SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

To Conformalise or not to Conformalise in Growth-at-Risk Analysis? Assessing the Empirical Performance of Calibrated Versus Non-Calibrated Quantile Prediction Models.

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Pietro Bogani, 10622041

Advisor:
Prof. Simone Vantini

Co-advisors:
Matteo Fontana
Luca Neri

Academic year:
2023-2024

Abstract: Accurate computation of robust estimates for extremal quantiles of empirical distributions is essential for a wide range of applications, including those in policymaking and the financial industry. These estimates are particularly critical for the calculation of risk measures such as Growth-at-Risk (GaR) and Value-at-Risk (VaR). This thesis presents an extensive simulation study and a real-world analysis of GaR to examine the benefits of using Conformal Prediction for quantile estimation. This work represents the first application of Conformal Prediction, traditionally used to generate prediction intervals, for the estimation of quantiles.
Our findings show that Conformal Prediction methods consistently enhance the calibration and robustness of quantile estimates across all levels, especially at extreme quantiles, which are critical for risk assessment and where traditional methods often fall short. Additionally, we introduce a novel property that guarantees coverage under the exchangeability assumption, providing a valuable tool for managing risks by quantifying and controlling the likelihood of future extreme observations.

Key-words: Conformal Prediction, Quantile Estimation, Quantile Regression, Growth at Risk

1. Introduction

Economic forecasts for GDP growth have traditionally focused on providing point estimates. Growth-at-Risk (GaR) refers to a framework that quantifies the potential downside risks to future economic growth, considering the distribution of possible outcomes rather than focusing solely on central forecasts. By assessing the likelihood of adverse economic conditions and their potential impact, GaR provides a more comprehensive understanding of economic vulnerability and resilience. GaR was introduced in 2017 by the IMF in its Global Financial Stability Report [1] and has since become a popular tool for a variety of economic actors. The importance of GaR has been magnified in recent years due to several converging factors. Firstly, the global economy has faced heightened uncertainty and volatility, driven by events such as the COVID-19 pandemic, geopolitical tensions, and climate-related issues. These events have underscored the need for robust risk assessment tools that can capture the tail risks and systemic vulnerabilities that traditional forecasting models may overlook, as they focus merely on the central forecast. GaR is valuable for a variety of stakeholders. Central banks and policymakers can use GaR to better assess economic risks and make informed decisions on monetary policy, while financial regulators leverage it to monitor systemic risks and enhance financial stability. Investment managers and financial institutions benefit from GaR by improving risk management practices and optimizing portfolio strategies under different economic scenarios. Additionally, government agencies and analysts utilize GaR to anticipate the impact of economic shocks on public finances.

In this context, it becomes crucial to provide well-calibrated estimations of the quantiles of the future GDP growth distribution. Standard algorithms, such as quantile regression [1, 2] or quantile random forest, are commonly used for these purposes, but they can produce poorly calibrated results. To address these challenges, we adopt a conformal prediction framework, enhancing the reliability of quantile estimates by providing valid probabilistic guarantees regardless of the underlying distribution. Conformal prediction methods, originally designed to compute prediction intervals, have not yet been explored for estimating quantiles. In our work, we adapt and extend a conformal algorithm to create a conformalised quantile estimation procedure. Our method offers several advantages: it does not rely on stringent assumptions about the data, adapts dynamically to varying levels of uncertainty, and provides theoretical guarantees on the calibration of the estimated quantiles (under certain conditions). To the best of our knowledge, this is the first study to apply conformalisation to quantile estimation, as all the existing literature focuses on prediction intervals.

In addition to its macroeconomic applications (such as GaR), our conformalised quantile estimation procedure can be applied in other fields and in any context where precise quantile estimates are needed, such as in risk management. A notable example is the calculation of Value-at-Risk (VaR), a critical measure used in financial risk management. VaR estimates the maximum potential loss of an investment or portfolio over a specified time period, given a certain confidence level. It focuses on the lower tail of the loss distribution, where extreme, low-probability losses occur. Accurate VaR calculations are crucial for financial institutions to assess their exposure to risk and determine the necessary capital reserves to cover potential losses.

In the next sections, we will test our model with generated data that are common in financial economics practice, such as time series data with the presence of ill-behaved distributions, characterized by fat tails, and very high dimensional models, as well as a real-world case focused on the computation of GaR for the USA. By doing so, we aim to demonstrate how our procedure can enhance the calibration of quantiles estimations, providing a more comprehensive and reliable tool for stakeholders across the financial and economic sectors. Results show that the robustness of non-conformal algorithms in terms of calibration is questioned in several scenarios, while conformalised estimates are far more robust.

In detail, the thesis is structured as follows: Section 2 presents the methodology, introducing the Conformalized Quantile Regression algorithm and its modification and adaptation for the estimation of quantiles. Section 3 is dedicated to the simulation study with generated data. Section 4 is reserved for the real case study, estimating future GDP growth in the USA.

2. Methodology

Conformal Prediction (CP) is a powerful, model-agnostic framework for constructing prediction intervals that provide a guaranteed coverage probability, regardless of the underlying data distribution. Unlike traditional approaches that rely on specific distributional assumptions, CP leverages the concept of nonconformity scores to measure the "strangeness" of new observations relative to a given dataset. Given a dataset $\mathcal{D} = \{(X_t, Y_t)\}_{t=1}^{n}$, a nonconformity score function $E(x, y)$ is defined to quantify how unusual is each observation (X_t, Y_t).

In this manuscript, we focus on a specific CP method using a split conformal prediction approach, where the dataset is partitioned into two disjoint subsets: one for training the predictive model and another for calibration. The nonconformity scores are computed on the calibration set, and an empirical quantile of these scores is used to construct prediction intervals. We propose a novel modification to extend the applicability of this method from constructing prediction intervals to directly estimating quantiles. To the best of our knowledge, this is the first work in the literature that uses a conformal prediction framework to estimate quantiles directly rather than to construct prediction intervals. This enhancement broadens the utility of CP, making it suitable for applications where understanding the entire distribution is crucial.

2.1. Conformalized Quantile Regression

Conformalized Quantile Regression (CQR) is an algorithm introduced in Romano et al. [3]. It's an advanced statistical technique that combines conformal prediction methods with quantile regression to provide robust prediction intervals with valid coverage guarantees in finite samples.
The algorithm begins with a split of the dataset in two disjoint subsets, T_1 and T_2. T_1 is used to fit two conditional quantile functions $\hat{Q}(\alpha/2, x)$ and $\hat{Q}(1-\alpha/2, x)$, where α is the miscoverage rate (i.e. prediction intervals of level $1-\alpha$ will be computed). The algorithm is flexible regarding how these quantile functions are computed, popular choices may include quantile regression and quantile random forest. Using the second subset T_2, the non-conformity scores are computed as

$$E_t = \max\left\{\hat{Q}(\alpha/2, X_t) - Y_t,\, Y_t - \hat{Q}(1-\alpha/2, X_t)\right\}$$

for each $(X_t, Y_t) \in T_2$. These non-conformity scores are then used to adjust the previously computed quantile functions to produce a new prediction interval:

$$C(x) = \left[\hat{Q}(\alpha/2, x) - Q_E(1-\alpha),\, \hat{Q}(1-\alpha/2, x) + Q_E(1-\alpha)\right],$$

where $Q_E(1-\alpha)$ is the $(1-\alpha)(1+1/|T_2|)$-th empirical quantile of E_t.
As in [3], the following guarantees about coverage hold:

Theorem 2.1. *If (X_i, Y_i), $i = 1, \dots, n+1$ are exchangeable, then the prediction interval $C(X_{n+1})$ constructed by the split CQR algorithm satisfies:*

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

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

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \leq 1-\alpha + \frac{1}{|T_2|+1}. \tag{2}$$

Some modifications are needed in order to produce quantiles estimation with this algorithm. There are two different approaches we can take to achieve our goal. In the first method, we start by producing only one quantile function of level $1-\alpha$: $\hat{Q}(1-\alpha, x)$. It can be interpreted that the second quantile function is of level 0, resulting in $\hat{Q}(0, x) = -\infty$. Therefore, the computation of non-conformity scores is simply

$$E_t = \max\left\{-\infty,\, Y_t - \hat{Q}(1-\alpha, X_t)\right\} = Y_t - \hat{Q}(1-\alpha, X_t),$$

and the resulting estimate of the quantile of level $1-\alpha$ follows from the prediction interval:

$$C(x) = \left(-\infty,\, \hat{Q}(1-\alpha, x) + Q_E(1-\alpha)\right].$$

In our quantile setting, (1) becomes:

$$\mathbb{P}\{Y_{n+1} \leq \hat{Q}(1-\alpha, x) + Q_E(1-\alpha)\} \geq 1-\alpha.$$

The quantile estimate will exceed the next observation at least $1-\alpha$ % of the time. However, in the context of GaR, we would be more interested in a different inequality, one that allows for controlling the frequency of observations falling below a specified threshold. To address this need, we introduce a second procedure for quantile estimation. We again produce only one quantile function but of level α: $\hat{Q}(\alpha, x)$. The other quantile function can be seen as $\hat{Q}(1, x) = \infty$:

$$E_t = \max\left\{\hat{Q}(\alpha, X_t) - Y_t,\, -\infty\right\} = \hat{Q}(\alpha, X_t) - Y_t,$$

and consequently

$$C(x) = \left[\hat{Q}(\alpha, x) - Q_E(1-\alpha),\, +\infty\right).$$

And now, according to (1), we conclude

$$\mathbb{P}\{Y_{n+1} \geq \hat{Q}(\alpha, x) - Q_E(1-\alpha)\} \geq 1 - \alpha,$$

correspondingly,

$$\mathbb{P}\{Y_{n+1} \leq \hat{Q}(\alpha, x) - Q_E(1-\alpha)\} \leq \alpha. \tag{3}$$

This inequality is exactly what is needed to control the frequency of extreme adverse outcomes in GDP growth in the future. This formulation enables for example central banks to establish a predefined threshold for GDP growth (e.g. 2%), and to quantify the maximum probability of experiencing an adverse or worse outcome. Conversely, the inequality can also be utilized to identify the specific GDP growth value associated with a given risk level α. This dual application is crucial for policy formulation, as it provides a robust framework for managing economic risks by aligning GDP growth projections with acceptable risk thresholds.

