Conformalized Composite Quantile Regression

Suin Kim[1] and Yoonsuh Jung[1*]

[1*]Department of Statistics, Korea University, Anam-ro 145, Seoul, 02841, South Korea.

*Corresponding author(s). E-mail(s): yoons77@korea.ac.kr;
Contributing authors: ksi2002@korea.ac.kr;

Abstract

Conformal prediction is a method for quantifying uncertainty in predictive modeling. It is particularly valuable when applied to black-box models that lack the inherent ability to generate prediction intervals. Conformalized quantile regression has recently emerged as one of the leading conformal prediction methods, which is constructed using estimates of conditional quantiles at the lower and upper tails. However, tail quantile estimates tend to be highly variable, often resulting in conservative and wide prediction intervals. To address these limitations, we propose conformalized composite quantile regression, a method that reduces variance by aggregating the calibration procedure across multiple quantile levels. Our method maintains coverage guarantees without unnecessarily increasing a width of prediction interval. It is computationally efficient as it requires only a single model to estimate multiple quantiles. We conduct extensive experiments on both synthetic and real-world data, including medical imaging tasks using the fastMRI dataset. Empirical results show that the proposed method generally achieves comparable or superior performance to state-of-the-art methods in both coverage and interval width.

Keywords: conformal prediction, ensemble method, multiple quantile regression, uncertainty quantification

1 Introduction

Conformal prediction is a distribution-free method for constructing a prediction interval or set that contains the true response with specified probability, even in finite-sample settings and without strong model assumptions. Given any target coverage level $1-\alpha \in (0,1)$, conformal prediction guarantees that the resulting interval will include the true response with probability at least $1-\alpha$ (Vovk et al. 1999; Lei et al. 2013). It is especially useful in high-stakes domains such as finance, healthcare, and engineering, where accurate uncertainty quantification is critical for informed decision-making (Vovk and Bendtsen 2018).

Conformalized quantile regression (CQR), introduced by Romano et al. (2019), has become one of the most widely used methods in conformal prediction. Conventional methods typically build prediction intervals centered on the conditional mean, often resulting in symmetric intervals that fail to account for heteroscedasticity and distributional asymmetry. In contrast, CQR uses conditional quantile estimates to produce intervals that adapt to the underlying distribution of the response. As a result, CQR provides more reliable prediction intervals, particularly under heteroscedasticity.

One major drawback of CQR is its reliance on tail quantile estimates, namely the $\alpha/2$ and $1-\alpha/2$ levels, for achieving nominal $1-\alpha$ coverage. Quantile estimators become increasingly unstable as the quantile level approaches 0 or 1 (Koenker and Bassett 1978; Koenker 2005). This form of uncertainty, referred to as epistemic uncertainty, affects the conformal prediction process, often leading to more conservative calibration adjustments to ensure valid coverage. As a result, the intervals become overly conservative, particularly in small samples or when the quantile regression model is unstable.

To overcome these limitations, we propose conformalized composite quantile regression (CCQR). Our method improves the stability of prediction intervals by incorporating information from multiple quantile levels. Previous studies have shown that

aggregating multiple quantile estimates can reduce variance in estimation tasks (Zou and Yuan 2008; Bloznelis et al. 2019). To the best of our knowledge, CCQR is the first approach to extend this idea to conformal prediction. Instead of relying solely on tail quantiles, CCQR jointly estimates and calibrates quantile functions across multiple levels within a single model. By reducing variance, CCQR can yield smoother and more stable interval boundaries than those derived from estimates at tail quantiles.

In contrast to ensemble-based conformal methods that rely on multiple models or resampling, CCQR achieves comparable variance reduction at lower computational cost. For example, Yang and Kuchibhotla (2025) aggregates prediction intervals from independently training models, and Carlsson et al. (2014) applies conformal prediction to resampled or cross-validated subsets and merges the results. Since CCQR is based on joint estimates obtained from single model, it requires no additional training procedures. Therefore, our method retains the simple process of standard conformal prediction while benefiting from the ensemble effect.

We concisely illustrate the advantages of our method through visualization. Figure 1 shows a toy example drawn from the synthetic dataset used in our sensitivity analysis. We compare standard CQR with our proposed method using two base models: quantile regression forest (QRF; Meinshausen (2006)) and monotone composite quantile regression neural network (MCQRNN; Cannon (2018)). We compare the prediction intervals produced by CQR and CCQR for each base model. While both methods attain the target coverage, CQR yields wider intervals in certain regions. This stems from the high variance of quantile estimates near the tail of distribution. In contrast, the proposed method yields tighter and more stable intervals due to the variance reduction gained from aggregating across multiple quantile levels.

The remainder of this paper is organized as follows. Section 2 reviews the foundational concepts and theoretical background underlying our approach. In Section 3,

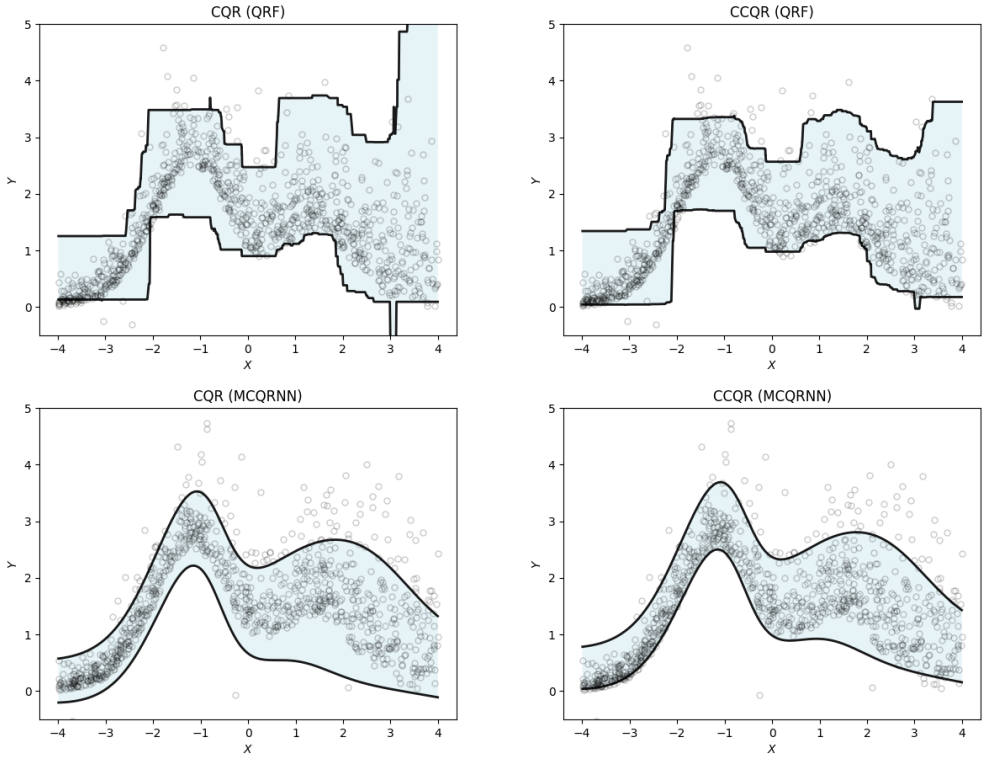

Fig. 1: 90% prediction intervals produced by CQR (left panels) and CCQR (right panels).

we present our proposed method along with several variants. We establish the theoretical coverage guarantee and provide a detailed explanation of how our method achieves variance reduction. Section 4 provides an empirical evaluation across synthetic and real-world datasets. The source code for our implementation is available at https://github.com/suinkim96/ccqr.

2 Background

Before introducing our proposed methodology, we first outline several key concepts.

Quantile regression: Introduced by Koenker and Bassett (1978), quantile regression (QR) estimates the τ-th conditional quantile of a response variable Y given a vector

of covariates $X = x$. The τ-th conditional quantile function is formally defined as $q_\tau(x) = \inf\{y \in \mathbb{R} : \mathbb{P}(Y \leq y \mid X = x) \geq \tau\}$, where $\tau \in (0, 1)$.

The quantile function estimator $\hat{q}_\tau(x)$ is obtained by minimizing the check loss: $L_\tau(u) = u(\tau - \mathbb{I}(u < 0))$. Here, $\mathbb{I}(\cdot)$ denotes the indicator function. Asymptotic consistency for quantile estimation have been established for linear, kernel-based, and spline-based models (Koenker and Bassett 1978; Takeuchi et al. 2006; He and Shi 1994). Popular machine learning models such as QRF (Meinshausen 2006), boosting-based models (Zheng 2012), and quantile deep neural networks (Xu et al. 2017; Moon et al. 2021) have demonstrated strong empirical performance in quantile estimation.

Models that support simultaneous estimation of multiple quantiles are particularly well-suited for our proposed method, which relies on multiple quantile estimates. For example, QRF can estimates all quantile levels with a single model by retaining leaf weights during training process (Meinshausen 2006; Wang et al. 2022). Similarly, neural networks can estimate several quantiles at once by setting multiple output nodes and optimizing a loss function that combines the corresponding check losses (Xu et al. 2017; Cannon 2018). While our method is more efficient when paired with models that produce multiple quantiles simultaneously, it is compatible with models estimating individual quantiles.

Conformal prediction: Conformal prediction (CP) is a methodology of constructing prediction intervals for regression or prediction sets for classification. It offers a finite-sample coverage guarantee with minimal assumptions on the underlying distribution of data (Vovk et al. 1999, 2005; Toccaceli 2022). Under the assumption that the data $(X_i, Y_i)_{i=1}^n$ and a new observation (X_{n+1}, Y_{n+1}) are exchangeable, CP constructs a set or interval, $\hat{C}(X_{n+1})$, for a new input X_{n+1} which satisfies

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

Here, $\alpha \in (0, 1)$ denotes the user-defined miscoverage level. This property is known as the *marginal coverage guarantee*.

A common strategy for ensuring the coverage guarantee in (1) is the split CP method (also known as inductive CP), which divides the data into a training set and a calibration set (Papadopoulos et al. 2002; Papadopoulos 2008). Let $\mathcal{I}_1$ and $\mathcal{I}_2$ denote the index sets for the training and calibration data, respectively, such that $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$ and $\mathcal{I}_1 \cup \mathcal{I}_2 = \{1, \ldots, n\}$. Split CP involves two stages: a predictive model is first trained on the data indexed by $\mathcal{I}_1$, followed by the computation of nonconformity scores using the observations in $\mathcal{I}_2$. Here, nonconformity scores measure the degree to which each calibration point deviates from the predictions of the trained model. The final prediction interval for a new data point is obtained by calibrating the initial prediction using nonconformity scores computed on the calibration set. Under the exchangeability, this procedure can guarantee the marginal coverage even in finite-sample settings.

Throughout the study, we use split CP as our primary tool for calibration. However, our method is fully compatible with other conformal variants, including cross-conformal prediction (Vovk 2015), Jackknife+ (Gupta et al. 2022), and out-of-bag method (Barber et al. 2021).

Various CP methods differ mainly in the way they define and compute nonconformity scores. A commonly used score is the absolute deviation between the observed response and the prediction (Vovk et al. 1999; Lei et al. 2018). Moreover, it can be scaled by estimates of local variance to account for heteroscedasticity (Papadopoulos 2008; Papadopoulos et al. 2011). These methods typically rely on estimators of the conditional mean, $\mathbb{E}(Y|X)$. Although they maintain valid marginal coverage, the resulting intervals are often unnecessarily conservative due to their symmetry around a single point estimate.

