A comparison of some conformal quantile regression methods

Matteo Sesia1 | Emmanuel J. Candès1,2

1Department of Statistics, Stanford University, Stanford, CA 94305, USA
2Department of Mathematics, Stanford University, Stanford, CA 94305, USA

Correspondence
Matteo Sesia, Department of Statistics, 390 Jane Stanford Way, Stanford University, Stanford, CA 94305, USA. Email: msesia@stanford.edu

Funding information
Army Research Office, Grant/Award Number: W911NF-17-1-0304; Division of Mathematical Sciences, Grant/Award Number: DMS 1712800

KEYWORDS
conformal inference, neural networks, quantile regression, random forests

1 | INTRODUCTION

1.1 | Background and motivation

Given a set of $n$ points $\{(X_i, Y_i)\}_{i=1}^n$, with $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^d$, we consider the problem of constructing a prediction interval for a new point $Y_{n+1}$ based on the observed value of $X_{n+1}$, assuming only that $\{(X_i, Y_i)\}_{i=1}^n$ are drawn exchangeably from some common distribution $P_{XY}$. There exists a vast selection of statistical and machine learning algorithms that can provide approximate answers to this question (Papadopoulos, Edwards, & Murray, 2001; Wager, Hastie, & Efron, 2014). However, the uncertainty in any of their predictions cannot be quantified without making strong assumptions and large-sample asymptotic approximations that may not be easily justifiable in applications. Conformal inference (Vovk, Gammerman, & Saunders, 1999; Vovk, Gammerman, & Shafer, 2005; Vovk, Nouretdinov, & Gammerman, 2009; Papadopoulos, Proedrou, Vovk, & Gammerman, 2002; Papadopoulos, Vovk, & Gammerman, 2007; Papadopoulos, Gammerman, & Vovk, 2008; Papadopoulos, 2008; Papadopoulos, Vovk, & Gammerman, 2011; Lei, G’Sell, Rinaldo, Tibshirani, & Wasserman, 2018) addresses this problem by constructing an exact marginal prediction interval $\hat{C}_\alpha(X_{n+1})$ such that

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

while relying only on the exchangeability of the $n + 1$ points. This interval is said to be marginal because all variables in (1) are treated as random, including $(X_{n+1}, Y_{n+1})$ and the data used to train $\hat{C}$. Therefore, it is not guaranteed that the interval will cover $Y_{n+1}$ conditional on a particular observed value of $X_{n+1}$, or a fixed prediction model $\hat{C}$. Despite this limitation, conformal prediction intervals are attractive because their coverage is guaranteed on average regardless of the distribution of the data.

Early work on conformal prediction focused on estimating a mean regression function for $Y|X$ and building a fixed-width band around it (Vovk et al., 1999; Vovk et al., 2005; Vovk et al., 2009; Lei et al., 2018). Even though this strategy produces valid marginal prediction intervals regardless of $P_{Y|X}$, it is clearly designed with a homoscedastic regression model in mind, and it may lead to unnecessarily wide intervals in other cases. Locally adaptive conformal prediction (Papadopoulos et al., 2007; Papadopoulos et al., 2008; Papadopoulos et al., 2011; Lei et al., 2018) goes a step beyond this model by weighting the residuals according to a local estimate of their variance. However, in general, this still cannot achieve the ideal goal of constructing prediction intervals that are as narrow as possible while maintaining coverage. This goal can be stated precisely as follows. Let us denote by $q_a(X_{n+1})$ the $a$th quantile of the conditional distribution of $Y$ given $X_{n+1} = x_{n+1}$. Then a desirable oracle prediction interval would be

