Conformalized Quantile Regression

Yaniv Romano
Department of Statistics
Stanford University

Evan Patterson
Department of Statistics
Stanford University

Emmanuel J. Candès
Departments of Mathematics and of Statistics
Stanford University

Abstract

Conformal prediction is a technique for constructing prediction intervals that attain valid coverage in finite samples, without making distributional assumptions. Despite this appeal, existing conformal methods can be unnecessarily conservative because they form intervals of constant or weakly varying length across the input space. In this paper we propose a new method that is fully adaptive to heteroscedasticity. It combines conformal prediction with classical quantile regression, inheriting the advantages of both. We establish a theoretical guarantee of valid coverage, supplemented by extensive experiments on popular regression datasets. We compare the efficiency of conformalized quantile regression to other conformal methods, showing that our method tends to produce shorter intervals.

1 Introduction

In many applications of regression modeling, it is important not only to predict accurately but also to quantify the accuracy of the predictions. This is especially true in situations involving high-stakes decision making, such as estimating the efficacy of a drug or the risk of a credit default. The uncertainty in a prediction can be quantified using a prediction interval, giving lower and upper bounds between which the response variable lies with high probability. An ideal procedure for generating prediction intervals should satisfy two properties. First, it should provide valid coverage in finite samples, without making strong distributional assumptions, such as Gaussianity. Second, its intervals should be as short as possible at each point in the input space, so that the predictions will be informative. When the data is heteroscedastic, getting valid but short prediction intervals requires adjusting the lengths of the intervals according to the local variability at each query point in predictor space. This paper introduces a procedure that performs well on both criteria, being distribution-free and adaptive to heteroscedasticity.

Our work is heavily inspired by *conformal prediction*, a general methodology for constructing prediction intervals [1–6]. Conformal prediction has the virtue of providing a nonasymptotic, distribution-free coverage guarantee. The main idea is to fit a regression model on the training samples, then use the residuals on a held-out validation set to quantify the uncertainty in future predictions. The effect of the underlying model on the length of the prediction intervals, and attempts to construct intervals with locally varying length, have been studied in numerous recent works [6–16]. Nevertheless, existing methods yield conformal intervals of either fixed length or length depending only weakly on the predictors, as argued in [6, 15, 17].

Quantile regression [18] offers a different approach to constructing prediction intervals. Take any algorithm for *quantile regression*, i.e., for estimating conditional quantile functions from data. To obtain prediction intervals with, say, nominal 90% coverage, simply fit the conditional quantile function

at the 5% and 95% levels and form the corresponding intervals. Even for highly heteroscedastic data, this methodology has been shown to be adaptive to local variability [19–25]. However, the validity of the estimated intervals is guaranteed only for specific models, under certain regularity and asymptotic conditions [22–24].

In this work, we combine conformal prediction with quantile regression. The resulting method, which we call *conformalized quantile regression* (CQR), inherits both the finite sample, distribution-free validity of conformal prediction and the statistical efficiency of quantile regression.[1] On one hand, CQR is flexible in that it can wrap around any algorithm for quantile regression, including random forests and deep neural networks [26–29]. On the other hand, a key strength of CQR is its rigorous control of the miscoverage rate, independent of the underlying regression algorithm.

Summary and outline

Suppose we are given n training samples $\{(X_i, Y_i)\}_{i=1}^{n}$ and we must now predict the unknown value of Y_{n+1} at a test point X_{n+1}. We assume that all the samples $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are drawn exchangeably—for instance, they may be drawn i.i.d.—from an arbitrary joint distribution P_{XY} over the feature vectors $X \in \mathbb{R}^p$ and response variables $Y \in \mathbb{R}$. We aim to construct a *marginal distribution-free prediction interval* $C(X_{n+1}) \subseteq \mathbb{R}$ that is likely to contain the unknown response Y_{n+1}. That is, given a desired miscoverage rate α, we ask that

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

for any joint distribution P_{XY} and any sample size n. The probability in this statement is marginal, being taken over all the samples $\{(X_i, Y_i)\}_{i=1}^{n+1}$.

To accomplish this, we build on the method of conformal prediction [2, 3, 8]. We first split the training data into two disjoint subsets, a proper training set and a calibration set.[2] We fit two quantile regressors on the proper training set to obtain initial estimates of the lower and upper bounds of the prediction interval, as explained in Section 2. Then, using the calibration set, we conformalize and, if necessary, correct this prediction interval. Unlike the original interval, the conformalized prediction interval is guaranteed to satisfy the coverage requirement (1) regardless of the choice or accuracy of the quantile regression estimator. We prove this in Section 4.

Our method differs from the standard method of conformal prediction [3, 15], recalled in Section 3, in that we calibrate the prediction interval using conditional quantile regression, while the standard method uses only classical, conditional mean regression. The result is that our intervals are adaptive to heteroscedasticity whereas the standard intervals are not. We evaluate the statistical efficiency of our framework by comparing its miscoverage rate and average interval length with those of other methods. We review existing state-of-the-art schemes for conformal prediction in Section 5 and we compare them with our method in Section 6. Based on extensive experiments across eleven datasets, we conclude that conformal quantile regression yields shorter intervals than the competing methods.

2 Quantile regression

The aim of conditional quantile regression [18] is to estimate a given quantile, such as the median, of Y conditional on X. Recall that the *conditional distribution function* of Y given $X = x$ is

$$F(y \mid X = x) := \mathbb{P}\{Y \leq y \mid X = x\},$$

and that the αth *conditional quantile function* is

$$q_\alpha(x) := \inf\{y \in \mathbb{R} : F(y \mid X = x) \geq \alpha\}.$$

Fix the lower and upper quantiles to be equal to $\alpha_{lo} = \alpha/2$ and $\alpha_{hi} = 1 - \alpha/2$, say. Given the pair $q_{\alpha_{lo}}(x)$ and $q_{\alpha_{hi}}(x)$ of lower and upper conditional quantile functions, we obtain a conditional prediction interval for Y given $X = x$, with miscoverage rate α, as

$$C(x) = [q_{\alpha_{lo}}(x), q_{\alpha_{hi}}(x)]. \tag{2}$$

---

[1]Source code implementing CQR is available online at `https://github.com/yromano/cqr`.
[2]Like conformal regression, CQR has a variant that does not require data splitting.

Figure 1: Visualization of the pinball loss function in (6), where $y \neq \hat{y}$.

By construction, this interval satisfies

$$P\{Y \in C(X) \mid X = x\} \geq 1 - \alpha. \tag{3}$$

Notice that the length of the interval $C(X)$ can vary greatly depending on the value of X. The uncertainty in the prediction of Y is naturally reflected in the length of the interval. In practice we cannot know this ideal prediction interval, but we can try to estimate it from the data.

Estimating quantiles from data

Classical regression analysis estimates the conditional mean of the test response Y_{n+1} given the features $X_{n+1} = x$ by minimizing the sum of squared residuals on the n training points:

$$\hat{\mu}(x) = \mu(x; \hat{\theta}); \qquad \hat{\theta} = \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu(X_i; \theta))^2 + R(\theta). \tag{4}$$

Here θ are the parameters of the regression model, $\mu(x; \theta)$ is the regression function, and R is a potential regularizer.

