Lecture 18 — May 26, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Amber Hu*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and scribe notes written by Zhaomeng Chen and Disha Ghandwani.*

Outline

Agenda: Conformal Prediction

1. Split conformal
2. Jackknife+/CV+
3. Better conformity scores: Conformalized quantile regression

18.1 Split conformal

18.1.1 Review of full conformal

In Lecture 17, we discussed the problem of full conformal prediction to produce valid prediction intervals. To summarize, suppose we have training data $(X_1, Y_1), \cdots, (X_n, Y_n)$ and a test point $(X_{n+1}, ?)$ for which the true label is unknown. The data are assumed to be exchangeable, e.g. i.i.d. from some distribution P_{XY}. Our goal is to construct a **marginal distribution-free prediction interval** $C(X_{n+1})$ where

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha \tag{18.1}$$

for any (unknown) distribution P_{XY} and any sample size n.

To construct this prediction interval, we introduced a **non-conformity score function** $\mathcal{S}$ with two arguments, a point (x, y) and a multiset Z. A common example of the score function is given by $\mathcal{S}((x, y), Z) = |y - \hat{\mu}(x)|$, where $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ is a regression function fitted by running an algorithm $\mathcal{A}$ on (x, y) and Z.

Now suppose we are given a test point x and we want to calculate the prediction interval $\hat{C}_n(x)$. Let $Z_i = (X_i, Y_i)$ for $i = 1, ..., n$. Define

$$V_i^{(x,y)} = \mathcal{S}(Z_i, Z_{-i} \cup (x, y)),\ i = 1, \cdots, n, \quad \text{and } V_{n+1}^{(x,y)} = \mathcal{S}((x, y), Z_{1:n})$$

Now for a value y, we include y in $\hat{C}_n(x)$ if $V_{n+1}^{(x,y)} \leq \text{Quantile}(1 - \alpha; V_{1:n}^{(x,y)} \cup \{\infty\}))$, where $V_{1:n}^{(x,y)} = \{V_1^{(x,y)}, \cdots, V_n^{(x,y)}\}$. We showed that $\hat{C}_n$ indeed satisfies the condition in Eq. 18.1.

18.1.2 Motivation for split conformal

Overall, the full conformal approach is computationally expensive. For every given value of y, we need to refit the model to decide if we want to include the value y to our prediction interval. One solution to this issue is to use the **split conformal**, which we introduce in the next section.

Roughly speaking, the split conformal operates by training $\hat{\mu}$ once in total on an independent dataset. Our guarantees will be true conditional on $\hat{\mu}$. One drawback of split conformal is that it is more wasteful with the data than full conformal, as it uses only a portion of the data to fit $\hat{\mu}$ and the remaining portion of data to calibrate conformity scores.

18.1.3 Split conformal procedure

In split conformal, we have two data sets: a proper training set and a calibration set. First, we compute the score function $\mathcal{S}$ (i.e. we fit the model $\hat{\mu}$) on the proper training set. Then, on the calibration set we compute the scores

$$V_i = \mathcal{S}(X_i, Y_i), i = 1, \cdots, n, \text{ and } V_{n+1} = \mathcal{S}(x, y).$$

Finally, we construct conformal intervals as before.

The same result as in full conformal also holds for split conformal. This is again a simple consequence of the quantile lemma.

Theorem 1. If the calibration and the test point (X_{n+1}, Y_{n+1}) are exchangeable, then the split conformal prediction interval $\hat{C}(X_{n+1})$ satisfies

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

Note that although the split conformal is less computationally expensive, full conformal usually produces tighter intervals since it is more efficient with the amount of data given.

18.2 Jackknife+/CV+

We have seen that while split conformal is computationally efficient, it comes with the statistical cost of using only a portion of the data for calibration. When data is scarce, we want a method that re-uses data for both model fitting and calibration (in the spirit of cross-validation) and still has rigorous coverage guarantees. One method that accomplishes this is the Jackknife+/CV+ [1].

18.2.1 Jackknife+

Suppose we want to find a prediction interval such that

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 90\%$$

There are a few approaches we could try:

1. **Naive approach:** Compute the residuals on the training data,

$$R_i = |Y_i - \hat{\mu}(X_i)|$$

Then consider the prediction interval

$$\hat{\mu}(X_{n+1}) \pm 90\text{th percentile of residuals } R_i$$

However, due to the overfitting of $\hat{\mu}$ on the training data, the residuals R_i on the training set are typically smaller than the residual on the test point. Thus, this prediction interval will cover Y_{n+1} with probability less than 90%.

