcal{S}_n = \{j : I_j = 1\}.$$

In general: $\quad (I_j, (\boldsymbol{X}_j, Y_j)) \sim P = P_I \times P_{(\boldsymbol{X},Y)|I}, \quad j = 1, \dots, N.$

CP in Official Statistics

Design-based CP [Wieczorek, 2024]

$$\mathbb{P}_I(Y_{j^*} \in \mathcal{C}(\boldsymbol{X}_{j^*})) \geq 1 - \alpha, \quad j^* \notin \mathcal{S}_n.$$

- Easy to handle with SRS designs: units are exchangeable
- Requires *ad hoc* corrections with more general sampling schemes (more on this later)

CP in Official Statistics

Design-based CP [Wieczorek, 2024]

$$\mathbb{P}_I(Y_{j^*} \in \mathcal{C}(\boldsymbol{X}_{j^*})) \geq 1-\alpha, \quad j^* \notin \mathcal{S}_n.$$

- Easy to handle with SRS designs: units are exchangeable
- Requires *ad hoc* corrections with more general sampling schemes (more on this later)

Model-based CP

$$\mathbb{P}_{(\boldsymbol{X},Y)}(Y_{j^*} \in \mathcal{C}(\boldsymbol{X}_{j^*})) \geq 1-\alpha, \quad j^* \notin \mathcal{S}_n.$$

Can provide great advantages:

- can mitigate the model-misspecification problem
- can produce narrower prediction intervals
- Bayes–Frequentist compromise

Simulations

apipop data (R package survey); $N = 6194$

Data description: Academic Performance Index (API)

- Response variable of interest

  Y :: `api00` Numeric response variable representing the API score in 2000, covering all California schools with at least 100 students (range: 200 to 1000)

- A set of auxiliary variables: we only consider

  X_1 :: `stype` Categorical variable representing the school type (elementary, middle, high)

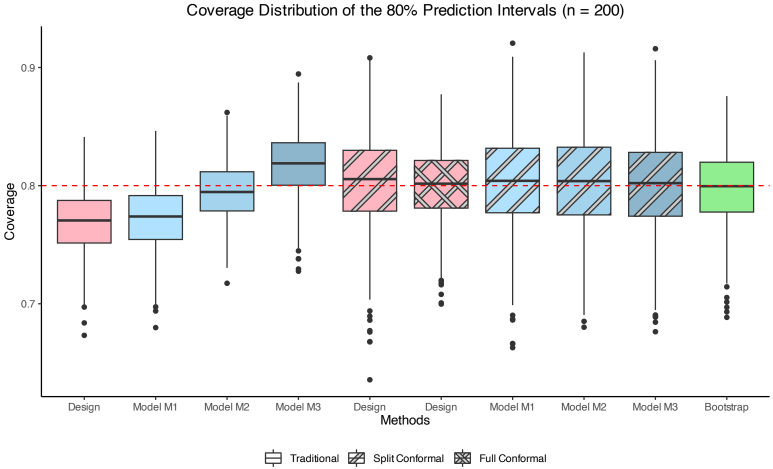
  X_2 :: `ell` Numeric variable given by the percentage of English Language Learners

  X_3 :: `meals` Numeric variable being the percentage of students eligible for subsidized meals

  X_4 :: `mobility` Numeric variable for the percentage of first-year students at the school

Simulations

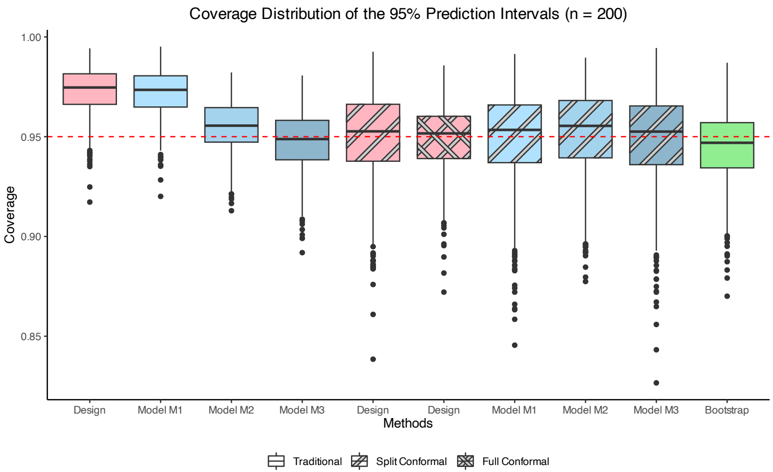
A comparison between traditional and CP methods



Expected coverage for a target $1 - \alpha = 0.8$ (red dashed line). $M = 1000$ independent SRS-WR with $n = 200$ from the apipop dataset with population size $N = 6194$.