Conformalized quantile regression: Romano et al. (2019) introduced CQR, a method that mitigates miscalibration under heteroscedasticity by constructing prediction intervals using conditional quantile estimators. CQR improves performance under both heteroscedasticity and skewness of data by separately estimating the lower and upper conditional quantiles. Ideally, if the true quantile functions, $q_{\alpha/2}(X)$ and $q_{1-\alpha/2}(X)$, are known, the interval $(q_{\alpha/2}, q_{1-\alpha/2})$ achieves exact coverage: $\mathbb{P}\{q_{\alpha/2}(X_{n+1}) \leq Y_{n+1} \leq q_{1-\alpha/2}(X_{n+1})\} = 1-\alpha$. Accordingly, the interval $(q_{\alpha/2}, q_{1-\alpha/2})$ is often referred to as the oracle interval. Using estimated quantiles, CQR approximates the oracle interval and applies calibration to guarantee valid coverage.

Within the split CP framework, a quantile regression model is fitted on the training set $\mathcal{I}_1$ to estimate the conditional quantiles at levels τ_ℓ and τ_u. The corresponding estimates are denoted by $\hat{q}_{\tau_\ell}$ and $\hat{q}_{\tau_u}$, respectively. A common choice of quantile levels is $\tau_\ell = \alpha/2$ and $\tau_u = 1-\alpha/2$. Next, CQR computes nonconformity scores on the calibration set $\mathcal{I}_2$. For each $j \in \mathcal{I}_2$, the score is defined as

$$E_j = \max\left\{\hat{q}_{\tau_\ell}(X_j) - Y_j,\ Y_j - \hat{q}_{\tau_u}(X_j)\right\}, \tag{2}$$

that quantifies the extent to which the observed response Y_j falls outside the uncalibrated prediction interval $[\hat{q}_{\tau_\ell}(X_j), \hat{q}_{\tau_u}(X_j)]$. Let $Q_E(\alpha)$ be the $(1-\alpha)(1+1/|\mathcal{I}_2|)$-th empirical quantile of the scores $\{E_j : j \in \mathcal{I}_2\}$, where $|\cdot|$ denotes cardinality of set. Given a new input X_{n+1}, CQR constructs the prediction interval as

$$\hat{C}(X_{n+1}) = \left[\hat{q}_{\tau_\ell}(X_{n+1}) - Q_E(\alpha),\ \hat{q}_{\tau_u}(X_{n+1}) + Q_E(\alpha)\right]. \tag{3}$$

Under the assumption of exchangeability, the interval defined in (3) satisfies the finite-sample marginal coverage guarantee.

Theorem 1 (Romano et al. (2019)) *Suppose* $\{(X_i, Y_i)\}_{i=1}^{n+1}$ *are exchangeable. The conformalized interval* $\hat{C}(X_{n+1})$ *defined in* (3) *satisfies*

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

Moreover, if the nonconformity scores E_i *for* $i = 1, \ldots, n+1$ *are almost surely distinct, then*

$$\mathbb{P}\{Y_{n+1} \in \hat{C}(X_{n+1})\} \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2| + 1}.$$

Several extensions of CQR include orthogonal quantile regression (Feldman et al. 2021), unconditional quantile regression (Alaa et al. 2023), and domain-specific applications such as medical image translation (Akrami et al. 2024).

Although CQR is broadly applicable, it often depends on tail quantile estimates at levels near 0 or 1, where sampling variability tends to be high. While finite-sample coverage is still maintained, calibration may inflate the width of interval due to high variance. It makes the resulting intervals conservative. The problem becomes especially pronounced when the sample size is small or the distribution of response variable has heavy tails. In response to this limitation, we present our proposed methodology in the next section.

3 Conformalized composite quantile regression

In this section, we explain the detailed algorithm of our suggested method, conformalized composite quantile regression (CCQR). The core idea is that we can reduce variance by aggregating information across multiple quantile estimates.

We first introduce the notation employed in subsequent sections. Let $\{(\ell_k, u_k)\}_{k=1}^{K}$ be a collection of K quantile pairs satisfying $0 < \ell_1 < \ell_2 < \cdots < \ell_K < 0.5$ and $\ell_k + u_k = 1$ for all $k = 1, \ldots, K$. These symmetry and ordering assumptions can be relaxed if needed. For notational convenience, we assume without loss of generality that the first pair corresponds to $(\ell_1, u_1) = (\alpha/2, 1 - \alpha/2)$. Given a trained quantile

regression model $\hat{q}_\tau(\cdot)$, define the averaged lower and upper estimates as $\bar{\ell}(X) := \frac{1}{K}\sum_{k=1}^{K}\hat{q}_{\ell_k}(X)$ and $\bar{u}(X) := \frac{1}{K}\sum_{k=1}^{K}\hat{q}_{u_k}(X)$. Let $E^{(k)} = \{E_j^{(k)} : j \in \mathcal{I}_2\}$ be the set of nonconformity scores, computed as defined in (2), for the k-th quantile pair.

To motivate our method, we begin by illustrating a naive strategy that averages the endpoints of multiple CQR intervals and explaining why it fails to achieve marginal coverage. For each quantile pair (ℓ_k, u_k), CQR yields a prediction interval that satisfies the marginal coverage guarantee, as established in Theorem 1. By taking the average of the K individual intervals, we obtain the following:

$$\left[\frac{1}{K}\sum_{k=1}^{K}\big(\hat{q}_{\ell_k}(X_{n+1}) - Q_{E^{(k)}}(\alpha)\big),\ \frac{1}{K}\sum_{k=1}^{K}\big(\hat{q}_{u_k}(X_{n+1}) + Q_{E^{(k)}}(\alpha)\big)\right], \tag{4}$$

Nevertheless, interval (4) does not ensure marginal coverage because it overlooks dependencies among the individual intervals. Establishing the validity of the interval would require additional assumptions, which conflicts with the distribution-free foundation of CP. Selecting the outermost bounds among the K intervals satisfies the marginal coverage condition in (1), but yields unnecessarily conservative prediction intervals. Therefore, an alternative aggregation strategy is needed. Here, we propose averaging the nonconformity scores, rather than averaging the calibrated intervals themselves.

3.1 Methodology

Our method aggregates nonconformity scores across K quantile pairs. Let $\bar{E} = \{\frac{1}{K}\sum_{k=1}^{K}E_j^{(k)}; j \in \mathcal{I}_2\}$ be the set of averaged scores on the calibration set. For a new observation X_{n+1} and a candidate response value y, define $\bar{E}_{n+1}(y) = \frac{1}{K}\sum_{k=1}^{K}E_{n+1}^{(k)}(y)$, where $E_{n+1}^{(k)}(y) = \max\{\hat{q}_{\ell_k}(X_{n+1}) - y, y - \hat{q}_{u_k}(X_{n+1})\}$. The prediction set is constructed by including all candidate responses whose averaged

nonconformity scores fall below a threshold determined by the calibration process:

$$\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1}) = \{y : \bar{E}_{n+1}(y) \leq Q_{\bar{E}}(\alpha)\}, \tag{5}$$

where $Q_{\bar{E}}(\alpha)$ denotes the $(1-\alpha)(1+1/|\mathcal{I}_2|)$-th empirical quantile of $\bar{E}$. By comparison, the standard CQR interval, defined in (3), can be similarly expressed as $\{y : E^{(k)}_{n+1}(y) \leq Q_{E^{(k)}}(\alpha)\}$. Since the set is defined via the average of exchangeable scores, it inherits the marginal coverage guarantee (see Section 3.2 for details).

The set $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})$ takes the form of a single interval. Since each $E^{(k)}_{n+1}(y)$ is convex in y, their average $\bar{E}_{n+1}(y)$ remains convex. Thus, the CCQR set has two unique endpoints, which are the solutions to $\bar{E}_{n+1}(y) = Q_{\bar{E}}(\alpha)$. Although these solutions cannot be expressed in closed-form, for any new point satisfying the mild condition described in the next subsection, the set can be written as:

$$\hat{C}_{\mathrm{CCQR}}(X_{n+1}) = \big[\bar{\ell}(X_{n+1}) - Q_{\bar{E}}(\alpha), \bar{u}(X_{n+1}) + Q_{\bar{E}}(\alpha)\big]. \tag{6}$$

Although it does not exactly correspond to (5), it serves as an effective approximation. Empirical evidence shows that the required condition holds in most cases, allowing $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})$ to be well approximated by this expression. The exact set can be computed by applying numerical root-finding only to those instances where our condition is not met.

Algorithm 1 presents a construction of approximated CCQR set using the split CP approach. The exact CCQR procedure, outlined in Algorithm 2, includes a verification step to determine whether the condition holds. For points where the condition fails, the interval endpoints are computed using a numerical procedure.

Algorithm 1 Approximated CCQR Interval

Input: Data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$; new feature vector X_{n+1}; quantile pairs $\{(\ell_1, u_1), (\ell_2, u_2), \ldots, (\ell_K, u_K)\}$; miscoverage rate α; quantile regression algorithm $\mathcal{M}$

Output: Prediction interval $\hat{C}_{\text{CCQR}}(X_{n+1})$ for the new data point

Partition $\mathcal{D}$ into training set $\mathcal{I}_1$ and calibration set $\mathcal{I}_2$.
Train a quantile regression model $\{\hat{q}_{\ell_k}, \hat{q}_{u_k}\}_{k=1}^K \leftarrow \mathcal{M}\{(X_i, Y_i) \in \mathcal{I}_1\}$.
for $i \in \mathcal{I}_2$ **do**
    for $k = 1$ to K **do**
        Compute nonconformity score: $E_i^{(k)} \leftarrow \max\{\hat{q}_{\ell_k}(X_i) - Y_i, Y_i - \hat{q}_{u_k}(X_i)\}$.
    end for
    Compute averaged nonconformity score: $\bar{E}_i \leftarrow \frac{1}{K}\sum_{k=1}^K E_i^{(k)}$.
end for
Compute empirical quantile: $Q_{\bar{E}}(\alpha) \leftarrow$ the $(1-\alpha)(1+\frac{1}{|\mathcal{I}_2|})$-th empirical quantile of $\{\bar{E}_i : i \in \mathcal{I}_2\}$
Compute lower bound: $c_{\text{lo}} \leftarrow \bar{\ell}(X_{n+1}) - Q_{\bar{E}}(\alpha)$
Compute upper bound: $c_{\text{hi}} \leftarrow \bar{u}(X_{n+1}) + Q_{\bar{E}}(\alpha)$
return $\hat{C}_{\text{CCQR}}(X_{n+1}) = [c_{\text{lo}}, c_{\text{hi}}]$

Algorithm 2 Exact CCQR Interval

Input: Same as in Algorithm 1

Output: Prediction interval $\hat{\mathcal{S}}_{\text{CCQR}}(X_{n+1})$

Partition $\mathcal{D}$ into training set $\mathcal{I}_1$ and calibration set $\mathcal{I}_2$.
Train a quantile regression model $\{\hat{q}_{\ell_k}, \hat{q}_{u_k}\}_{k=1}^K \leftarrow \mathcal{M}\{(X_i, Y_i) \in \mathcal{I}_1\}$.
Compute $m_k(X_{n+1}) \leftarrow \frac{1}{2}\left(\hat{q}_{\ell_k}(X_{n+1}) + \hat{q}_{u_k}(X_{n+1})\right)$ for $k = 1, \ldots, K$
Compute $\bar{\ell}(X_{n+1}) \leftarrow \frac{1}{K}\sum_{k=1}^K \hat{q}_{\ell_k}(X_{n+1})$, and $\bar{u}(X_{n+1}) \leftarrow \frac{1}{K}\sum_{k=1}^K \hat{q}_{u_k}(X_{n+1})$
Compute threshold $Q_{\bar{E}}(\alpha)$ as in Algorithm 1
if $Q_{\bar{E}}(\alpha) \geq \max\left(\bar{\ell}(X_{n+1}) - \min_k m_k(X_{n+1}),\ \max_k m_k(X_{n+1}) - \bar{u}(X_{n+1})\right)$ **then**
    $\hat{\mathcal{S}}_{\text{CCQR}}(X_{n+1}) \leftarrow [\bar{\ell}(X_{n+1}) - Q_{\bar{E}}(\alpha),\ \bar{u}(X_{n+1}) + Q_{\bar{E}}(\alpha)]$
else
    Define function $E_{n+1}(y) \leftarrow \frac{1}{K}\sum_{k=1}^K \max\{\hat{q}_{\ell_k}(X_{n+1}) - y,\ y - \hat{q}_{u_k}(X_{n+1})\}$
    Numerically solve for endpoints:

- Find c_{lo} such that $E_{n+1}(c_{\text{lo}}) = Q_{\bar{E}}(\alpha)$ (search left)
- Find c_{hi} such that $E_{n+1}(c_{\text{hi}}) = Q_{\bar{E}}(\alpha)$ (search right)

    Set $\hat{\mathcal{S}}_{\text{CCQR}}(X_{n+1}) \leftarrow [c_{\text{lo}},\ c_{\text{hi}}]$
end if
return $\hat{\mathcal{S}}_{\text{CCQR}}(X_{n+1})$

3.2 Theoretical coverage guarantee

We now establish the marginal coverage guarantee of CCQR. First, we provide a finite-sample guarantee for $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})$. Then, we further specify a necessary and sufficient condition under which the closed-form approximation $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$ attains the same guarantee.