2. **Jackknife approach:** Compute leave-one-out residuals on the training data,

$$R_i^{\text{LOO}} = |Y_i - \hat{\mu}_{-i}(X_i)|$$

Then consider the prediction interval

$$\hat{\mu}(X_{n+1}) \pm 90\text{th percentile of LOO residuals } R_i^{\text{LOO}}$$

or equivalently

$$[10\text{th perc. } \{\hat{\mu}(X_{n+1}) - R_i^{\text{LOO}}\},\ 90\text{th perc. } \{\hat{\mu}(X_{n+1}) + R_i^{\text{LOO}}\}]$$

Intuitively, this approach should yield a valid prediction interval, since the leave-one-out residuals represent the typical error in predicting at a new test point. However, due to potential instability in fitting $\hat{\mu}$, this prediction interval may lose coverage.

3. **Jackknife+ approach:** Consider the prediction interval

$$[10\text{th perc. } \{\hat{\mu}_{-i}(X_{n+1}) - R_i^{\text{LOO}}\},\ 90\text{th perc. } \{\hat{\mu}_{-i}(X_{n+1}) + R_i^{\text{LOO}}\}]$$

By excluding the ith training point in fitting $\hat{\mu}$, we correct for potential instability of $\hat{\mu}$. We will see that the Jackknife+ has rigorous coverage guarantees.

Figure 18.1 compares prediction intervals for the Jackknife and Jackknife+.

Formally, we define the Jackknife+ prediction interval as

$$C^{\text{Jackknife+}}(X_{n+1}) = [q^-(\alpha, \{\hat{\mu}_{-i}(X_{n+1}) - R_i^{\text{LOO}}\}), q^+(1-\alpha, \{\hat{\mu}_{-i}(X_{n+1}) - R_i^{\text{LOO}}\})]$$

The following theorem provides a guarantee on the coverage rate of $C^{\text{Jackknife+}}(X_{n+1})$.

Theorem 2. If (X_i, Y_i), $i = 1, ..., n+1$ are exchangeable, then

$$\mathbb{P}(Y_{n+1} \in C^{\text{Jackknife+}}(X_{n+1})) \geq 1 - 2\alpha$$

In contrast, the Jackknife coverage can be zero, i.e. we can have

$$\mathbb{P}(Y_{n+1} \in C^{\text{Jackknife}}(X_{n+1})) = 0$$

To demonstrate the difference between Jackknife and Jackknife+, consider the following simple example. Suppose we have $n = 100$ samples and $d = 100$ features, and $Y \mid X$ follows a linear model. We fit a linear regression using least squares (minimal ℓ_2 norm solution). Note that since $n = d$ this solution is highly unstable. Table 18.1 shows average empirical coverage rates over 50 trials for the two methods.

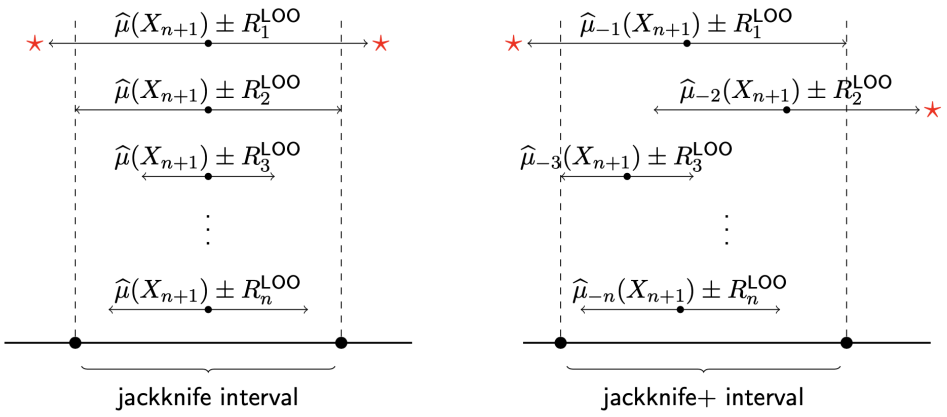

Figure 18.1. Comparison between Jackknife and Jackknife+. Intervals are such that on either side the boundary is exceeded by a sufficiently small proportion of the two sided arrows (marked with a star).

