1 Conformalized Quantile Regression

This chapter introduces *Conformalized Quantile Regression (CQR)* as the first of two main Post-Processing procedures which are implemented in the `postforecasts` package.

Section 1.1 explains the original Conformalized Quantile Regression algorithm as proposed by Romano, Patterson, and Candès (2019). There, we highlight potential limitations of the traditional implementation that could potentially be diminished by more flexible variants of CQR that are discussed in **??** and **??**.

1.1 Traditional CQR

All derivations in this section are taken from the original paper (Romano, Patterson, and Candès 2019). The authors motivate Conformalized Quantile Regression by stating two criteria that the 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 narrow 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 Theorem 1.1. The term *conformity scores* is defined in Section 1.1.2.

Theorem 1.1. *If* $(X_i, Y_i), i = 1, \ldots, n+1$ *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 (validation) 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 $\mathcal{A}$, compute the original lower and upper quantile predictions on the training set:

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

Note that the algorithm does *not* make any assumptions about the structural form of $\mathcal{A}$ which, in theory, could be a highly nonlinear function like a Deep Neural Network.

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

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

Thus, 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 score vector E in the calibration set. For small sample sizes and small quantiles α this procedure might result in quantiles greater than 1. In this case we simply select the maximum value of the score vector.

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

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

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

1.1.3 Results

We now investigate how well the algorithm performs for post-processing Covid-19 forecasts. Thereby we start UK Covid-19 Forecasting Challenge data set and briefly point out recurrent trends within CQR adjustments. Then, we continue with a more detailed discussion of the findings on the larger European Forecast Hub data.

One common characteristic that applies to almost all feature combinations is that CQR *expands* the original forecast intervals. As stated in **step 5** of Section 1.1.2 it moves the original lower and upper bounds in a *symmetric* way either inwards or outwards by using the *same* margin. This implies that the interval *midpoint* remains unchanged when applying the traditional CQR algorithm.

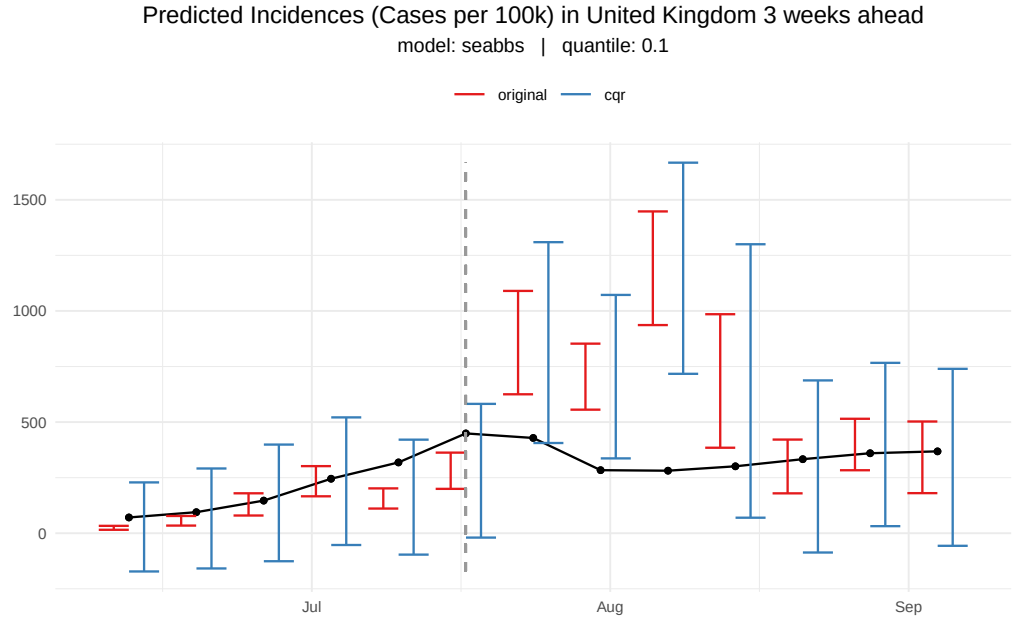

Figure 1: CQR tends to make prediction intervals larger, here for the `seabbs` forecasting model