In the preceding step, we assume the continuity of $\hat{Q}(\alpha, x)$. For a fixed x, $\hat{Q}(\alpha, x)$ corresponds to the inverse cumulative distribution function of y given x, which is guaranteed to be continuous when the model includes a continuous error term, as is the case in all simulations conducted in this and most studies. For a fixed α, the continuity of $\hat{Q}(\alpha, x)$ with respect to x depends on the smoothness of the relationship between y and x. In all our simulations, and in most practical scenarios, this condition is satisfied. Due to its useful coverage guarantee, this second procedure is employed throughout this thesis.

Lastly, it is important to highlight the main limitation of this property: the exchangeability assumption. The assumption is naturally violated in time series data. In the absence of exchangeability, the result stated in (3) does not hold. Since exchangeability is a crucial assumption in many CP algorithms, new techniques and methods are being developed to address this limitation. However, these techniques won't be explored in this manuscript; instead, we focus on assessing whether this property holds, even approximately, in both simulated and real data when data are not exchangeable.

3. Simulation Study

In order to stress and test the robustness of standard methods for quantile estimation, as well as to assess the effectiveness of conformal methods, we conduct a comprehensive simulation study, using synthetic data. Since the main focus of our work is the economic and financial sector, the simulations will involve time series with autoregressive components, with fat-tailed distributions and/or high dimensionality. Simulations are all built in a similar fashion. A time series $\{(X_i, Y_i)\}_{i=1}^{n+100}$ is generated for $n \in \{98, 198, 998\}$. The first n observations will be used to train (and calibrate) the models, while the last 100 observations will be used to evaluate the models' performance. Performances are then averaged over 100 iterations, which corresponds to generating 100 different time series as just described. In each iteration, the models are trained to predict the entire distribution of the target variable Y_i through the prediction of 20 quantile levels $L_i \in \{0.01, 0.05, 0.10, 0, 15, ..., 0.95\}$. For each quantile level in each iteration the unconditional empirical coverage $\hat{P}(\hat{Q}_i)$ is calculated, which is simply the proportion of test data that falls below the empirical quantile value $\hat{Q}_i$ estimated by our model. In a perfectly calibrated model, the coverages are equal to their associated quantile level: the estimated quantile of level 0.9 should be greater than 90% of test data.

Four different models are implemented and tested: Quantile Regression (QR), Quantile Random Forest (QRF) and their CQR versions. QR and QRF offer complementary strengths and differ in their

underlying assumptions, interpretability, and flexibility. QR is a parametric technique that allows for the estimation of specific quantiles of the dependent variable, given a set of predictors. It assumes a linear relationship between the predictors and the quantiles of the response. Its interpretability makes it a popular tool in the economic field. On the other hand, QRF extends the traditional Random Forest algorithm to provide non-parametric estimates of conditional quantiles. Unlike QR, QRF does not assume a linear relationship between predictors and the response, allowing it to capture complex, non-linear interactions within the data. Given these differences, employing both Quantile Regression and Quantile Random Forest allows for a comprehensive comparison of linear versus non-linear effects, interpretability versus flexibility, and parametric versus non-parametric methods. QR and QRF then will also be used to fit a conditional quantile function $\hat{Q}(\alpha, x)$ to perform the CQR algorithm, obtaining the CQR QR and the CQR QRF models.
Performances are evaluated estimating the mean absolute error between the estimated coverages and the quantile level:

$$\text{MAE} := \frac{1}{2000} \sum_{i=1}^{20} \sum_{j=1}^{100} \left| \hat{P}(\hat{Q}_{ij}) - L_i \right|. \tag{4}$$

The quantity is averaged across all the iterations and quantiles. Perfectly calibrated models have $\text{MAE} = 0$ and a low value of MAE is associated with a well calibrated model. However, MAE alone is not sufficient to fully evaluate a model's performance, as it does not distinguish whether the error is due to $\hat{P}(\hat{Q}_i) > L_i$ or to a conservative coverage $\hat{P}(\hat{Q}_i) < L_i$. To address this limitation, we construct a 95% binomial proportion confidence interval, calculated using the Wilson score, centred around the empirical coverage $\hat{P}(\hat{Q}_i)$. This procedure allows us to construct a confidence interval for the real quantile level L_i. In each of the simulation sections below, a table summarises these results. These tables provide the percentage of quantile levels L_i that fall within, below, or above the confidence interval constructed around their empirical coverage. "Within CI" refers to the percentage of quantile levels L_i that fall within this confidence interval, and we consider these estimates as perfectly calibrated. "Below CI" and "Above CI" represent the percentages of quantile levels L_i that fall below or above, respectively, the constructed confidence interval. If (3) holds, we expect for CQR algorithms a small % in "Below CI", lower when compared to their respective standard algorithms. Additionally, we include calibration curves to provide an immediate and insightful visual representation. Two different data generating processes are explored in the following subsections, while a third is available in Appendix B.1.

3.1. AR(2) with Cauchy distributed errors

As a first simulation, the following Data Generating Process (DGP) is implemented:

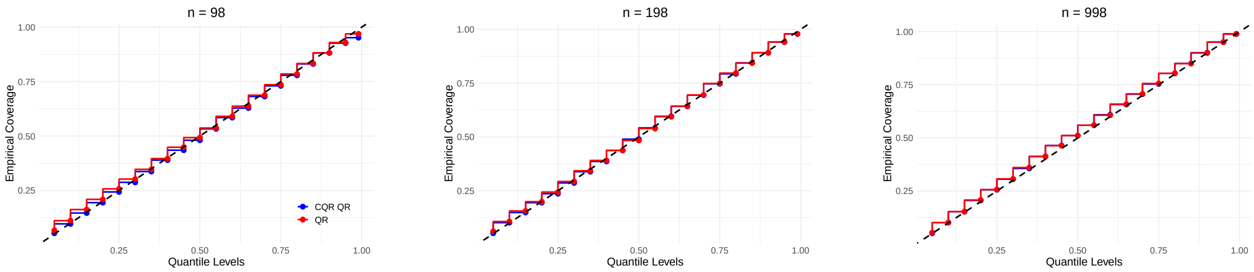
$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$$

where

$$\phi_1 = 0.5, \quad \phi_2 = -0.2, \quad \epsilon_t \sim \text{Cauchy}(0, 1).$$

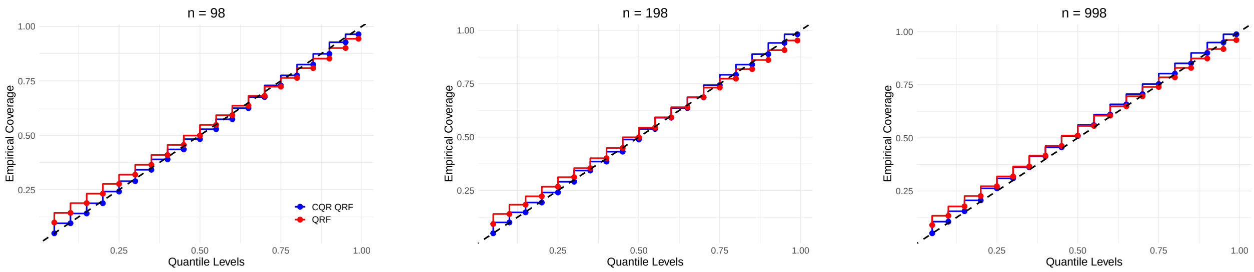
The introduction of a Cauchy error term significantly impacts the characteristics of an AR(2) model. Specifically, it renders the moments (such as mean, variance, skewness) undefined, thereby posing substantial challenges to standard prediction models. The heavy-tailed nature of the Cauchy distribution, characterized by frequent outliers, makes it particularly suitable for modeling financial data. This is especially relevant in the field of risk management (GaR, for example), where accurately capturing the probability of extreme losses induced by outliers is crucial.
The results presented in Table 1 show that QR reaches a satisfactory calibration even without conformalisation, while the latter significantly improves the calibration for QRF. More importantly, the table shows that property (3) is valid across almost all quantile levels for CQR algorithms. This was not guaranteed because the exchangeability hypothesis is not met in this model. Table 2 further confirms an improvement in performance for CQR QRF, showing a decrease in MAE, while the performances of QR and CQR QR are overall similar. The calibration curves in Figures 1 and 2 highlight which

quantile levels L_i are better calibrated. As previously noted, QR estimations are accurate even without conformalisation. In contrast, the QRF algorithm yields poorly calibrated estimates, particularly at extreme quantile levels, such as 0.05 or 0.95. Since GaR specifically targets the lowest quantile of the GDP growth distribution, the use of conformalisation techniques becomes even more critical.



