1 Conformalized Quantile Regression

This chapter introduces *Conformalized Quantile Regression (CQR)* as the first of two main post-processing procedures that we implemented in the `postforecasts` package.

Section 1.1 explains the original Conformalized Quantile Regression algorithm as proposed in Romano, Patterson, and Candès (2019). There, we highlight potential limitations of the traditional implementation that could potentially be fixed by more flexible modifications, which are discussed in Section 1.2 and Section 1.3.

1.1 Traditional CQR

All derivations in this sections are taken from the original paper (Romano, Patterson, and Candès 2019). The authors motivate Conformalized Quantile Regression by stating two criteria that an ideal procedure for generating prediction intervals should satisfy:

- It should provide valid coverage in finite samples without making strong distributional assumptions
- The resulting intervals should be as short as possible at each point in the input space

According to the authors, CQR performs well on both criteria while being *distribution-free* and adaptive to *heteroscedasticity*.

1.1.1 Statistical Validity

The algorithm that CQR is build upon is statistically supported by the following Theorem:

Theorem 1.1. *If $(X_i, Y_i), i = 1, \dots, n+1$ are exchangeable, then the $(1-\alpha)\cdot 100\%$ prediction interval $C(X_{n+1})$ constructed by the CQR algorithm satisfies*

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

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

$$P\left(Y_{n+1} \in C(X_{n+1})\right) \leq 1 - \alpha + \frac{1}{|I_2| + 1},$$

where I_2 denotes the calibration set.

Thus, the first statement of Theorem 1.1 provides a coverage *guarantee* in the sense that the adjusted prediction interval is *lower-bounded* by the desired coverage level. The second statement adds an *upper-bound* to the coverage probability which gets tighter with increasing sample size and asymptotically converges to the desired coverage level $1-\alpha$ such that lower bound and upper bound are asymptotically identical.

1.1.2 Algorithm

The CQR algorithm is best described as a multi-step procedure.

Step 1:
Split the data into a training and validation (here called *calibration*) set, indexed by I_1 and I_2, respectively.

Step 2:
For a given quantile α and a given quantile regression algorithm $\mathcal{A}$, calculate lower and upper interval bounds on the training set:

$$\{\hat{q}_{\alpha,low},\ \hat{q}_{\alpha,high}\} \leftarrow \mathcal{A}\left(\{(X_i, Y_i) : i \in I_1\}\right).$$

Step 3:
Compute *conformity scores* on the calibration set:

$$E_i := \max\left\{\hat{q}_{\alpha,low}(X_i) - Y_i,\ Y_i - \hat{q}_{\alpha,high}(X_i)\right\} \quad \forall\, i \in I_2$$

For each i, the corresponding score E_i is *positive* if Y_i is *outside* the interval $[\hat{q}_{\alpha,low}(X_i),\ \hat{q}_{\alpha,high}(X_i)]$ and *negative* if Y_i is *inside* the interval.

Step 4:
Compute the *margin* $Q_{1-\alpha}(E, I_2)$ given by the $(1-\alpha)(1+\frac{1}{1+|I_2|})$-th empirical quantile of the scores E_i in the calibration set. For small sample sizes and small quantiles α the quantile above can be greater than 1 in which case it is simply set to 1 such that the maximum value of the score vector is selected.

Step 5:
On the basis of the original prediction interval bounds $\hat{q}_{\alpha,low}(X_i)$ and $\hat{q}_{\alpha,high}(X_i)$, the new *post-processed* prediction interval for Y_i is given by

$$C(X_{n+1}) = [\hat{q}_{\alpha,low}(X_i) - Q_{1-\alpha}(E, I_2),\ \hat{q}_{\alpha,high}(X_i) + Q_{1-\alpha}(E, I_2)]\,.$$

Note that the *same* margin $Q_{1-\alpha}(E, I_2)$ is subtracted from the original lower quantile prediction and added to the original upper quantile prediction. This limitation is adressed in Section 1.2.