Analogously, quantile regression estimates a conditional quantile function q_{α} of Y_{n+1} given $X_{n+1} = x$. This can be cast as the optimization problem

$$\hat{q}_{\alpha}(x) = f(x; \hat{\theta}); \qquad \hat{\theta} = \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \rho_{\alpha}(Y_i, f(X_i; \theta)) + R(\theta), \tag{5}$$

where $f(x; \theta)$ is the quantile regression function and the loss function ρ_{α} is the "check function" or "pinball loss" [18, 24], defined by

$$\rho_{\alpha}(y; \hat{y}) := \begin{cases} \alpha(y - \hat{y}) & \text{if } y - \hat{y} > 0; \\ (1 - \alpha)(\hat{y} - y) & \text{otherwise} \end{cases} \tag{6}$$

and illustrated in Figure 1. The simplicity and generality of this formulation makes quantile regression widely applicable. As in classical regression, one can leverage the great variety of machine learning methods to design and learn f [19–21, 23, 30].

All this suggests an obvious strategy to construct a prediction band with nominal miscoverage rate α: estimate $\hat{q}_{lo}(x)$ and $\hat{q}_{hi}(x)$ using quantile regression, then output $\hat{C}(X_{n+1}) = [\hat{q}_{lo}(X_{n+1}); \hat{q}_{hi}(X_{n+1})]$ as an estimate of the ideal interval $C(X_{n+1})$ from equation (2). This approach is widely applicable and often works well in practice, yielding intervals that are adaptive to heteroscedasticity. However, it is not guaranteed to satisfy the coverage statement (3) when $C(X)$ is replaced by the estimated interval $\hat{C}(X_{n+1})$. Indeed, the absence of any finite sample guarantee can sometimes be disastrous. This worry is corroborated by our experiments, which show that the intervals constructed by neural networks can substantially undercover.

Under certain regularity conditions and for specific models, estimates of conditional quantile functions via the pinball loss are known to be asymptotically consistent [23, 24]. Related methods that do not minimize the pinball loss, such as quantile random forests [22], are also asymptotically consistent. But to get valid coverage in finite samples, we must draw on a different set of ideas, from conformal prediction.

3 Conformal Prediction

We now describe how conformal prediction [1, 3] constructs prediction intervals that satisfy the finite-sample coverage guarantee (3). To be carried out exactly, the original, full, conformal procedure

3

effectively requires the regression algorithm to be invoked in nitely many times. In contrast, the method of split, or inductive, conformal prediction [2, 8] avoids this problem, at the cost of splitting the data. While our proposal is applicable to both versions of conformal prediction, in the interest of space we will restrict our attention to split conformal prediction and refer the reader to [3, 15] for a more detailed comparison between the two methods.

Under the assumptions of Section 1, the split conformal method begins by splitting the training data into two disjoint subsets: a proper training set $\{(X_i, Y_i) : i \in I_1\}$ and calibration set $\{(X_i, Y_i) : i \in I_2\}$. Then, given any regression algorithm A,[3] a regression model is t to the proper training set:

$$\hat{\mu}(x) \leftarrow A(\{(X_i, Y_i) : i \in I_1\}) :$$

Next, the absolute residuals are computed on the calibration set, as follows:

$$R_i = |Y_i - \hat{\mu}(X_i)|, \qquad i \in I_2: \tag{7}$$

For a given level α, we then compute a quantile of the empirical distribution[4] of the absolute residuals,

$$Q_{1-\alpha}(R; I_2) := (1 - \alpha)(1 + 1/|I_2|)\text{-th empirical quantile of } \{R_i : i \in I_2\}:$$

Finally, the prediction interval at a new point X_{n+1} is given by

$$C(X_{n+1}) = [\hat{\mu}(X_{n+1}) - Q_{1-\alpha}(R; I_2); \hat{\mu}(X_{n+1}) + Q_{1-\alpha}(R; I_2)] : \tag{8}$$

This interval is guaranteed to satisfy (1), as shown in [3]. For related theoretical studies, see [15, 31].

A closer look at the prediction interval (8) reveals a major limitation of this procedure: the length of $C(X_{n+1})$ is xed and equal to $2Q_{1-\alpha}(R; I_2)$, independent of X_{n+1}. Lei et al [15] observe that the intervals produced by the full conformal method also vary only slightly with X_{n+1}, provided the regression algorithm is moderately stable. This brings us to our proposal, which offers a principled approach to constructing variable-width conformal prediction intervals.

4 Conformalized quantile regression (CQR)

In this section we introduce our procedure, beginning with a small experiment on simulated data to show how it improves upon standard conformal prediction. Figure 2 compares the prediction intervals produced by (a) the split conformal method, (b) its locally adaptive variant (described later in Section 5), and (c) our method, conformalized quantile regression (CQR). The heteroskedasticity of the data is evident, as the dispersion of Y varies considerably with X. The data also contains outliers, shown in Figure 7 from Appendix B. For all three methods, we construct 90% prediction intervals on the test data. From Figures 2a and 2d, we see that the lengths of the split conformal intervals are xed and equal to 2.91. The prediction intervals of the locally weighted variant, shown in Figure 2b, are partially adaptive, resulting in slightly shorter intervals, of average length 2.86. Our method, shown in Figure 2c, is also adaptive, but its prediction intervals are considerably shorter, of average length 1.99, due to better estimation of the lower and upper quantiles. We refer the reader to Appendix B for additional information about this experiment.

We now describe CQR itself. As in split conformal prediction, we begin by splitting the data into a proper training set, indexed by I_1, and a calibration set, indexed by I_2. Given any quantile regression algorithm A, we then t two conditional quantile functions $\hat{q}_{lo}$ and $\hat{q}_{hi}$ on the proper training set:

$$\{\hat{q}_{lo}, \hat{q}_{hi}\} \leftarrow A(\{(X_i, Y_i) : i \in I_1\}):$$

In the essential next step, we compute conformity scores that quantify the error made by the plug-in prediction interval $\hat{C}(x) = [\hat{q}_{lo}(x); \hat{q}_{hi}(x)]$. The scores are evaluated on the calibration set as

$$E_i := \max\{\hat{q}_{lo}(X_i) - Y_i, Y_i - \hat{q}_{hi}(X_i)\} \tag{9}$$

for each $i \in I_2$. The conformity score E_i has the following interpretation. If Y_i is below the lower endpoint of the interval, $Y_i < \hat{q}_{lo}(X_i)$, then $E_i = |Y_i - \hat{q}_{lo}(X_i)|$ is the magnitude of the error

---

[3] In full conformal prediction, the regression algorithm must treat the data exchangeably, but no such restrictions apply to split conformal prediction.

[4] The explicit formula for empirical quantiles is recalled in Appendix A.

(a) Split: Avg. coverage 91.4%; Avg. length 2.91. (b) Local: Avg. coverage 91.7%; Avg. length 2.86.

(c) CQR: Avg. coverage 91.06%; Avg. length 1.99.   (d) Length of prediction intervals.