$$C_{\text{oracle}}(X_{n+1}) = [q_{a_1}(X_{n+1}), q_{a_2}(X_{n+1})]. \quad (2)$$

We compare two recent methods that combine conformal inference with quantile regression to produce locally adaptive and marginally valid prediction intervals under sample exchangeability (Romano, Patterson, & Candès, 2019, arXiv:1905.03222; Kivaranovic, Johnson, & Leeb, 2019, arXiv:1905.10634). First, we prove that these two approaches are asymptotically efficient in large samples, under some additional assumptions. Then we compare them empirically on simulated and real data. Our results demonstrate that the method of Romano et al. typically yields tighter prediction intervals in finite samples. Finally, we discuss how to tune these procedures by fixing the relative proportions of observations used for training and conformalization. Our empirical results suggest that using between 70% and 90% of the data for training often achieves a good balance between minimizing the average width of the predictions intervals and the variability in their practical coverage.
By construction, this is the narrowest symmetric prediction interval that has valid coverage conditional on the value of $X_{n+1}$. Here, we say that a prediction interval is symmetric if $Y_{n+1}$ is equally likely to be smaller or larger than predicted. Unfortunately, the oracle interval in (2) is unachievable in practice because we do not know $P_{Y|X}$. Conformal quantile regression (Romano, Patterson, & Candès, 2019) constructs a practical prediction interval $\hat{C}$ that estimates (2) as closely as possible while satisfying (1) exactly. In this work, we compare theoretically and empirically the method from Romano et al. (2019), which has already been shown to outperform earlier methods in practice (Romano et al., 2019), with a similar approach that was proposed independently in Kivaranovic, Johnson, and Leeb (2019).

1.2 Conformal quantile regression

Throughout this paper, we follow the split-conformal approach to conformal inference (Papadopoulos et al., 2002; Papadopoulos, 2008; Lei et al., 2018) adopted in Romano et al. (2019) and Kivaranovic et al. (2019), since it is computationally feasible even with large data. The first step of the conformal quantile regression method in Romano et al. (2019) is to split the data samples into two disjoint subsets, $I_1$ and $I_2$. Lower and upper quantile regression functions, namely, $\hat{q}_{\alpha/2}$, $\hat{q}_{1-\alpha/2} : \mathbb{R}^d \to \mathbb{R}$, are fitted on the observations in $I_1$. Any algorithm can be employed for this purpose; for example, one may rely on linear regression (Koenker & Bassett Jr, 1978), neural networks (Taylor, 2000), or random forests (Meinshausen, 2006). In any case, this algorithm is treated as a black box. The estimated quantile functions are used to compute a conformity score for each $i \in I_2$:

$$E_{i}^{CQR} = \max \left\{ \hat{q}_{\alpha/2}(X_i) - Y_i, Y_i - \hat{q}_{1-\alpha/2}(X_i) \right\}.$$  

(3)

Then, with $\hat{Q}_{\alpha}(E_{CQR}; I_2)$ defined as the $(\lfloor (1 - \alpha) |I_2| + 1 \rfloor)$-th largest element$^1$ of $\{ E_i \}_{i \in I_2}$, the conformal prediction interval for $X_{n+1}$ is given by

$$\hat{C}_{CQR}^{\alpha}(X_{n+1}) = \left[ \hat{q}_{\alpha/2}(X_{n+1}) - \hat{Q}_{\alpha}(E_{CQR}; I_2), \hat{q}_{1-\alpha/2}(X_{n+1}) + \hat{Q}_{\alpha}(E_{CQR}; I_2) \right].$$  

(4)

This method is summarized in Algorithm 1 as CQR. It is shown in Romano et al. (2019) that $\hat{C}_{CQR}^{\alpha}(X_{n+1})$ has marginal coverage at level $1 - \alpha$.

