Predictive inference with the jackknife+

Rina Foygel Barber, Emmanuel J. Candés, Aaditya Ramdas, Ryan J. Tibshirani

Kiho Park

parkkiho@uchicago.edu

1 Introduction

Suppose that we have i.i.d. training data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \cdots, n$, and a new test point (X_{n+1}, Y_{n+1}) drawn independently from the same distribution. We would like to construct a prediction interval $\widehat{C}_{n,\alpha}$ such that $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha$. To avoid the overfitting problem, the jackknife prediction method computes a margin of error with a leave-one-out construction:

(1) For each $i = 1, \cdots, n$, fit the regression function $\widehat{\mu}_{-i}$ to the training data with the ith point removed, and compute the corresponding leave-one-out residual, $|Y_i - \widehat{\mu}_{-i}(X_i)|$.

(2) Fit the regression function $\widehat{\mu}$ to the full training data, and output the prediction interval

$$\widehat{\mu}(X_{n+1}) \pm (\text{the } (1-\alpha) \text{ quantile of } |Y_1 - \widehat{\mu}_{-1}(X_1)|, \cdots, |Y_n - \widehat{\mu}_{-n}(X_n)|)$$

Intuitively, this method should have the right coverage properties on average since it avoids overfitting. However, the jackknife procedure does not have any universal theoretical guarantees. In this paper, we introduce a new method, the jackknife+, that provides non-asymptotic coverage guarantees under no assumptions beyond the training and test data being exchangeable.

1.1 Background

The idea of resampling or subsampling from the available data, in order to assess the accuracy of our parameter estimates or predictions, has a rich history in the statistics literature. Jackknife, bootstrap, and cross-validation are developed. The performance of the jackknife was demonstrated, but relatively little theoretical properties have been known.

In contrast to cross-validation methods, we can instead consider a simple validation or holdout method, which was studied under the name split conformal prediction. This method is also computationally very cheap, whereas jackknife and cross-validation methods require running the regression many times. However, these benefits come at a statistical cost with large training data. Finally, Vovk (2015) proposed the cross-conformal prediction method, which is closely related to the jackknife+.

1.2 Notation

For any values v_i indexed by $i = 1, \cdots, n$, define $\widehat{q}^{+}_{n,\alpha}\{v_i\}$ as the $\lceil (1-\alpha)(n+1) \rceil$-th smallest value of $v_1, \cdots, v_n$, and $\widehat{q}^{-}_{n,\alpha}\{v_i\}$ as the $\lfloor \alpha(n+1) \rfloor$-th smallest value of $v_1, \cdots, v_n$ $(= -\widehat{q}^{+}_{n,\alpha}\{-v_i\})$. The naive preidction interval can be defined as

$$\widehat{C}^{\text{navie}}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}^{+}_{n,\alpha}\{Y_i - \widehat{\mu}(X_i)\}$$

We assume the algorithm $\mathcal{A} : \cup_{m \geq 1} (\mathbb{R}^d \times \mathbb{R})^m \to \{f \mid f : \mathbb{R}^d \to \mathbb{R}\}$ satisfies a symmetry condition. Let

$$\widehat{\mu}_{-i} = \mathcal{A}\left((X_1, Y_1), \cdots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \cdots, (X_n, Y_n)\right)$$

Then, the jackkife prediction interval can be written as

$$\begin{aligned}\widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1}) &= \widehat{\mu}(X_{n+1}) \pm \widehat{q}^{+}_{n,\alpha}\{R^{\text{LOO}}_i\} \\ &= \left[\widehat{q}^{-}_{n,\alpha}\left\{\widehat{\mu}(X_{n+1}) - R^{\text{LOO}}_i\right\}, \widehat{q}^{+}_{n,\alpha}\left\{\widehat{\mu}(X_{n+1}) + R^{\text{LOO}}_i\right\}\right]\end{aligned}$$

where $R^{\text{LOO}}_i = |Y_i - \widehat{\mu}_{-i}(X_i)|$ denotes the ith leave-one-out residual.