Simulations

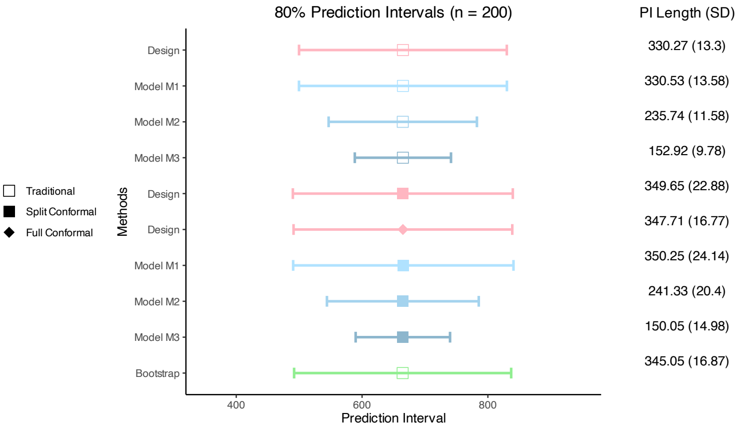
A comparison between traditional and CP methods



Expected coverage for a target $1 - \alpha = 0.95$ (red dashed line). $M = 1000$ independent SRS-WR with $n = 200$ from the apipop dataset with population size $N = 6194$.

Simulations

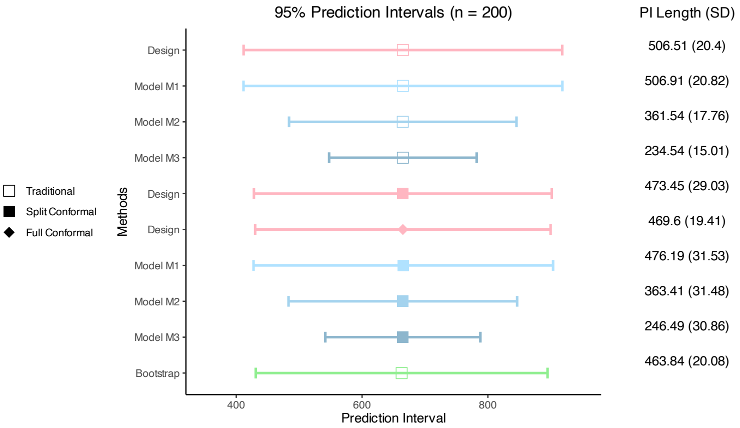
A comparison between traditional and CP methods



Expected prediction interval, length, and SD for a target $\alpha = 0.2$. Average across $M = 1000$ independent SRS-WR with $n = 200$ from the `apipop` data.

Simulations

A comparison between traditional and CP methods



Expected prediction interval, length, and SD for a target $\alpha = 0.05$. Average across $M = 1000$ independent SRS-WR with $n = 200$ from the `apipop` data.

Design-based and Model-based CP

Advantages when compared with alternative methods

Design-based CP

- *versus* Linearization: finite-sample guarantees & model-free (no need for *ad hoc* calculations)
- *versus* Bootstrap and other Resampling methods: finite-sample guarantees & less computationally demanding (at least for Split CP)

Design-based and Model-based CP

Advantages when compared with alternative methods

Design-based CP

- *versus* Linearization: finite-sample guarantees & model-free (no need for *ad hoc* calculations)
- *versus* Bootstrap and other Resampling methods: finite-sample guarantees & less computationally demanding (at least for Split CP)

Model-based CP

- The combination of CP and the correct model provides the optimal intervals, both in terms of coverage and length
- A poor model specification can cause an increase in length but does not undermine coverage
- Coverage is guaranteed for finite sample sizes

Design-based and Model-based CP

Advantages when compared with alternative methods

Design-based CP

- *versus* Linearization: finite-sample guarantees & model-free (no need for *ad hoc* calculations)
- *versus* Bootstrap and other Resampling methods: finite-sample guarantees & less computationally demanding (at least for Split CP)

Model-based CP

- The combination of CP and the correct model provides the optimal intervals, both in terms of coverage and length
- A poor model specification can cause an increase in length but does not undermine coverage
- Coverage is guaranteed for finite sample sizes

↪ In general, given the exact coverage, one can simply choose *among* alternative CP approaches, either design-based or model-based, in terms of the average length of the resulting prediction intervals