Method	Coverage
Jackknife	0.475
Jackknife+	0.913

Table 18.1. Empirical coverage results in linear regression setting where $n = d$.

18.2.2 CV+

CV+ is a generalization of Jackknife+ that uses folds of data to compute out-of-sample residuals, instead of computing leave-one-out residuals. The procedure is as follows. We split the data into K equal-sized folds (e.g. $K = 10$). Then, we compute out-of-sample residuals,

$$R_i^{\mathrm{CV}} = |Y_i - \hat{\mu}_{-S_{k(i)}}(X_i)|, \qquad i = 1, ..., n$$

where $\hat{\mu}_{-S_{k(i)}}$ denotes the model fit to the data excluding the fold containing the ith data point. The CV+ prediction interval is

$$C^{\mathrm{CV+}}(X_{n+1}) = [q^-(\alpha, \{\hat{\mu}_{-i}(X_{n+1}) - R_i^{\mathrm{CV}}\}), q^+(1-\alpha, \{\hat{\mu}_{-i}(X_{n+1}) - R_i^{\mathrm{CV}}\})]$$

Note that when $K = n$, CV+ reduces to Jackknife+. The result of Theorem 2 also holds for CV+, which provides a guarantee on the coverage rate.

18.2.3 Predictive sets: classification setting

In classification problems, we can use similar approaches to produce **prediction sets** instead of prediction intervals on a test data point. For example, suppose our labels are

$$Y_i \in \mathcal{Y} = \{\text{“red”}, \text{“blue”}, \text{“green”}\}$$

where $\mathcal{Y}$ is a discrete and unordered set. Given (X_i, Y_i) for $i = 1, ..., n$, our goal is to construct a prediction set such that

$$\mathbb{P}(Y_{n+1} \in \hat{\mathcal{S}}(X_{n+1})) \geq 90\%$$

We take a Jackknife+ approach as follows. Our input is an algorithm $\hat{\pi}(y \mid x)$ which estimates the probability that $Y = y$ given $X = x$ (e.g. the output of a neural network's softmax layer). Then we construct the prediction set

$$\hat{\mathcal{S}}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \sum_{i=1}^{n} \mathbb{1}\{\hat{\pi}_{-i}(y \mid X_{n+1}) \geq \hat{\pi}_{-i}(Y_i \mid X_i)\} \geq \alpha(n+1) \right\}$$

Intuitively, $\hat{\mathcal{S}}(X_{n+1})$ contains elements $y \in \mathcal{Y}$ such that the predicted probability of (X_{n+1}, y) is in the top 90% of LOO probabilities.

The following theorem [4] provides a coverage guarantee for Jackknife+ prediction sets.

Theorem 3. If (X_i, Y_i), $i = 1, ..., n+1$ are exchangeable, then

$$\mathbb{P}(Y_{n+1} \in \hat{\mathcal{S}}(X_{n+1})) \geq 1 - 2\alpha.$$

In [4], experiments were performed on the Fashion-MNIST and MNIST datasets, with 10 classes, 50 features (after PCA), and 1000 training examples. Figure 18.2 displays the empirical coverage and set size of full conformal, split conformal, and Jackknife+ averaged over 20 train/test splits. Note that all three methods guarantee 90% coverage, and the Jackknife+ produces the smallest prediction sets.

Classifier	Method	Fashion-MNIST		MNIST	
		Coverage	Set Size	Coverage	Set Size
Logistic	Full conformal	0.897	1.623	0.900	1.248
	Split conformal	0.893	1.552	0.898	1.387
	Jackknife+	0.897	**1.407**	0.897	**1.177**
Random Forests	Full conformal	0.895	1.497	0.901	1.324
	Split conformal	0.895	1.651	0.901	1.590
	Jackknife+	0.903	**1.473**	0.909	**1.288**
Kernel SVM	Full conformal	0.898	1.901	0.904	1.098
	Split conformal	0.894	1.382	0.899	1.092
	Jackknife+	0.897	**1.266**	0.898	**0.966**
Neural Net	Full conformal	0.899	3.942	0.898	2.733
	Split conformal	0.893	1.818	0.897	1.270
	Jackknife+	0.915	**1.499**	0.913	**1.041**

Figure 18.2. Comparison between full conformal, split conformal, and Jackknife+ in producing prediction sets for Fashion-MNIST and MNIST.