(a) (b) (c)

Figure 1: Calibration curves for QR and CQR QR for the 3 values of n.



(a) (b) (c)

Figure 2: Calibration curves for QRF and CQR QRF for the 3 values of n.

	Within CI	**Below CI**	**Above CI**
QR	58.33%	16.67%	25.00%
QRF	25.00%	38.34%	36.66%
CQR QR	55.00%	5.00%	40.00%
CQR QRF	40.00%	10.00%	50.00%

Table 1: The table contains the percentage of quantile levels L_i which falls within, above or below the 95% binomial proportion confidence interval calculated using Wilson Score, built around the empirical coverage $\hat{P}(\hat{Q}_i)$. Results are calculated over the 3 values of n.

MAE	98	198	998
QR	0.011	0.008	0.006
QRF	0.026	0.021	0.018
CQR QR	0.015	0.008	0.005
CQR QRF	0.017	0.009	0.006

Table 2: MAE (4) for the 4 models and the 3 values of n

3.2. AR(2) with Exogenous variables

As a second simulation, the following DGP is implemented:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \boldsymbol{\beta}^\top \mathbf{X}_t + \epsilon_t$$

where

$$\phi_1 = 0.5, \quad \phi_2 = -0.2, \quad \epsilon_t \sim t_2,$$

$$\beta_i \sim U(0,1) \quad \text{for each} \quad i = 1, \ldots, p, \quad p/n \in \{0.1, 0.2, 0.3, 0.4\}$$

$$\mathbf{X}_t \sim \mathcal{N}_p(\mathbf{m}, \mathbf{v})$$

$$\mathbf{m} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}), \quad \mathbf{v} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \end{bmatrix}, \quad \sigma_i \sim U(0, 10)$$

In this second simulation, we again use a fat-tailed error in the DGP, this time a t-Student distribution with two degrees of freedom. This error has heavy tails similar to the Cauchy distribution, but unlike the Cauchy, it has a finite mean, although its variance remains undefined. We also introduce a term related to exogenous variables to test the models' calibration when the number of covariates is high. In particular, we will train models for 4 different ratios of the number of covariates and observations p/n : $0.1, 0.2, 0.3, 0.4$. Given the complexity of 21st-century economies, identifying the drivers of growth, even retrospectively, is a challenging task that often results in datasets containing a vast number of macroeconomic and financial variables. Although efforts have been made to consolidate many of these variables into a single index, such as the National Financial Conditions Index (NFCI) [4], the ability to manage a large number of covariates is a crucial feature for a model designed for GaR analysis.

In this simulation, even more than in the first, it is clear that conformalisation significantly enhances model calibration. From the MAE in Table 4 and the calibration curves in Figure 3 and 4, we note that QR provides very poor quantile estimations, especially for extreme quantile levels, and calibration worsens as the ratio p/n increases.

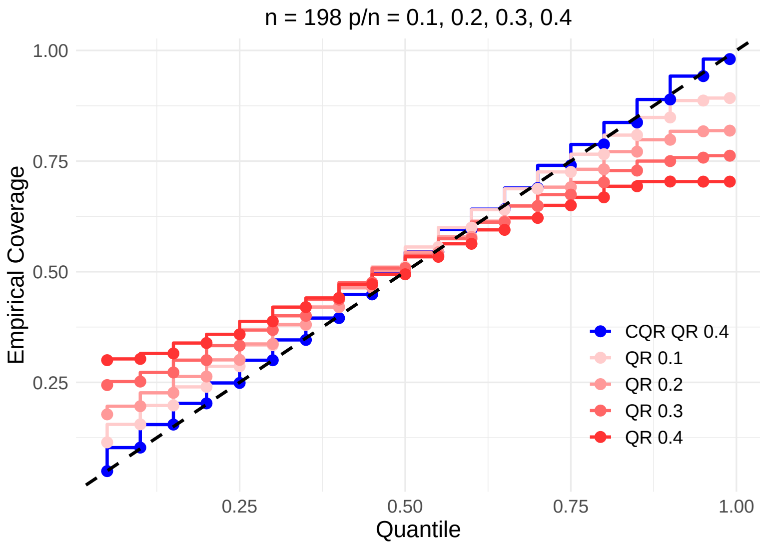
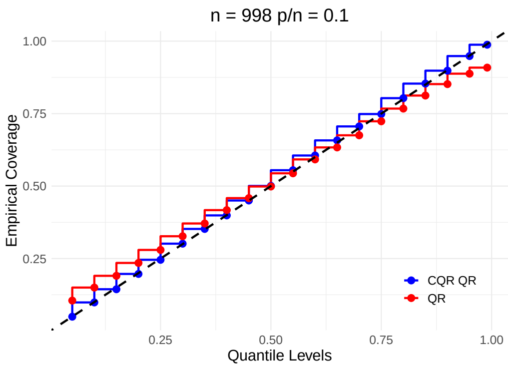
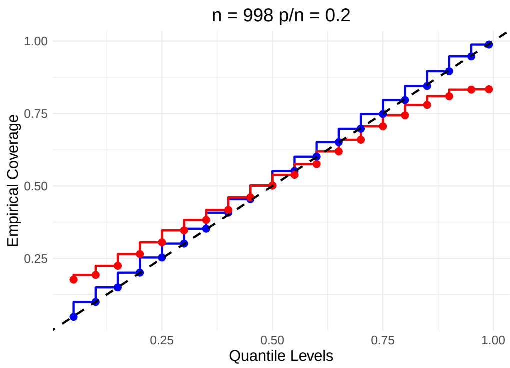


Figure 3: This plot is intended to provide a quick and intuitive visualization of the calibration of QR models when the ratio p/n increases, compared to a single CQR QR model with $p/n = 0.4$, that is very well calibrated. The individual plots can be seen below in Figure 4 and in Appendix A.1.

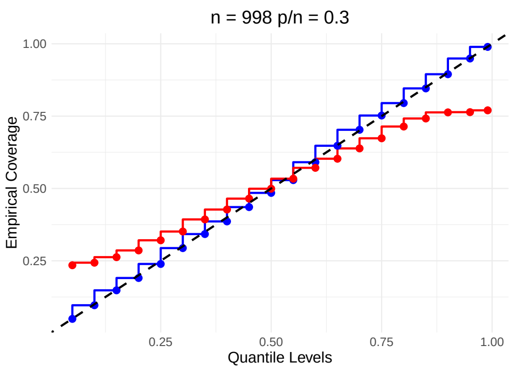
QRF estimations are more precise than QR estimations, but they nevertheless benefit from conformalisation. CQR methods are also better calibrated for an increasing value of n, while improvement for QR and QRF is either less evident or entirely questionable. The guarantee (3) holds once again, and even more convincingly, with all the positive consequences that have already been discussed (Table 3).



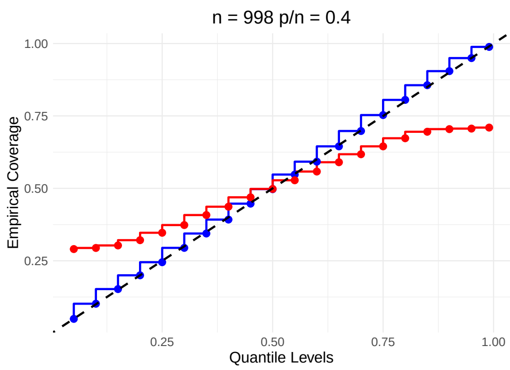
(a)



(b)



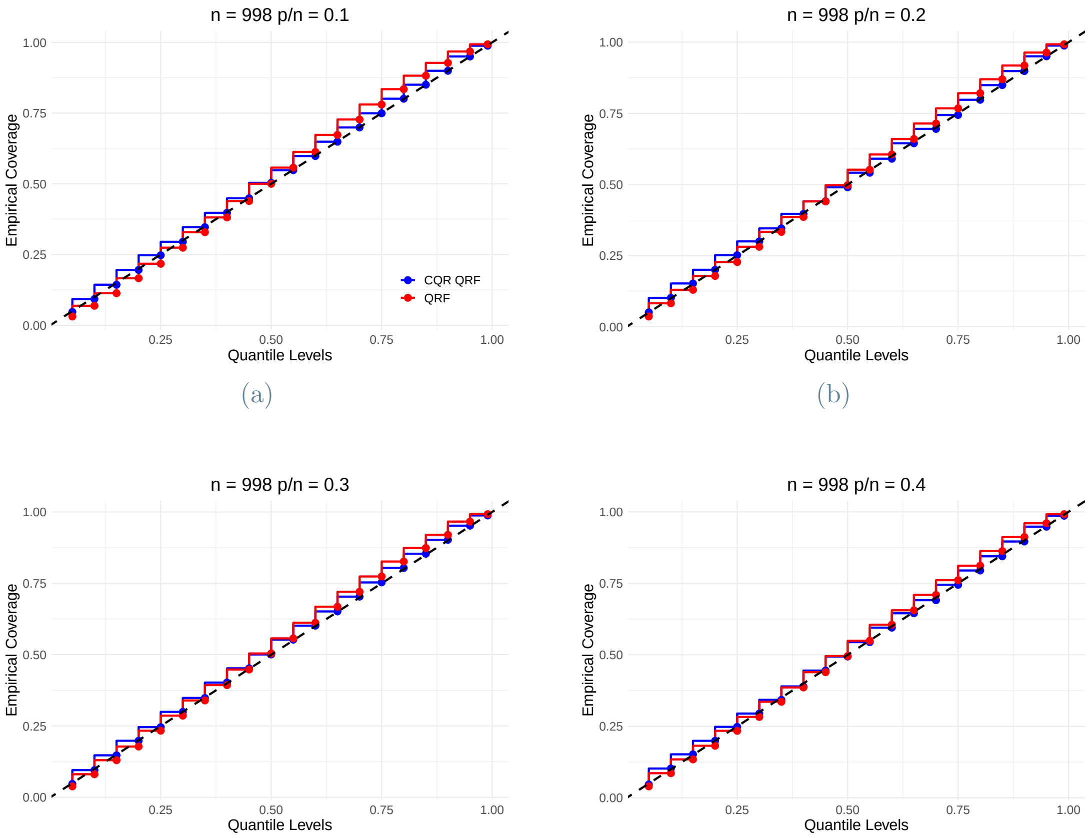
(c)