Marginal coverage of the CCQR set is guaranteed by a standard argument involving exchangeable nonconformity scores. We invoke the following lemma from Romano et al. (2019), which provides a probability bound for an exchangeable random variable.

Lemma 2 (Romano et al. (2019)) *Let $Z_1, \ldots, Z_{n+1}$ be exchangeable random variables and let $Q_Z(\alpha)$ be an empirical $(1-\alpha)$-th quantile of $\{Z_i\}_{i=1}^n$. Then*

$$\mathbb{P}\{Z_{n+1} \leq Q_Z(\alpha)\} \geq 1-\alpha,$$

and if the Z_i are almost surely distinct,

$$\mathbb{P}\{Z_{n+1} \leq Q_Z(\alpha)\} \leq 1-\alpha+\frac{1}{n+1}.$$

Consequently, we obtain the marginal coverage guarantee of CCQR, as stated in the following proposition.

Proposition 3 *Under exchangeability of $\{(X_i, Y_i)\}_{i=1}^{n+1}$, $\hat{\mathcal{S}}_{\mathrm{CCQR}}$ satisfies*

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

If all averaged nonconformity scores $\{\bar{E}_j\}_{j=1}^{n+1}$ are almost surely distinct, then

$$\mathbb{P}\big(Y_{n+1} \in \hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})\big) \leq 1-\alpha+\frac{1}{|\mathcal{I}_2|+1}.$$

Proof of Proposition 3 The result follows by applying standard arguments for exchangeable nonconformity scores. Since the nonconformity scores $\{E_j^{(k)}\}_{j=1}^{n+1}$ are exchangeable for each

k, their averages are also exchangeable. Applying Lemma 2 to $\{\bar{E}_j\}_{j=1}^{n+1}$ yields the desired coverage bounds. □

Recall that $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})$ is not equal to $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$, but is instead a subset: $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1}) \subseteq \hat{C}_{\mathrm{CCQR}}(X_{n+1})$. By construction, $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$ is equivalent to the set $\{y : M(y) \leq Q_{\bar{E}}(\alpha)\}$, where $M(y) = \max\{\bar{\ell}(X_{n+1}) - y, y - \bar{u}(X_{n+1})\}$. Since both $\hat{q}_{\ell_k}(X_{n+1}) - y$ and $y - \hat{q}_{u_k}(X_{n+1})$ are affine in y, their pointwise maximum, $M(y)$, is convex. Jensen's inequality implies that $M(y) \leq \bar{E}_{n+1}(y)$ for all $y \in \mathbb{R}$, which confirms that $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})$ is a subset of $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$. Accordingly, $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$ shares the same lower coverage bound as $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})$: $\mathbb{P}(Y_{n+1} \in \hat{C}_{\mathrm{CCQR}}(X_{n+1})) \geq 1 - \alpha$. An upper bound on the coverage can be established for the majority of cases, since $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$ and $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})$ are identical for new observations under a mild condition. We formalize this condition in Theorem 4.

Theorem 4 *Suppose the data $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable. For a new feature vector X_{n+1}, define the midpoint of the lower and upper quantile estimates at the k-th quantile pair as $m_k(X_{n+1}) = (\hat{q}_{\ell_k}(X_{n+1}) + \hat{q}_{u_k}(X_{n+1}))/2$ for $k = 1, \ldots, K$. Then, $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$ is identical to $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})$ if and only if the following condition holds:*

$$Q_{\bar{E}}(\alpha) \geq \max\left(\bar{\ell}(X_{n+1}) - \min_k\{m_k(X_{n+1})\},\ \max_k\{m_k(X_{n+1})\} - \bar{u}(X_{n+1})\right). \tag{7}$$

Proof of Theorem 4 By definition, $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1})$ is constructed as $\{y : \bar{E}_{n+1}(y) \leq Q_{\bar{E}}(\alpha)\}$. Since $\bar{E}_{n+1}(y)$ is a convex function of y, this set is a closed interval whose endpoints are the two solutions of the equation $\bar{E}_{n+1}(y) = Q_{\bar{E}}(\alpha)$. Meanwhile, the interval $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$ is defined by the endpoints $y_L = \bar{\ell}(X_{n+1}) - Q_{\bar{E}}(\alpha)$ and $y_U = \bar{u}(X_{n+1}) + Q_{\bar{E}}(\alpha)$. The equivalence $\hat{\mathcal{S}}_{\mathrm{CCQR}}(X_{n+1}) = \hat{C}_{\mathrm{CCQR}}(X_{n+1})$ holds if and only if y_L and y_U are the two unique solutions of $\bar{E}_{n+1}(y) = Q_{\bar{E}}(\alpha)$. Thus, the identity requires the simultaneous satisfaction of two conditions:

$$\bar{E}_{n+1}(y_L) = Q_{\bar{E}}(\alpha) \quad \text{and} \quad \bar{E}_{n+1}(y_U) = Q_{\bar{E}}(\alpha).$$

We proceed by deriving the necessary and sufficient conditions for each equality to hold. First, we examine the condition for the lower endpoint, y_L. $\bar{E}_{n+1}(y_L)$ is defined as the average of K nonconformity scores:

$$\bar{E}_{n+1}(y_L) = \frac{1}{K}\sum_{k=1}^{K} \max\{\hat{q}_{\ell_k}(X_{n+1}) - y_L,\ y_L - \hat{q}_{u_k}(X_{n+1})\}.$$

The equation $\bar{E}_{n+1}(y_L) = Q_{\bar{E}}(\alpha)$ holds if and only if the maximum in each term always selects its first argument. It is formally stated as

$$\max\{\hat{q}_{\ell_k}(X_{n+1}) - y_L, y_L - \hat{q}_{u_k}(X_{n+1})\} = \hat{q}_{\ell_k}(X_{n+1}) - y_L, \quad \text{for all } k \in \{1,\dots,K\}. \tag{8}$$

It holds if and only if $y_L \le m_k(X_{n+1})$ for all k, which is equivalent to $y_L \le \min_k m_k(X_{n+1})$. Substituting y_L with its defined value and rearranging the terms yields the first necessary condition:

$$Q_{\bar{E}}(\alpha) \ge \bar{\ell}(X_{n+1}) - \min_k m_k(X_{n+1}). \tag{9}$$

An analogous argument applies to the upper endpoint, y_U. The equality $\bar{E}_{n+1}(y_U) = Q_{\bar{E}}(\alpha)$ holds if and only if the maximum in (8) consistently selects its second argument for all k, which requires $y_U \ge \max_k m_k(X_{n+1})$. The requirement leads to the second necessary condition:

$$Q_{\bar{E}}(\alpha) \ge \max_k m_k(X_{n+1}) - \bar{u}(X_{n+1}). \tag{10}$$

The result follows directly by combining the necessary and sufficient conditions in inequalities (9) and (10). □

For test points satisfying condition (7), Corollary 5 shows that our approximated interval $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$ achieves the same upper coverage bound as the exact set of CCQR.

Corollary 5 *Suppose the test input X_{n+1} satisfies condition (7) in Theorem 4, then $\hat{C}_{\mathrm{CCQR}}(X_{n+1})$ has the upper coverage bound*

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}_{\mathrm{CCQR}}(X_{n+1})\big) \le 1 - \alpha + \frac{1}{|\mathcal{I}_2| + 1}.$$

Condition (7) is satisfied in most cases, largely due to our construction of the quantile pairs. Recall that the outermost pair $(\ell_1, u_1) = (\alpha/2, 1-\alpha/2)$ is always included, and the remaining pairs $(\ell_k, u_k)_{k=2}^K$ lie strictly within this range. Therefore, the interval $(\bar{\ell}(X_{n+1}), \bar{u}(X_{n+1}))$ often leads to undercoverage relative to the target level $1-\alpha$, necessitating a strictly positive calibration term $Q_{\bar{E}}(\alpha)$. In condition (7), $Q_{\bar{E}}(\alpha)$ should be greater than two quantities: $\bar{\ell}(X_{n+1}) - \min_k m_k(X_{n+1})$ and $\max_k m_k(X_{n+1}) - \bar{u}(X_{n+1})$. These quantities measure the deviation of the averaged estimates from the corresponding midpoints. Since each midpoint m_k lies between its corresponding lower and upper quantile estimates, $\hat{q}_{\ell_k}$ and $\hat{q}_{u_k}$, the resulting deviations are typically negative or only slightly positive. Therefore, condition (7) is usually satisfied.

To support this statement, we assess its satisfaction on the `bike` dataset, one of the benchmarks examined in Section 4. We construct CCQR prediction intervals using $K = 9$ quantile pairs, with lower quantiles $(\ell_1, \ldots, \ell_K) = (0.05, 0.075, \ldots, 0.20)$, and impose symmetry by setting $\ell_k + u_k = 1$ for all k. We consider two base models: QRF and MCQRNN. Among the 250 test observations, condition (7) is satisfied in 98.7% of cases under the QRF model and in 100% of cases under the MCQRNN model. We observe the similar pattern consistently across all real and simulated datasets evaluated in Section 4.

3.3 Variance reduction and quantile level selection

We now highlight a central strength of our method, its variance reduction capability, and provide a detailed explanation of the mechanism behind it. The degree of variance reduction depends heavily on how far the innermost pair (ℓ_K, u_K) deviate from the outermost pair $(\alpha/2, 1-\alpha/2)$. We explore an empirical rule for choosing quantile pairs and propose a data-driven tuning procedure.

We begin by revisiting the challenge of tail quantile estimation which limits the effectiveness of CQR. It constructs prediction intervals by estimating the conditional quantiles at the $\alpha/2$ and $1-\alpha/2$ levels, which lie in the tails of the distribution. Estimating these quantiles is known to suffer from high sampling variability, especially in small samples or when outliers exist (Koenker 2005). The variability propagates into the conformal calibration step, making the nonconformity scores unstable and leading to miscalibrated prediction intervals.

Our method builds on the statistical principle that averaging reduces variance. CCQR addresses the limitation of CQR by averaging nonconformity scores across K quantile pairs, yielding a more stable and reliable calibration. The averaged scores generally have lower variance than the score derived from the outermost quantiles, $\{E_j^{(1)}; j \in \mathcal{I}_2\}$. Thus, the averaging makes a calibration step less sensitive to outliers and helps avoid unnecessarily wide prediction intervals.