|Algorithm 1 Conformal quantile regression|
|---|
|**Input:**|
|data $(\{X_i, Y_i\})_{n \in \mathbb{N}}$, covariates for new sample $X_{n+1}$;|
|proportion of data for training $\gamma \in (0, 1)$;|
|quantile regression algorithm $\hat{q}$;|
|conformalization method $\vartheta \in \{\text{CQR}, \text{CQR-m}, \text{CQR-r}\}$;|
|coverage level $\alpha \in (0, 1)$;|
|**Procedure:**|
|randomly split $\{1, \ldots, n\}$ into $I_1$, $I_2$, of size $|I_1| = \gamma n$, $|I_2| = n - |I_1|$;|
|fit the quantile regression functions $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ on $(\{X_i, Y_i\})_{i \in I_1}$;|
|If $\vartheta = \text{CQR}$:
|compute the conformity scores $E_{i}^{CQR}$ for each $i \in I_2$, as in (3);|
|compute $\hat{Q}_{\alpha}(E_{CQR}; I_2)$;|
|compute the prediction interval $\hat{C}_{\vartheta}(X_{n+1})$, as in (4).|
|If $\vartheta = \text{CQR-m}$:
|fit the median regression function $\hat{q}_{\vartheta}(\cdot)$ on $(\{X_i, Y_i\})_{i \in I_1}$;|
|compute the conformity scores $E_{i}^{CQR-m}$ for each $i \in I_2$, as in (5);|
|compute $\hat{Q}_{\vartheta}(E_{CQR-m}; I_2)$;|
|compute the prediction interval $\hat{C}_{\vartheta}(X_{n+1})$, as in (6).|
|If $\vartheta = \text{CQR-r}$:
|compute the conformity scores $E_{i}^{CQR-r}$ for each $i \in I_2$, as in (7);|
|compute $\hat{Q}_{\vartheta}(E_{CQR-r}; I_2)$;|
|compute the prediction interval $\hat{C}_{\vartheta}(X_{n+1})$, as in (8).|
|**Output:** A prediction interval $\hat{C}_{\vartheta}(X_{n+1})$.|

The method described in Kivaranovic et al. (2019) differs from CQR in the choice of the conformity scores, as outlined in Algorithm 1 as CQR-m. Instead of (3), one computes$^2$

$$E_{i}^{CQR-m} = \max \left\{ \frac{\hat{q}_{\alpha/2}(X_i) - Y_i}{\hat{q}_{\alpha/2}(X_i) - \hat{q}_{1-\alpha/2}(X_i)}, \frac{Y_i - \hat{q}_{1-\alpha/2}(X_i)}{\hat{q}_{1-\alpha/2}(X_i) - \hat{q}_{\alpha/2}(X_i)} \right\}.$$  

(5)

$^1$For a motivation of this definition, see the result on the empirical quantiles of exchangeable random variables (Lemma 1) in Romano et al. (2019).

$^2$Note that we present CQR-m with a slightly different notation than in Kivaranovic et al. (2019) to facilitate the comparison.
where \( \hat{q}_{1/2} \) indicates an estimated median regression function obtained with the same black box algorithm as \( q_{1/2} \) and \( \hat{q}_{1-\alpha/2} \). Then the conformal prediction interval for \( X_{n+1} \) is given by

\[
\hat{C}^{\text{CQR-m}}_{\alpha}(X_{n+1}) = \left[ \hat{q}_{1/2}(X_{n+1}) - \Delta^{\text{CQR-m}}_{\text{lo}}, \hat{q}_{1-\alpha/2}(X_{n+1}) + \Delta^{\text{CQR-m}}_{\text{up}} \right],
\]

where

\[
\Delta^{\text{CQR-m}}_{\text{lo}} = \hat{Q}_{1\text{-}\alpha}(E^{\text{CQR-m}}; I_2) \left[ \hat{q}_{1/2}(X_{n+1}) - \hat{q}_{1-\alpha/2}(X_{n+1}) \right],
\]

\[
\Delta^{\text{CQR-m}}_{\text{up}} = \hat{Q}_{1\text{-}\alpha}(E^{\text{CQR-m}}; I_2) \left[ \hat{q}_{1-\alpha/2}(X_{n+1}) - \hat{q}_{1/2}(X_{n+1}) \right].
\]