CP Challenges in Official Statistics

(A) Conditional Coverage and Adaptivity: domain-restricted predictions

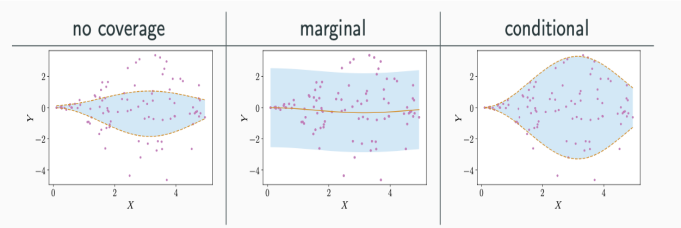
(B) Beyond Exchangeability: covariate shift, time series data, complex designs

(C) Classification: here prediction sets are discrete and different methods are necessary, based on the cumulative likelihood [Romano et al., 2020]

(D) Combining prediction intervals (i.e. (sub)-population size estimation)

(A) Marginal and Conditional Coverage

- *Marginal* coverage: $P\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}(\boldsymbol{X}_{n+1})\right) \geq 1 - \alpha$
  $\hookrightarrow$ errors may differ across regions of the covariate space
- *Conditional* coverage: $P\left(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}(\boldsymbol{x}) | \boldsymbol{X}_{n+1} = \boldsymbol{x}\right) \geq 1 - \alpha$
  $\hookrightarrow$ conditional coverage implies adaptiveness



- Alert! Conditional coverage is stronger than marginal coverage but, in general (e.g. for a continuous X), not attainable using nonparametric methods [Lei and Wasserman, 2014].

Achieving Adaptivity in CP

Standard mean-regression CP is not adaptive ...

- However, it is not reasonable to have a constant width! Uncertainty quantification depends on the amount of data at given $\boldsymbol{x}$...
- Simple solution: use a *studentized* conformity score

$$S_i(\boldsymbol{x}_i, y_i) = \frac{R_i(\boldsymbol{x}_i, y_i)}{\hat{\sigma}(\boldsymbol{x}_i)} = \frac{|y_i - \hat{f}_{n/2}(\boldsymbol{x}_i)|}{\hat{\sigma}(\boldsymbol{x}_i)}$$

  with

$$\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}) = \left[\hat{f}_{n/2}(\boldsymbol{x}_i) \pm \hat{\sigma}(\boldsymbol{x}) Q_{1-\alpha}(S)\right]$$

- More complex alternative: conformalized quantile regression [Romano et al., 2019]

Conformalized Quantile Regression

Romano et al. [2019]

The algorithm

1. Randomly split the training data into a proper training set (size n_T) and a calibration set (size n_C)
2. Fit the lower ($\hat{Q}_{\alpha/2}$) and upper ($\hat{Q}_{1-\alpha/2}$) quantile by training a suitable algorithm on the proper training set $\text{Data}_{n_T}^{\texttt{Train}}$
3. Compute the n_C conformity scores:

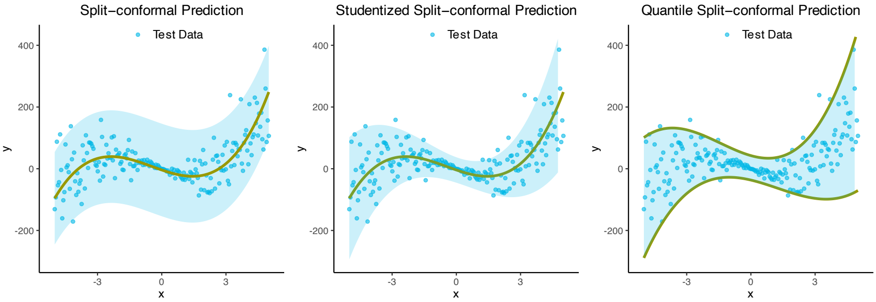
$$S_i = \max\left(\hat{Q}_{\frac{\alpha}{2}}(X_i) - Y_i,\ Y_i - \hat{Q}_{1-\frac{\alpha}{2}}(X_i)\right), \quad i \in \text{Data}_{n_C}^{\texttt{Cal}}$$

4. Compute $q_{n,1-\alpha} = S_{(\lceil (n_C+1)(1-\alpha)\rceil)}$
5. For a new (test) point X_{n+1}, set

$$\mathcal{C}_{n,1-\alpha}(X_{n+1}) = \left[\hat{Q}_{\frac{\alpha}{2}}(X_{n+1}) - q_{n,1-\alpha}; \hat{Q}_{1-\frac{\alpha}{2}}(X_{n+1}) + q_{n,1-\alpha}\right]$$