The variance reduction introduces a modest bias. We briefly explain this bias–variance tradeoff relationship in the context of the approximated CCQR interval, $\hat{C}_{\text{CCQR}}$. For all $k > 1$, the interval $(\hat{q}_{\ell_k}(X), \hat{q}_{u_k}(X))$ is typically narrower than that of the outermost pair unless quantile crossing occurs. Note that many simultaneous quantile regression models incorporate mechanisms to prevent such violations (Bondell et al. 2010; Liu and Wu 2011; Cannon 2018). Averaging the inner quantile estimates leads to a systematic shift in the resulting interval. To better understand the source of this discrepancy, we decompose the difference between $\bar{\ell}(X)$ and the true lower quantile $q_{\alpha/2}(X)$ as follows:

$$\left(\bar{\ell}(X) - \mathbb{E}[\bar{\ell}(X)]\right) + \left(\mathbb{E}[\bar{\ell}(X)] - \frac{1}{K}\sum_k q_{\ell_k}(X)\right) + \left(\frac{1}{K}\sum_k q_{\ell_k}(X) - q_{\alpha/2}(X)\right).$$

The first term corresponds to estimation variance when squared, the second reflects bias, and the third term, referred to as the aggregation gap, captures the shift introduced by averaging over inner quantile levels. A similar decomposition applies to the difference between $\bar{u}(X)$ and $q_{1-\alpha/2}(X)$.

The approximated CCQR interval, $\hat{C}_{\text{CCQR}}$, is formed by expanding $[\bar{\ell}(X), \bar{u}(X)]$ symmetrically by the calibration constant $Q_{\bar{E}}(\alpha)$. In CQR, the calibration constant corrects a finite-sample variation of the interval. The calibration constant in our method addresses two issues: it accounts for estimation variability and corrects for the undercoverage caused by the aggregation gap. Since calibration applies a uniform adjustment across all new observations, it effectively corrects the aggregation gap on average. However, it does not account for localized variations in the quantile functions, thereby introducing an additional source of bias in CCQR. Our empirical results indicate that stabilizing the score to reduce variance often leads to narrower and more reliable intervals. This suggests that the benefits of variance reduction generally outweigh the impact of the induced bias.

Local discrepancies that contribute to the aggregation gap are closely linked to the degree of heteroscedasticity in the data. When the inner and tail quantile curves are parallel across all values of X, in a homogeneous setting, the calibration constant $Q_{\bar{E}}(\alpha)$ offsets most of the aggregation gap. However, in the presence of pronounced heteroscedasticity, the constant adjustment by $Q_{\bar{E}}(\alpha)$ is insufficient to fully address the gap, potentially limiting the effectiveness of CCQR. Nevertheless, the aggregation gap can be mitigated by selecting quantile pairs (ℓ_k, u_k) that are closer to the tails of distribution, highlighting the critical role of the quantile level selection.

Quantile selection: A key consideration in implementing CCQR is the selection of quantile pairs $\{(\ell_k, u_k)\}_{k=1}^{K}$. Instead of selecting each quantile individually, which can be computationally demanding, we adopt a parameterized approach that adjusts the overall range of combined quantiles.

We use a single hyperparameter, $d > 0$, which controls the range of the quantile levels averaged. With $\ell_1 = \alpha/2$ and $u_1 = 1 - \alpha/2$ fixed, the remaining lower quantiles are uniformly distributed over the interval $[\ell_1, \ell_1 + d]$, and the corresponding upper quantiles are determined symmetrically. The full set of quantile levels is given by

$$\ell_k = \ell_1 + (k-1)\frac{d}{K}, \quad u_k = u_1 - (k-1)\frac{d}{K}, \quad k = 1, \ldots, K.$$

Critical considerations for selecting d include the properties of the underlying base model and the extent to which quantile functions vary across levels as a result of heteroscedasticity. Based on the empirical findings in Section 4.2, a safe choice for d lies between 0.1 and 0.2, as this range consistently yields robust performance across diverse settings.

Although a heuristic choice of d may perform adequately in many scenarios, it is not universally optimal. Performance can often be enhanced by tuning this parameter. A common approach involves selecting the value of d that minimizes the average width of the prediction intervals on the calibration set. Specifically, we can select d as follows:

$$\hat{d} = \arg\min_{d} \sum_{i \in \mathcal{I}_2} \left[\left(\bar{u}_d(X_i) + Q_{\bar{E}_d}(\alpha) \right) - \left(\bar{\ell}_d(X_i) - Q_{\bar{E}_d}(\alpha) \right) \right], \tag{11}$$

where both the quantile estimates and the calibration term vary with d. Since Proposition 3 guarantees marginal coverage for any fixed value of d, the tuning procedure does not hinder its validity. We empirically determine the value of $\hat{d}$ via a grid search, as detailed in Section 4.2.

3.4 Variants of CCQR

CCQR is adaptable to different scoring schemes, as long as the nonconformity scores satisfy the exchangeability condition. Below, we present several representative variants.

Asymmetric CCQR: Inspired by Romano et al. (2019), this variant allows for separated calibration of the lower and upper bounds. It is particularly advantageous when Y given X has a skewed distribution, allowing asymmetric allocation of the coverage probability. Define the prediction set as:

$$\left\{y : \bar{E}^{(\ell)}_{n+1}(y) \leq Q_{\bar{E}^{(\ell)}}(\alpha/2) \text{ and } \bar{E}^{(u)}_{n+1}(y) \leq Q_{\bar{E}^{(u)}}(\alpha/2)\right\},$$

where $\bar{E}^{(\ell)}_{n+1}(y) = \frac{1}{K}\sum_{k=1}^{K}(\hat{q}_{\ell_k}(X_{n+1}) - y)_+$, $\bar{E}^{(u)}_{n+1}(y) = \frac{1}{K}\sum_{k=1}^{K}(y - \hat{q}_{u_k}(X_{n+1}))_+$. Here, $(a)_+ := \max(a, 0)$ denotes the hinge function. $Q_{\bar{E}^{(\ell)}}(\alpha/2)$ denotes the $(1 - \alpha/2)(1 + 1/|\mathcal{I}_2|)$-th empirical quantile of $\{\bar{E}^{(\ell)}_j = \frac{1}{K}\sum_{k=1}^{K}(\hat{q}_{\ell_k}(X_j) - Y_j)_+;\ j \in \mathcal{I}_2\}$, and $Q_{\bar{E}^{(u)}}(\alpha/2)$ represents the corresponding quantile of $\{\bar{E}^{(u)}_j = \frac{1}{K}\sum_{k=1}^{K}(Y_j - \hat{q}_{u_k}(X_j))_+;\ j \in \mathcal{I}_2\}$. Since the scores for the lower and upper bounds are exchangeable, the variant retains the marginal coverage guarantee. Although the asymmetric variant provides more flexible intervals than standard CCQR in (5), it tends to produce wider and more conservative prediction intervals. Empirical results are presented in Appendix A.

Locally adaptive CCQR: Recent work in CP has increasingly focused on establishing conditional coverage guarantee, defined as $\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})|X_{n+1} = x) \geq 1-\alpha$. Although achieving exact conditional coverage is infeasible without distributional assumptions (Foygel Barber et al. 2020), several approximation strategies have been proposed in the literature, including conditional group-based approaches (Gibbs et al. 2025; Ding et al. 2023).

While achieving conditional coverage may demand significant methodological changes, our method can still be tailored to account for local heteroscedasticity. We can assign weights to nonconformity scores using the width of the prediction interval at each data point. Our design is motivated by locally weighted conformal prediction (Papadopoulos et al. 2011; Kivaranovic et al. 2020).

For each calibration point (X_j, Y_j) and each quantile pair (ℓ_k, u_k), define the weighted nonconformity score as:

$$\tilde{E}_j^{(k)} = \frac{1}{w_{j,k}} \max\left\{ Y_j - \hat{q}_{u_k}(X_j),\ \hat{q}_{\ell_k}(X_j) - Y_j \right\}, \quad j \in \mathcal{I}_2,\ k = 1, \ldots, K,$$

where $w_{j,k} = \hat{q}_{u_k}(X_j) - \hat{q}_{\ell_k}(X_j)$ denotes the local width of the k-th prediction interval at X_j. For each calibration point $j \in \mathcal{I}_2$, the locally adaptive score is obtained by averaging the weighted scores over all K quantile pairs: $\bar{E}_j^{(\text{adapt})} = \frac{1}{K} \sum_{k=1}^{K} \tilde{E}_j^{(k)}$. Given a new input X_{n+1} and a candidate response y, define the prediction set as $\left\{ y : \bar{E}_{n+1}^{(\text{adapt})}(y) \leq Q_{\text{adapt}}(\alpha) \right\}$, where $Q_{\text{adapt}}(\alpha)$ is the $(1-\alpha)(1+1/|\mathcal{I}_2|)$-th empirical quantile of the calibration scores $\{\bar{E}_j^{(\text{adapt})};\ j \in \mathcal{I}_2\}$.

To address heteroscedasticity, we scale nonconformity scores by the associated local interval widths $w_{j,k}$. Observations with wide predictive intervals, which indicate greater uncertainty, receive less weight than those with narrower intervals. By adapting to local uncertainty, the calibration step enhances conditional coverage.

While our default choice for $w_{j,k}$ is $\hat{q}_{u_k}(X_j) - \hat{q}_{\ell_k}(X_j)$, this can be replaced with other measures of local variability as needed. For instance, the tail interval, $\hat{q}_{1-\alpha/2}(X_j) - \hat{q}_{\alpha/2}(X_j)$, captures a broader range of variability but tends to be more volatile in small-sample settings. A more stable alternative is the aggregated width, $\sum_{k=1}^{K} [\hat{q}_{u_k}(X_j) - \hat{q}_{\ell_k}(X_j)]$, which smooths over individual quantile fluctuations at the cost of reduced sensitivity to local variability.

CCQR for image-to-image regression: We extend the CCQR framework to image-to-image regression tasks, including image denoising and super-resolution. Both the inputs X and outputs Y are images in the space $\mathcal{X} = [0,1]^{U \times V}$. For notational simplicity, we assume that X and Y have the same dimensions. Our goal is to construct pixelwise prediction intervals $\{\hat{C}(X)_{u,v} \subseteq [0,1] : u = 1,\ldots,U,\ v = 1,\ldots,V\}$, such that the expected fraction of miscovered pixels is controlled below a predefined risk level α. We formalize our objective using the risk-controlling prediction set (RCPS) framework introduced by Bates et al. (2021).

Definition 1 (Risk-Controlling Prediction Set) Let $\hat{C} : \mathcal{X} \to (2^{[0,1]})^{U \times V}$ be a possibly randomized mapping from each input image $X \in \mathcal{X}$ to a grid of $U \times V$ prediction intervals. Define the per-image loss function as

$$L(\hat{C}(X), Y) = 1 - \frac{1}{UV} \sum_{u=1}^{U} \sum_{v=1}^{V} \mathbb{I}(Y_{u,v} \in \hat{C}(X)_{u,v}),$$

which measures the fraction of pixels for which the true value $Y_{u,v}$ falls outside the prediction interval. Then, $\hat{C}$ is an (α, δ)-risk-controlling prediction set if

$$\mathbb{P}\left\{\mathbb{E}[L(\hat{C}(X), Y)] > \alpha\right\} \leq \delta.$$

Constructing the RCPS is conceptually similar to building prediction intervals for tabular data under CP frameworks. Angelopoulos et al. (2022) proposed using CQR in this context. First, the model is trained to produce per-pixel conditional quantiles at levels $\alpha/2$ and $1 - \alpha/2$, along with a central point estimate such as the conditional mean or median. These outputs are denoted by $\hat{q}_{\alpha/2}(X)_{u,v}$, $\hat{q}_{1-\alpha/2}(X)_{u,v}$, and $\hat{f}(X)_{u,v}$, respectively. Following Angelopoulos et al. (2022), the prediction interval at pixel (u, v) for a test image X_{n+1} is given by

$$\left[\hat{f}(X_{n+1})_{u,v} - \lambda \hat{q}_{\alpha/2}(X_{n+1})_{u,v}, \hat{f}(X_{n+1})_{u,v} + \lambda \hat{q}_{1-\alpha/2}(X_{n+1})_{u,v}\right],$$

where $\lambda \geq 0$ is a calibration parameter controlling the overall risk. λ is selected to ensure that the empirical miscoverage rate on the calibration set remains below the target risk level α.