(d)

Figure 4: Calibration curves for QR and CQR QR for $n = 998$ and p/n values 0.1, 0.2, 0.3, and 0.4. Plots are similar for the other two values on n and can be seen in Appendix A.1

.



(a) (b) (c) (d)

Figure 5: Calibration curves for QRF and CQR QRF for $n = 998$ and p/n values 0.1, 0.2, 0.3, and 0.4. Plots are similar for the other two values on n and can be seen in Appendix A.1

.

	Within CI	**Below CI**	**Above CI**
QR	7.92%	43.33%	48.75%
QRF	18.75%	41.25%	40.00%
CQR QR	59.16%	3.75%	37.08%
CQR QRF	70.00%	0.83%	29.17%

Table 3: The table contains the percentage of quantile levels L_i which falls within, below or above the 95% binomial proportion confidence interval calculated using Wilson Score, built around the empirical coverage $\hat{P}(\hat{Q}_i)$. Results are calculated over the 3 values of n and the 4 values of p/n.

MAE	98	198	998
QR	0.08	0.077	0.073
QRF	0.019	0.02	0.015
CQR QR	0.012	0.007	0.004
CQR QRF	0.013	0.005	0.003

Table 4: MAE (4) for the 4 models and the 3 values of n, averaged on the 4 values of p/n.

It is interesting to note that due to the design of CQR algorithm, only half of the training set is actually used to train a quantile prediction model, while the other half is reserved for calibration. Thus, in practice, CQR QR and CQR QRF models are trained on a effective ratio up to 0.8. This fact further highlights the effectiveness of the conformalisation step is in such a scenario. Notably, all four models under consideration demonstrate resilience to misspecification. There is no significant difference in performance when the DGP introduced in this section is modeled with incorrect autoregressive terms, such as AR(1) and AR(3). Real-world models are often imperfectly specified due to the complexity of economic systems and data limitations, especially when predicting aggregate data like GDP. By intentionally testing models that deviate from the true underlying process, we aim to evaluate the sensitivity of quantile forecasts to such misspecifications. Detailed results can be found in Appendix A.2.

4. Real Case Study

In this section, we analyze the performance of conformalised and non-conformalised methods when applied to real-world data. The dataset consists of quarterly values of NFCI [4] and GDP growth in the USA from 1973 to 2016, derived from Adrian et al. [2]. We use their framework and findings as a benchmark for assessing our models' performances. Adrian et al.'s research centers on predicting GDP growth using an autoregressive model that incorporates the NFCI, which captures both financial and macroeconomic conditions. One of their key findings is that a decline in the NFCI (indicating deteriorating economic conditions) not only shifts the mean of the GDP growth distribution to the left, as one might expect, but also increases the leftward skewness of the distribution. This heightened skewness suggests a greater probability of extreme negative GDP growth outcomes. Interestingly, when the NFCI improves, the same degree of rightward skewness is not observed. This asymmetry underscores how downside risks to GDP growth tend to intensify more than upside risks over time. Scenarios like this, where the behavior of the target variable distribution, or specific quantiles within it, is of interest, provide an ideal context to assess the effectiveness of conformalisation. In these cases, calibration becomes a critical metric, ensuring that the predicted probabilities across all quantiles accurately reflect the true distribution of future GDP growth.
It is important to note a methodological difference between how Adrian et al. measured calibration and how it was measured in our simulation study, even though this difference does not lead to any practical discrepancies in the results. In our simulation study, we calculate the unconditional empirical coverage $\hat{P}(\hat{Q}_i)$, which is simply the proportion of test data that falls below the empirical quantile value $\hat{Q}_i$ estimated by our model. In contrast, Adrian et al. employed the Probability Integral Transform (PIT). Specifically, if X is a random variable with a continuous cumulative distribution function $F_X(x)$, then the PIT is defined as:

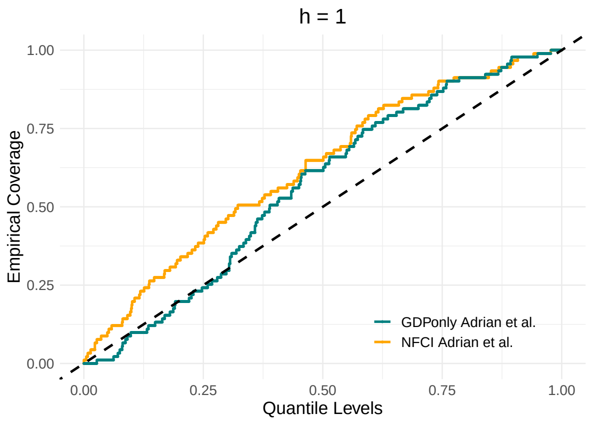
$$U = F_X(X)$$

Here, U is a new random variable that, under the assumption that X follows the distribution described by F_X, will be uniformly distributed on the interval $[0, 1]$. When applying the PIT to a set of observations, the transformed values resulting should follow a uniform distribution if the assumed model (represented by F_X) is correct. Another key difference from our approach in our earlier simulation study is that Adrian et al. retrain their model at each time step, incorporating new information to predict the next observation. To ensure the most direct comparison between the performance of our models and those in the original study, we adopt both the PIT approach and the procedure of retraining the model at each time step in this section.

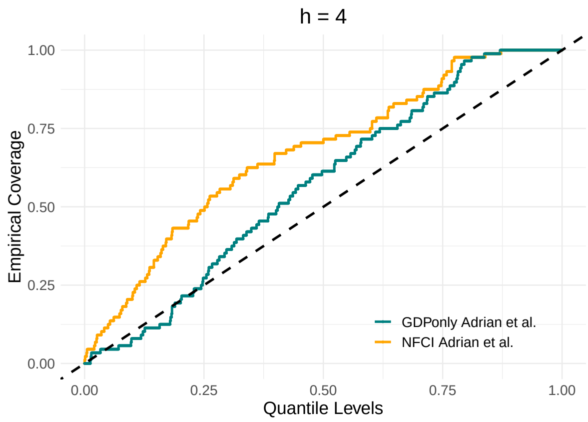
4.1. Quantile Regression

We begin by focusing on QR model, which is the same model selected by Adrian et al. in their study. In the final step of their estimation process, Adrian et al. fit a parametric distribution to the estimated quantiles to construct a continuous distribution of future GDP growth. However, because a final smoothing procedure would compromise the properties of CQR methods, we don't implement any smoothing but estimate a much finer grid of quantiles instead. An important consequence is that models which interpolate a continuous distribution tend to be better calibrated at the tails of the distribution compared to models that rely solely on estimating a large number of discrete quantiles without smoothing. This smoothness allows the model to provide more accurate and realistic probability estimates for extreme events, specifically at our two most extreme quantiles: below 0.01 and above 0.99.

In Figure 6, we present the calibration curves obtained by Adrian et al. The values $h = 1$ and $h = 4$ correspond to one-quarter-ahead and one-year-ahead estimates of GDP growth, respectively. These estimates are generated by QR models that use either the NFCI combined with the previous GDP growth value as predictors, or only the previous GDP growth value as the sole predictor. As shown, there is room for improving calibration. Our attention will focus only on the model using NFCI.