One extreme example of this behaviour is shown in Figure 1. Since the `seabbs` forecasts are submitted by a single individual, we find evidence for the hypothesis of **??** that humans tend to be too confident in their

Table 1: WIS improvement by CQR for one particular feature combination on the validation set.

method	model	target_type	horizon	quantile	interval_score	dispersion
cqr	seabbs	Cases	3	0.1	157.32	87.49
original	seabbs	Cases	3	0.1	210.62	38.14

Table 2: Overall WIS improvement by CQR on the validation set.

method	interval_score	dispersion
cqr	62.15	24.1
original	65.74	12.0

own predictions resulting in too narrow uncertainty bounds. By extending the intervals symmetrically CQR maintains *pointwise* information from the original forecasts while simultaneously increasing interval coverage.

Yet, the pure effect of increasing coverage does not automatically imply that the Weighted Interval Score has improved as well due to the trade-off between coverage and precision. Thus, we explicitly compute the WIS, once for the specific covariate combination of Figure 1 in Table 1 and once aggregated over all *models*, *target types*, *horizons* and *quantiles* in Table 2.

Both tables confirm the visual impression of Figure 1: CQR improves the WIS by increasing the *dispersion* value, a measure for the interval *spread*. This effect is particularly strong in case of the `seabbs` model but still applies to a more moderate extent to most of the other forecasting models.

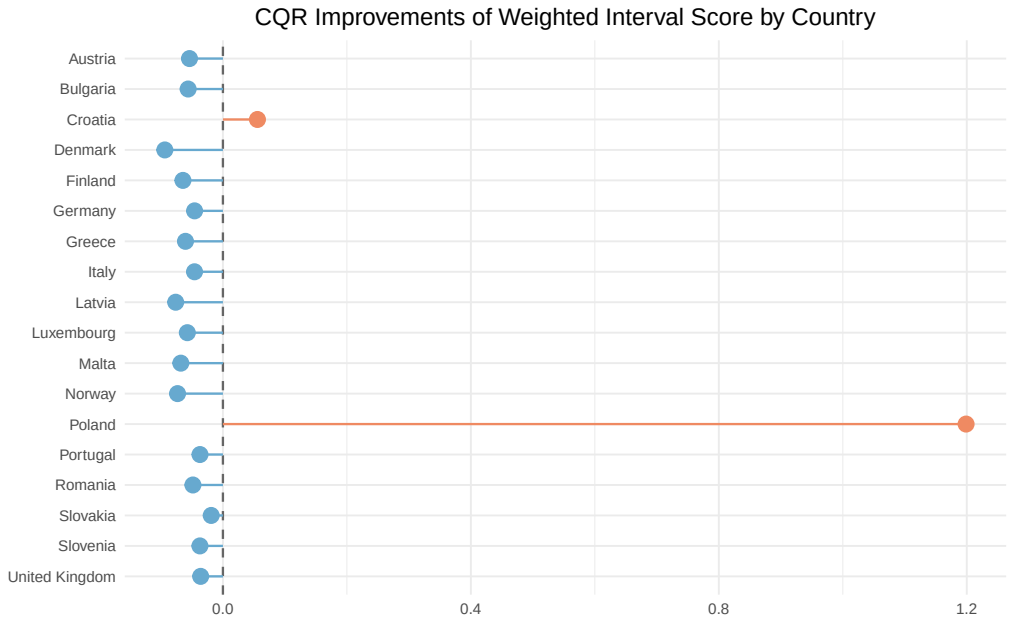

Figure 2: CQR proves to be beneficial for the vast majority of countries, with the major exception of Poland

Since many of the general findings for traditional CQR coincide between the UK data and the EU Forecast Hub data, we jump straight to the latter for the following analysis. First, we investigate if CQR is equally effective across all countries. Figure 2 indicates that this is clearly *not* the case: CQR is beneficial on out of sample data in almost all of the 18 selected countries. The largest effect size, however, is linked to Poland in *negative* direction.