18.3 Better conformity scores

Now, we will discuss how to find better conformity scores. An important learning goal is to estimate the quantiles of the conditional distribution of Y given x, as in Figure 18.3. If we

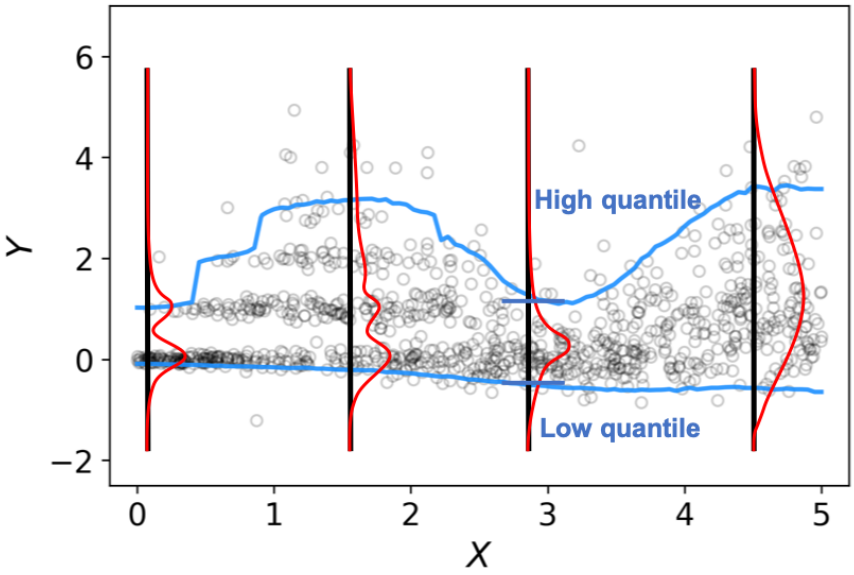

Figure 18.3. Estimated and true quantiles for $Y \mid X$.

knew that

$$\mathbb{P}(q_1(x) \leq Y \leq q_2(x) \,|\, X = x) \geq 1 - \alpha,$$

then the band $[q_1(x), q_2(x)]$ would be a $(1-\alpha)$th prediction interval for Y. However, we do not have sufficient data to capture the conditional distribution of $Y|X = x$ for each value of x, so we have to try a few other methods.

18.3.1 Quantile estimation as a learning task

We can estimate quantiles of the conditional distribution $Y \mid X$ by $\hat{f}$ which minimizes the following loss function over a class of functions $\mathcal{F}$:

$$\hat{f}(.) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_i \rho_\alpha(Y_i - f(X_i)) + \mathcal{R}(f),$$

where $\mathcal{R}(\cdot)$ is a possible regularizer and ρ_α is the pinball loss given by

$$\rho_\alpha(y) = y(\alpha - \mathbb{I}_{(y<0)}).$$

However, the upper and lower quantiles found by the above method would not give valid predictive range for future data points, as they are uncalibrated. This is shown in Figure 18.4. We would need to calibrate the intervals by conformal inference to get valid coverage.

18.3.2 Conformal Quantile Regression

Conformal Quantile Regression (CQR) [3] is a method to construct prediction intervals that attain valid coverage with finite samples. The CQR method consists of five steps.

1. Split the data into a proper training set and a calibration set.

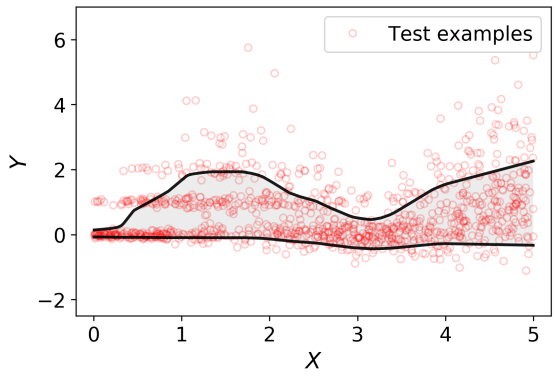

Figure 18.4. Estimated quantiles from quantile regression on a test set. These intervals are not valid: the target coverage is 90%, while the actual coverage is around 72%.