Figure 2: Prediction intervals on simulated heteroscedastic data with outliers (see Figure 7 for a full range display): (a) the standard split conformal method, (b) its locally adaptive variant, and (c) CQR (our method). The length of the interval as a function of X is shown in (d). The target coverage rate is 90%. The broken black curve in (a) and (b) is the pointwise prediction from the random forest estimator. In (c), we show two curves, representing the lower and upper quantile regression estimates based on random forests [22]. Observe how in this example the quantile regression estimates closely match the adjusted estimates—the boundary of the blue region—obtained by conformalization.

incurred by this mistake. Similarly, if Y_i is above the upper endpoint of the interval, $Y_i > \hat{q}_{hi}(X_i)$, then $E_i = |Y_i - \hat{q}_{hi}(X_i)|$. Finally, if Y_i correctly belongs to the interval $\hat{q}_{lo}(X_i) \leq Y_i \leq \hat{q}_{hi}(X_i)$, then E_i is the larger of the two non-positive numbers $\hat{q}_{lo}(X_i) - Y_i$ and $Y_i - \hat{q}_{hi}(X_i)$ and so is itself non-positive. The conformity score thus accounts for both undercoverage and overcoverage.

Finally, given new input data X_{n+1}, we construct the prediction interval for Y_{n+1} as

$$C(X_{n+1}) = [\hat{q}_{lo}(X_{n+1}) - Q_{1-\alpha}(E; I_2); \hat{q}_{hi}(X_{n+1}) + Q_{1-\alpha}(E; I_2)]; \qquad (10)$$

where

$$Q_{1-\alpha}(E; I_2) := (1-\alpha)(1 + 1 = |I_2|)\text{-th empirical quantile of } \{E_i : i \in I_2\} \qquad (11)$$

conformalizes the plug-in prediction interval.

For ease of reference, the CQR procedure is summarized in Algorithm 1. We now prove that its prediction intervals satisfy the marginal, distribution-free coverage guarantee (1).

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

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

5

Algorithm 1: Split Conformal Quantile Regression.

Input:

Data $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$; $1 \le i \le n$.

Miscoverage level $\alpha \in (0, 1)$.

Quantile regression algorithm $\mathcal{A}$.

Process:

Randomly split $\{1, \ldots, n\}$ into two disjoint sets I_1 and I_2.

Fit two conditional quantile functions: $\{\hat{q}_{lo}, \hat{q}_{hi}\} \leftarrow \mathcal{A}(\{(X_i, Y_i) : i \in I_1\})$.

Compute E_i for each $i \in I_2$, as in equation (9).

Compute $Q_{1-\alpha}(E, I_2)$, the $(1-\alpha)(1 + 1/|I_2|)$-th empirical quantile of $\{E_i : i \in I_2\}$.

Output:

Prediction interval $C(x) = [\hat{q}_{lo}(x) - Q_{1-\alpha}(E, I_2), \ \hat{q}_{hi}(x) + Q_{1-\alpha}(E, I_2)]$ for unseen input $X_{n+1} = x$.

Moreover, if the conformity scores E_i are almost surely distinct, then the prediction interval is nearly perfectly calibrated:

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \le 1 - \alpha + \frac{1}{|I_2| + 1}.$$

Proof. The result even holds, and we will prove it, conditionally on the proper training set.

Let E_{n+1} be the conformity score (9) at the test point (X_{n+1}, Y_{n+1}). By the construction of the prediction interval, we have

$$Y_{n+1} \in C(X_{n+1}) \quad \text{if and only if} \quad E_{n+1} \le Q_{1-\alpha}(E, I_2);$$

and, in particular,

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1}) \mid (X_i, Y_i) : i \in I_1\} = \mathbb{P}\{E_{n+1} \le Q_{1-\alpha}(E, I_2) \mid (X_i, Y_i) : i \in I_1\}. \quad (12)$$

Since the original pairs (X_i, Y_i) are exchangeable, so are the calibration variables E_i for $i \in I_2$ and $i = n + 1$. Therefore, by Lemma 2 on inflated empirical quantiles (stated in Appendix A),

$$\mathbb{P}\{E_{n+1} \le Q_{1-\alpha}(E, I_2) \mid (X_i, Y_i) : i \in I_1\} \ge 1 - \alpha; \quad (13)$$

and, under the additional assumption that the E_i's are almost surely distinct,

$$\mathbb{P}\{E_{n+1} \le Q_{1-\alpha}(E, I_2) \mid (X_i, Y_i) : i \in I_1\} \le 1 - \alpha + \frac{1}{|I_2| + 1}. \quad (14)$$

The result follows by taking expectations over the proper training set in (12), (13), and (14). □

Practical considerations and extensions

Conformalized quantile regression can accommodate a wide range of quantile regression methods [18–23, 25, 30] to estimate the conditional quantile functions q_{lo} and q_{hi}. The estimators can be even be aggregates of different quantile regression algorithms. Recently, new deep learning techniques have been proposed [26] for constructing prediction intervals. These methods could be wrapped by our framework and would then immediately enjoy rigorous coverage guarantees. In our experiments, we focus on quantile neural networks [20] and quantile regression forests [22].

Because the underlying quantile regression algorithm may process the proper training set in arbitrary ways, our framework affords broad flexibility in hyper-parameter tuning. Consider, for instance, the tuning of typical hyper-parameters of neural networks, such as the batch size, the learning rate, and the number of epochs. The hyperparameters may be selected, as usual, by cross validation, where we minimize the average interval length over the folds.

In this vein, we record two specific implementation details that we have found to be useful.

1. Quantile regression is sometimes too conservative, resulting in unnecessarily wide prediction intervals. In our experience, quantile regression forests [22] are often overly conservative

and quantile neural networks [20] are occasionally so. We can mitigate this problem by tuning the nominal quantiles of the underlying method as additional hyper-parameters in cross validation. Notably, this tuning does not invalidate the coverage guarantee, but it may yield shorter intervals, as our experiments confirm.

2. To reduce the computational cost, instead of fitting two separate neural networks to estimate the lower and upper quantile functions, we can replace the standard one-dimensional estimate of the unknown response by a two-dimensional estimate of the lower and upper quantiles. In this way, most of the network parameters are shared between the two quantile estimators. We adopt this approach in the experiments of Section 6.

Another avenue for extension is the conformalization step. The conformalization implemented by equations (10) and (11) allows coverage errors to be spread arbitrarily over the left and right tails. By using a method reminiscent of [32], we can control the left and right tails independently, resulting in a stronger coverage guarantee.

Theorem 2. Define the prediction interval

$$C(X_{n+1}) := [\hat{q}_{lo}(X_{n+1}) - Q_{1-\alpha_{lo}}(E_{lo}; I_2); \hat{q}_{hi}(X_{n+1}) + Q_{1-\alpha_{hi}}(E_{hi}; I_2)];$$

where $Q_{1-\alpha_{lo}}(E_{lo}; I_2)$ is the $(1-\alpha_{lo})$-th empirical quantile of $\{\hat{q}_{lo}(X_i) - Y_i : i \in I_2\}$ and $Q_{1-\alpha_{hi}}(E_{hi}; I_2)$ is the $(1-\alpha_{hi})$-th empirical quantile of $\{Y_i - \hat{q}_{hi}(X_i) : i \in I_2\}$. If the samples $(X_i; Y_i)$, $i = 1; \ldots; n+1$ are exchangeable, then

$$P\{Y_{n+1} \geq \hat{q}_{lo}(X_{n+1}) - Q_{1-\alpha_{lo}}(E_{lo}; I_2)\} \geq 1 - \alpha_{lo} \tag{15}$$

and

$$P\{Y_{n+1} \leq \hat{q}_{hi}(X_{n+1}) + Q_{1-\alpha_{hi}}(E_{hi}; I_2)\} \geq 1 - \alpha_{hi}: \tag{16}$$

Consequently, assuming $\alpha = \alpha_{lo} + \alpha_{hi}$, we also have $P\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$.