1.2 Asymmetric CQR

As noted in Section 1.1 this section suggests a first extension to the original algorithm. Instead of limiting ourselves to choosing the *same* margin $Q_{1-\alpha}(E, I_2)$ for adjusting the original lower and upper quantile predictions, we allow for individual and, thus, generally different margins $Q_{1-\alpha,low}(E, I_2)$ and $Q_{1-\alpha,high}(E, I_2)$ such that the post-processed prediction interval is given by

$$C(X_{n+1}) = [\hat{q}_{\alpha,low}(X_i) - Q_{1-\alpha,low}(E_{low}, I_2),\ \hat{q}_{\alpha,high}(X_i) + Q_{1-\alpha,high}(E_{high}, I_2)]\,.$$

This asymmetric version additionally requires a change in the computation of the conformity scores. Instead of considering the elementwise maximum of the differences between observed values Y_i and original bounds, we simply compute two separate score vectors:

$$\begin{aligned} E_{i,low} &:= \hat{q}_{\alpha,low}(X_i) - Y_i \quad \forall\, i \in I_2 \\ E_{i,high} &:= Y_i - \hat{q}_{\alpha,high}(X_i) \quad \forall\, i \in I_2 \end{aligned}$$

1.3 Multiplicative CQR

Theory

On top of the asymmetric CQR version introduced in Section 1.2, we can extend the CQR algorithm further. So far, the adjustments to the original prediction interval were always chosen in *additive* form. It may be useful to leverage the *magnitude* of the original bounds more explicitly by using *relative* or *multiplicative* adjustments.

Hence, we again compute separate margins $Q_{1-\alpha,low}(E, I_2)$ and $Q_{1-\alpha,high}(E, I_2)$ which are now *multiplied* with the existing forecasts. The post-processed prediction interval is then given by

$$C(X_{n+1}) = [\hat{q}_{\alpha,low}(X_i) \cdot Q_{1-\alpha,low}(E_{low}, I_2),\ \hat{q}_{\alpha,high}(X_i) \cdot Q_{1-\alpha,high}(E_{high}, I_2)]\,.$$

Just like the asymmetric version, the computation of the score vectors is changed accordingly to respect the new multiplicative relationship:

$$\begin{aligned} E_{i,low} &:= \frac{Y_i}{\hat{q}_{\alpha,low}(X_i)} \quad \forall\, i \in I_2 \\ E_{i,high} &:= \frac{Y_i}{\hat{q}_{\alpha,high}(X_i)} \quad \forall\, i \in I_2, \end{aligned}$$

where we have to exclude original predictions with the value 0. Since in our application of Covid-19 Cases and Deaths all values are non-negative, we threshold the scores at zero such that $E_{i,low}$ equals 0 whenever $\hat{q}_{\alpha,low}(X_i) \leq 0$.

Regularization

While the idea of multiplicative correction terms is appealing, it turns out that the approach above is flawed in two ways:

1. Recall that the (lower) margin $Q_{1-\alpha,low}(E, I_2)$ basically *picks* a value of the score vector E_{low} at a given quantile level. The score vectors are computed for each combination of *location*, *model*, *target type*, *horizon* and *quantile*, i.e. the number of values in the score vector is identical to the number of distinct time points in the training set. For short time series such as our small UK data set, the margin selects the *largest* value in the score vector for small levels of α such as 0.01 or 0.05, where each such value represents a *ratio* of observed Y_i and original prediction $\hat{q}_{\alpha,low}(X_i)$.

   As one might guess, these factors frequently get very large for small initial quantile predictions $\hat{q}_{\alpha,low}(X_i)$ such that the selected margin for post-processing is unreasonably large. In fact, the margin can remain huge if there exists a *single* outlier in the score vector. In particular, this naive multiplicative version frequently adjusts the lower quantile prediction to a higher value than its upper quantile counterpart.

   We counteract this extreme sensitivity to outliers by *reducing the spread* inside of the score vector to make it more well behaved. Since we deal with multiplicative factors it makes no sense to standardize them to zero mean and unit variance. Instead, we regularize the score vector by pulling all values closer to 1, while keeping all values nonnegative and respecting their *directions*, i.e. values smaller than 1 that reduce the interval width keep doing so but to a lesser extent than before and, analogously, prior values greater than one remain to be greater than 1.