(a)

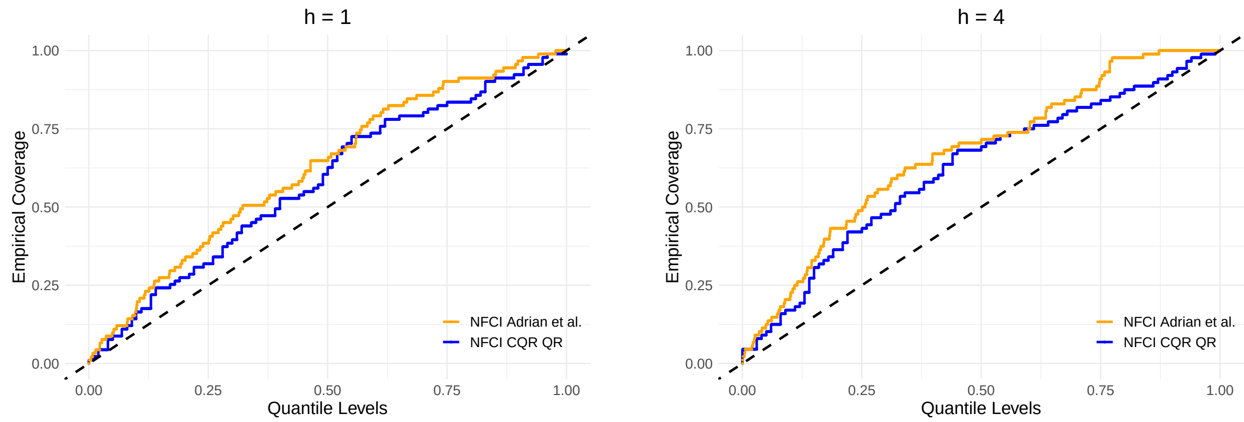


(b)

Figure 6: The calibration curves are from QR models that use either the NFCI combined with the previous GDP growth value as predictors, or only the previous GDP growth value as the sole predictor. The values $h = 1$ and $h = 4$ correspond to one-quarter-ahead and one-year-ahead estimates of GDP growth, respectively.
The same plot can be found in Panel C and Panel D in Figure 11 in Adrian et al. [2]

.

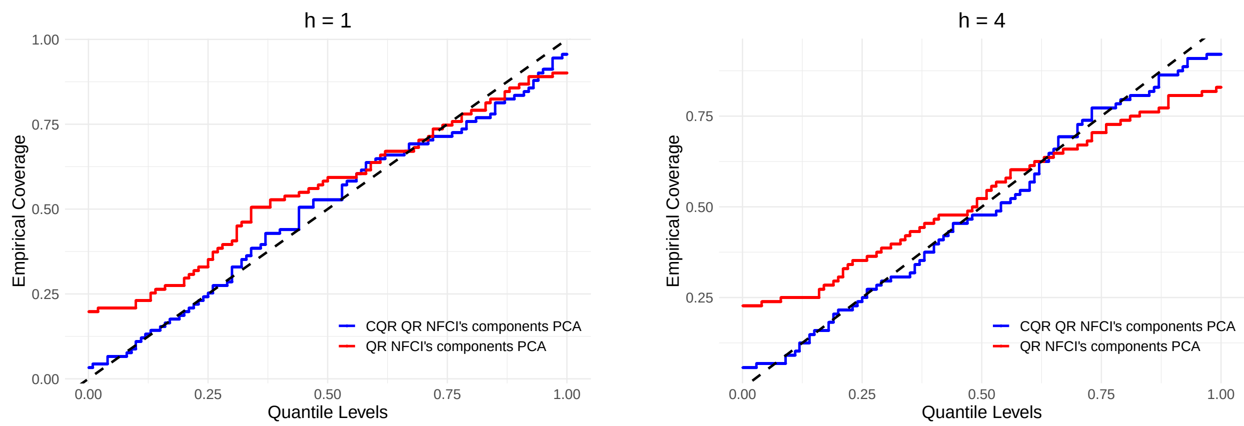
To begin, we immediately apply our CQR algorithm and compare it to the version used by Adrian et al. The results show that CQR QR model provides a noticeable, though not particularly significant, improvement in calibration (Figure 7).



(a) (b)

Figure 7: NFCI Adrian et al. refers to the original work. NFCI CQR QR is our estimate applying the CQR QR algorithm to the same data. The values $h = 1$ and $h = 4$ correspond to one-quarter-ahead and one-year-ahead estimates of GDP growth, respectively.

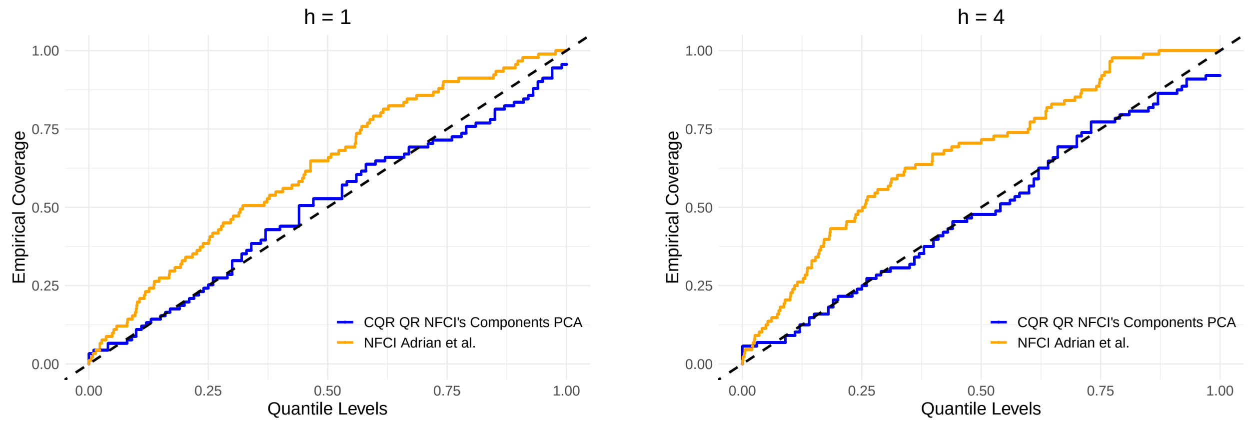
Simulations suggest that the CQR algorithm performs very well when estimating with a significant number of covariates. The NFCI is an index that tracks financial markets by aggregating around 100 different data series and indexes. Given how effective CQR QR is at handling a large number of covariates, and considering that using individual components can provide more detailed information for predicting future GDP growth than a single aggregated index, we have removed the NFCI from the model and have instead used its individual components. Since the number of covariates now exceeds the length of the time series, we have performed a Principal Component Analysis (PCA), retaining the components that explain 90% of the total variance. The CQR algorithm is now much better calibrated (Figure 8). We also observe that the calibration curves for the QR model tend to flatten at the extreme quantiles, where calibration is poor. A similar issue was observed during the simulation with exogenous variables (Figure 3).



(a) (b)

Figure 8: PCA components of NFCI are used in QR and CQR QR models as predictors for GDP growth. The values $h = 1$ and $h = 4$ correspond to one-quarter-ahead and one-year-ahead estimates of GDP growth, respectively.

Lastly, in Figure 9 we compare the calibration of the original model by Adrian et al. and that of the CQR QR model with the components of NFCI.



(a) (b)

Figure 9: Comparing curves from Figure 6 and 8

.

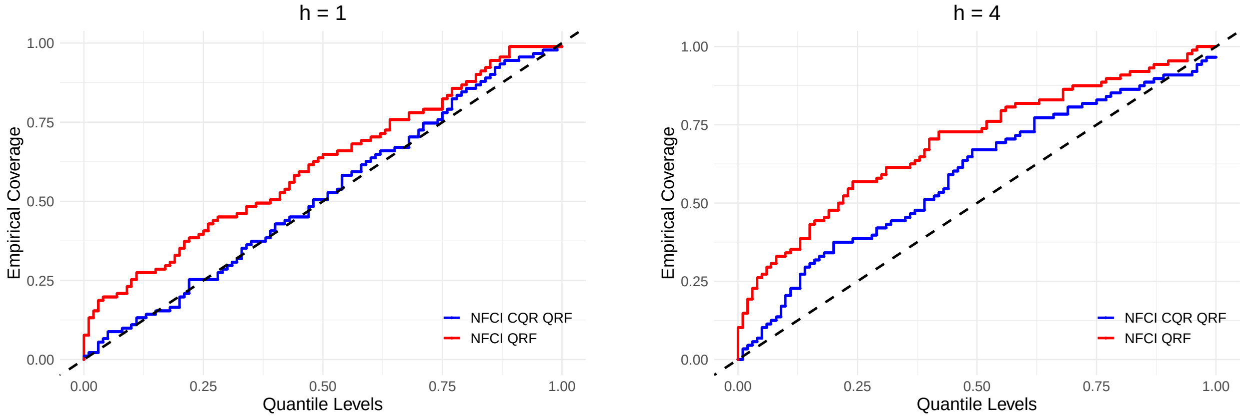
By also examining the MAE results in Table 5, we can draw several conclusions. Using the NFCI components instead of the aggregated index enables the QR and CQR QR models to better capture the distribution of future GDP growth. However, although the QR model reduces its MAE when using NFCI components instead of NFCI itself, its inability to correctly calibrate the extreme quantiles highlights the importance of the calibration step.

MAE	**QR**	**CQR QR**
NFCI h = 1	0.125	0.075
NFCI h = 4	0.177	0.115
PCA on NFCI components h = 1	0.075	0.028
PCA on NFCI components h = 4	0.079	0.023

Table 5: MAE (4) of all the models and time steps h. NFCI QR is the original work of Adrian et al. [2]

4.2. Quantile Random Forest

We now switch our focus to QRF and CQR QRF. Since the original work by Adrian et al. does not include QRF, we constructed both QRF and CQR QRF models to predict future GDP growth using the NFCI and past growth values (Figure 9). Our findings indicate that the conformalisation procedure significantly improves calibration, particularly for $h = 1$.