Proof. The two events inside the probabilities (15) and (16) are equivalent to $\hat{q}_{lo}(X_{n+1}) - Y_{n+1} \leq Q_{1-\alpha_{lo}}(E_{lo}; I_2)$ and $Y_{n+1} - \hat{q}_{hi}(X_{n+1}) \leq Q_{1-\alpha_{hi}}(E_{hi}; I_2)$, respectively. We can thus apply Lemma 2 twice, in the same manner as in the proof of Theorem 1. □

As we will see in Section 6, the price paid for the stronger coverage guarantee is slightly longer intervals.

5 Related work: locally adaptive conformal prediction

Locally adaptive split conformal prediction, first proposed in [7, 9] and later studied in [15], is an earlier approach to making conformal prediction adaptive to heteroskedascity. Like our method, it starts from the observation that one can replace the absolute residuals in equation (7) by any other loss function that treats the data exchangeably. In this case, the absolute residuals R_i are replaced by the scaled residuals

$$\tilde{R}_i := \frac{|Y_i - \hat{\mu}(X_i)|}{\hat{\sigma}(X_i)} = \frac{R_i}{\hat{\sigma}(X_i)}; \qquad i \in I_2;$$

where $\hat{\sigma}(X_i)$ is a measure of the dispersion of the residuals at X_i. Usually $\hat{\sigma}(x)$ is an estimate of the conditional mean absolute deviation (MAD) of $|Y - \hat{\mu}(x)|$ given $X = x$. Finally, the prediction interval at a new point X_{n+1} is computed as

$$C(X_{n+1}) = \left[\hat{\mu}(X_{n+1}) - \hat{\sigma}(X_{n+1})Q_{1-\alpha}(\tilde{R}; I_2); \hat{\mu}(X_{n+1}) + \hat{\sigma}(X_{n+1})Q_{1-\alpha}(\tilde{R}; I_2)\right]:$$

Both $\hat{\mu}$ and $\hat{\sigma}$ are fit only on the proper training set. Consequently, $\hat{\mu}$ and $\hat{\sigma}$ satisfy the assumptions of conformal prediction and, hence, locally adaptive conformal prediction inherits the coverage guarantee of standard conformal prediction.

In practice, locally adaptive conformal prediction requires fitting two functions, in sequence, on the proper training set. (Thus it is more computationally expensive than standard conformal prediction.) First, one fits the conditional mean function $\hat{\mu}(x)$, as described in Section 3. Then one fits $\hat{\sigma}(x)$ to the pairs $\{(X_i; R_i) : i \in I_1\}$, using a regression model that predicts the residuals R_i given the inputs

7

X_i. As an example, the intervals in Figure 2b above are created by locally adaptive split conformal prediction, where both $\hat{\mu}$ and $\hat{\sigma}$ are random forests.

Practitioners employ various tweaks to improve the method's numerical stability and statistical performance. Following [10], we add a hyper-parameter $\epsilon > 0$ as a constant offset to the scale estimator $\hat{\sigma}(x)$. The scaled residuals then become

$$\tilde{R}_i = \frac{R_i}{\hat{\sigma}(X_i) + \epsilon}: \tag{17}$$

Limitations of locally adaptive conformal prediction

Locally adaptive conformal prediction is limited in several ways, some more important than others. A first limitation, already noted in [15], appears when the data is actually homoskedastic. In this case, the locally adaptive method suffers from inflated prediction intervals compared to the standard method. This is presumably due to the extra variability introduced by estimating $\hat{\sigma}$ as well as $\hat{\mu}$.

The locally adaptive method faces a more fundamental statistical limitation. There is an essential difference between the residuals on the proper training set and the residuals on the calibration set: the former are biased by an optimization procedure designed to minimize them, while the latter are unbiased. Because it uses the proper training residuals (as it must to ensure valid coverage), the locally adaptive method tends to systematically underestimate the prediction error. In general, this forces the correction constant $Q_{1-\alpha}(\tilde{R}; I_2)$ to be large and the intervals to be less adaptive than they could be.

To press this point further, suppose the conditional mean function $\hat{\mu}$ is a deep neural network. It is well attested in the deep learning literature that, given enough training samples, the best prediction error is attained by "over-fitting" to the training data, in the sense that the training error is nearly zero. The training residuals are then very poor estimates of the test residuals, resulting in severe loss of adaptivity. The original training objective of our method, in contrast, is to estimate the lower and upper conditional quantiles, not the conditional mean. Having sufficient training data, the fitted network is expected to provide reasonable approximations of these two quantile functions, which are used to construct adaptive prediction intervals.

6   Experiments

In this section we systematically compare our method, conformalized quantile regression, to the standard and locally adaptive versions of split conformal prediction. Among preexisting conformal prediction algorithms, we select leading variants that use random forests [10] and neural networks [33] for conditional mean regression. Likewise, we configure our method to use quantile regression algorithms based on random forests [22] and neural networks [20]. As a baseline, we also include conformal ridge regression [2] in the comparison. A detailed description of each of the methods is given below.

We conduct the experiments on eleven popular benchmark datasets for regression, listed in Section 6.3. In each case, we standardize the features to have zero mean and unit variance and we rescale the response by dividing it by its mean absolute value.[5] The performance metrics are averaged over 20 different training-test splits. 80% of the examples are used for training and the remaining 20% for testing. The proper training and calibration sets for split conformal prediction have equal size. Throughout the experiments the nominal miscoverage rate is fixed and equal to $\alpha = 0.1$.

6.1   Methods

In more detail, we compare the following methods related to conformal prediction. We evaluate the original version of split conformal prediction (Section 3) using the following three regression algorithms.

---

[5]In the experiments, we compute the needed sample means and variances only on the proper training set. This ensures that if the original data is exchangeable, then the rescaled data remains so. That being said, we could also rescale using sample means and variances computed on the test data, because it would preserve exchangeability even while it destroys independence.

Ridge: We include ridge regression as a baseline. The regularization parameter is tuned by cross validation.

Random Forests We use the implementation of (conditional mean) random forest regression in the Python package sklearn . The hyper-parameters are the package defaults, except for the total number of trees in the forest, which we set to 1000.

Neural Net: Our neural network architecture consists of three fully connected layers, with ReLU nonlinearities between layers. The first layer takes as input the p-dimensional feature vector X and outputs 64 hidden variables. The second layer follows the same template, outputting another 64 hidden variables. Finally, a linear output layer returns a pointwise estimate of the response variable Y. The parameters of the network are fit by minimizing the quadratic loss function (4). We use the stochastic optimization algorithm Adam [34], with fixed learning rate of 5×10^{-4}, minibatches of size 64, and weight decay parameter equal to 10^{-6}. We employ dropout regularization [35], with the probability of retaining a hidden unit equal to 0.1. To avoid overfitting, we found that early stopping performs well; we tune the number of epochs by cross validation, with an upper limit of 1000 epochs.