Table 3: Weighted Interval Score for Poland by Model on Training and Validation Set

method	model	training score	validation score
cqr	epiforecasts-EpiNow2	22.84	1.94
original	epiforecasts-EpiNow2	23.71	1.32
cqr	EuroCOVIDhub-ensemble	29.44	2.34
original	EuroCOVIDhub-ensemble	31.37	0.96
cqr	IEM_Health-CovidProject	56.54	4.68
original	IEM_Health-CovidProject	62.04	0.91

At first sight this finding seems like a data entry error, there is no obvious reason why such a general algorithm like Conformalized Quantile Regression might not work for one specific location. The large negative effect is also interesting in light of Theorem 1.1: We know that CQR *always* improves the forecast intervals on the training set which, of course, applies to Poland as well. We can confirm this theoretical guarantee empirically by evaluating the Weighted Interval Score for Poland on the training set only. Table 3 collects the training and validation scores for three selected forecasting models separately.

Indeed, CQR improves the WIS for all three models in-sample whereas the out-of-sample performance drops dramatically. This finding provides evidence that the observations used for the initial training phase must be *fundamentally different* to those encountered during the Cross Validation process. More specifically, it suggests a *distribution shift* of the true observed values and/or the original quantile predictions right at the split of training and validation phase. Further, the *scale* of training and validation scores is quite different, which usually stems from different magnitudes of the observed incidences within each stage.