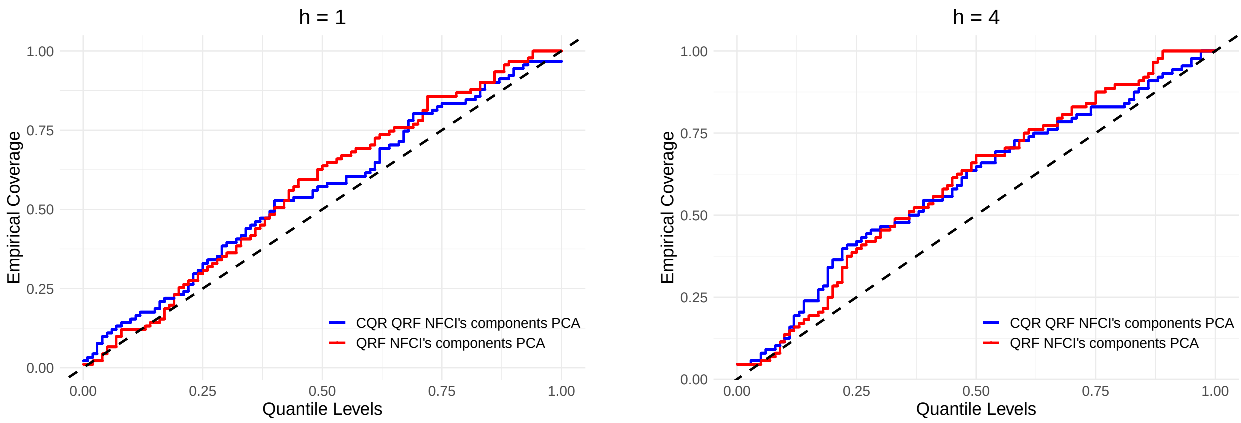


(a) (b)

Figure 10: QRF and CQR QRF calibration curves using NFCI combined with the previous GDP growth value as predictors. The values $h = 1$ and $h = 4$ correspond to one-quarter-ahead and one-year-ahead estimates of GDP growth, respectively.

.

We proceed as detailed in the previous subsection, removing NFCI and using its components. Based on our simulation study, we anticipated the calibration of CQR QRF would not be significantly better than that of QRF when disaggregating NFCI into its components, as QRF alone is already effective at handling a significant number of covariates (Figure 5). The results of QRF and CQR QRF with NFCI divided into its components are presented in Figure 11. CQR only slighty improves the performances of the standard method, and the two models are largely equivalent. Furthermore, the benefits of using NFCI's components are less evident this time; for example, at $h = 1$, using the aggregated NFCI itself appears to be preferable.



(a) (b)

Figure 11: The NFCI is broken down into its components, a PCA is performed, retaining the principal components capturing 90% of total variance. Then they are used in QRF and CQR QRF models as predictors for GDP growth.
The values $h = 1$ and $h = 4$ correspond to one-quarter-ahead and one-year-ahead estimates of GDP growth, respectively.

.

MAE	QRF	CQR QRF
NFCI h = 1	0.104	0.017
NFCI h = 4	0.192	0.091
PCA on NFCI components h = 1	0.064	0.051
PCA on NFCI components h = 4	0.094	0.086

Table 6: MAE (4) of all the models and time steps h.

5. Conclusions

In this thesis, we explored the application of CQR in the context of economic forecasting, particularly for GaR analysis. Our primary goal was to assess the empirical performance of conformalised versus non-conformalised quantile prediction models and to determine whether conformalisation enhances the reliability and robustness of quantile estimates.

Our results from extensive simulation studies and the real-world case study provide several key insights. Firstly, we found that the QR and QRF models offer complementary strengths in different scenarios. When one model performs poorly, the other often compensates by providing better results. For example, QR performed better than QRF in scenarios with data drawn from distributions without finite moments, such as the Cauchy distribution. In contrast, QRF shows superior performance in high-dimensional environments. This complementarity suggests that the combined use of both methods, possibly through model averaging or ensemble techniques, could provide a more robust approach to quantile estimation across diverse data contexts, even without conformalisation. Moreover, the simulation study demonstrated that CQR methods offer substantial, consistent improvements in calibration compared to standard QR and QRF models. Even in contexts where the standard models performed well, their CQR counterpart consistently matched or exceeded their calibration performance. The only exception is nearly unit root time series, where all models—conformalised or not—struggled to provide reliable estimates. In such cases, the inherent persistence and (near-)random-walk behavior of the series led to poor model performance, highlighting a limitation of these quantile estimation techniques in handling processes with strong autocorrelation and persistence. Also in the real case study, which focused on predicting US GDP growth using financial and macroeconomic data, CQR methods demonstrated their value by improving the accuracy and calibration. Our analysis also highlighted that when calibration is poor, the primary issues often arise at extreme quantile levels, which are of the greatest interest for GaR analysis. Even models that exhibit a relatively low MAE can still be problematic if they fail to accurately estimate these extreme quantiles. Since these quantiles are crucial for assessing the risk of adverse economic outcomes, poor calibration in these areas undermines the reliability of the model's predictions and limits its utility for GaR. However, CQR methods demonstrate a clear advantage in this regard by consistently providing better calibration at these critical extreme quantiles. In addition, our results demonstrate that both conformalised and non-conformalised models are remarkably robust to misspecifications in the data generating process, such as incorrect autoregressive terms or variations in error distributions. This resilience is particularly relevant in real-world applications where the true underlying structure of the data is often unknown, reinforcing the practical utility of these methods in economic forecasting and risk management. More research is needed to fully understand the extent of this robustness across various settings and to explore potential losses in model performance under different forms of misspecification.

A limitation of our findings arises from the assumption of exchangeability, which is not only crucial for general conformal prediction methods but is also central to the coverage guarantee (3) of our specific approach. In the simulated scenarios, CQR methods exhibit calibration properties that closely align with theoretical coverage guarantees. This alignment is particularly important because it directly

supports the ability to control the probability of extreme adverse outcomes in future GDP growth. By ensuring that the probability of exceeding a specified threshold is kept within a desired risk limit, our approach enables economic policymakers, such as central banks, to set and manage risk levels effectively. However, in the real-world case study, there is no evidence that the coverage property holds, as indicated by calibration curves that frequently lie above the diagonal line, reflecting a tendency to overestimate coverage rather than underestimate it. The time dependence inherent in the GDP growth rate series significantly violates the exchangeability assumption, compromising the coverage guarantee. Despite this, the performance of CQR methods remained strong, consinstently exceeding that of the standard methods. Addressing the challenges posed by non-exchangeable data or time series in conformal prediction is an area already explored in the literature, as in Barber et al. [5]
Overall, this thesis contributes to the literature by providing the first application of conformal prediction methods to quantile estimation, offering a novel perspective on how these methods can be adapted and utilized for economic and financial forecasting. Our findings underscore the potential of CQR in enhancing the robustness of risk assessments. Future research could explore methods to overcome the limitations posed by the exchangeability assumption in time-series data, as well as adapt other conformal prediction techniques for quantile estimation. Additionally, the proposed method could be applied to other economic and financial contexts, such as Value-at-Risk estimation, and in any setting where accurate quantile estimation is crucial.

References

[1] International Monetary Fund. Global financial stability report: Is growth at risk? Technical report, International Monetary Fund, Washington, DC, 2017.

[2] Tobias Adrian, Nina Boyarchenko, and Domenico Giannone. Vulnerable Growth. *American Economic Review*, 2019.

[3] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized Quantile Regression. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

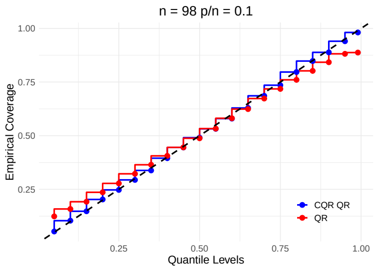
[4] Federal Reserve Bank of Chicago. National Financial Conditions Index (NFCI), 2024.

[5] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal Prediction Beyond Exchangeability. *The Annals of Statistics*, 2023.

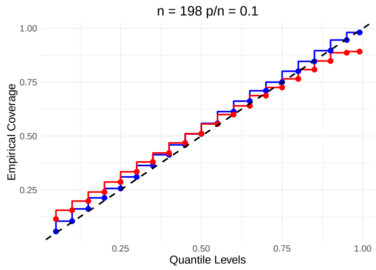
[6] James H. Stock and Mark W. Watson. Testing for common trends. *Journal of the American Statistical Association*, 1988.