We evaluate locally adaptive conformal prediction (Section 5) using the same three underlying regression algorithms. We set the hyper-parameter in equation (17) to 1, which improves performance considerably compared to $\gamma = 0$.

Ridge Local: The conditional mean estimator $\hat{\mu}$ is fit by ridge regression, as described above, and the mean absolute deviation (MAD) estimator $\hat{\sigma}$ is a k-nearest neighbors with $k = 11$.

Random Forests Local: Both $\hat{\mu}$ and $\hat{\sigma}$ are random forests with the hyper-parameters described above.

Neural Net Local: Both $\hat{\mu}$ and $\hat{\sigma}$ are neural networks, with the network architecture, hyper-parameters, and training algorithm described above.

For our own proposal, conformalized quantile regression (Algorithm 1), we evaluate two variants:

CQR Random Forests We use CQR with quantile regression forests [22]. To ensure a fair comparison, the hyper-parameters of the quantile regression forests are made identical to those of the random forests in the previous methods. Quantile regression forests have two additional parameters that control the coverage rate on the training data. We tune them using cross validation, as explained in Section 4.

CQR Neural Net: We apply CQR using neural networks for quantile regression [20]. The network architecture is the same as above, except that the output of the quantile regression network is a two-dimensional vector, representing the lower and upper conditional quantiles. The training algorithm is also the same, except that the cost function is now the pinball loss in equation (5) instead of the quadratic loss.

Finally, for the sake of comparison, we also include the previous two quantile regression algorithms, but without any conformalization:

Quantile Random Forests We use quantile regression forests with hyperparameters as in the CQR procedure, except that the upper and lower levels are exactly 0.05 and 0.95.

Quantile Neural Net: We use quantile regression neural networks with exactly the same architecture and training algorithm as in the CQR procedure.

Unlike the preceding methods, the last two methods do not need a calibration set and do not have a finite-sample coverage guarantee. We fit them on the entire training set.

6.2   Summary of results

Table 1 summarizes our 2,200 experiments, showing the average performance across all the datasets and training-test splits. On average, our method achieves shorter prediction intervals than both standard and locally adaptive conformal prediction. It may seem surprising that our method also outperforms non-conformalized quantile regression, which is permitted more training data. There are

Method	Avg. Length	Avg. Coverage
Ridge	3.06	90.03
Ridge Local	2.94	90.13
Random Forests	2.24	89.99
Random Forests Local	1.82	89.95
Neural Net	2.16	89.92
Neural Net Local	1.81	89.95
CQR Random Forests	1.41	90.33
CQR Neural Net	1.40	90.05
*Quantile Random Forests	*2.23	*92.62
*Quantile Neural Net	*1.49	*88.51

Table 1: Length and coverage of prediction intervals ($\alpha = 0.1$) constructed by various methods, averaged across 11 datasets and 20 random training-test splits. Our methods are shown in bold font. The methods marked by an asterisk are not supported by finite-sample coverage guarantees.

several possible explanations for this. First, the non-conformalized methods sometimes overcover, but that is mitigated by our signed conformity scores (9). In addition, by using CQR, we can tune the quantiles of the underlying quantile regression algorithms using cross-validation (Section 4). Interestingly, CQR selects quantiles below the nominal level.

Turning to the issue of valid coverage, all methods based on conformal prediction successfully construct prediction bands at the nominal coverage rate of 90%, as the theory suggests they should. One of the non-conformalized methods, based on random forests, is slightly conservative, while the other, based on neural networks, tends to undercover. In fact, other authors have shown that the coverage of quantile neural networks depends greatly on the tuning of the hyper-parameters, with, for instance, the actual coverage in [25, Figure 3] ranging from the 95% nominal level in that paper to well below 50%. Such volatility demonstrates the importance of the conformal prediction's finite-sample guarantee.

When estimating a lower and an upper quantile by two separate quantile regressions, there is no guarantee that the lower estimate will actually be smaller than the upper estimate. This is known as the quantile crossing problem [36]. Quantile crossing can affect quantile neural networks, but not quantile regression forests. When the two quantiles are far apart, as in the 5% and 95% quantiles, we should expect the estimates to cross very infrequently and that is indeed what we find in the experiments. Nevertheless, we also evaluated a post-processing method to eliminate crossings [37]. It yields a slight improvement in performance: the average interval length of the CQR neural networks drops from 1.40 to 1.35 and the average interval length of the unconformalized quantile neural networks drops from 1.49 to 1.41, with the coverage rates remaining about the same.

As expected, adopting the two-tailed, asymmetric conformalization proposed in Theorem 2 causes an increase in average interval length compared to the symmetric conformalization of Theorem 1. Specifically, the average length for CQR neural networks increases from 1.40 to 1.58, while the coverage rate stays about the same. The average length for the CQR random forests increases from 1.41 to 1.57, accompanied by a slight increase in the average coverage rate, from 90.33 to 90.99.

6.3 Performance on individual datasets

In a series of figures, we break down the performance of the different methods on each of the benchmark datasets. Figure 3 summarizes our experiments on the datasets: medical expenditure panel survey number 19 (MEPS_19) [38], number 20 (MEPS_20) [39], and number 21 (MEPS_21) [40]. Figure 4 shows the results for: blog feedback (blog_data) [41]; physicochemical properties of protein tertiary structure (bio) [42]; and bike sharing (bike) [43]. Figure 5 shows the results for: community and crimes (community) [44]; Tennessee's student teacher achievement ratio (STAR) [45]; and concrete compressive strength (concrete) [46]. Lastly, Figure 6 shows the results for: Facebook comment volume, variants one (facebook_1) and two (facebook_2) [47, 48].

The performance on individual datasets confirms the overall trend in Table 1. Locally adaptive conformal prediction generally outperforms standard conformal prediction, and, on ten out of eleven

datasets, conformalized quantile regression outperforms both. The CQR random forests are overly conservative on the two Facebook datasets. This is consistent with the theory, because in this case there are ties among the conformity scores and so the upper bound in Theorem 1 does not apply.

7 Conclusion

Conformal quantile regression is a new way of constructing prediction intervals that combines the advantages of conformal prediction and quantile regression. It provably controls the miscoverage rate in nite samples, under the mild distributional assumption of exchangeability, while adapting the interval lengths to heteroskedasticity in the data.

We expect the ideas behind conformal quantile regression to be applicable in the related setting of conformal predictive distributions [49]. In this extension of conformal prediction, the aim is to estimate a predictive probability distribution, not just an interval. We see intriguing connections between our work and a very recent, independently written paper on conformal distributions [17].

References

[1] Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*, pages 444–453, 1999.

[2] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive con dence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

[3] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[4] Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590, 2009.

[5] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

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

[7] Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. In *International Conference on Arti cial Intelligence and Applications*, pages 64–69, 2008.

[8] Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in arti cial intelligence*. IntechOpen, 2008.

[9] Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Arti cial Intelligence Research*, 40:815–840, 2011.

[10] Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine Learning*, 97(1-2):155–176, 2014.