   This goal is achieved by a *root transformation*. Since a greater spread of the score vector should lead to larger regularization we settled on the corrections

$$E_{i,low}^{reg} = E_{i,low}^{\left(\frac{1}{\sigma_{E_{low}}}\right)}, \quad E_{i,high}^{reg} = E_{i,high}^{\left(\frac{1}{\sigma_{E_{high}}}\right)},$$

   where σ_E denotes the standard deviation of the corresponding score vector.

2. Chances are high that at least *one* of the original true values Y_i is larger than its corresponding lower quantile prediction $\hat{q}_{\alpha,low}(X_i)$ such that the maximum of the (regularized) score vector is still larger than 1. Thus, the lower bound for small quantiles α is almost *always* pushed upwards. The same logic applies to the upper bound in which case the entire interval is shifted to the top. This behaviour is usually not desired.

   To prevent interval shifts, we add the additional constraint that the lower and upper margin must multiply to 1, i.e.

$$Q_{1-\alpha,low} \cdot Q_{1-\alpha,high} \overset{!}{=} 1.$$

   Hence, when the lower bound is adjusted upwards ($Q_{1-\alpha,low} > 1$), the upper bound *must* decrease ($Q_{1-\alpha,high} < 1$) and the interval becomes smaller. Similarly, when the upper bound is adjusted upwards ($Q_{1-\alpha,high} > 1$), the lower bound must decrease ($Q_{1-\alpha,low} < 1$) leading to larger intervals overall after post-processing.

Results

As noted in the previous section, *naive* multiplicative Conformalized Quantile Regression without any regularization is useless for post-processing quantile predictions. Typically, one can observe strong overfitting on the training set such that the training performance indicates promising effects, yet the scores on the validation set are *much* worse than the original forecasts. Further, the adjusted intervals are shifted upwards and usually too large.

Before numerically evaluating the performance of *regularized* CQR, it is instructive to look at a visual comparison of the original and post-processed forecasts of all three CQR modifications for one specific feature combination, which is shown in Figure 1.