2 The jackknife+

The jackknife+ prediction interval is given by:

$$\widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1}) = \left[\widehat{q}^{-}_{n,\alpha}\left\{\widehat{\mu}_{-i}(X_{n+1}) - R^{\text{LOO}}_i\right\}, \widehat{q}^{+}_{n,\alpha}\left\{\widehat{\mu}_{-i}(X_{n+1}) + R^{\text{LOO}}_i\right\}\right]$$

While the jackknife interval $\widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})$ is defined as a symmetric interval around the prediction $\widehat{\mu}(X_{n+1})$ for the test point, the jackknife+ interval can be interpreted as an interval around the median prediction, $\text{Median}\left(\widehat{\mu}_{-1}(X_{n+1}), \cdots, \widehat{\mu}_{-n}(X_{n+1})\right)$ which is guaranteed to lie inside $\widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})$ for $\alpha \leq \frac{1}{2}$.

2.1 Assumption-free guarantees

Theorem 1. *The jackknife+ prediction interval satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - 2\alpha$$

Theorem 2. *For any sample size $n \geq 2$, any $\alpha \in \left[\frac{1}{n+1}, 1\right]$, and any dimension $d \geq 1$, there exists a distribution on $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ and a regression algorithm $\mathcal{A}$, such that the predictive coverage of the naive prediction interval and the jackknife prediction interval satisfy*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{naive}}_{n,\alpha}(X_{n+1})\right\} = \mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})\right\} = 0$$

Furthermore, if $\alpha \leq \frac{1}{2}$, there exists a distribution on $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ and a regression algorithm $\mathcal{A}$, such that the predictive coverage of jackknife+ satisfies

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\right\} \leq 1 - 2\alpha + 6\sqrt{\frac{\log n}{n}}$$

2.2 The jackknife-minimax method

We define the jackknife-minimax method as follows:

$$\widehat{C}^{\text{jack-mm}}_{n,\alpha}(X_{n+1}) = \left[\min_{i=1,\cdots,n} \widehat{\mu}_{-i}(X_{n+1}) - \widehat{q}^{+}_{n,\alpha}\left\{R^{\text{LOO}}_i\right\}, \max_{i=1,\cdots,n} \widehat{\mu}_{-i}(X_{n+1}) + \widehat{q}^{+}_{n,\alpha}\left\{R^{\text{LOO}}_i\right\}\right]$$

This interval is strictly more conservative than jackknife+, i.e.,

$$\widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1}) \subseteq \widehat{C}^{\text{jack-mm}}_{n,\alpha}(X_{n+1})$$

Theorem 3. *The jackknife-minimax prediction interval satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jack-mm}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha$$

3 CV+ for K-fold cross-validation

Suppose that we split the training sample into K disjoint subsets $S_1, \cdots, S_K$ each of size $m = n/K$. Let

$$\widehat{\mu}_{-S_k} = \mathcal{A}\left((X_i, Y_i) : i \in \{1, \cdots, n\} \setminus S_k\right)$$

We wconsider the residuals from this K-fold process:

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

where $k(i) \in \{1, \cdots, K\}$ satisfies $i \in S_{k(i)}$. Then, we can define the CV+ prediction interval as

$$\widehat{C}^{\mathrm{CV+}}_{n,K,\alpha}(X_{n+1}) = \left[\widehat{q}^{-}_{n,\alpha}\left\{\widehat{\mu}_{-S_{k(i)}}(X_{n+1}) - R_i^{\mathrm{CV}}\right\}, \widehat{q}^{+}_{n,\alpha}\left\{\widehat{\mu}_{-S_{k(i)}}(X_{n+1}) + R_i^{\mathrm{CV}}\right\}\right]$$

Of course, jackknife+ can be viewed as a special case of CV+, by setting $K = n$.