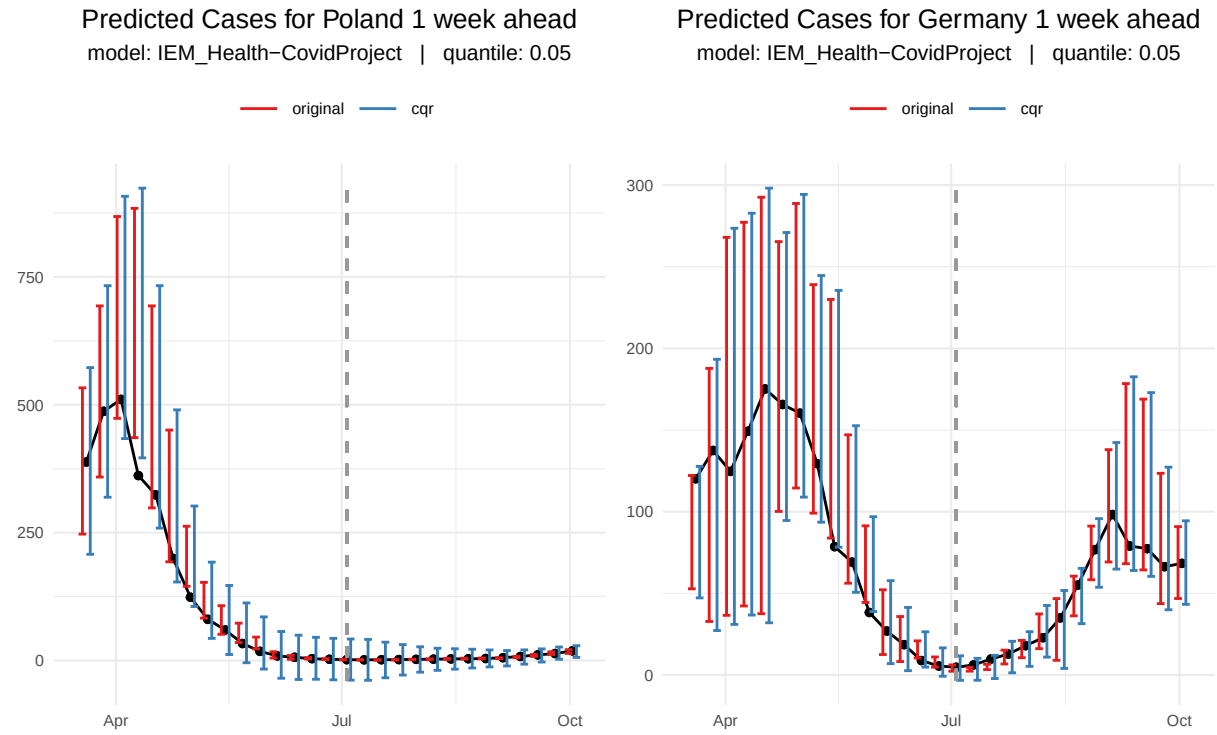

Figure 3: Development of Covid-19 Cases in 2021 in Poland and Germany

Figure 3 confirms our hypothesis for 1 week-ahead forecasts of 90% prediction intervals for Covid-19 Cases. The left plot displays the development of observed and predicted values for the outlier Poland compared to the same setting for Germany where CQR performs just fine. A few weeks before the training-validation

split, which is highlighted by the grey dashed line, the true incidences plummeted in Poland. In strong contrast to Germany, where the Covid-19 situation relaxed during the summer months of 2021 as well, the incidences *remain* low until late autumn in Poland (according to the collected data of the European Forecast Hub). Thus, the incidences are indeed much lower on average in the validation set which explains the scale discrepancy in Table 3.

The consistently low incidences are connected to decreased uncertainty margins of the original forecasts that were submitted only one week in advance. The forecasters were well aware of the current Covid-19 situation and were able to quickly react with lower point forecasts and narrower prediction intervals. CQR is not capable of competing with this flexibility and requires a long time span to adapt to irregular behaviour. The reasons for these slow adjustments, which reveal a major downside of CQR, follow immediately from the underlying statistical theory and are explained in detail in **??**.

Lastly, we summarize the performance of vanilla CQR across different *quantiles*, *target types* and *horizons*. To obtain more informative visual illustrations we exclude Poland from the further analysis.

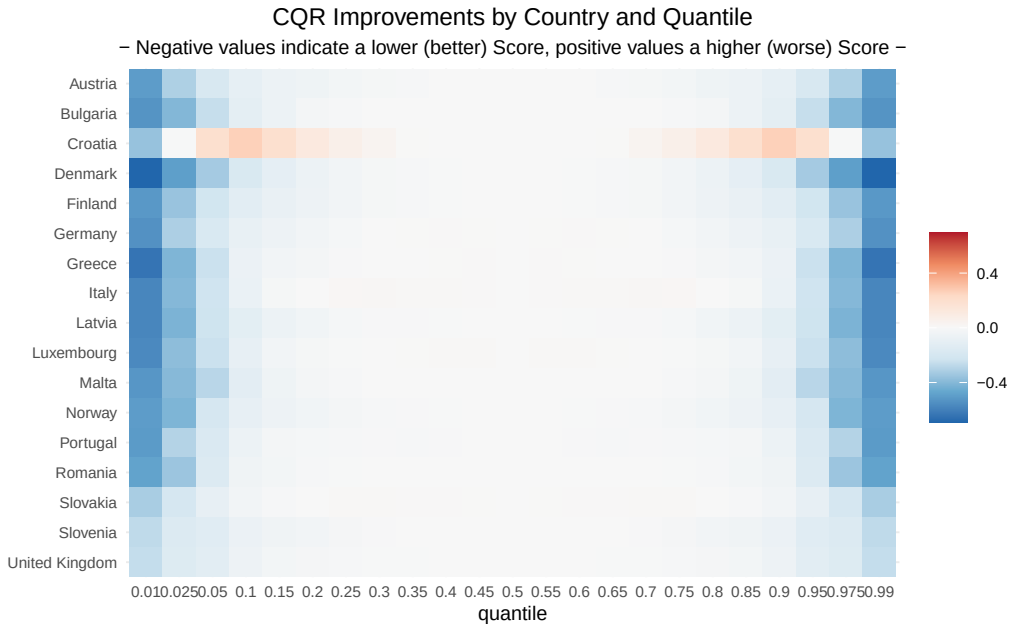

Figure 4: CQR is most beneficial for extreme quantiles

Figure 4 shows the performance of CQR for all 23 quantile levels in the data set. Although the effect size varies by country, the general trend holds unanimously: Extreme quantiles in the tails of the predictive distribution benefit most from post-processing with a gradual decline towards centered quantiles. The same trend can be observed to an even larger extent for non-expert forecasts in the UK Covid-19 Forecasting Challenge data set.

