Journal of Econometrics 236 (2023) 105490

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Out-of-sample tests for conditional quantile coverage an application to Growth-at-Risk☆

Valentina Corradi [a,*], Jack Fosten [b,c], Daniel Gutknecht [d]

[a] School of Economics, University of Surrey, GU2 7XH, Guildford, UK
[b] King's Business School, King's College London, WC2B 4BG, London, UK
[c] Data Analytics in Finance and Macro (DAFM) Research Centre, King's College London, UK
[d] Faculty of Economics and Business, Goethe University Frankfurt, 60629 Frankfurt am Main, Germany

ARTICLE INFO

Article history:
Received 12 October 2022
Received in revised form 3 May 2023
Accepted 25 June 2023
Available online 27 July 2023

JEL classification:
C01
C12
C22
C53

Keywords:
Interval prediction
Quantile regression
Multiple hypothesis testing
Weak moment inequalities
Wild bootstrap
Growth-at-Risk

ABSTRACT

This paper proposes tests for out-of-sample comparisons of interval forecasts based on parametric conditional quantile models. The tests rank the distance between actual and nominal conditional coverage with respect to the set of conditioning variables from all models, for a given loss function. We propose a pairwise test to compare two models for a single predictive interval. The set-up is then extended to a comparison across multiple models and/or intervals. The limiting distribution varies depending on whether models are strictly non-nested or overlapping. In the latter case, degeneracy may occur. We establish the asymptotic validity of wild bootstrap based critical values across all cases. An empirical application to Growth-at-Risk (GaR) uncovers situations in which a richer set of financial indicators are found to outperform a commonly-used benchmark model when predicting downside risk to economic activity.

1. Introduction

Interval predictions based on quantile forecasts play an important role in economics and finance. Conditional quantile predictions are routinely used in the finance literature, where measures like Value-at-Risk (VaR) are computed as a tool for risk management (see for instance Escanciano and Olmo, 2010, and Engle and Manganelli, 2004). More recently, a similar approach has gained momentum in macroeconomics where Growth-at-Risk (GaR) predictions are used to monitor downside risk to future GDP growth (Adrian et al., 2019; Reichlin et al., 2020; Plagborg-Møller et al., 2020).[1] The use of

☆ We are grateful to the Editor, Serena Ng, to the Associate Editor and to two anonymous referees as well as to Christian Brownlees, Marcelo Fernandes, Silvia Gonçalves, Jesus Gonzalo, Kirill Evdokimov, David Hendry, Fabian Krüger, Sebastien Laurent, Katerina Petrova, Marc Pohle, Barbara Rossi, Esther Ruiz and Giovanni Urga. We would also like to thank seminar participants at Cologne University, Universitat Pompeu Fabra, the 24th Dynamic Econometrics Conference, the IAAE Meeting 2021 in Rotterdam, the 11th Time Series Symposium Zaragoza, the NASMES 2021 in Montreal, the 9th ICEEE conference, Surrey University, the 2020 Conference of the Verein für Socialpolitik, and the 2019 Conference in Computational and Financial Econometrics for insightful comments and discussions. The paper has previously been circulated under the title 'Conditional Quantile Coverage: An Application to Growth-at-Risk'.

* Corresponding author.
E-mail addresses: V.Corradi@surrey.ac.uk (V. Corradi), jack.fosten@kcl.ac.uk (J. Fosten), Gutknecht@wiwi.uni-frankfurt.de (D. Gutknecht).

[1] As a complementary concept, Gonzalez-Rivera et al. (2019) introduced the notion of Growth-in-Stress, which is the prediction interval (or quantile) of future growth, conditional on underlying factors being in the tails of their distribution.

https://doi.org/10.1016/j.jeconom.2023.105490

GaR is now part of the standard toolkit of institutions like the International Monetary Fund (Prasad et al., 2019). The common feature of both VaR and GaR is that they are defined to quantify the expected drop in the target variable for a pre-determined probability, which is usually set by the regulator or a policymaking institution. This gives strong motivation for statistical tests based on coverage (Christoffersen, 1998; Escanciano and Velasco, 2010) when evaluating quantile forecasts for VaR and GaR, instead of other measures of quantile forecast accuracy such as (average) length or check loss (e.g., Manzan, 2015; Brownlees and Souza, 2021).