[11] Ulf Johansson, Cecilia Sönströd, Henrik Linusson, and Henrik Boström. Regression trees for streaming data with local performance guarantees. In *IEEE International Conference on Big Data*, pages 461–470. IEEE, 2014.

[12] Ulf Johansson, Cecilia Sönströd, and Henrik Linusson. Ef cient conformal regressors using bagged neural nets. In *IEEE International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2015.

[13] Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Arti cial Intelligence*, 74(1-2):9–28, 2015.

[14] Henrik Boström, Henrik Linusson, Tuve Löfström, and Ulf Johansson. Accelerating dif culty estimation for conformal regression forests. *Annals of Mathematics and Arti cial Intelligence*, 81(1-2):125–144, 2017.

[15] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[16] Wenyu Chen, Kelli-Jean Chun, and Rina Foygel Barber. Discretized conformal prediction for ef cient distribution-free inference. *Stat*, 7(1):e173, 2018.

[17] Vladimir Vovk, Ivan Petej, Paolo Toccaceli, and Alex Gammerman. Conformal calibrators. *arXiv preprint arXiv:1902.06579*, 2019.

[18] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[19] David R. Hunter and Kenneth Lange. Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000.

[20] James W. Taylor. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311, 2000.

[21] Roger Koenker and Kevin F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, 2001.

[22] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

[23] Ichiro Takeuchi, Quoc V. Le, Timothy D. Sears, and Alexander J. Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

[24] Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

[25] Natasa Tagasovska and David Lopez-Paz. Frequentist uncertainty estimates for deep learning. *arXiv preprint arXiv:1811.00908*, 2018.

[26] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[27] Cheng Lian, Zhigang Zeng, Wei Yao, Huiming Tang, and Chun Lung Philip Chen. Landslide displacement prediction with uncertainty based on neural networks with random hidden weights. *IEEE Transactions on Neural Networks and Learning Systems*, 27(12):2683–2695, 2016.

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[29] Tim Pearce, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International Conference on Machine Learning*, pages 6473–6482, 2018.

[30] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[31] Wenyu Chen, Zhaokai Wang, Wooseok Ha, and Rina Foygel Barber. Trimmed conformal prediction for high-dimensional models. *arXiv preprint arXiv:1611.09933*, 2016.

[32] Henrik Linusson, Ulf Johansson, and Tuve Löfström. Signed-error conformal regression. In *Paci c-Asia Conference on Knowledge Discovery and Data Mining*, pages 224–236. Springer, 2014.

[33] Harris Papadopoulos and Haris Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011.

[34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdi-nov. Dropout: A simple way to prevent neural networks from over tting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[36] Gilbert Bassett Jr and Roger Koenker. An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, 77(378):407–415, 1982.

[37] Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

[38] Medical expenditure panel survey, panel 19. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181. Accessed: January, 2019.

[39] Medical expenditure panel survey, panel 20. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181. Accessed: January, 2019.

[40] Medical expenditure panel survey, panel 21. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192. Accessed: January, 2019.

[41] BlogFeedback data set. https://archive.ics.uci.edu/ml/datasets/BlogFeedback. Accessed: January, 2019.

[42] Physicochemical properties of protein tertiary structure data set. https://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure. Accessed: January, 2019.

[43] Bike sharing dataset data set. https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset. Accessed: January, 2019.

[44] Communities and crime data set. http://archive.ics.uci.edu/ml/datasets/communities+and+crime. Accessed: January, 2019.

[45] C.M. Achilles, Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. Tennessee's Student Teacher Achievement Ratio (STAR) project, 2008.

[46] Concrete compressive strength data set. http://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength. Accessed: January, 2019.

[47] Facebook comment volume data set. https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset. Accessed: January, 2019.

[48] Kamaljot Singh, Ranjeet Kaur Sandhu, and Dinesh Kumar. Comment volume prediction using neural networks and decision trees. In *IEEE International Conference on Computer Modelling and Simulation*. IEEE, 2015.

[49] Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, pages 1–30, 2017.

[50] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction under covariate shift. *arXiv preprint arXiv:1904.06019*, 2019.