A. Appendix A

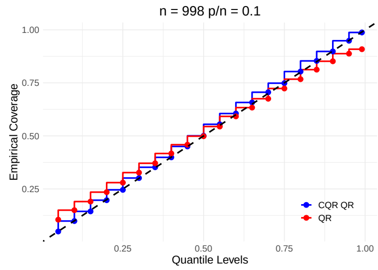
A.1. Calibration curves for the AR(2) Exogeneus model



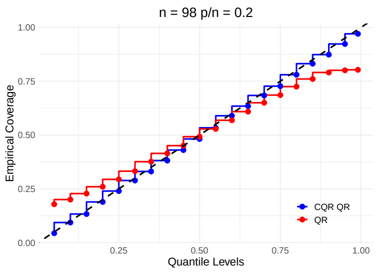
(a)



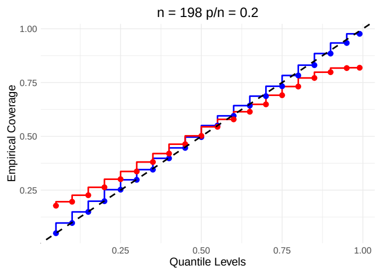
(b)



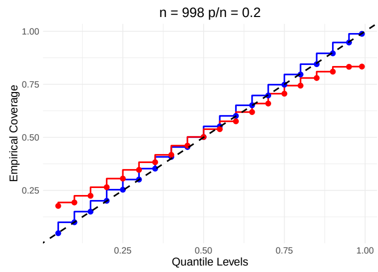
(c)



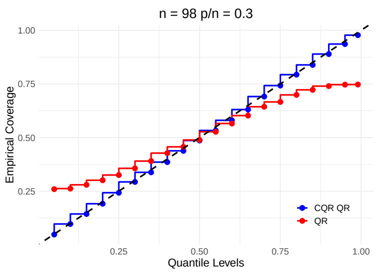
(d)



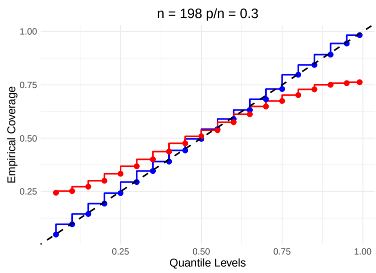
(e)



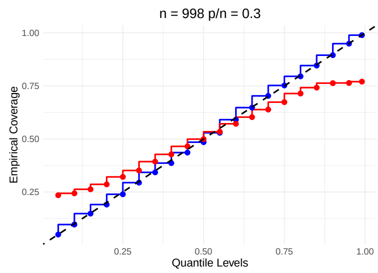
(f)



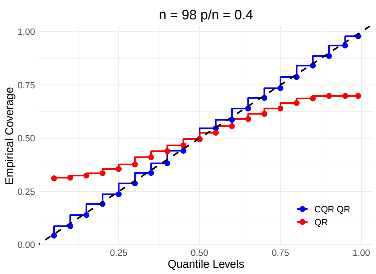
(g)



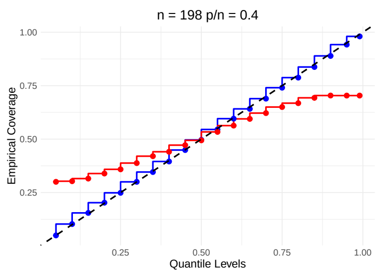
(h)



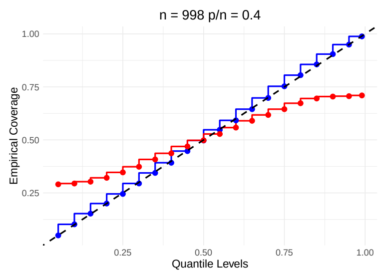
(i)



(j)

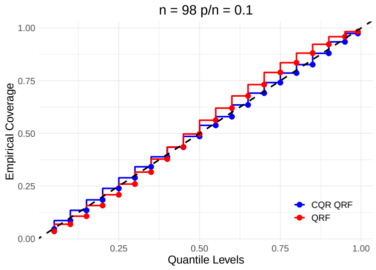


(k)

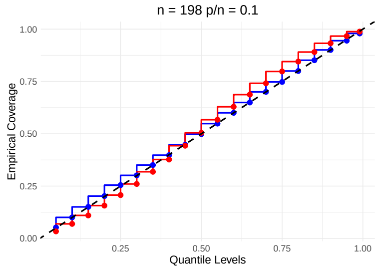


(l)

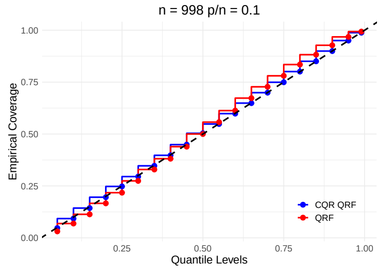
Figure 12: Calibration curves for QR and CQR QR for the 3 values of n and $p/n = 0.1$ (first row), 0.2 (second row), 0.3 (third row), and 0.4 (fourth row).



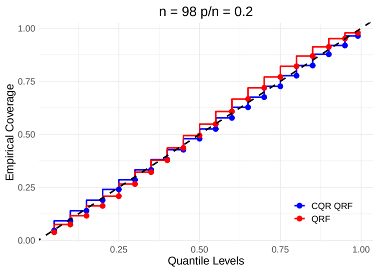
(a)



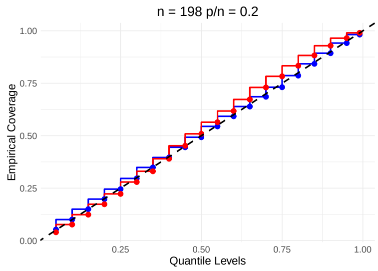
(b)



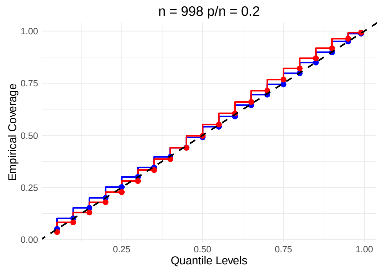
(c)



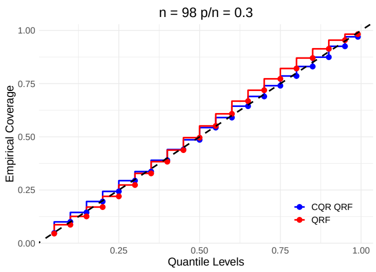
(d)



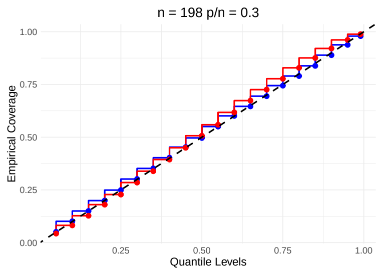
(e)



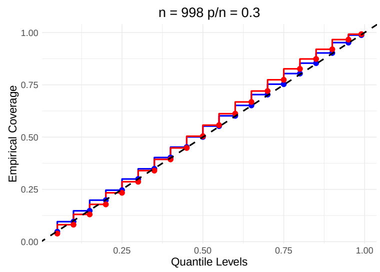
(f)



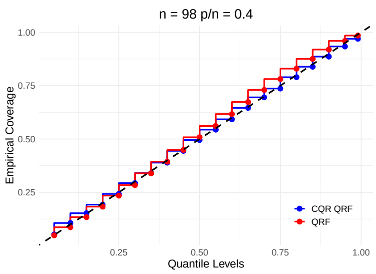
(g)



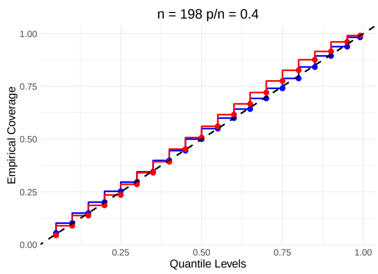
(h)



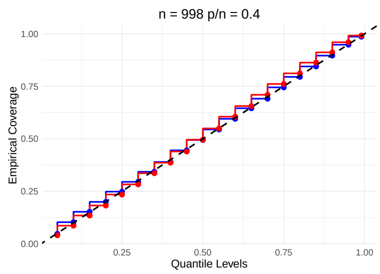
(i)



(j)



(k)



(l)

Figure 13: Calibration curves for QRF and CQR QRF for the 3 values of n and $p/n = 0.1$ (first row), 0.2 (second row), 0.3 (third row), and 0.4 (fourth row).

A.2. Resilience to Misspecification for the AR(2) Exogeneus model

We assess the predictive accuracy and robustness of each model by fitting it to data using three different autoregressive specifications: AR(1), AR(2), and AR(3), while keeping the exogenous component correctly specified. By doing so, we intentionally introduce misspecification to determine how well each model can adapt to errors in the assumed data structure and maintain reliable quantile estimates. The AR(1) and AR(3) models represent misspecified alternatives that either simplify or overfit the true AR(2) process, while the correctly specified AR(2) model serves as a benchmark.

In Table 7, all four models exhibit similar performance across the different autoregressive specifications. This consistency indicates that the models' accuracy and calibration are relatively unaffected by misspecification of the autoregressive component.

MAE	AR1	AR2	AR3
QR	0.074	0.077	0.078
QRF	0.018	0.018	0.018
CQR QR	0.006	0.008	0.007
CQR QRF	0.007	0.007	0.007

Table 7: MAE (4) for the 4 models for AR(1), AR(2) and AR(3), averaged on the 4 values of p/n and on the 3 values of n.

Given the resilience of AR(1) and AR(3) models, we further test by modifying the DGP error using $\epsilon_t \sim \mathcal{N}(0,1)$ instead of $\epsilon_t \sim t_2$. A possibility is that a fat-tailed error masks the effect of the autoregressive component, making its impact on the target variable neglectable. Even with this additional modification, the models' performances are indistinguishable. (Table 8)

MAE	AR1	AR2	AR3
QR	0.075	0.077	0.079
QRF	0.019	0.019	0.019
CQR QR	0.007	0.008	0.008
CQR QRF	0.007	0.007	0.007

Table 8: MAE (4) of the new DGP with normal error, for the 4 models for AR(1), AR(2) and AR(3), averaged on the 4 values of p/n and on the 3 values of n.