CCQR can be extended to image-to-image regression by incorporating multiple quantile levels instead of only the tail quantiles $\alpha/2$ and $1-\alpha/2$. Now, the base model is fitted on training set to produce $2K+1$ outputs per pixel: a point estimate $\hat{f}(X)_{u,v}$ and K pairs of quantile estimates, $\{\hat{q}_{\ell_k}(X)_{u,v}, \hat{q}_{u_k}(X)_{u,v} : k = 1, \ldots, K\}$. We select the quantile pairs $\{(\ell_k, u_k)\}_{k=1}^{K}$ following the procedure outlined in Section 3.3. For a test image X_{n+1}, the CCQR prediction interval at pixel (u, v) is constructed as follows:

$$\hat{C}^{\lambda}_{\mathrm{CCQR}}(X_{n+1})_{u,v} = \left[\hat{f}(X_{n+1})_{u,v} - \lambda\bar{\ell}(X_{n+1})_{u,v}, \hat{f}(X_{n+1})_{u,v} + \lambda\bar{u}(X_{n+1})_{u,v}\right], \quad (12)$$

where $\bar{\ell}(X)_{u,v} = \frac{1}{K}\sum_k \hat{q}_{\ell_k}(X)_{u,v}$ and $\bar{u}(X)_{u,v} = \frac{1}{K}\sum_k \hat{q}_{u_k}(X)_{u,v}$ represent the averaged quantile estimates.

We now show that $\hat{C}^{\lambda}_{\mathrm{CCQR}}(X_{n+1}) = \{\hat{C}^{\lambda}_{\mathrm{CCQR}}(X_{n+1})_{u,v} : u = 1, \ldots, U; v = 1, \ldots, V\}$ is an (α, δ)-RCPS with an appropriately chosen λ. We empirically select λ using the calibration set. For any candidate value $\lambda \geq 0$, denote the empirical miscoverage rate by

$$\hat{R}(\lambda) = \frac{1}{|\mathcal{I}_2|}\sum_{i\in\mathcal{I}_2} L(\hat{C}^{\lambda}_{\mathrm{CCQR}}(X_i), Y_i). \quad (13)$$

This is an estimate of the population risk $R(\lambda) = \mathbb{E}_{X,Y}[L(\hat{C}^{\lambda}_{\mathrm{CCQR}}(X), Y)]$. Since $L(\cdot,\cdot)$ is bounded in $[0, 1]$, we can bound $R(\lambda)$ by applying Hoeffding's inequality (Hoeffding 1963). Following Bates et al. (2021) and Angelopoulos et al. (2022), we derive that

$$\mathbb{P}\left\{R(\lambda) > \hat{R}(\lambda) + \sqrt{\frac{1}{2|\mathcal{I}_2|}\log\frac{1}{\delta}}\right\} \leq \delta. \quad (14)$$

Denote $\hat{R}^{+}(\lambda) = \hat{R}(\lambda) + \sqrt{\frac{1}{2|\mathcal{I}_2|}\log(\frac{1}{\delta})}$. We search for the smallest λ for which the empirical miscoverage remains below the target risk level α: $\hat{\lambda} = \inf\{\lambda \geq 0 : \hat{R}^{+}(\lambda) \leq$

$\alpha\}$. Because $\hat{R}(\lambda)$, $R(\lambda)$, and $\hat{R}^+(\lambda)$ are all monotonically decreasing in λ, and since $\bar{\ell}(X)_{u,v}$ and $\bar{u}(X)_{u,v}$ are nonnegative, the resulting prediction sets are nested. That is, $\hat{C}^{\lambda_1}_{\text{CCQR}}(X) \subseteq \hat{C}^{\lambda_2}_{\text{CCQR}}(X)$ whenever $0 \leq \lambda_1 < \lambda_2$. This structure allows $\hat{\lambda}$ to be efficiently selected via grid search.

Similar to the calibration constant $Q_{\bar{E}}(\alpha)$ in the original CCQR, $\hat{\lambda}$ is determined from held-out calibration data. Together with (14), $\mathbb{P}[R(\hat{\lambda}) \leq \hat{R}^+(\hat{\lambda})] \geq 1 - \delta$ holds. Therefore, $\hat{C}^{\hat{\lambda}}_{\text{CCQR}}(X_{n+1})$ is an (α, δ)-RCPS. Additional implementation details are provided in Section 4.3, while the theoretical results are discussed by Angelopoulos et al. (2022).

4 Experiments

4.1 Benchmarking experiments

We compare our method with four established CP approaches:

- Conformal histogram regression (CHR): estimates conditional histograms and selects prediction intervals with the narrowest bin ranges that satisfy marginal coverage (Sesia and Romano 2021).
- Distributional conformal prediction (DCP): uses empirical cumulative distribution functions to construct prediction intervals (Chernozhukov et al. 2021).
- Lasso conformal predictor (LASSO): applies standard CP to conditional mean estimates obtained via an $\mathcal{L}_1$-regularized linear model.
- Conformalized quantile regression (CQR): serves as our primary baseline and forms the basis for CCQR (Romano et al. 2019).

Among the methods compared, only LASSO estimates the conditional mean; all others are based on conditional quantile estimation. We use QRF (Meinshausen 2006), MCQRNN (Cannon 2018), and quantile extremely randomized trees (QERT; Geurts et al. (2006)) as base models for quantile-based methods. All three base models are

capable of estimating multiple quantiles simultaneously. For CCQR, the number of quantile pairs is fixed at $K = 9$, and the quantile range parameter d is tuned on the calibration set by minimizing the average width of prediction interval. All experiments are conducted using `Python 3.11.9`.

We evaluate the performance of the method using two criteria: empirical coverage and average width of interval. The empirical coverage, defined as the proportion of test observations where the true outcome falls within the prediction interval, measures whether the method achieves the desired coverage level. We target a coverage level of 90%. The average width of interval indicates a precision of uncertainty quantification. In general, narrower intervals suggest a more precise and informative assessment of uncertainty.

We conduct experiments on several regression benchmarks, including the `Abalone`, `Bike`, `Bio`, `Blog_data`, `Concrete`, `Community`, and `Popularity` datasets from the UCI Machine Learning Repository (Nottingham et al. 2023), as well as the Student–Teacher Achievement Ratio (`STAR`) dataset from the U.S. State Department of Education (Achilles et al. 2008). To reduce scale-dependent effects, the response variable in each dataset is normalized by its mean absolute value. The dataset is randomly partitioned into training, calibration, and test sets in proportions of 35%, 35%, and 30%, respectively. Prediction intervals are computed on the test set over 100 iterations. Additional details, including hyperparameter tuning procedures and model-specific results, are provided in Appendix A.

Figure 2 shows empirical coverage and average prediction interval widths, averaged across the three base models. Detailed results for each model are available in the appendix. To facilitate comparison, interval widths are rescaled within each dataset and iteration, setting the widest interval to 1 and expressing all others as proportions of this maximum. CCQR consistently achieves empirical coverage near the nominal level across datasets, whereas competing methods occasionally fall short. It also produces

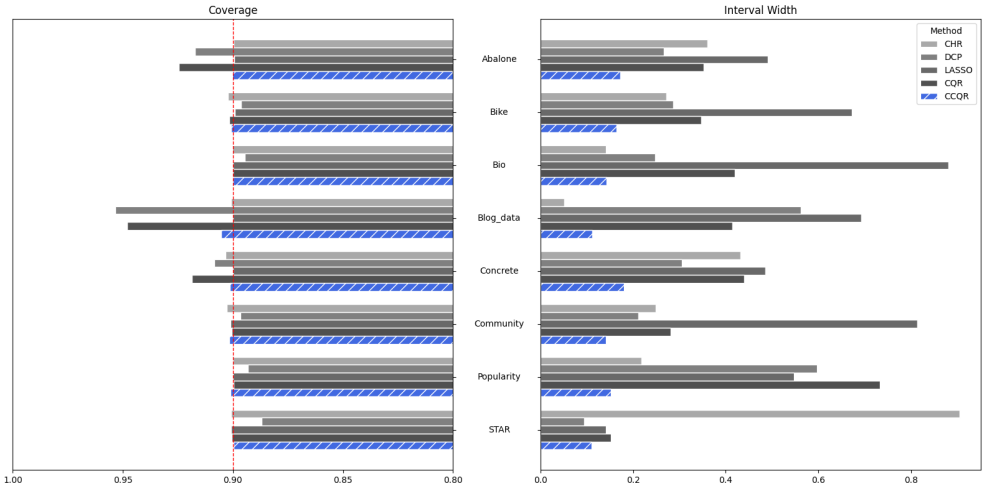

Fig. 2: Results from 100 iterations comparing conformal prediction methods. Empirical coverage (left) and interval width (right) are displayed. The red dashed line indicates the target coverage level.

narrower prediction intervals, particularly on small to medium-sized datasets such as `Abalone`, `Bike`, `Concrete`, `Community`, and `Popularity`.

LASSO performs poorly under distributional asymmetry, as it produces symmetric intervals centered on conditional mean estimates. Other methods struggle to estimate tail quantiles reliably; their high variance can lead to undercoverage, overcoverage, or unnecessarily wide intervals. In contrast, CCQR uses multiple quantile estimates to construct more stable and robust intervals, even with the same base model. While the variance reduction is less pronounced for large datasets such as `Bio` ($n = 30{,}000$) and `Blog` ($n = 50{,}000$), where tail quantiles can be reliably estimated, CCQR still achieves the second narrowest intervals, with CHR performing slightly better.

4.2 Sensitivity analysis

The performance of CCQR depends on two key parameters: d, which determines the range of aggregated quantile levels, and K, the number of quantile pairs. We investigate the impact of these parameters through a synthetic experiment.

We first investigate how the heteroscedasticity of the data influences the optimal choice of d. We generate data from the following heteroscedastic regression model:

$$y = \sin(2\pi x) + \sigma_0^2 \cdot (1 + hx)\epsilon, \tag{15}$$

where $\epsilon \sim N(0, 1)$ is i.i.d. standard normal noise, $\sigma_0^2 = 0.1$ is the base variance, and $h \in \{0, 10, 100, 1000\}$ controls the degree of heteroscedasticity. We allocate 500 samples for training and 500 for calibration, then assess prediction intervals on the test set of 5000 samples. For each value of h, we construct prediction intervals using values of d ranging from 0.05 to 0.40 in increments of 0.05. We use three base models which are employed in the previous subsection. Across all values of d and base models, CQR serves as the baseline, with its interval width normalized to 1. We then report the relative interval widths of our method for each d.

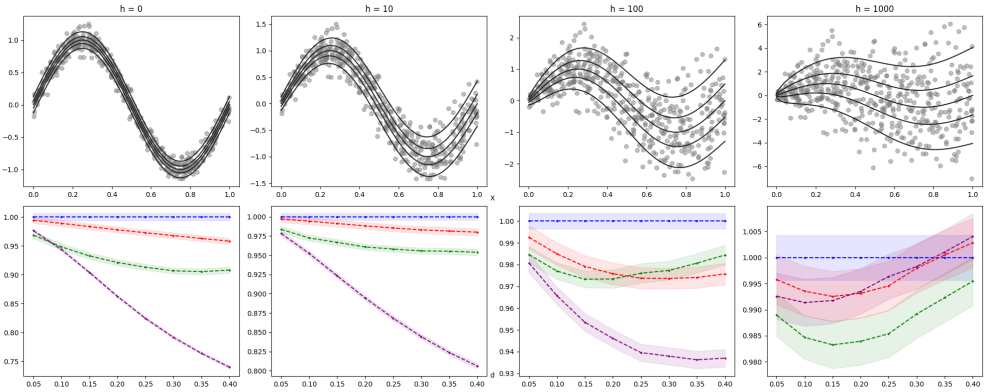

Fig. 3: Relative interval widths under the setting described in (15), with d ranging from 0.05 to 0.40. The top panels show examples of the generated data along with the true quantile functions ($\tau = 0.1, 0.3, \ldots, 0.9$). The bottom panels display interval widths of CCQR relative to CQR across three base models: QRF (green), MCQRNN (red), and QERT (purple). The blue dashed line represents the normalized CQR baseline, and shaded areas represent ± 1 standard error.