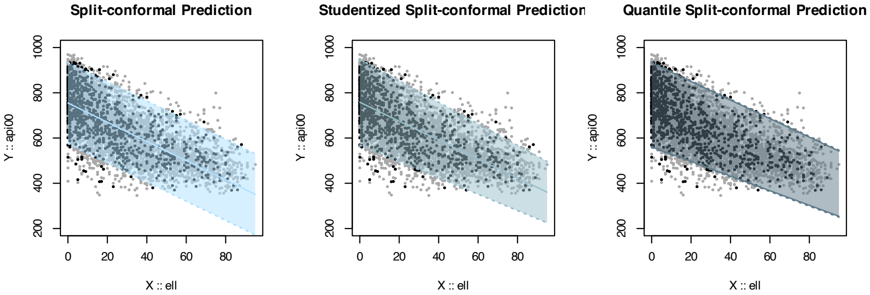
Adaptivity: A comparison of different methods

Initial illustrative example



Adaptivity: A comparison of different methods

apipop data



(B) Beyond Exchangeability

- Exchangeability is the main requirement for using CP
- Conformal measure computed on the test unit can be considered
- This might not be the case in survey sampling where observed values in the sample $\mathcal{S}_n$ may be the result of a complex sampling design, while units for which we need to make a prediction might be generated by a different system.

A weighted version

Sample units adhere to a specific sampling design which is not necessarily shared by non-sample units
This problem has been considered in Tibshirani et al. [2019], who adopted a weighted version of the conformal scores. More in detail, assume that while the original sample data were generated by a model

$$(\boldsymbol{X}_j, Y_j) \overset{\text{i.i.d}}{\sim} P = P_I \times P_{Y|\boldsymbol{X}} \times P_{\boldsymbol{X}}, \quad j \in \mathcal{S}_n$$

the new observation comes from a different marginal distribution of $\boldsymbol{X}$, say

$$(\boldsymbol{X}_{j^*}, Y_{j^*}) \overset{\text{i.i.d}}{\sim} P^* = P_I \times P_{Y|\boldsymbol{X}} \times P^*_{\boldsymbol{X}}, \quad j^* \notin \mathcal{S}_n.$$

(B) Beyond Exchangeability

Covariate Shift Solution

- The problem is solved by weighting the original conformal scores of the observations $(x_1, x_2, \ldots, x_n)$ using the likelihood ratio

$$w(x_j) = \mathrm{d}P^*(x_j)/\mathrm{d}P(x_j),$$

which plays a "weight" role.

- Consider, for simplicity, a full CP setup where the calibration scores are computed for the full sample dataset $\text{Data}_n^{\texttt{Sample}}$ and the augmented candidate y. Under a weighted version, the new set of empirical conformal scores will then be $(R_1 p_1(x), \ldots, R_n p_n(x), R_{j^*} p_{j^*})$,where

$$p_j(x) = \frac{w(\boldsymbol{X}_j)}{\sum_{i=1}^n w(\boldsymbol{X}_i) + w(x)}, \quad j \in \mathcal{S}_n,$$
$$p_{j^*}(x) = \frac{w(x)}{\sum_{i=1}^n w(\boldsymbol{X}_i) + w(x)}, \quad j^* \notin \mathcal{S}_n.$$

CP as a calibrated Bayes approach

A new line of research?

Bayes–Frequentist compromise?

- One of the main criticisms regarding model-based techniques in survey sampling is the potential dependence on the assumed model
- Also, the frequentist performance of Bayesian methods can be jeopardized by the use of the prior
- The conformal modification of the estimates produced via a full model-based Bayesian approach is then a promising way to obtain a calibration of Bayesian estimates
- Idea: combine all the information sources via an HB model-based approach and take as the *natural conformity measure the posterior predictive distribution*, both in a Full- or in a Split-CP scenario. See Bersson and Hoff [2024] for an example in Small Area Estimation.