Similar to quantiles there exist obvious tendencies for different forecast *horizons* as well. Figure 5 shows the performance of CQR across horizons, again stratified by country. Although the effects are more diverse compared to Figure 4, CQR generally works better for larger forecast horizons. Exceptions of this rule are Croatia, which is the only country besides Poland with a negative effect of CQR, and Malta, where the trend is actually reversed and CQR corrections are most beneficial for short-term forecasts.

Both of the previous figures suggest that post-processing with Conformalized Quantile Regression is worthwhile whenever the uncertainty is comparably high, which is the case for both quantiles in the tails and large forecast horizons.

Lastly, Table 4 aggregates Weighted Interval Scores on the validation set by *target type*. Interestingly, the effect directions disagree for the first time: While forecasts for Covid-19 Cases benefit significantly, forecasts

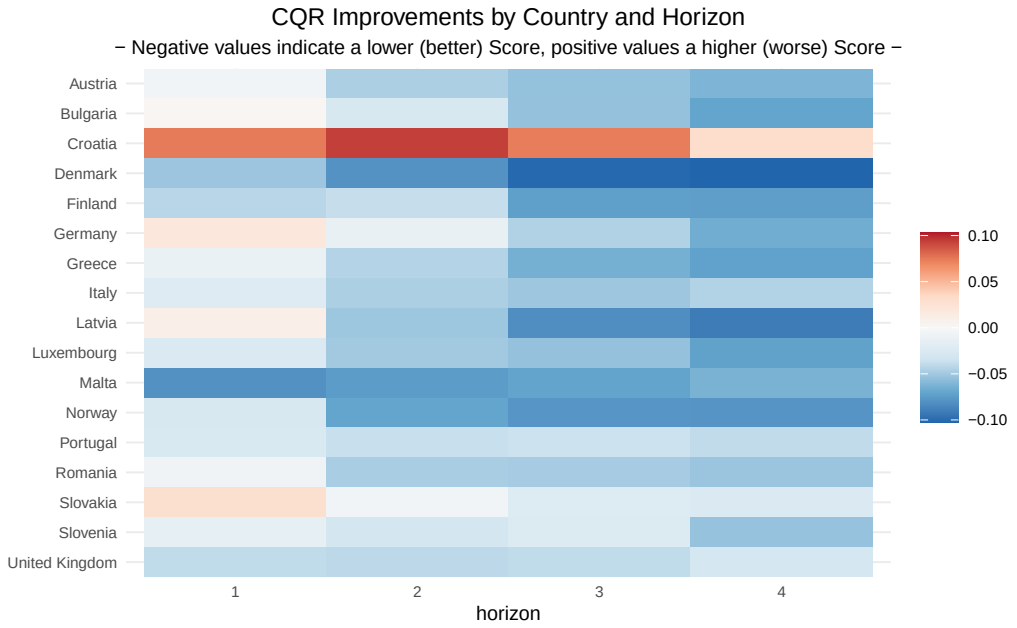

Figure 5: CQR is most beneficial for large Forecast Horizons

Table 4: CQR Improvements by Target Type for European Forecast Hub Data excluding Poland

method	**target_type**	**interval_score**	**dispersion**
cqr	Cases	59.26	19.55
original	Cases	62.69	15.49
cqr	Deaths	0.32	0.15
original	Deaths	0.32	0.13

for Deaths become slightly worse through CQR adjustments.

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