2. Fit two conditional quantile functions, $\text{lower}(x)$ and $\text{upper}(x)$, on the proper training set.

3. For the ith point in the calibration set, compute

$$S_i = \max\{\text{lower}(X_i) - Y_i, Y_i - \text{upper}(X_i)\}$$

   Note that S_i is a signed distance to the boundary, where S_i is negative if $\text{lower}(X_i) \leq Y_i \leq \text{upper}(X_i)$ and positive otherwise.

4. Compute Q as the $(1-\alpha)$th quantile of the S_i's. Intuitively, Q is positive if the initial intervals are too narrow.

5. Define the prediction interval as

$$C(x) = [\text{lower}(x) - Q, \text{upper}(x) + Q]$$

Figure 18.5 illustrates the steps of the CQR method.

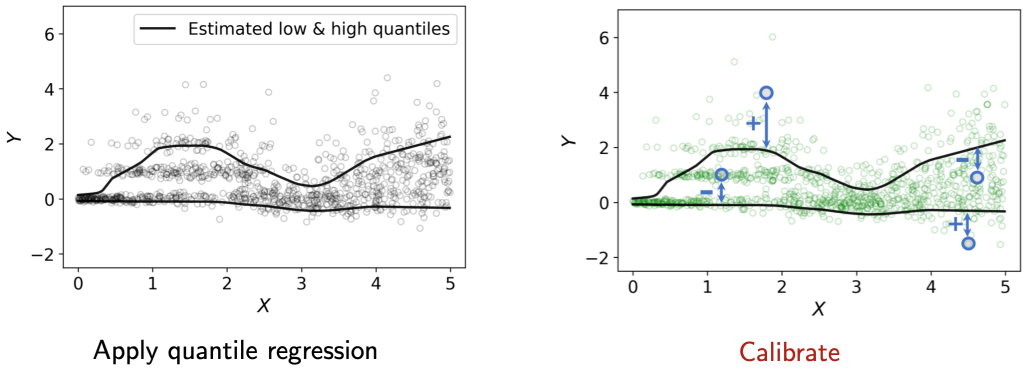

Figure 18.5. An illustration of the CQR method.

The following result [3] provides coverage guarantees for the CQR prediction interval.

Theorem 4. If $(X_i, Y_i), i = 1, \cdots, n+1$, are exchangeable, then the prediction interval $C(X_{n+1})$ constructed by the split CQR algorithm satisfies

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

18.3.3 CQR versus Split Conformal

Recall the split conformal method we discussed in Lecture 9, where the prediction interval for Y_{n+1} was given by

$$C(X_{n+1}) = [\hat{\mu}(X_{n+1}) - q, \hat{\mu}(X_{n+1}) + q]$$

CQR performs better than split conformal in the following settings.

In case of heteroskedasticity, CQR will perform better than split conformal, as split conformal produces intervals of constant width while CQR produces adaptive intervals, as in Figure 18.6. Observe that both have similar average coverage (as guaranteed by the theorems), but the average length of CQR is much smaller than that of split conformal.

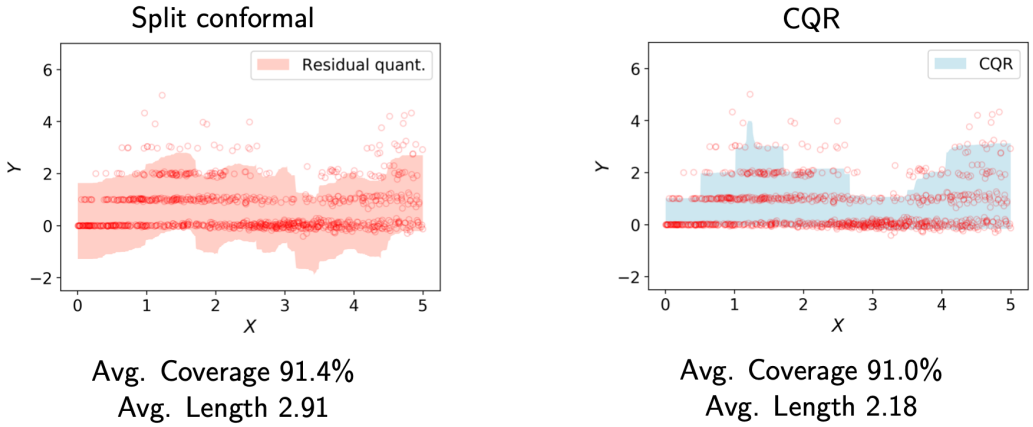

Figure 18.6. Comparison of fixed vs. adapted prediction intervals.