3.1 Assumption-free guarantee for CV+

Theorem 4. *The K-fold CV+ prediction interval satisfies the following coverage guarantees:*

$$\begin{aligned}\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\mathrm{CV+}}_{n,K,\alpha}(X_{n+1})\right\} &\geq 1 - 2\alpha - \min\left\{\frac{2(1-1/K)}{n/K+1}, \frac{1-K/n}{K+1}\right\} \\ &\geq 1 - 2\alpha - \sqrt{2/n}\end{aligned}$$

Holdout or split conformal method is equivalent to using only a single subset S_1. This method offers an assumption-free guarantee of $1-\alpha$ coverage, but this comes at the cost of higher variance due to the single split. In contrast, CV+ reduces variance by averaging over all K splits, but at the cost of a weaker theoretical guarantee.

3.2 Related method: cross-conformal predictors

Our proposed CV+ prediction interval is related to the cross-conformal prediction method of Vovk (2015), which returns the predictive set

$$\begin{aligned}&\widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1}) \\ &= \left\{y \in \mathbb{R} : \frac{\tau + \sum_{i=1}^{n} \mathbb{1}\left\{\left|y - \widehat{\mu}_{-S_{k(i)}}(X_{n+1})\right| < R_i^{\mathrm{CV}}\right\} + \tau \mathbb{1}\left\{\left|y - \widehat{\mu}_{-S_{k(i)}}(X_{n+1})\right| = R_i^{\mathrm{CV}}\right\}}{n+1} > \alpha\right\}\end{aligned}$$

where $\tau \sim \mathrm{Unif}[0,1]$. We can verify that

$$\widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1}) \subseteq \widehat{C}^{\mathrm{CV+}}_{n,K,\alpha}(X_{n+1})$$

3.3 An alternative method: conformal prediction

The final related method we present is conformal prediction. The full conformal prediction method outputs a prediction set constructed as follows:

$$\widehat{C}^{\mathrm{conf}}_{n,\alpha}(X_{n+1}) = \{y \in \mathbb{R} : |y - \widehat{\mu}^{y}(X_{n+1})| \leq \widehat{q}^{+}_{n,\alpha}\{Y_i - \widehat{\mu}^{y}(X_i)\}\}$$

where

$$\widehat{\mu}^{y} = \mathcal{A}\left((X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, y)\right)$$

The exchangeability of the test and training data ensures that

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

4 Summary of coverage guarantees and computational costs

Table: Summary of theoretical results and typical empirical performance

Method	Assumption-free theory	Typical empirical coverage
Naive	No guarantee	$< 1-\alpha$
Split conf. (holdout)	$\geq 1-\alpha$ coverage	$\approx 1-\alpha$
Jackknife	No guarantee	$\approx 1-\alpha$, or $< 1-\alpha$ if $\widehat{\mu}$ unstable
Jackknife+	$\geq 1-2\alpha$ coverage	$\approx 1-\alpha$
Jackknife-minimax	$\geq 1-\alpha$ coverage	$> 1-\alpha$
Full conformal	$\geq 1-\alpha$ coverage	$\approx 1-\alpha$, or $> 1-\alpha$ if $\widehat{\mu}$ overfits
K-fold CV+	$\geq 1-2\alpha$ coverage	$\gtrsim 1-\alpha$
K-fold cross-conf.	$\geq 1-2\alpha$ coverage	$\gtrsim 1-\alpha$

We recommend the jackknife+ as a practical alternative to the usual jackknife predictive intervals. On the other hand, the methods achieving $1-\alpha$ coverage guarantees are either less statistically efficient in the sense of producing wider intervals or suffer from computational infeasibility.

Table: Summary of computational costs

Method	Model training cost	Model evaluation cost
Naive	1	$n + n_{\text{test}}$
Split conf. (holdout)	1	$n + n_{\text{test}}$
Jackknife	n	$n + n_{\text{test}}$
Jackknife+	n	$n_{\text{test}} \cdot n$
Jackknife-minimax	n	$n_{\text{test}} \cdot n$
K-fold CV+	K	$n + n_{\text{test}} \cdot K$
K-fold cross-conf.	K	$n + n_{\text{test}} \cdot K$
Full conformal	$n_{\text{test}} \cdot n_{\text{grid}}$	$n_{\text{test}} \cdot n_{\text{grid}} \cdot n$

In most settings, the model training cost is dominant. If the training sample size n is so large, we may instead prefer to use K-fold cross-validation. In contrast, the full conformal method must train $\mathcal{A}$ many more times.