Figure 3 illustrates that the degree of heteroscedasticity in the data is closely related to the optimal choice of d. When $h = 0$, corresponding to a homogeneous setting, CCQR achieves the the narrowest intervals at the largest tested value of d, consistently across all base models. As h increases from 0 to 1000, heteroscedasticity becomes more pronounced, and the conditional quantile functions vary more substantially across quantile levels. Therefore, CCQR performs better with smaller values of d. In the most heteroscedastic setting, $h = 1000$, values of $d = 0.10$ or $d = 0.15$ are preferred, since larger d can introduce bias due to misalignment in the shape of the quantile functions.

Several well-established statistical methods exist for testing heteroscedasticity in data (Koenker and Bassett 1982; Dette and Munk 1998; Shinkyu 2025). These tests can offer prior insight into the degree of heteroscedasticity, which may inform the choice of d. When such information is unavailable, one can examine the estimated quantile functions, paying attention to their shape differences rather than just their nominal distance, to gain practical guidance for selecting an appropriate value of d.

To further examine the sensitivity of CCQR to d, we design four simulation scenarios that encompass both univariate and multivariate settings. The four scenarios are defined as follows:

(i) $y = \text{pois}(\sin^2(x) + 0.1) + 0.03x \cdot \epsilon_1 + 25\mathbb{I}(u < 0.01)\epsilon_2$,
where $x \sim U(0, 5)$ and $\epsilon_1, \epsilon_2 \sim N(0, 1)$ (adapted from Romano et al. (2019));

(ii) $y = \left(1 - x + 2x^2\right) e^{-0.5x^2} + (0.2 + 0.04x)\,\epsilon_1 + 15\mathbb{I}(u < 0.01)\epsilon_2$,
where $x \sim U(-4, 4)$, $\epsilon_1 \sim \chi^2(3)$, and $\epsilon_2 \sim N(0, 1)$ (adapted from Cannon (2018));

(iii) $y = 2\sin(\pi\beta'X) + \pi\beta'X + \epsilon_1\sqrt{1 + (\beta'X)^2} + 5\mathbb{I}(u < 0.01)\epsilon_2$,
where $\beta = [1, 1, 1, 1, 1, 0, \ldots, 0]'$, $X \sim U(0, 1)^{50}$, and $\epsilon_1, \epsilon_2 \sim N(0, 1)$ (adapted from Sesia and Candès (2020));

(iv) $y = \text{pois}\left(\sin^2(X_1) + \cos^4(X_2) + 0.01\right) + 0.03X_1\epsilon_1 + 25\mathbb{I}(u < 0.01)\epsilon_2$,
where $X \sim U(0,1)^{100}$ and $\epsilon_j \sim t(3) \times (1 + \sqrt{X_1^{2j} + X_2^{2j}})$ for $j = 1, 2$ (following Yang and Kuchibhotla (2025)).

Each scenario includes 1% of outliers, generated using a random variable $u \sim U(0,1)$. The sample sizes are set to 500 for training, 500 for calibration, and 5000 for test. We set the number of quantile pairs to $K = 9$ and vary d over the set $\{0.05, 0.10, \ldots, 0.40\}$. Again, we use three base predictive models: QRF, MCQRNN, QERT. We generate a 90% prediction interval using CCQR on the test set and evaluate its empirical coverage and relative width against CQR.

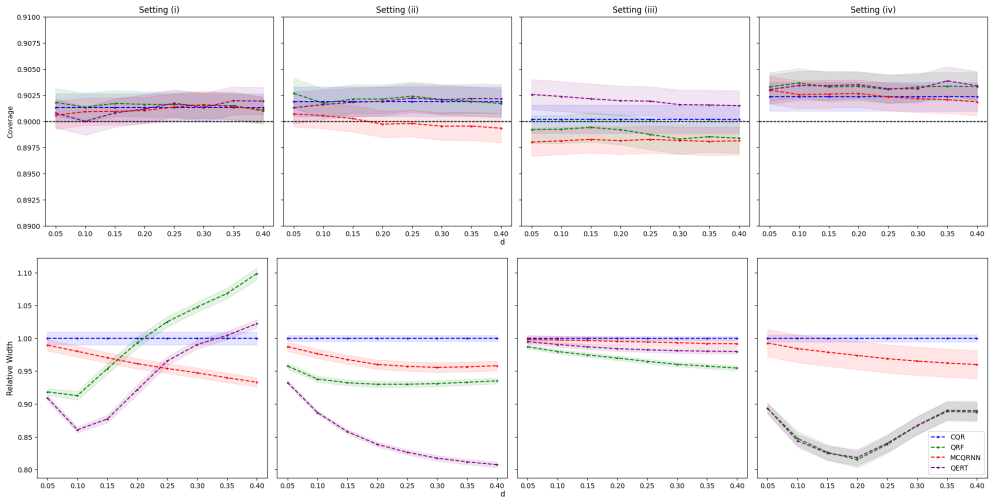

Fig. 4: Simulation results across four scenarios. The top panels report empirical coverage rates, and the bottom panel presents interval widths as a function of d. Results are shown for baseline CQR (blue) and CCQR implemented with QRF (green), MCQRNN (red), and QERT (purple). Shaded regions represent ± 1 standard error. In the bottom-right panel, the QRF and QERT curves nearly overlap due to their similar performance.

Figure 4 presents the results averaged over 100 iterations for each scenario. Although empirical coverage shows little variation with d, the corresponding interval

width is highly sensitive to this parameter. Optimal values of d seem to vary depending on the data-generating scenario and the base model employed. While MCQRNN achieves the best performance at $d = 0.40$ across all scenarios, two tree models, QRF and QERT, exhibit a similar pattern: they perform better with smaller values of d in settings (i) and (iv), and with larger values in settings (ii) and (iii). This may be attributed to the fact that MCQRNN produces similarly shaped quantile functions across levels, because they share most of network weights. In such cases, using a larger d can lead to optimal performance, unless the data exhibit severe heteroscedasticity. In contrast, the quantile estimates produced by tree-based models tend to differ more substantially across quantile levels and known to be biased (Tung et al. 2014; Nguyen et al. 2015). As a result, QRF and QERT are more sensitive to the choice of d.

Based on our results, we recommend using $d \in (0.10, 0.20)$ as a default when no prior information is available, since it consistently produces narrower intervals than CQR across all scenarios. However, for optimal performance, we suggest tuning d on the calibration set using the procedure described in Section 3.3.

Next, we investigate the impact of the number of quantile pairs, K, on CCQR performance, using the four previously introduced data-generating scenarios. The quantile range is fixed at $d = 0.15$, while K varies from 2 to 20.

Figure 5 shows that empirical coverage remains almost consistent across all tested values of K, confirming that the marginal coverage guarantee holds regardless of the number of quantiles. The average width of the prediction interval generally remains constant or decreases slightly as K increases, but the rate of improvement diminishes beyond a certain threshold. These findings indicate that the choice of K is less sensitive than that of d. Choosing K between 5 and 10 generally suffices to ensure reliable performance.

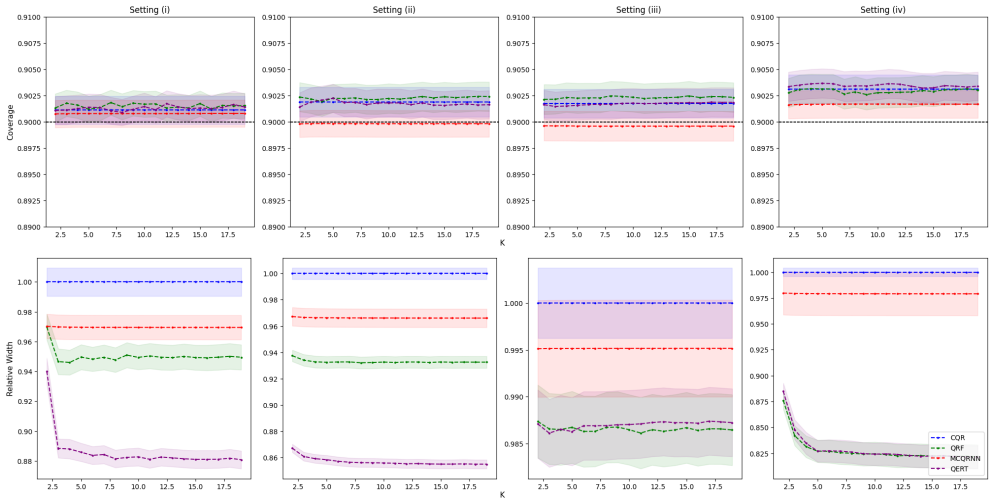

Fig. 5: Simulation results across four scenarios. The top row displays empirical coverage rates, and the bottom row shows interval widths as a function of K. Curves correspond to CQR (blue) and CCQR implemented with QRF (green), MCQRNN (red), and QERT (purple). Shaded regions indicate ± 1 standard error.

4.3 image-to-image regression with FastMRI dataset

In this subsection, we present empirical results from applying CCQR to image-to-image regression, as described in Section 3.4. We apply our method to accelerated MRI reconstruction using the FastMRI dataset (Zbontar et al. 2018). Our experimental design follows the setup proposed by Angelopoulos et al. (2022), with minor adaptations. We generate undersampled input images X by reducing k-space data by a factor of four along the phase-encoding dimension and applying an inverse Fourier transform. The fully sampled MRI images, Y, are used as the ground truth. The dataset includes 27,993 coronal slices (320×320 resolution) derived from 10,000 clinical knee MRI volumes (3T or 1.5T) for training, along with 3,474 slices each for calibration and testing.

We compare the 90% pixelwise prediction intervals produced by CQR and CCQR. Both methods employ the same predictive model based on the U-Net architecture (Ronneberger et al. 2015), trained with a batch size of 10. The model produces 19

outputs per pixel: a point estimate and pairs of quantile predictions, determined by parameters $K = 9$ and $d = 0.2$. CCQR constructs prediction intervals using all outputs, whereas CQR uses only the two tail quantile estimates at $\ell_1 = 0.05$ and $u_1 = 0.95$.

Figure 6 presents uncertainty visualizations for both conformal methods. We generate uncertainty heatmaps that highlight regions of high predictive uncertainty, defined as pixels whose interval widths are greater than the 70th percentile of all widths within the image. The highlighted region is shown in red, indicating areas where the image reconstruction is more uncertain. The second-to-last column in Figure 6 shows a heatmap that visualizes pixel-level differences in prediction interval width between the two methods. Regions in red denote where CCQR yields shorter prediction intervals, whereas light sky blue, primarily in uninformative background areas, indicates tighter intervals from CQR. CCQR yields narrower prediction intervals than CQR in high-uncertainty regions, especially around tissue interfaces and anatomically intricate areas where precise uncertainty quantification is essential. These findings suggest that CCQR provides more informative uncertainty visualizations than CQR.

5 Conclusion

Conformalized composite quantile regression constructs prediction intervals by averaging nonconformity scores across multiple quantile levels, improving stability without added complexity. Our method reduces variance and improves interval sharpness through ensemble averaging, while maintaining finite-sample coverage guarantees. CCQR is particularly effective for the data with small to moderate sample sizes, where estimating tail quantiles tends to introduce high variance and reduce reliability.