One can show that \( \hat{C}^{\text{CQR-m}}_{\alpha}(X_{n+1}) \) also has marginal coverage at level \( 1 - \alpha \) (Kivaranovic et al., 2019). We also find it interesting to consider a modified version of CQR-m that does not require estimating the regression median. This third approach, listed in Algorithm 1 as CQR-r, is based on the following conformity scores:

\[
E^{\text{CQR-r}}_{\alpha} = \max \left\{ \frac{\hat{q}_{1/2}(X_i) - Y_i}{\hat{q}_{1-\alpha/2}(X_i) - \hat{q}_{1/2}(X_i)}, \frac{Y_i - \hat{q}_{1-\alpha/2}(X_i)}{\hat{q}_{1-\alpha/2}(X_i) - \hat{q}_{1/2}(X_i)} \right\}. \tag{7}
\]

The CQR-r prediction intervals are

\[
\hat{C}^{\text{CQR-r}}_{\alpha}(X_{n+1}) = \left[ \hat{q}_{1/2}(X_{n+1}) - \Delta^{\text{CQR-r}}_{\text{lo}}, \hat{q}_{1-\alpha/2}(X_{n+1}) + \Delta^{\text{CQR-r}}_{\text{up}} \right], \tag{8}
\]

\[
\Delta^{\text{CQR-r}}_{\text{lo}} = \hat{Q}_{1\text{-}\alpha}(E^{\text{CQR-r}}; I_2) \left[ \hat{q}_{1/2}(X_{n+1}) - \hat{q}_{1-\alpha/2}(X_{n+1}) \right],
\]

\[
\Delta^{\text{CQR-r}}_{\text{up}} = \hat{Q}_{1\text{-}\alpha}(E^{\text{CQR-r}}; I_2) \left[ \hat{q}_{1-\alpha/2}(X_{n+1}) - \hat{q}_{1/2}(X_{n+1}) \right].
\]

It is easy to show that \( \hat{C}^{\text{CQR-r}}_{\alpha}(X_{n+1}) \) also attains marginal coverage at level \( 1 - \alpha \). A proof is omitted because it would be identical to those in Romano et al. (2019) and Kivaranovic et al. (2019). CQR-r is similar in spirit to CQR-m, but it has a more direct interpretation: the conformity scores of CQR-r in (7) weight the distance of \( Y \) from the corresponding prediction interval by the inverse width of the interval. Therefore, the conformalization expands or contracts the black box prediction bands proportionally to their width, instead of adding a constant shift as in CQR. Since it is not clear how the regression median \( q_{1/2} \) should generally be related to the upper and lower \( \alpha \)-quantiles of \( P_{XY} \), we find this approach slightly more intuitive than CQR-m.

2 \| THEORETICAL ANALYSIS

We show that the output of Algorithm 1 converges to the oracle bands in (2) as \( n \) grows, if the black box quantile regression estimates are consistent and a few additional assumptions hold. This can be established for any of the three alternative types of conformity scores discussed in this paper, which are therefore asymptotically equivalent in this sense.

**Assumption 1** (i.i.d.). The points \( (X_i, Y_i)_{i=1}^{n+1} \) are drawn i.i.d. from some distribution \( P_{XY} \).

**Assumption 2** (Regularity). The probability density of the conformity scores, either in (3), (5), or (7), depending on the conformalization method in Algorithm 1, is bounded away from zero in an open neighbourhood of zero.

**Assumption 3** (Consistency). For simplicity, denote by \( n \) the size of the training data set \( I_1 \) used to fit the quantile regression functions \( \hat{q} \).

Let \( X \) be a new observation independent of \( I_1 \). Then the assumption is that, for \( n \) large enough,

\[
P[|E(\hat{q}_{1/2}(X) - q_{1/2}(X))| \leq \eta_n] \geq 1 - \rho_n,
\]

\[
P[|E(\hat{q}_{1-\alpha/2}(X) - q_{1-\alpha/2}(X))| \leq \eta_n] \geq 1 - \rho_n,
\]

for some sequences \( \eta_n = o(1) \) and \( \rho_n = o(1) \), as \( n \to \infty \).

Assumption 3 is similar to that used in Lei et al. (2018) for mean regression estimators, and it is weaker than requiring \( \hat{q}_{1/2}(X) \overset{L^2}{\longrightarrow} q_{1/2}(X) \) and \( \hat{q}_{1-\alpha/2}(X) \overset{L^2}{\longrightarrow} q_{1-\alpha/2}(X) \) as \( n \to \infty \), by Markov’s inequality.