5 Guarantees under stability assumptions

We consider how adding stability assumptions - conditions that ensure that the fitted regression function $\widehat{\mu}$ is not too sensitive to perturbations of the training data set - can improve the theoretical guarantees of the jackknife and its variants.

5.1 In-sample and out-of-sample stability

Fix any $\epsilon \geq 0$, $\nu \in [0,1]$, any sample size $n \geq 2$, and any distribution P on (X,Y). We say that a regression algorithm $\mathcal{A}$ satisfies (ϵ,ν) out-of-sample stability with respect to the distribution P and sample size n if, for all $i \in \{1,\cdots,n\}$,

$$\mathbb{P}\left\{|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})| \leq \epsilon\right\} \geq 1-\nu$$

Similarly, $\mathcal{A}$ satisfies (ϵ, ν) in-sample stability with respect to the distribution P and sample size n if, for all $i \in \{1, \cdots, n\}$,

$$\mathbb{P}\left\{|\widehat{\mu}(X_i) - \widehat{\mu}_{-i}(X_i)| \leq \epsilon\right\} \geq 1 - \nu$$

In a scenario where the model fitting algorithm suffers from strong overfitting, we would expect in-sample stability to be very poor, while out-of-sample stability may still hold.

5.2 Summary of stability results

Table: Summary of predictive coverage guarantees, with and without stability assumptions

Method	Assumption-free theory	Out-of-sample stability	In-sample and out-of-sample stability
Naive	No guarantee	No guarantee	$\approx 1 - \alpha$
Jackknife	No guarantee	$\approx 1 - \alpha$	$\approx 1 - \alpha$
Jackknife+	$1 - 2\alpha$	$\approx 1 - \alpha$	$\approx 1 - \alpha$
Jackknife-minimax	$1 - \alpha$	$1 - \alpha$	$1 - \alpha$

5.3 Out-of-sample stability and the jackknife

Define ϵ-inflated versions of the jackknife and jackknife+ predictive intervals:

$$\widehat{C}_{n,\alpha}^{\text{jackknife},\epsilon}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \left(\widehat{q}^{+}_{n,\alpha}\{R_i^{\text{LOO}}\} + \epsilon\right)$$

$$\widehat{C}_{n,\alpha}^{\text{jackknife+},\epsilon}(X_{n+1}) = \left[\widehat{q^{-}}_{n,\alpha}\left\{\widehat{\mu}_{-i}(X_{n+1}) - R_i^{\text{LOO}}\right\} - \epsilon, \widehat{q}^{+}_{n,\alpha}\left\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{\text{LOO}}\right\} + \epsilon\right]$$

Theorem 5. *Suppose that the regression algorithm satisfies the (ϵ, ν) out-of-sample stability property with respect to the data distribution P and the sample size n. Then the ϵ-inflated jackknife prediction interval satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{jackknife},\epsilon}(X_{n+1})\right\} \geq 1 - \alpha - 2\sqrt{\nu}$$

Similarly, the 2ϵ-inflated jackknife+ prediction interval satisfies

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{jackknife+},2\epsilon}(X_{n+1})\right\} \geq 1 - \alpha - 4\sqrt{\nu}$$

We remark that if we additionally assume that, in the data distribution, $Y \mid X$ has a bounded conditional density, then the result of Theorem 5 is sufficient to ensure that the (non-inflated) jackknife and jackknife+ intervals achieve close to target coverage.