Figure 3: Average length (left) and coverage (right) of prediction intervals (0:1), averaged over 20 random (80%/20%) training/test splits. The numbers in the colored boxes are the average lengths, shown in red for split conformal, in gray for locally adapative split conformal, and in light blue for our method. The name of the dataset is located at the top of each plot.
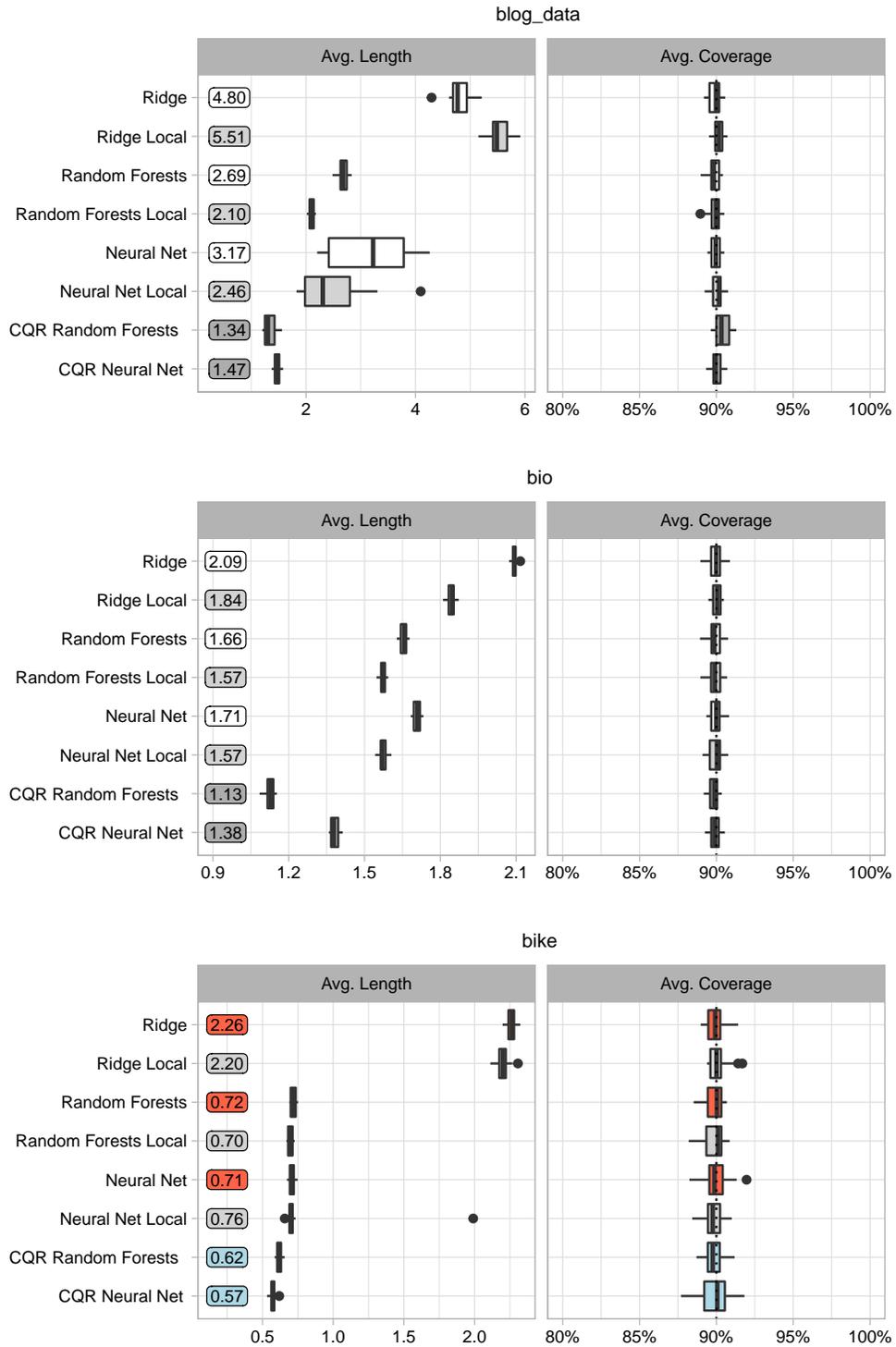
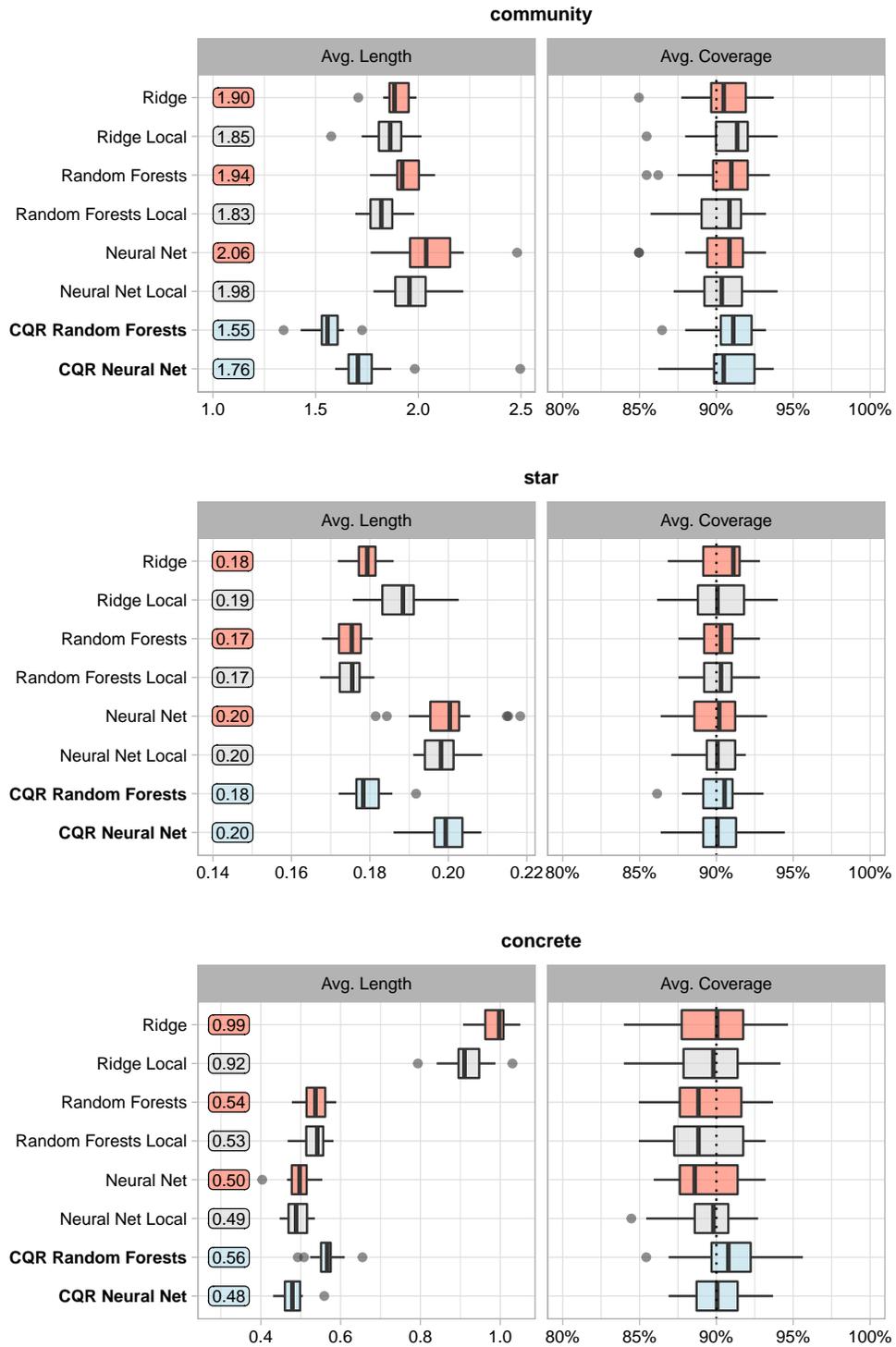Figure 4: Refer to the caption of Figure 3 for details.

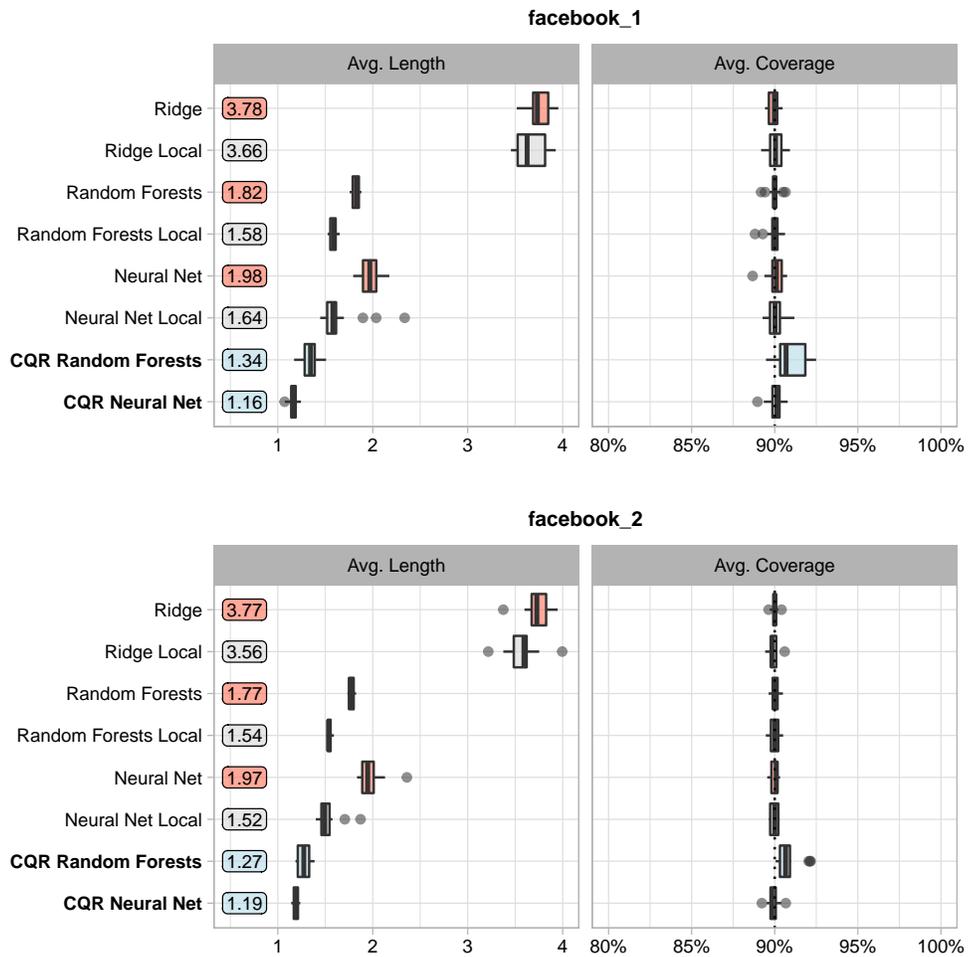Figure 5: Refer to the caption of Figure 3 for details.