**Theorem 1.** Under Assumptions 1–3, the conformal quantile regression bands \( \hat{C}_\alpha \) obtained with Algorithm 1 satisfy

\[
L \left( \hat{C}_\alpha(X_{n+1}) \bigtriangleup C^{\text{oracle}}_\alpha(X_{n+1}) \right) = o_P(1),
\]

as \( |I_1|, |I_2| \to \infty \). Here, \( L(A) \) indicates the Lebesgue measure of the set \( A \), and \( \bigtriangleup \) is the symmetric difference operator, i.e., \( A \bigtriangleup B = (A \setminus B) \cup (B \setminus A) \).

The proof of Theorem 1 can be found in the Supporting Information and is inspired by that of Theorem 3.4 in Lei et al. (2018), although the oracle and the conformalization methods considered here are different. Theorem 1 establishes a stronger form of statistical efficiency for conformal quantile regression compared with the result in Lei et al. (2018), which assumes \( Y = \mu(X) + \epsilon \), for some regression function \( \mu \), and homoscedastic noise \( \epsilon \). In general, the conformal prediction intervals described in Lei et al. (2018) will not converge to those of our oracle if

---

3This was first suggested by Yaniv Romano through personal communication.
the noise is heteroscedastic, regardless of the consistency of the black box regression estimator \( \hat{\mu} \). By contrast, conformal quantile regression is efficient in the sense that, under Theorem 1, the prediction bands converge to those of the oracle, which are the narrowest possible bands achieving conditional coverage. Finally, the asymptotic consistency assumption may be verified theoretically for some specific algorithms under certain conditions, for example, random forests (Meinshausen, 2006). In any case, our result provides some theoretical backing to conformal quantile regression even if the assumptions cannot be verified in practice. It also follows as a corollary that conformal quantile regression has asymptotic conditional coverage, defined as in Lei et al. (2018); the proof of this corollary can be found in the Supporting Information.

**Definition 1** (Asymptotic conditional coverage). We say that a sequence \( \hat{C}_n \) of random prediction bands has asymptotic conditional coverage at the level \( 1 - \alpha \) if there exists a sequence of random sets \( \Lambda_n \subseteq \mathbb{R}^d \) such that \( \mathbb{P}[X \in \Lambda_n] = 1 - \alpha(1) \) and

\[
\sup_{x \in \Lambda_n} \mathbb{P}[Y \in \hat{C}_n(x)|X = x] - (1 - \alpha) = o_{\mathbb{P}}(1).
\]

**Corollary 1** (Asymptotic conditional coverage). Under the assumptions of Theorem 1, assuming also for simplicity that \( Y \) has a continuous density with respect to the Lebesgue measure, conformal quantile regression bands satisfy the asymptotic conditional coverage property in Definition 1.

Despite being asymptotically efficient under Assumptions 1–3, the three conformalization methods in Algorithm 1 typically perform differently with finite data, as discussed next.

### 3 | EMPIRICAL COMPARISON

#### 3.1 | Black box quantile regression

In the following, we utilize two alternative black box quantile regressors, implemented and trained as in Romano et al. (2019). The first procedure is based on quantile regression forests (Meinshausen, 2006). We fit 1,000 trees and set the other tuning parameters equal to their default values. The second black box is a neural network (Taylor, 2000) with three fully connected layers and ReLU non-linearities. We have chosen this design, which is slightly different from that in Kivaranovic et al. (2019), because it leads to conformal prediction intervals that are tighter than those reported in Kivaranoovic et al. (2019). If the estimated lower and upper quantiles overlap, which may sometimes occur, we swap them. The nominal level of the black boxes is tuned so that their empirical coverage, estimated by cross-validation, is approximately equal to \( 1 - 2\alpha \).