5.4 In-sample stability and overfitting

The ϵ-inflated version of the naive prediction interval is defined as

$$\widehat{C}_{n,\alpha}^{\text{naive},\epsilon}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \left(\widehat{q}^{+}_{n,\alpha}\{|Y_i - \widehat{\mu}(X_i)|\} + \epsilon\right)$$

Theorem 6. *Suppose that the regression algorithm $\mathcal{A}$ satisfies both the (ϵ, ν) in-sample stability property and the (ϵ, ν) out-of-sample stability property with respect to the data distribution P and the sample size n. Then the 2ϵ-inflated naive prediction interval satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{naive},2\epsilon}(X_{n+1})\right\} \geq 1 - \alpha - 4\sqrt{\nu}$$

5.5 Example: K-nearest-neighbors

To give an illustrative example, consider a K-nearest-neighbor (K-NN) method. Given a training data set $(X_1, Y_1), \cdots, (X_n, Y_n)$ and a new test point x, our prediction is

$$\widehat{\mu}(x) = \frac{1}{K} \sum_{i \in N(x)} Y_i$$

where $N(x) \subset \{1, \cdots, n\}$ is the set of the K indices i giving the smallest values of $\|X_i - x\|_2$. Then,

$$i \notin N(X_{n+1}) \iff N(X_{n+1}) = N_{-i}(X_{n+1}) \implies \widehat{\mu}(X_{n+1}) = \widehat{\mu}_{-i}(X_{n+1})$$

Therefore,

$$\mathbb{P}\left\{|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})| = 0\right\} \geq \mathbb{P}\{i \notin N(X_{n+1})\} = 1 - \frac{K}{n}$$

It means that K-NN method satisfies (ϵ, ν) out-of-sample stability with $\epsilon = 0$ and $\nu = K/n$. Applying the conclusion of Theorem 5 to this setting, we see that K-NN leads to a coverage rate at lest

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha - 2\sqrt{K/n}$$

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha - 4\sqrt{K/n}$$

5.6 Comparison to existing results

The theory obtained in previous work is asymptotic, and relies also on distributional assumptions. In contrast, our guarantee, in Theorem 5, offers a simple finite-sample coverage guarantee with no distributional assumptions, requiring only algorithm stability.

6 Proof of Theorem 1

suppose for a moment that we have access to the test point (X_{n+1}, Y_{n+1}) as well as the training data. For any indices $i, j \in \{1, \cdots, n+1\}$ with $i \neq j$, let $\widetilde{\mu}_{-(i,j)}$ define the regression function fitted on the training plus test data, with points i and j removed. Note that $\widetilde{\mu}_{-(i,j)} = \widetilde{\mu}_{-(j,i)}$ for any $i \neq j$, and that $\widetilde{\mu}_{-(i,n+1)} = \widehat{\mu}_{-i}$ for $i = 1, \cdots, n$. We define a matrix of residuals, $R \in \mathbb{R}^{(n+1)\times(n+1)}$, with entries

$$R_{ij} = \begin{cases} +\infty & i = j \\ |Y_i - \widetilde{\mu}_{-(i,j)}(X_i)| & i \neq j \end{cases}$$

We also define a comparison matrix, $A \in \{0,1\}^{(n+1)\times(n+1)}$, with entries $A_{ij} = \mathbb{1}\{R_{ij} > R_{ji}\}$. Naturally we see that $A_{ij} = 1$ implies $A_{ji} = 0$, for any i, j. We will define a set $\mathcal{S}(A) \subseteq \{1, \cdots, n+1\}$ of "strange" points,

$$\mathcal{S}(A) = \{i \in \{1, \cdots, n+1\} : A_{i\bullet} \geq (1-\alpha)(n+1)\}$$

where $A_{i\bullet} = \sum_{j=1}^{n+1} A_{ij}$ is the ith row sum of A.

From this point on, the proof proceeds as follows:

Step 1: we estabilish deterministically that $|\mathcal{S}(A)| \leq 2\alpha(n+1)$.

Step 2: using the exchangeablity, we show that the probability that the test point $n+1$ is strange is therefore bounded by 2α.

Step3: finally, we verify that the jackknife+ interval can only fail to cover the test response value Y_{n+1} if $n+1$ is strange point.