B. Appendix B

B.1. AR(1) with nearly unit root coefficients

The following DGP is implemented:

$$Y_t = \phi_1 Y_{t-1} + \epsilon_t$$

where

$$\phi_1 \in \{0.95, 1, 1.05\}\,, \quad \epsilon_t \sim \mathcal{N}(0,1).$$

Financial and economic time series frequently exhibit behaviors close to a unit root ($\phi_1 \approx 1$) [6]. These characteristics pose significant challenges for traditional forecasting models. In this section, we explore the potential benefits of CQR in addressing these challenges. Through 3 values of ϕ_1, we examine different scenarios:

- $\phi_1 < 1$ (Stationary): Shocks have temporary effects; the series reverts to a mean over time.
- $\phi_1 = 1$ (Unit Root): Shocks have permanent effects; the series is non-stationary and follows a random walk.
- $\phi_1 > 1$ (Explosive): Shocks amplify over time; the series exhibits exponential growth or decay.

From Table 9 and Figures 14 and 15, we observe distinct differences in the behavior of the QR and QRF models. The QR model achieves the highest calibration performance among all models, and its conformalised counterpart (CQR QR) does not provide any further improvement in this regard; in fact, it results in a slight decrease in performance. In contrast, the QRF model faces greater challenges in calibration but shows improvement with the application of conformalisation. All models improves as the number of observations n increases. For the most difficult case, where $\phi_1 = 1.05$, none of the four models successfully captures the dynamics of the data generating process. In conclusion, conformalisation does not appear a very effective solution for dealing with nearly unit root autoregressive time series.

MAE	**0.95**	**1**
QR	0.007	0.015
QRF	0.075	0.112
CQR QR	0.012	0.026
CQR QRF	0.023	0.094

Table 9: MAE (4) for the four models with $\phi_1 \in \{0.95, 1\}$, averaged across three different values of n. Results for $\phi_1 = 1.05$ are not reported because both models completely fail to estimate the DGP.

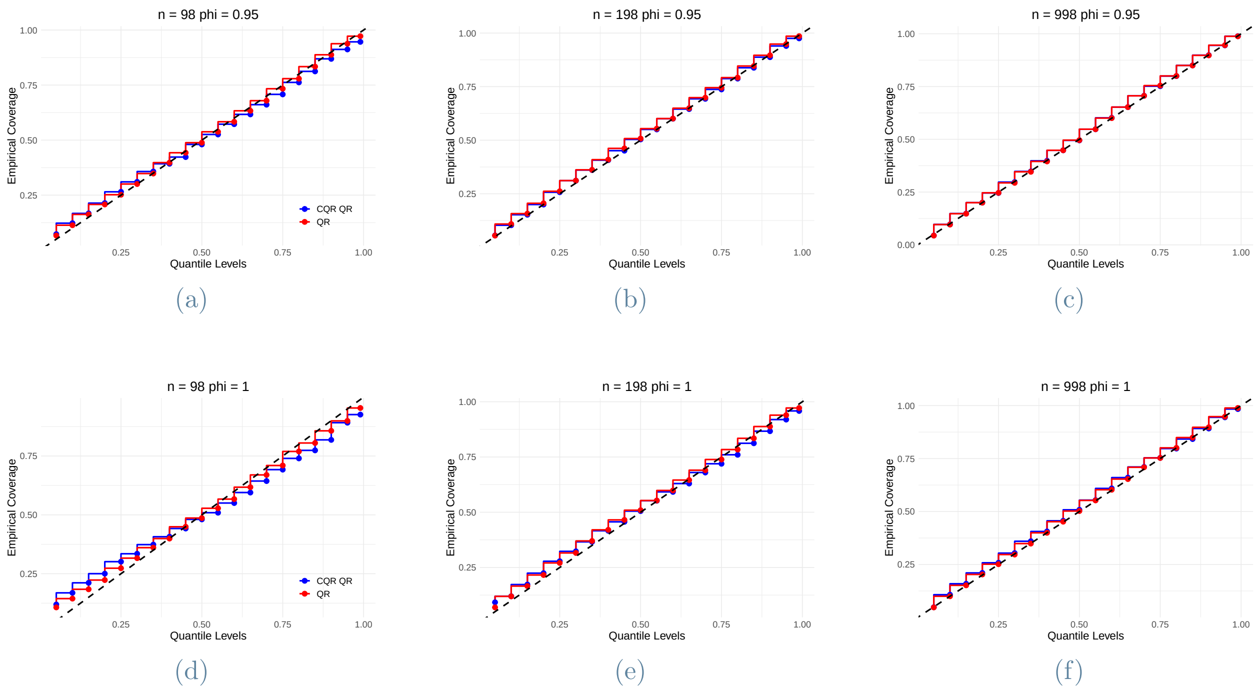


Figure 14: Calibration curves for QR and CQR QR for the 3 values of n and $\phi_1 = 0.95$ (first row), 1 (second row). Results for $\phi_1 = 1.05$ are not reported because both models completely fail to estimate the DGP.

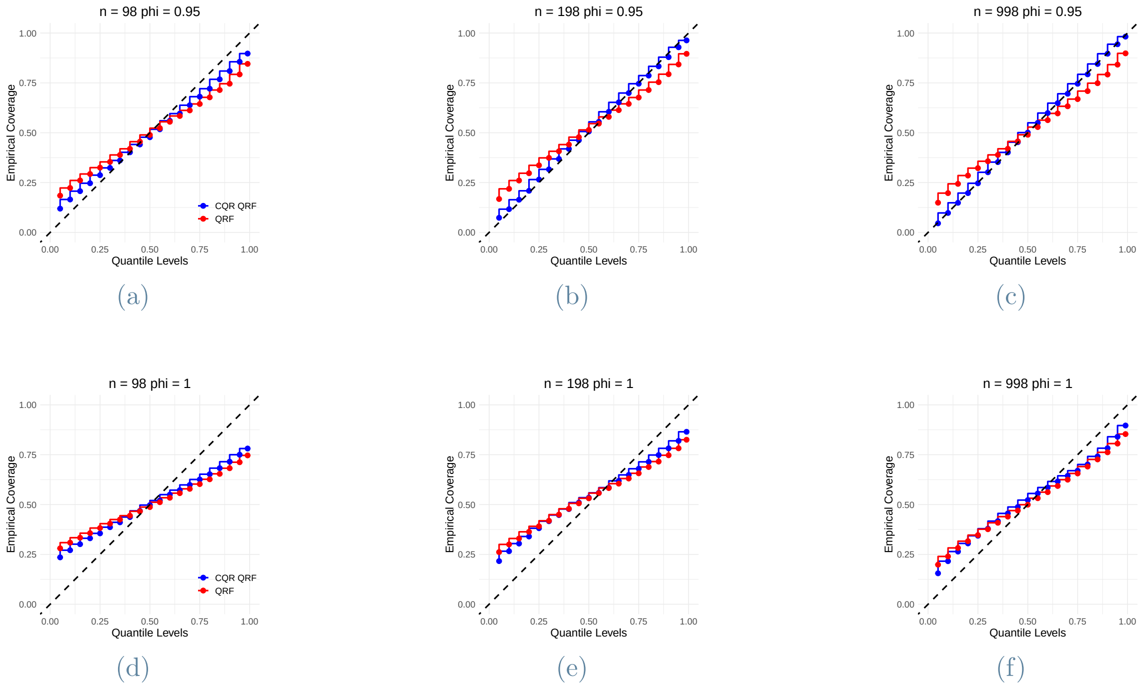


Figure 15: Calibration curves for QRF and CQR QRF for the 3 values of n and $\phi_1 = 0.95$ (first row), 1 (second row). Results for $\phi_1 = 1.05$ are not reported because all models completely fail to estimate the DGP.

Abstract in lingua italiana

Stimare con precisione quantili estremi è un compito cruciale con significative implicazioni empiriche. I quantili estremi sono essenziali nel policymaking e nell'industria finanziaria, ad esempio per calcolare misure di rischio come il Growth-at-Risk (GaR) e il Value-at-Risk (VaR). In questa tesi, conduciamo un ampio studio di simulazione e un'analisi reale di GaR per studiare i possibili benefici dell'implementazione di Conformal Prediction alla stima dei quantili. Questo lavoro rappresenta la prima applicazione di Conformal Prediction, tradizionalmente utilizzata per generare intervalli di previsione, alla stima dei quantili.
I nostri risultati dimostrano che i metodi di Conformal Prediction migliorano costantemente la calibrazione e la robustezza delle stime dei quantili a tutti i livelli quantilici, ma in particolare per i quantili estremi, dove i metodi tradizionali spesso falliscono. Inoltre, introduciamo una nuova proprietà che garantisce una copertura sotto l'ipotesi di exchangeability, fornendo uno strumento prezioso per gestire il rischio, quantificando e controllando la probabilità di future osservazioni estreme.

Parole chiave: Conformal Prediction, Stima Quantilica, Regressione Quantilica, Growth at Risk

Acknowledgements

I would like to express my sincere gratitude to my advisor, Prof. Simone Vantini, and my co-advisors, Matteo Fontana and Luca Neri, for their invaluable guidance and support throughout this research. I am also deeply thankful to my family for their unwavering encouragement, patience, and love, which have been a constant source of strength and inspiration during this journey.