We have observed that this heuristic generally leads to tighter conformal intervals compared with those obtained by directly requesting the black boxes to estimate \( q_{\alpha/2} \) and \( q_{1-\alpha/2} \); for empirical evidence, see Section 3.3 and Figures S2–S11. Throughout this section, we set \( \alpha = 0.1 \), while the results of additional experiments exploring different values of \( \alpha \) are shown in Figure S1.

#### 3.2 | Experiments with artificial data

We begin by considering the same experiment based on artificial data as in Kivaranovic et al. (2019). We simulate \( X \sim \text{Unif}(0,1)^d \), for \( d = 100 \), and \( Y \in \mathbb{R} \) from

\[
Y = f(\beta'X) + \epsilon \sqrt{1 + (\beta'X)^2},
\]

where \( f(x) = 2\sin(\pi x) + \pi x, \beta' = (1,1,1,1,0,\ldots,0) \) and \( \epsilon \) is independent standard Gaussian noise. Here, we have access to a natural benchmark: the oracle that knows \( P_{Y|X} \) exactly. It follows from (9) that the expected width of the oracle prediction bands is

\[
\mathbb{E}[q_{1-\alpha/2}(X) - q_{\alpha/2}(X)] = 2 \mathbb{E}\sqrt{1 + (\beta'X)^2} Q_{1-\alpha/2}(\epsilon) \approx 8.91,
\]

where \( Q_{\alpha}(\epsilon) \) is the \( \alpha \)-quantile of the standard Gaussian distribution.

The performances of CQR, CQR-m, and CQR-r are compared in Figure 1 as a function of the number of data points \( n \). The proportion of observations used to train the black box is 3/4, as in Kivaranovic et al. (2019). The coverage and average width of the prediction bands is evaluated on an independent test set of size 20,000. The experiment is repeated for 100 independent realizations of the data and of the test set. The width and coverage of the conformal prediction bands approach those of the oracle as the sample size increases. This suggests that the estimated black box quantiles may be approximately consistent. However, CQR typically produces narrower bands compared with the other methods.

#### 3.3 | Experiments with real data

We now apply Algorithm 1 on the same data analysed in Romano et al. (2019) and Kivaranovic et al. (2019).4 Some details about these data sets and information on the corresponding sources are summarized in Table 1. For all data sets except homes, we randomly hold out 20% of the samples for testing. Then we divide the remaining observations into two disjoint sets, \( I_1 \) and \( I_2 \), to train the black box and conformalize the prediction bands, respectively. The response variables \( Y \) are standardized as in Romano et al. (2019) and Kivaranovic et al. (2019). We explore

---

4 We have excluded the X-ray data in Kivaranovic et al. (2019) because we are unsure of how to replicate the pre-processing.
The performances obtained with different black boxes and values of $\gamma$ are reported in Figure 2 for the community data and in Figures S2–S11 for the other data sets. The results are shown as a function of $\gamma$, which affects the average width of the prediction intervals as well as their variability. If $\gamma$ is small, the prediction intervals are not sufficiently adaptive because the black box cannot estimate the regression quantiles accurately. Larger values of $\gamma$ may lead to tighter predictions on average but at the cost of increased variability in the conditional coverage. In this comparison because different values are used in Romano et al. (2019) and Kivaranovic et al. (2019): $\gamma = 0.5$ and $\gamma = 0.75$, respectively. In the case of the homes data, we follow in the footsteps of Kivaranovic et al. (2019): First, we randomly hold out 3,613 test samples; then we train the black box on 15,000 samples and conformalize on the remaining 3,000. All experiments are repeated 10 times, starting from the data splitting.

The test-set performances of CQR, CQR-m, and CQR-r are summarized in Tables 2 and 3. These quantities correspond to the optimal choices of black box and value of the hyperparameter $\gamma$, defined as those leading to the shortest intervals on average, separately for each algorithm. The CQR method consistently produces the narrowest valid prediction bands, while CQR-m and CQR-r are often comparable.