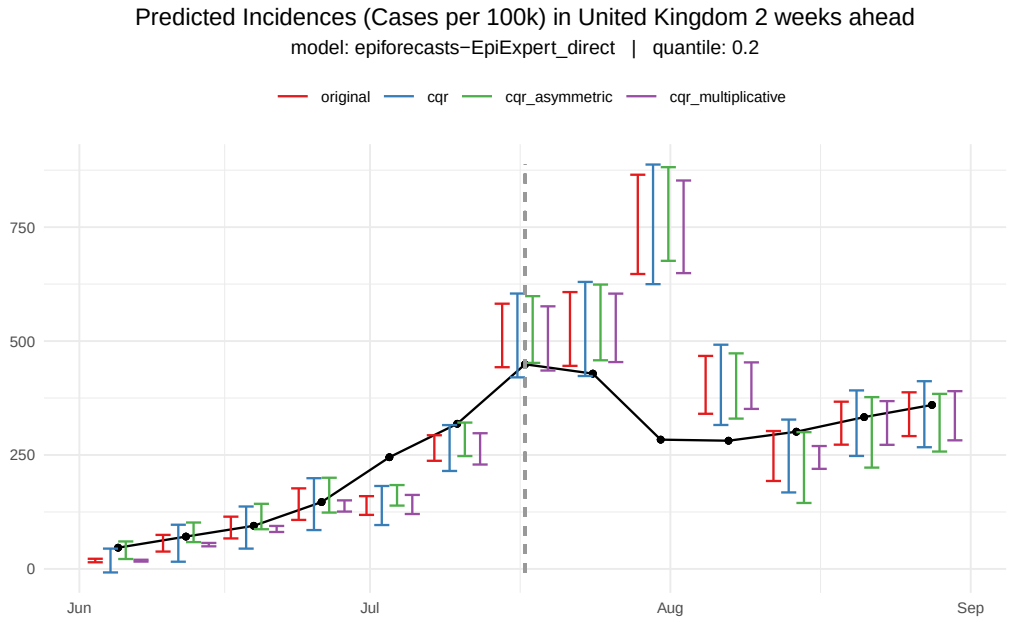

Figure 1: Comparison of CQR variations on the UK data set

Table 1: Performance of Multiplicative CQR on the Training Set

method	interval_score	dispersion	underprediction	overprediction
cqr_multiplicative	24.49	5.98	18.01	0.51
original	23.62	4.16	18.88	0.58

The effect of scaling the score vectors in step 1 of the regularization procedure and constraining lower and upper margins in the second step can be detected immediately: Similar to vanilla CQR, the corrected intervals are now centered around the same midpoint as the original forecasts. In strong contrast to the additive CQR versions, however, the issue of interval explosion has not only been diminished by downscaling the scores, but rather *reversed* such that the interval widths now actually *decreased* at most time points and generally appear too narrow.

Moreover, we no longer have any theoretical guarantees of improved forecasts on the training set since Theorem 1.1 only applies to the original additive and symmetric version of CQR. This fact is confirmed empirically by Table 1 which shows the Weighted Interval Score aggregated over all categories of `model`, `target_type`, `horizon` and `quantile`. Indeed, the multiplicative adjustments result in a slightly worse Weighted Interval Score.

Recall that this behaviour is different from the unregularized version, which performed better than the original forecasts across almost all feature combinations. On the flipside, the performance on the validation set improved dramatically compared to the naive implementation, even though it does *not* lead to a score improvement in absolute terms as is shown in Table 2 separated by the forecasting `model`. Interestingly, multiplicative CQR indicates the strongest *relative* performance for the `EuroCOVIDhub-baseline` model where the additive CQR algorithms struggle the most. Overall the score differences across models appear to be smoothed out which also results from the regularization component that is unique to the multiplicative modification.

TODO: Add mixed effect on intervals size, some are increased others are decreased

The trend of Figure 1 can be observed in a similar fashion across many different feature combinations: Starting from huge intervals in the naive implementation, the effect of downscaling the score vectors tends to

Table 2: Performance of Multiplicative CQR for each Model on the Validation Set

method	model	interval_score	dispersion
cqr_multiplicative	epiforecasts-EpiExpert	71.60	10.05
original	epiforecasts-EpiExpert	67.74	12.07
cqr_multiplicative	epiforecasts-EpiExpert_direct	67.33	9.01
original	epiforecasts-EpiExpert_direct	66.30	11.62
cqr_multiplicative	epiforecasts-EpiExpert_Rt	89.03	15.48
original	epiforecasts-EpiExpert_Rt	79.61	14.37
cqr_multiplicative	EuroCOVIDhub-baseline	29.85	17.95
original	EuroCOVIDhub-baseline	29.61	5.92
cqr_multiplicative	EuroCOVIDhub-ensemble	61.24	15.48
original	EuroCOVIDhub-ensemble	56.07	14.00
cqr_multiplicative	seabbs	98.18	9.14
original	seabbs	95.11	14.03

be too extreme. As a consequence, we experimented with adding an additional *hyperparameter* λ to dampen the regularization. However, we could not find any values of λ within the trade-off between lower scaling effects and too large intervals and stronger regularization at the cost of too narrow intervals that consistently outperformed even the original forecasts.

Hence, we must conclude that the original CQR algorithm as described in (Romano, Patterson, and Candès 2019) can *not* be modified towards multiplicative margins in any straightforward way. For this reason, we do not extend the analysis of multiplicative CQR to the European Forecast Hub data set and do not include it in the detailed method comparison in **??**.

Romano, Yaniv, Evan Patterson, and Emmanuel J. Candès. 2019. "Conformalized Quantile Regression." *arXiv:1905.03222 [Stat]*, May. http://arxiv.org/abs/1905.03222.