Figure 6: Refer to the caption of Figure 3 for details.

A Lemmas about quantiles

Recall that the *quantile function* Q of a random variable Z, with cumulative distribution function $F(z) := P\{Z \le z\}$, is defined by the equivalence
$$Q(\alpha) \le z \quad \text{if and only if} \quad \alpha \le F(z)$$
for all $\alpha \in (0,1)$ and $z \in \mathbb{R}$. Dually, but less standardly, the *right quantile function* R of the random variable Z is defined by the equivalence
$$F^-(z) \le \alpha \quad \text{if and only if} \quad z \le R(\alpha);$$
where $F^-(z) := F(z^-) = P\{Z < z\}$. The quantile functions have the explicit formulas
$$Q(\alpha) = \inf\{z \in \mathbb{R} : \alpha \le F(z)\}; \qquad R(\alpha) = \sup\{z \in \mathbb{R} : F^-(z) \le \alpha\}.$$

As a special case, the *empirical quantile function* $\hat{Q}_n$ of random variables $Z_1, \ldots, Z_n$ is the quantile function with respect to the empirical CDF $\hat{F}_n(z) := \frac{1}{n}\sum_{i=1}^n 1_{Z_i \le z}$. Likewise, the *right empirical quantile function* $\hat{R}_n$ of $Z_1, \ldots, Z_n$ is the right quantile function with respect to $\hat{F}_n^-(z) = \frac{1}{n}\sum_{i=1}^n 1_{Z_i < z}$. They have the explicit formulas
$$\hat{Q}_n(\alpha) = Z_{(\lceil \alpha n \rceil)}; \qquad \hat{R}_n(\alpha) = Z_{(\lfloor \alpha n \rfloor + 1)};$$
where $Z_{(k)}$ denotes the kth smallest value in $Z_1, \ldots, Z_n$.

Variants of the following lemmas appear in the literature [3, 15, 50]. In the interest of clarity and a self-contained exposition, we state and prove them here.

Lemma 1 (Quantiles and exchangeability). *Suppose $Z_1, \ldots, Z_n$ are exchangeable random variables. For any $\alpha \in (0,1)$,*
$$P\{Z_n \le \hat{Q}_n(\alpha)\} \ge \alpha:$$
Moreover, if the random variables $Z_1, \ldots, Z_n$ are almost surely distinct, then also
$$P\{Z_n \le \hat{Q}_n(\alpha)\} \le \alpha + \frac{1}{n}:$$
In this statement, the probabilities are taken over all the variables $Z_1, \ldots, Z_n$.

Proof. By exchangeability and the symmetry of $\hat{Q}_n(\alpha)$ as a function of $Z_1, \ldots, Z_n$, the probability $P\{Z_i \le \hat{Q}_n(\alpha)\}$ is equal to $P\{Z_n \le \hat{Q}_n(\alpha)\}$ for every i. Therefore,
$$E\,\hat{F}_n(\hat{Q}_n(\alpha)) = \frac{1}{n}\sum_{i=1}^n P\{Z_i \le \hat{Q}_n(\alpha)\} = P\{Z_n \le \hat{Q}_n(\alpha)\}:$$

By the defining property of the quantile functions, $\hat{F}_n(\hat{Q}_n(\alpha)) \ge \alpha$ and $\hat{F}_n^-(\hat{R}_n(\alpha)) \le \alpha$. Moreover, if the samples $Z_1, \ldots, Z_n$ are distinct, then $\lVert\hat{F}_n - \hat{F}_n^-\rVert_1 \le \frac{1}{n}$, and since $\hat{Q}_n \le \hat{R}_n$, we have $\hat{F}_n(\hat{Q}_n(\alpha)) \le \hat{F}_n(\hat{R}_n(\alpha)) \le \hat{F}_n^-(\hat{R}_n(\alpha)) + \frac{1}{n} \le \alpha + \frac{1}{n}$. To complete the proof, take expectations of the inequalities $\hat{F}_n(\hat{Q}_n(\alpha)) \ge \alpha$ and $\hat{F}_n(\hat{Q}_n(\alpha)) \le \alpha + \frac{1}{n}$. $\qquad\square$

Lemma 2 (Inflation of quantiles). *Suppose $Z_1, \ldots, Z_{n+1}$ are exchangeable random variables. For any $\alpha \in (0,1)$,*
$$P\{Z_{n+1} \le \hat{Q}_n((1 + \tfrac{1}{n})\alpha)\} \ge \alpha:$$
Moreover, if the random variables $Z_1, \ldots, Z_{n+1}$ are almost surely distinct, then also
$$P\{Z_{n+1} \le \hat{Q}_n((1 + \tfrac{1}{n})\alpha)\} \le \alpha + \frac{1}{n}:$$

Proof. Let $Z_{(k;m)}$ denote the kth smallest value in $Z_1, \ldots, Z_m$. Then for any $0 \le k \le n$, we have
$$Z_{n+1} \le Z_{(k;n)} \quad \text{if and only if} \quad Z_{n+1} \le Z_{(k;n+1)}:$$
Indeed, if $Z_{n+1} \le Z_{(k;n)}$, then $Z_{(k;n+1)}$ is the larger of $Z_{(k-1;n)}$ and Z_{n+1}; in particular, $Z_{(k;n+1)} \ge Z_{n+1}$. Conversely, if $Z_{n+1} \le Z_{(k;n+1)}$ then also $Z_{n+1} \le Z_{(k;n)}$ because $Z_{(k;n+1)} \le Z_{(k;n)}$.

Thus, since $\hat{Q}_n((1 + \tfrac{1}{n})\alpha) = Z_{(\lceil (n+1)\alpha\rceil;n)}$ and $\hat{Q}_{n+1}(\alpha) = Z_{(\lceil (n+1)\alpha\rceil;n+1)}$, we have
$$Z_{n+1} \le \hat{Q}_n((1 + \tfrac{1}{n})\alpha) \quad \text{if and only if} \quad Z_{n+1} \le \hat{Q}_{n+1}(\alpha)$$
and, hence,
$$P\{Z_{n+1} \le \hat{Q}_n((1 + \tfrac{1}{n})\alpha)\} = P\{Z_{n+1} \le \hat{Q}_{n+1}(\alpha)\}:$$
To conclude the proof, apply Lemma 1 with n replaced by $n + 1$. $\qquad\square$