The performances obtained with different black boxes and values of $\gamma$ are reported in Figure 2 for the community data and in Figures S2–S11 for the other data sets. The results are shown as a function of $\gamma$, which affects the average width of the prediction intervals as well as their variability. If $\gamma$ is small, the prediction intervals are not sufficiently adaptive because the black box cannot estimate the regression quantiles accurately. Larger values of $\gamma$ may lead to tighter predictions on average but at the cost of increased variability in the conditional coverage. In fact, the conditional coverage for new observations given the data may be lower than the expected marginal level, especially when $\gamma$ is very close to one and the sample size is not very large. The empirical results in Figure 2 and in Figures S2–S11 suggest that fixing $\gamma \in [0.7, 0.9]$ achieves a
### Table 3

Average coverage (and standard deviation) of the prediction bands in Table 2

| Dataset    | Coverage         | Coverage         | Coverage         |
|------------|------------------|------------------|------------------|
|            | CQR              | CQR-r            | CQR-m            |
| bike       | 0.899 (0.012)    | 0.901 (0.012)    | 0.900 (0.012)    |
| bio        | 0.891 (0.012)    | 0.893 (0.016)    | 0.895 (0.008)    |
| blog       | 0.905 (0.003)    | 0.901 (0.007)    | 0.903 (0.004)    |
| community  | 0.896 (0.025)    | 0.899 (0.025)    | 0.902 (0.017)    |
| concrete   | 0.875 (0.061)    | 0.879 (0.061)    | 0.877 (0.059)    |
| facebook-1 | 0.901 (0.006)    | 0.898 (0.004)    | 0.902 (0.002)    |
| facebook-2 | 0.900 (0.003)    | 0.900 (0.002)    | 0.900 (0.002)    |
| homes      | 0.902 (0.009)    | 0.904 (0.009)    | 0.904 (0.009)    |
| meps-19    | 0.902 (0.008)    | 0.902 (0.007)    | 0.900 (0.011)    |
| meps-20    | 0.897 (0.004)    | 0.898 (0.004)    | 0.899 (0.006)    |
| meps-21    | 0.899 (0.008)    | 0.898 (0.009)    | 0.897 (0.009)    |
|            | 0.905 (0.024)    | 0.904 (0.024)    | 0.903 (0.020)    |

**FIGURE 2** Conformal prediction bands obtained with different black boxes and conformalization methods on the community data set from Table 1. The dotted line in the lower plots indicates the nominal coverage level (90%). A different black box is considered in each column. The vertical axis in the upper panels is discontinuous to facilitate the visualization of values on different scales.

A reasonable compromise for all data sets analysed in this paper. More explicitly, we see that larger values of $\gamma$ typically yield shorter intervals, but there is often little improvement between 0.7 and 0.9, while increasing $\gamma$ above 0.9 usually significantly increases the variance of the conditional coverage. Finally, we note that our recommendation is consistent with the choice in Kivaranovic et al. (2019).

The CQR-m method sometimes produces very wide intervals because the denominator in (5) can be close to zero (we added a small constant to prevent overflowing). An example is visible in the second plot in Figure 2, where some of the CQR-m prediction intervals based on a random forest black box are extremely large (hence the discontinuous vertical axis in Figure 2) when $\gamma \geq 0.9$. The CQR-r method is less susceptible to this problem because the denominator in the conformity scores in (7) is larger.
4  |  CONCLUSION
In this paper, we have strengthened the case for conformal quantile regression as a method to obtain valid marginal prediction intervals that are adaptive to heteroscedasticity, by proving that it is asymptotically efficient in large samples if the quantile regression estimates are consistent. The empirical comparison of three alternative conformity scores has shown that those proposed in Romano et al. (2019) are preferable because they typically lead to shorter prediction intervals in practice. Even though we have only explicitly considered symmetric intervals for simplicity, it is straightforward to generalize these methods to asymmetric intervals and conformity scores (Romano et al., 2019). Finally, we have highlighted a bias-variance trade-off in the choice of the proportion of data points used to train the black box quantile regressors. Our empirical results show that it is usually better to invest more of the available data (between 70% and 90%, indicatively) to train the black box than to conformalize the predictions. We hope that these results will be helpful to practitioners and may inspire others to develop even more powerful variations of conformal quantile regression.