CCQR achieves computational efficiency by producing multiple quantile estimates from a single predictive model, in contrast to other ensemble methods that require training multiple independent models. While exact CCQR endpoints require numerical

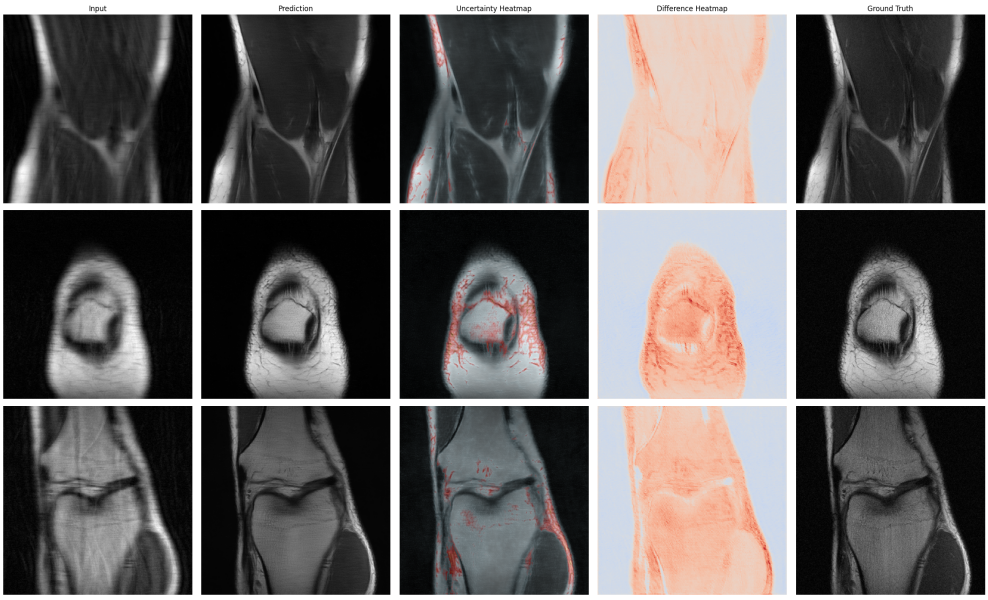

Fig. 6: Visualization of MRI reconstructions and uncertainty estimates. From left to right: (1) undersampled input, (2) predicted reconstruction, (3) uncertainty heatmap from CCQR, (4) difference in prediction interval width between CQR and CCQR, where red highlights regions where CCQR is narrower and light sky blue where CQR is tighter, and (5) ground truth.

computation, we establish the condition that allow closed-form expression in most cases, keeping the method computationally lightweight. Empirical results show that our method generally achieves the target coverage while producing tighter intervals than existing approaches.

Several directions remain open for future research on CCQR. First, it would be valuable to develop a principled, data-driven method for selecting the quantile range d, rather than relying on heuristics. Second, while we have explored some variants of CCQR, many additional extensions are possible by modifying the nonconformity score. For instance, replacing simple averaging with weighted aggregation of scores could improve performance. In such cases, defining and computing appropriate weights emerges as a key research challenge. Lastly, adapting CCQR to ensure conditional coverage aligns with growing interest in distribution-aware uncertainty quantification.

Declarations

Jung's work has been partially supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (No. RS-2022-NR068754).

Appendix

Appendix B offers additional empirical evaluations of the asymmetric and locally adaptive variants introduced in Section 3.4. Appendix A provides full implementation details, including hyperparameter configurations and results for each base model used in Section 4.1.

Appendix A Detail of experiments in Section 4.1

We provide additional details on the experiments presented in Section 4.1. First, we describe the base models which we use in the main text. We employ three base models: QRF, MCQRNN, and QERT. All three models support simultaneous estimation of multiple quantiles, making them particularly well-suited for CCQR. QRF extends standard random forests to estimate conditional quantiles by leveraging weights derived during model fitting. As a neural network model, MCQRNN mitigates quantile crossing by enforcing monotonicity through architectural constraints. QERT shares QRF's ability to estimate conditional quantiles but introduces additional randomness during training by selecting split points at random. Tree-based models rely on the `quantile-forest` package, while the MCQRNN implementation is translated from its original `R` package to `python` for this study.

For each dataset, we first set aside a test set and then split the remaining data into training and validation sets in a 7:3 ratio to tune hyperparameters of base models prior to the calibration step. For the tree-based models, QRF and QERT, we tune the following hyperparameters:

- `max_depth`: Maximum depth of each tree.
- `n_estimators`: Number of trees in the forest.
- `max_features`: Number of features considered at each split.
- `min_samples_split`: Minimum number of samples required to split an internal node.

For MCQRNN, we tuned three hyperparameters: the number of nodes in the hidden layers, the dropout rate, and the ℓ_1 regularization parameter.

Once tuning is complete, we re-split the data into training and calibration sets with equal sizes 5:5. The final model is trained on the training set, and conformal prediction methods are applied using the calibration set. For CCQR, the quantile range d is chosen from $\{0.05, 0.10, \ldots, 0.40\}$ to minimize the average width of prediction interval on the calibration set.

Figures A1, A2, and A3 present the results for each base model. While the performance of other conformal methods varies considerably across base models, CCQR generally achieves target coverage and tigher prediction intervals. This robustness demonstrates CCQR's reduced sensitivity to base model misspecification.

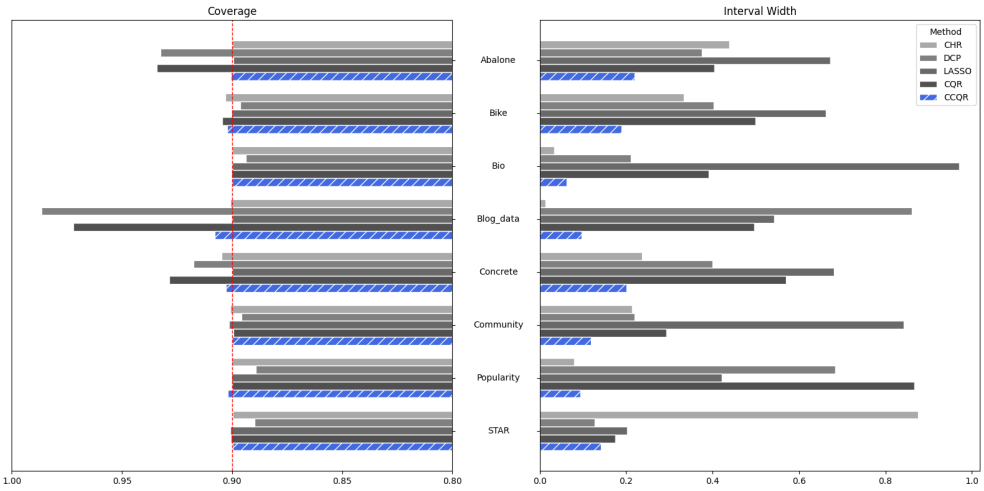

Fig. A1: Comparison of conformal prediction methods using the QRF base model

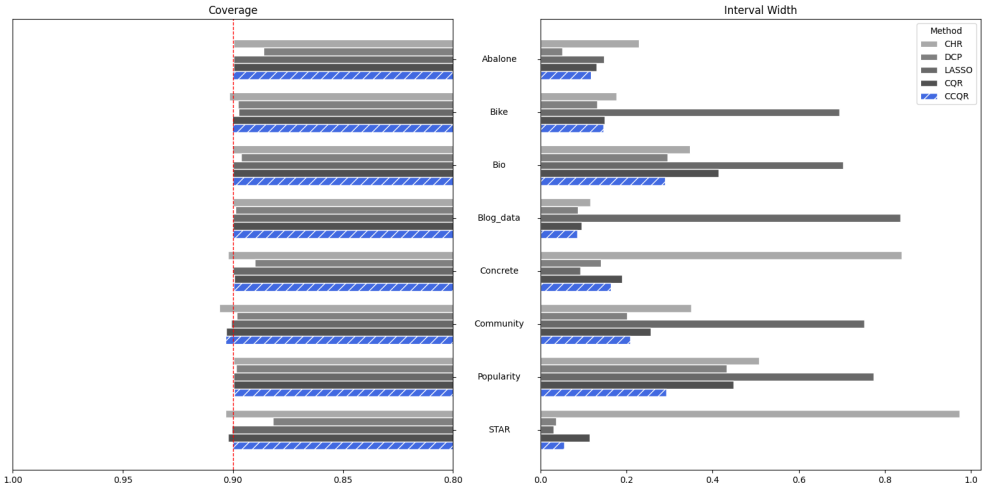

Fig. A2: Comparison of conformal prediction methods using the MCQRNN base model

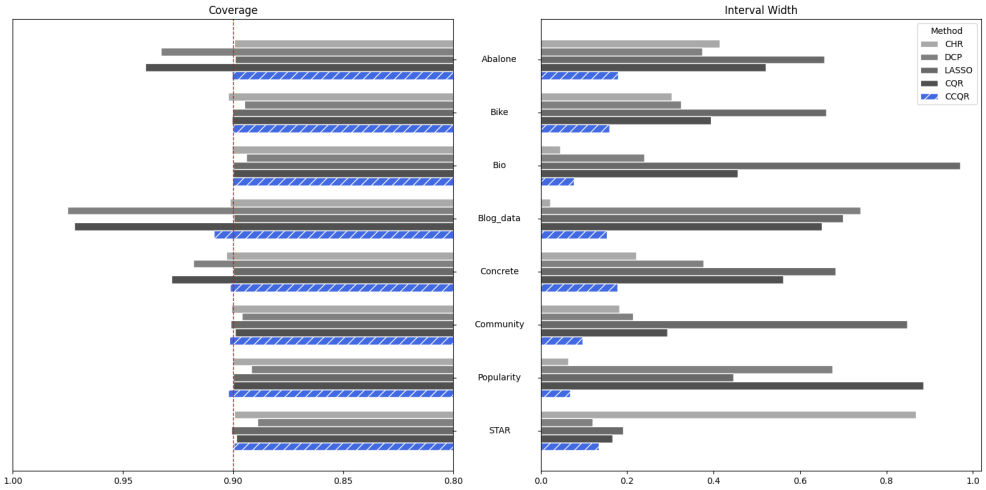

Fig. A3: Comparison of conformal prediction methods using the QERT base model

Appendix B The performance of CCQR variants

We conduct additional experiments to evaluate two CCQR variants: the asymmetric version (CCQR-A) and the locally adaptive version (CCQR-L). CCQR-A treats lower and upper quantiles separately, introducing two hyperparameters: the lower quantile

range d_1 and the upper quantile range d_2. We determine the optimal values of d_1 and d_2 via grid search over $\{0.05, 0.10, \ldots, 0.40\}$, selecting the combination that minimizes average interval width on the calibration set. Meanwhile, CCQR-L uses a weighted nonconformity score, where each weight corresponds to the interval wdith between the estimated k-th upper and lower quantiles at each calibration point: $\hat{q}_{u_k}(X_j) - \hat{q}_{\ell_k}(X_j)$ for $j \in \mathcal{I}_2$. The quantile range d for CCQR-L is tuned using the same procedure described earlier. The number of quantiles is set to $K = 9$ for both methods.

We use the `Bike` and `Concrete` datasets to evaluate original CCQR, CCQR-A, CCQR-L, and CQR. Two base models are considered: QRF and MCQRNN. We construct 90% prediction intervals and evaluate performance using empirical marginal coverage, average width of interval, and a conditional coverage metric. To assess conditional coverage, we follow the procedure in Section 4.3, grouping test points into quartiles based on the width of their prediction intervals. We report coverage within the narrowest (Q1) and widest (Q4) quartiles, denoted as Coverage-Q1 and Coverage-Q4, respectively.

