## 0.1 Multiplicative CQR

### 0.1.1 Theory

On top of the asymmetric CQR modification described in **??**, 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 $Q_{1-\alpha,low}(E_{low}, I_2)$ and $Q_{1-\alpha,high}(E_{high}, I_2)$ which are now *multiplied* with the existing forecasts. The post-processed prediction interval is thus given by

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

Similar to the asymmetric additive 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 all Covid-19 Cases and Deaths are non-negative, we threshold the scores at zero such that $E_{i,low}$ equals 0 whenever $\hat{q}_{\alpha,low}(X_i) \leq 0$.

Note that the actual limiting value

$$\lim_{\hat{q}_{\alpha,low}(X_i) \to 0} \frac{Y_i}{\hat{q}_{\alpha,low}(X_i)} = \infty$$

does *not* make sense here since infinite scores would cause infinite lower margins $Q_{1-\alpha,low}(E_{low}, I_2)$, which in return result in infinite updated lower bounds. Thus, the value 0 is deliberately chosen to minimize the influence of negative original forecasts and keep the updated lower quantile predictions always nonnegative.

### 0.1.2 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 $Q_{1-\alpha,low}(E_{low}, I_2)$ 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 $\alpha$ 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 computed margin $Q_{1-\alpha,low}(E_{low}, I_2)$ 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, leading to (an extreme form of) quantile crossing.

   We counteract this sensitivity to outliers by *reducing the spread* of the score vector. 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 remain smaller than 1 and prior values greater than 1 remain greater than 1.

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 |

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

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

where $\sigma_E$ denotes the standard deviation of the corresponding score vector.

*Remark*: We first restricted the scaling to the case $\sigma_{E_{low}}, \sigma_{E_{high}} > 1$, i.e. the spread of the score vector should only get reduced. However, the above correction empirically proved to be beneficial even for $\sigma_{E_{low}}, \sigma_{E_{high}} < 1$ in which case the score variance gets *increased*. Therefore we removed the original restriction and only handled the (unlikely) case of constant score vectors with $\sigma_{E_{low}} = 0$ or $\sigma_{E_{high}} = 0$ separately.

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 $\alpha$ 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.

### 0.1.3 Results

As noted in Section 0.1.2, *naive* multiplicative Conformalized Quantile Regression without any regularization is useless for updating quantile predictions. Typically, one would observe strong overfitting on the training set such that the training performance indicated promising effects, yet the scores on the validation set would be *much* worse than the original forecasts. Further, the adjusted intervals would be shifted upwards and usually be too large.

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

The effect of scaling the score vectors in step one of the regularization procedure and constraining lower and upper margins in the second step can be detected immediately: Similar to vanilla CQR, the multiplicatively 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 **??** only applies to the original additive and symmetric version of CQR. This fact is confirmed empirically by Table 1

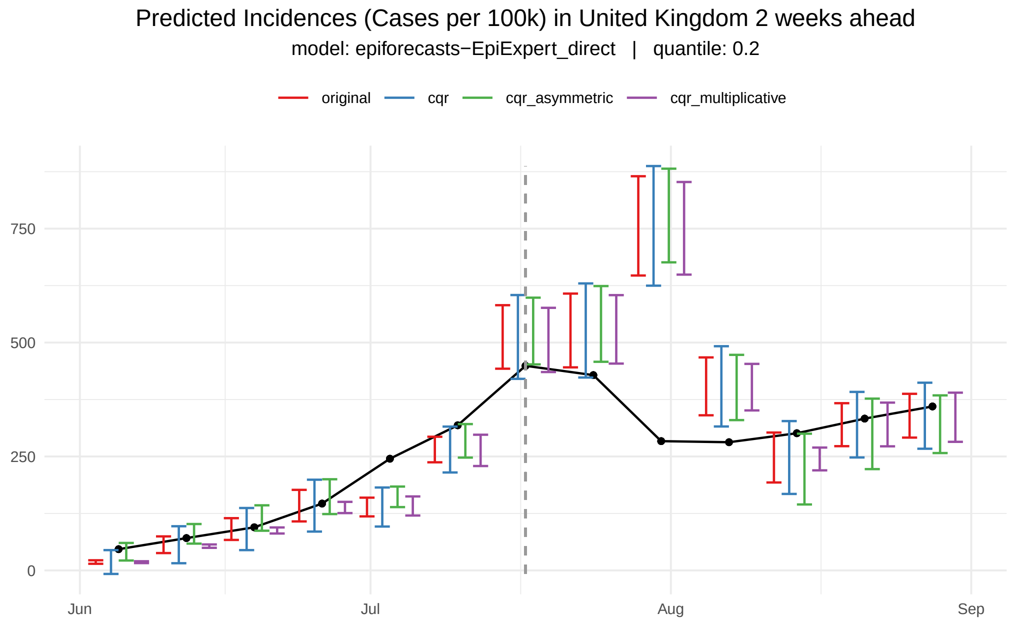

Figure 1: Comparison of CQR variations on the UK data set

Table 2: Performance of Multiplicative CQR by 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 | 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 |

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 WIS on the training set.

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

The impression of too narrow adjusted intervals does not generalize to the entire data set. The *dispersion* column in Table 2 shows that the intervals are downsized only for some models such as `epiforecasts-EpiExpert` whereas for others like `epiforecasts-ensemble` the distance between lower and upper bound gets larger on average.

Table 3 indicates a connection of the dispersion change by multiplicative CQR with the `quantile` level. Aggregated over all models, target types and horizons the dispersion value is increased by a large amount

Table 3: Dispersion of Multiplicative CQR by Quantile on the Validation Set

| **method** | **0.01** | **0.025** | **0.05** | **0.1** | **0.15** | **0.2** | **0.25** | **0.3** | **0.35** | **0.4** | **0.45** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cqr_multiplicative | 8.79 | 7.40 | 11.98 | 16.83 | 18.26 | 18.82 | 18.24 | 17.53 | 14.60 | 10.12 | 5.22 |
| original | 2.82 | 5.82 | 9.65 | 14.86 | 17.84 | 18.99 | 18.83 | 17.44 | 14.71 | 10.97 | 6.08 |

for extreme quantiles but remains in a similar range as before for quantiles in the center of the predictive distribution. This behaviour is in line with the previously seen additive correction methods and emphasizes that Figure 1 is not representative for the entire UK data set.

Overall, we must conclude that the original CQR algorithm as described by Romano, Patterson, and Candès (2019) can *not* be modified towards multiplicative margins in any straightforward way. For this reason, we neither extend the analysis of multiplicative CQR to the European Forecast Hub data set nor include it in the 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.