Conclusions and Perspectives

Advantages of CP in Official Statistics

- CP has finite-sample and distribution-free exact **marginal** coverage
- CP can be built on top of the preferred prediction strategy that has been used to impute missing values in the response variable
- CP also allows to quantify uncertainty also on predictions arising from *multiple* strategies [Gasparin and Ramdas, 2024]

Conclusions and Perspectives

Advantages of CP in Official Statistics

- CP has finite-sample and distribution-free exact **marginal** coverage
- CP can be built on top of the preferred prediction strategy that has been used to impute missing values in the response variable
- CP also allows to quantify uncertainty also on predictions arising from *multiple* strategies [Gasparin and Ramdas, 2024]

Challenges and Directions

- Exchangeability: does not hold for complex designs, requiring a more elaborated approach (e.g., covariate shift, and *adaptive* strategies)
- Conditional coverage: when interest is in sub-population statistics (e.g., class-conditional, label-conditional) this is not ensured with standard CP $\hookrightarrow$ Mondrian Conformal Classification [Vovk et al., 2003]
- Combination of prediction sets remains an open problem (e.g., population size estimation)

References I

E. Bersson and P. D. Hoff. Optimal conformal prediction for small areas. *Journal of Survey Statistics and Methodology*, page smae010, 2024.

M. Gasparin and A. Ramdas. Conformal online model aggregation. *arXiv preprint arXiv:2403.15527*, 2024.

J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

R. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

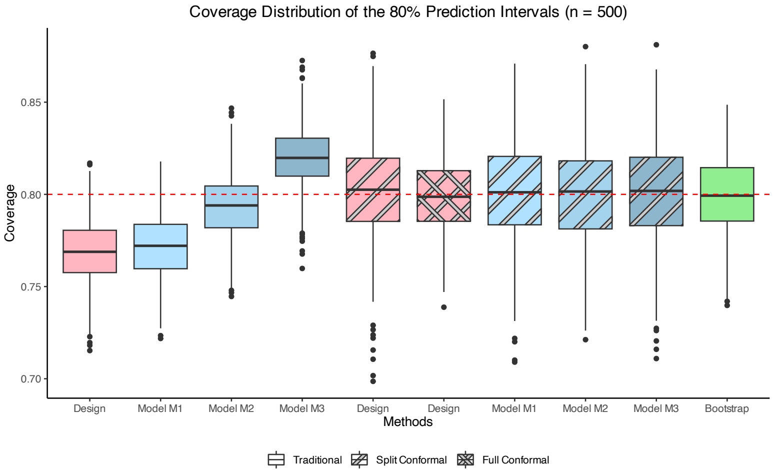
V. Vovk, D. Lindsay, I. Nouretdinov, and A. Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

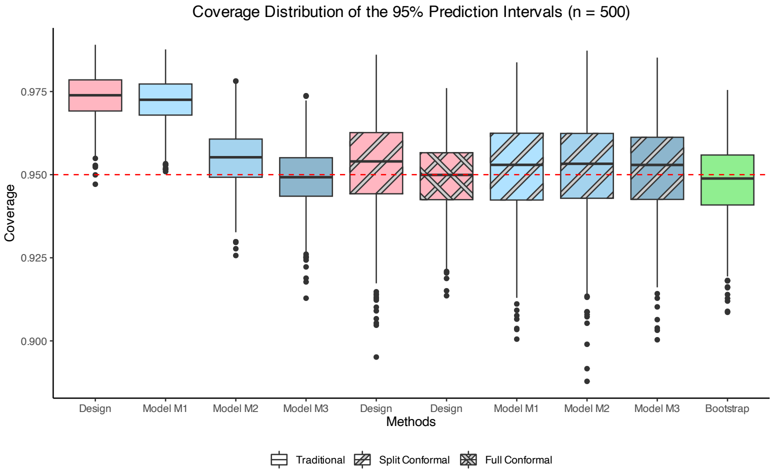
J. Wieczorek. Design-based conformal prediction. *Survey Methodology*, 49(2), 2024.

Appendix

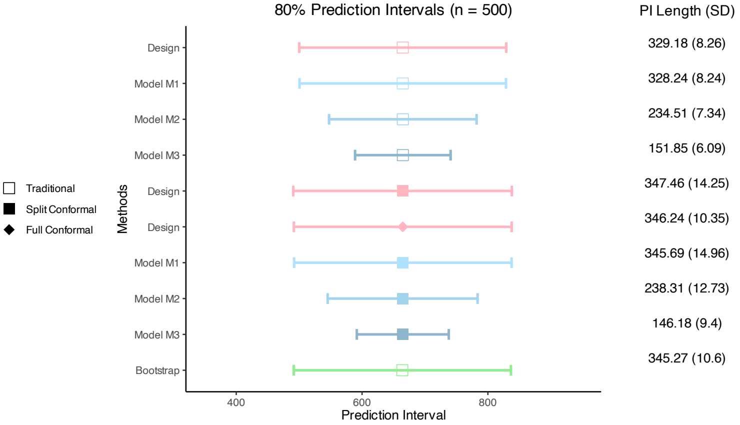
Simulations



Simulations



Simulations



Simulations