ACKNOWLEDGEMENTS
M.S. and E.C. are supported by the National Science Foundation under Grant DMS 1712800. E.C. is also supported by the Army Research Office under Grant W911NF-17-1-0304. We thank Yaniv Romano for helpful discussions, during which he suggested the CQR-r method.

DATA AVAILABILITY STATEMENT
The data used in this work are openly available from the sources listed in Table 1. The code to reproduce the numerical experiments in Section 3 can be found on https://github.com/msesia/cqr-comparison.

ORCID
Matteo Sesia  https://orcid.org/0000-0001-9046-907X

REFERENCES
Achilles, C. M., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., ..., & Word, E. (2008). Tennessee's student teacher achievement ratio (STAR) project data set. Harvard Dataverse, Accessed: July, 2019.
Bike sharing data set (2013). https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset, Accessed: July, 2019.
BlogFeedback data set (2014). https://archive.ics.uci.edu/ml/datasets/BlogFeedback, Accessed: July, 2019.
Communities and crime data set (2009). http://archive.ics.uci.edu/ml/datasets/communities+and+crime, Accessed: July, 2019.
Concrete compressive strength data set (2007). http://archive.ics.uci.edu/ml/datasets/concrete+compressive+strenght, Accessed: July, 2019.
Facebook comment volume data set (2016). https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset, Accessed: July, 2019.
House sales in King County data set (2016). https://www.kaggle.com/harlfoxem/housesalesprediction/metadata, Accessed: August, 2019.
Medical expenditure panel survey (panel 19) data set (2018). https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192, Accessed: July, 2019.
Medical expenditure panel survey (panel 20) data set (2018). https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192, Accessed: July, 2019.
Medical expenditure panel survey (panel 21) data set (2018). https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181, Accessed: July, 2019.
Meinshausen, N. (2006). Quantile regression forests. Journal of Machine Learning Research, 7, 983–999.
Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In Fritzsche, P. (Ed.), Tools in artificial intelligence. Rijeka: IntechOpen, 315–330.
Papadopoulos, G., Edwards, P. J., & Murray, A. F. (2001). Confidence estimation methods for neural networks: A practical comparison. IEEE Transactions on Neural Networks, 12(6), 1278–1287.
Papadopoulos, H., Gammerman, A., & Vovk, V. (2008). Normalized nonconformity measures for regression conformal prediction. In Proceedings of the 26th IASTED international conference on artificial intelligence and applications, AIA '08, ACTA Press, Anaheim, CA, USA, pp. 64–69.
Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. In European Conference on Machine Learning, Berlin, Heidelberg: Springer, pp. 345–356.
Papadopoulos, H., Vovk, V., & Gammerman, A. (2007). Conformal prediction with neural networks. In 19th IEEE international conference on tools with artificial intelligence (ICTAI), 2, Patras, Greece, pp. 388–395.
Papadopoulos, H., Vovk, V., & Gammerman, A. (2011). Regression conformal prediction with nearest neighbours. Journal of Artificial Intelligence Research, 40(1), 815–840.
Physicochemical properties of protein tertiary structure data set (2013). https://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure, Accessed: July, 2019.
Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. arXiv preprint arXiv:1905.03222.
Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4), 299–311.
Vovk, V., Gammerman, A., & Saunders, C. (1999). Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML ’99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 444–453.
Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Berlin, Heidelberg: Springer-Verlag.
Vovk, V., Nourdinov, I., & Gammerman, A. (2009). On-line predictive linear regression. *Annals of Statistics*, 37(3), 1566–1590.
Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1), 1625–1651.

**SUPPORTING INFORMATION**
Additional supporting information may be found online in the Supporting Information section at the end of the article.

---

**How to cite this article:** Sesia M, Candès E. A comparison of some conformal quantile regression methods. *Stat.* 2020;9:e261.  
https://doi.org/10.1002/sta4.261