It is intuitive that CQR will provide better conditional coverage than split conformal, as shown in Figure 18.7. As the intervals are not adaptive for split conformal, coverage will be high for some values of x but low for the other values of x.

As $n \to \infty$, prediction intervals provided by CQR are consistent with the quantiles of conditional distribution $Y \mid X$. Prediction intervals provided by split conformal are consistent only if $Y = f(X) + \epsilon$ and ϵ is symmetric.

18.3.4 Comparison of Split Conformal and CQR using Medical Services data

The Medical Expenditure Panel Survey (MEPS) 2015 comprises of information on 16,000 subjects on 140 features including age, marital status, race, poverty status, functional limitations, health status, and health insurance types. We want to predict health care system

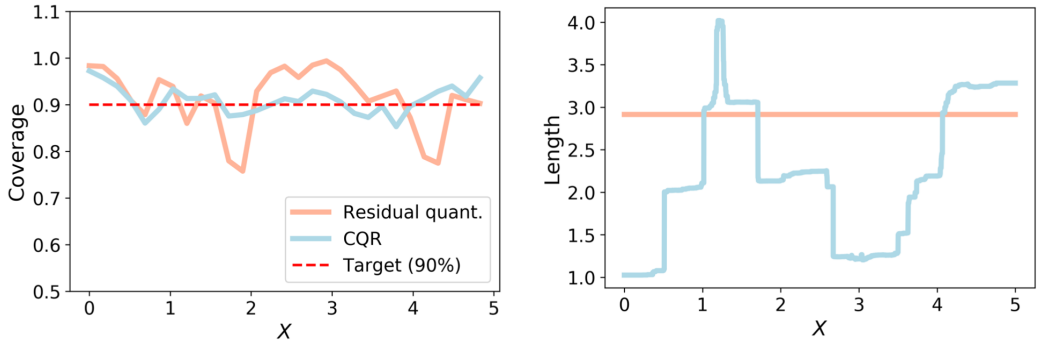

Figure 18.7. Approximate conditional coverage (left) and adapted length (right).

utilization (reflected by the number of visits to doctor's office and hospital, etc.) using the above covariates.

Results on MEPS data

Split conformal and CQR were performed on MEPS data. We calculate marginal coverage, conditional coverage (due to [2], conditional coverage is measured on the worst slab), and the length of the intervals for both CQR and split conformal, averaged over 20 random train-test (80%/20%) splits. Results are shown in Figure 18.8. We can see that CQR outperforms split conformal by having higher conditional coverage and smaller interval length.

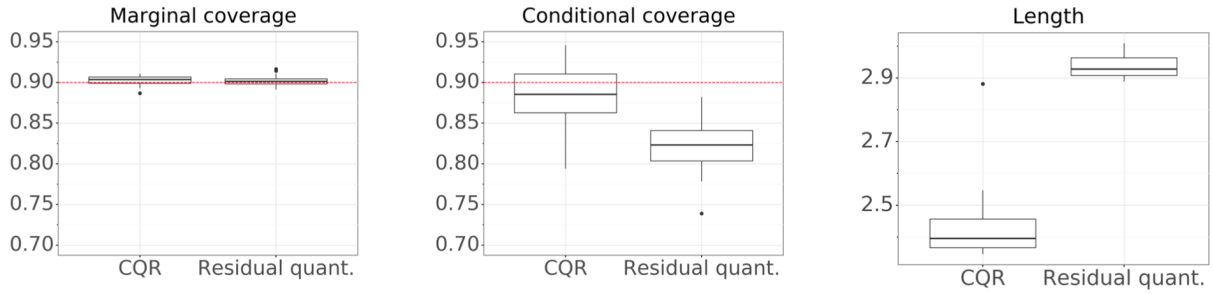

Figure 18.8. Results of split conformal and CQR applied to MEPS data.

Bibliography

[1] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

[2] Maxime Cauchois, Suyash Gupta, and John Duchi. Knowing what you know: valid confidence sets in multiclass and multilabel prediction. 2020.

[3] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[4] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.