7 Empirical results

In this section, we compare seven method - naive, jackknife, jackknife+, jackknife-minimax, CV+, split conformal, and full conformal - on simulated and real data.

7.1 Simulations

We first examine the performance of the various prediction intervals on a simulated example, using least squares as our regression method.

7.1.1 Data and methods

Fix $1 - \alpha = 0.9$. We use $n = 100$, and $d = 5, 10, \cdots, 200$, with i.i.d. data points (X_i, Y_i) generated as

$$X_i \sim \mathcal{N}(0, I_d) \text{ and } Y_i \mid X_i \sim \mathcal{N}(X_i^T \beta, 1)$$

where $\beta = \sqrt{10} \cdot u$ for a uniform random unit vector $u \in \mathbb{R}^d$. The regression method $\mathcal{A}$ is simply least squares. We then generate 100 test data points from the same distribution, and calculate the empirical probability of coverage and the average width of the prediction interval.

7.1.2 Results

Figure displays the results of the simulation, averaged over 50 trials. When $d < n$, the jackknife and jackknife+ show very similar performance. For $d \approx n$, the jackknife has substantial undercoverage, but the jackknife+ continues to show the right coverage level, at the cost of a slightly wider interval. For large d, the jackknife and jackknife+ again show very similar performance. In fact, this connects to recent work on interpolation methods that achieve zero training error.

As expected, the jackknife-minimax is over-conservative across all dimensions d, while the naive method drastically undercovers. When $d > n$, we note that full conformal prediction will always have infinite length intervals with coverage equal to one. Split conformal is the only method other than jackknife+ to maintain coverage at 0.9 throughout.

7.2 Real data

We next compare the various methods on three real data sets. We try three regression algorithms: ridge regression, random forests, and neural networks.

7.2.1 Data

The Communities and Crime data set contains information on 1994 communities, with $d = 99$ covariates and the goal of predicting a response variable defined as the per capita violent crime rate.

The BlogFeedback data set contains 52397 data points, with $d = 280$ covariates and the goal of predicting the response variable of the number of comments left on the blog post in the following 24 hours.

The Medical Expenditure Panel Survey 2016 data set contains 33005 data points, with $d = 107$ covariates and the goal of predicting the health care system utilization of each individual.

7.2.2 Methods

We randomly sample $n = 200$ data points from the full data set, to use as the training data. The remaining points form the test set. For each choice of $\mathcal{A}$, we construct six prediction intervals, and calculate their empirical coverage rate and their average width on the test set. We then repeat this procedure 20 times.

7.2.3 Results

Figure displays the results of the real data experiments. In general, the naive method undercovers while jackknife-minimax overcovers, and the remaining methods have well calibrated coverage. In terms of their interval lengths, we typically get the expected order: jackknife < jackknife+ < 10-fold CV+ < split conformal.

8 Summary

The jackknife+ differs from the jackknife in that it uses the quantiles of

$$\widehat{\mu}_{-i}(X_{n+1}) \pm R_i^{\mathrm{LOO}} = \widehat{\mu}(X_{n+1}) + (\widehat{\mu}_{-i}(X_{n+1}) - \widehat{\mu}(X_{n+1})) \pm R_i^{\mathrm{LOO}}$$

instead of those of $\widehat{\mu}(X_{n+1}) \pm R_i^{\mathrm{LOO}}$, to build predictive intervals. By applying the shifts $\widehat{\mu}_{-i}(X_{n+1}) - \widehat{\mu}(X_{n+1})$, the jackknife+ effectively accounts for the algorithm instability, yielding rigorous coverage guarantees. This makes it a better choice than the jackknife in practice. In cases where the jackknife+ is computationally prohibitive, K-fold CV+ offers an attractive alternative.

References

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):9-28, 2015.

9 Question

1. Because you are the author of the paper, I would like to ask one question about the process. How did you come up with the idea of jackknife+? Also, how did you get the idea of proof like the matrix of residuals?