Figure B4 shows results averaged over 100 iterations, with metrics aggregated across both datasets. All methods achieve the target marginal coverage across both base models. Conditional coverage analysis indicates that CCQR-A and CCQR-L provide modest improvements over standard CQR. However, these gains come at the cost of increased interval width. CCQR variants tend to produce wider intervals than standard CCQR, and even wider than CQR when MCQRNN is used as the base model.

References

Achilles CM, Bain HP, Bellott F, et al (2008) Tennessee's student teacher achievement ratio (star) project. https://doi.org/10.7910/DVN/SIWH9F

Akrami H, Zamzam O, Joshi A, et al (2024) Beta quantile regression for robust estimation of uncertainty in the presence of outliers. In: ICASSP 2024 - 2024 IEEE

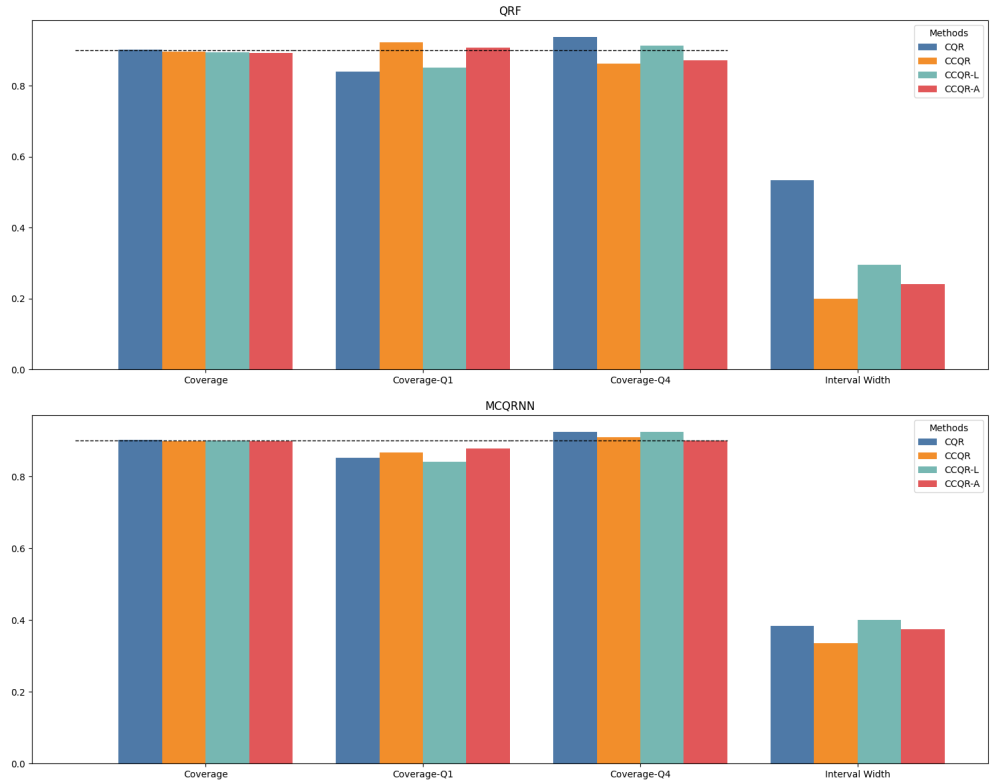

Fig. B4: 90% prediction intervals obtained from CQR and three variants of our method: standard (CCQR), asymmetric (CCQR-A), and locally adaptive (CCQR-L). The top panel presents results with QRF as the base model, while the bottom panel shows outcomes using MCQRNN.

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 7480–7484, https://doi.org/10.1109/ICASSP48485.2024.10445867

Alaa AM, Hussain Z, Sontag D (2023) Conformalized unconditional quantile regression. In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol 206. PMLR, pp 10690–10702

Angelopoulos AN, Kohli AP, Bates S, et al (2022) Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In: Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 162. PMLR, pp 717–730

Barber RF, Candès EJ, Ramdas A, et al (2021) Predictive inference with the jackknife+. The Annals of Statistics 49(1):486–507. https://doi.org/10.1214/20-AOS1965

Bates S, Angelopoulos A, Lei L, et al (2021) Distribution-free, risk-controlling prediction sets. Journal of the ACM 68(6):43. https://doi.org/10.1145/3478535

Bloznelis D, Claeskens G, Zhou J (2019) Composite versus model-averaged quantile regression. Journal of Statistical Planning and Inference 200:32–46. https://doi.org/10.1016/j.jspi.2018.09.003

Bondell HD, Reich BJ, Wang H (2010) Noncrossing quantile regression curve estimation. Biometrika 97(4):825–838. https://doi.org/10.1093/biomet/asq048

Cannon AJ (2018) Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. Stochastic Environmental Research and Risk Assessment 32(11):3207–3225. https://doi.org/10.1007/s00477-018-1573-6

Carlsson L, Eklund M, Norinder U (2014) Aggregated conformal prediction. In: Artificial Intelligence Applications and Innovations. Springer Berlin Heidelberg, pp 231–240

Chernozhukov V, Wüthrich K, Zhu Y (2021) Distributional conformal prediction. Proceedings of the National Academy of Sciences 118(48):e2107794118. https://doi.org/10.1073/pnas.2107794118

Dette H, Munk A (1998) Testing heteroscedasticity in nonparametric regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60(4):693–708. https://doi.org/10.1111/1467-9868.00149

Ding T, Angelopoulos A, Bates S, et al (2023) Class-conditional conformal prediction with many classes. In: Advances in Neural Information Processing Systems, pp 64555–64576

Feldman S, Bates S, Romano Y (2021) Improving conditional coverage via orthogonal quantile regression. In: Advances in Neural Information Processing Systems, vol 34. Curran Associates, Inc., pp 2060–2071

Foygel Barber R, Candès EJ, Ramdas A, et al (2020) The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA 10(2):455–482. https://doi.org/10.1093/imaiai/iaaa017

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Machine Learning 63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1

Gibbs I, Cherian JJ, Candès EJ (2025) Conformal prediction with conditional guarantees. Journal of the Royal Statistical Society Series B: Statistical Methodology p qkaf008. https://doi.org/10.1093/jrsssb/qkaf008

Gupta C, Kuchibhotla AK, Ramdas A (2022) Nested conformal prediction and quantile out-of-bag ensemble methods. Pattern Recognition 127:108496. https://doi.org/10.1016/j.patcog.2021.108496

He X, Shi P (1994) Convergence rate of b-spline estimators of nonparametric conditional quantile functions. Journal of Nonparametric Statistics 3(3-4):299–308. https://doi.org/10.1080/10485259408832589

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58(301):13–30. https://doi.org/10.2307/2282952

Kivaranovic D, Johnson KD, Leeb H (2020) Adaptive, distribution-free prediction intervals for deep networks. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol 108. PMLR, pp 4346–4356

Koenker R (2005) L-Statistics and Weighted Quantile Regression, Cambridge University Press, Cambridge, pp 151–172. Econometric Society Monographs

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46(1):33–50

Koenker R, Bassett G (1982) Robust tests for heteroscedasticity based on regression quantiles. Econometrica 50(1):43–61

Lei J, Robins J, Wasserman L (2013) Distribution-free prediction sets. Journal of the American Statistical Association 108(501):278–287. https://doi.org/10.1080/01621459.2012.751873

Lei J, G'Sell M, Rinaldo A, et al (2018) Distribution-free predictive inference for regression. Journal of the American Statistical Association 113(523):1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Liu Y, Wu Y (2011) Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. Journal of Nonparametric Statistics 23(2):415–437. https://doi.org/10.1080/10485252.2010.537336

Meinshausen N (2006) Quantile regression forests. Journal of Machine Learning Research 7(35):983–999

Moon S, Jeon J, Lee JS, et al (2021) Learning multiple quantiles with neural networks. Journal of Computational and Graphical Statistics 30(4):1238–1248. https://doi.org/10.1080/10618600.2021.1909601

Nguyen TT, Huang JZ, Nguyen TT (2015) Two-level quantile regression forests for bias correction in range prediction. Machine Learning 101(1):325–343. https://doi.org/10.1007/s10994-014-5452-1

Nottingham K, Longjohn R, Kelly M (2023) Uci machine learning repository

Papadopoulos H (2008) Inductive conformal prediction: Theory and application to neural networks. In: Tools in Artificial Intelligence. InTech

Papadopoulos H, Proedrou K, Vovk V, et al (2002) Inductive confidence machines for regression. In: Machine Learning: ECML 2002. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 345–356

Papadopoulos H, Vovk V, Gammerman A (2011) Regression conformal prediction with nearest neighbours. Journal of Artificial Intelligence Research 40:815–840

Romano Y, Patterson E, Candès E (2019) Conformalized quantile regression. In: Advances in Neural Information Processing Systems

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, pp 234–241, https://doi.org/10.1007/978-3-319-24574-4_28

Sesia M, Candès EJ (2020) A comparison of some conformal quantile regression methods. Stat 9(1):e261. https://doi.org/10.1002/sta4.261

Sesia M, Romano Y (2021) Conformal prediction using conditional histograms. In: Advances in Neural Information Processing Systems, vol 34. Curran Associates, Inc., pp 6304–6315

Shinkyu A (2025) Testing heteroskedasticity in high-dimensional linear regression. Econometrics and Statistics 35:120–134. https://doi.org/10.1016/j.ecosta.2023.10.003

Takeuchi I, Le QV, Sears TD, et al (2006) Nonparametric quantile estimation. Journal of Machine Learning Research 7(45):1231–1264

Toccaceli P (2022) Introduction to conformal predictors. Pattern Recognition 124:108507. https://doi.org/10.1016/j.patcog.2021.108507

Tung NT, Huang JZ, Nguyen TT, et al (2014) Bias-corrected quantile regression forests for high-dimensional data. In: Proceedings of the 2014 International Conference on Machine Learning and Cybernetics, pp 1–6, https://doi.org/10.1109/ICMLC.2014.7009082

Vovk V (2015) Cross-conformal predictors. Annals of Mathematics and Artificial Intelligence 74(1):9–28. https://doi.org/10.1007/s10472-013-9368-4

Vovk V, Bendtsen C (2018) Conformal predictive decision making. In: Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications, Proceedings of Machine Learning Research, vol 91. PMLR, pp 52–62

Vovk V, Gammerman A, Saunders C (1999) Machine-learning applications of algorithmic randomness. In: Proceedings of the Sixteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '99, pp 444–453

Vovk V, Gammerman A, Shafer G (2005) Algorithmic Learning in a Random World. Springer-Verlag, Berlin, Heidelberg

Wang D, Wang P, Wang C, et al (2022) Calibrating probabilistic predictions of quantile regression forests with conformal predictive systems. Pattern Recognition Letters 156:81–87. https://doi.org/10.1016/j.patrec.2022.02.003

Xu Q, Deng K, Jiang C, et al (2017) Composite quantile regression neural network with applications. Expert Systems with Applications 76:129–139. https://doi.org/https://doi.org/10.1016/j.eswa.2017.01.054

Yang Y, Kuchibhotla AK (2025) Selection and aggregation of conformal prediction sets. Journal of the American Statistical Association 120(549):435–447. https://doi.org/10.1080/01621459.2024.2304061

Zbontar J, Knoll F, Sriram A, et al (2018) fastmri: An open dataset and benchmarks for accelerated mri. arXiv preprint arXiv:181108839

Zheng S (2012) Qboost: Predicting quantiles with boosting for regression and binary classification. Expert Systems with Applications 39(2):1687–1697. https://doi.org/10.1016/j.eswa.2011.06.060

Zou H, Yuan M (2008) Composite quantile regression and the oracle model selection theory. The Annals of Statistics 36(3):1108–1126. https://doi.org/10.1214/07-AOS507