In this paper we propose novel tests to compare the out-of-sample performance of two or more competing models across single or multiple different predictive quantile interval(s) in terms of expected (conditional) coverage loss. The tests can be seen as a 'hybrid' between conditional and unconditional testing approaches since we use coverage probability with respect to a set of conditioning variables from *all* candidate quantile models as an input to an unconditional predictive ability test. Specifically, for each model we construct a sequence of conditional coverage probabilities for every out-of-sample evaluation point. For a given loss function chosen by the researcher (for example, quadratic or linear exponential loss), our tests then rank the models' forecast accuracy based on the expected loss of the conditional coverage errors, defined as the difference between the actual and the nominal coverage level.

The benefit of our hybrid form of the test is to combine the advantages of both the conditional and unconditional testing approaches. The use of conditional coverage as an input to the loss function contrasts to traditional unconditional coverage tests (Christoffersen, 1998; Escanciano and Olmo, 2010), which can give unsatisfactory conclusions in many circumstances. More recently, Li et al. (2022) and Horvath et al. (2022) have suggested conditional out-of-sample tests, where the benchmark model is favoured under the null hypothesis if it dominates its competitor(s) across all possible states of the world. By using expected loss, we instead choose the model that has more accurate (conditional) coverage on average, thus placing less weight onto extreme and rare events. The tests of this paper also differ from the conditional testing idea of Giacomini and White (2006), which incorporates parameter estimation into the null hypothesis. Their set-up, while popular in practice, has been documented to be restrictive in terms of the estimation schemes for model parameters (McCracken, 2020) and the time dependence structure (Zhu and Timmermann, 2020) admissible in a standard Diebold and Mariano (1995) test. The application of Giacomini and White (2006) in our context would rule out dynamically misspecified models, which we can accommodate.[2]

Our test takes a different approach to studies which compare quantile model predictions directly by using the check loss function from the Generalized Piecewise Linear (GPL) class (Gneiting, 2011). That approach can be used in the one-sided interval case though comparing predictions for two-sided or even multiple intervals would require the use of linear combinations of different GPL loss functions. This would be a disadvantage relative to our tests as it would make the interpretation of rejections in terms of the functional of interest (conditional coverage) rather difficult, if not impossible.

The tests we propose are very flexible and can be extended for use in a wide range of modelling scenarios which are relevant for empirical researchers. Firstly, the tests are applicable in the context of (strictly) nested and overlapping model comparisons alike, where the latter refers to a situation in which, for a given quantile level, the predictions of both quantile models are identical under the null hypothesis of equal expected predictive content with probability one. This definition of overlapping models comprises situations where the two quantile models are strictly nested in the usual sense with one model based on a strictly smaller set of conditioning variables than the other model (e.g., comparing the predictions of two linear quantile autoregressions of different lag orders), but is generally broader, see Section 3.2 for a more detailed discussion and examples.[3] This is important in the empirical GaR context as studies often compare possibly overlapping models where a quantile autoregression is augmented with different macrofinancial predictors that may or may not have predictive power. Specifically, if the macrofinancial predictors hold no predictive content, the corresponding quantile predictions will coincide in population with probability one which fits the definition of overlapping models. We therefore build on the literature of out-of-sample tests for strictly nested/overlapping models which have become well known for conditional mean forecasting (Clark and McCracken, 2001, 2014), but not in the case of quantiles. Secondly, our tests are applicable for testing multiple models and multiple quantile-based intervals. This allows us to provide robust inference when studies wish to compare GaR models using various different predictors (e.g., Brownlees and Souza, 2021) and when different GaR levels are used. Finally, we can also extend the test to a setting where forecasts are made at multiple different horizons. This builds on recent interest in multi-horizon forecast evaluation in the conditional mean forecasting context, such as Quaedvlieg (2021).

We begin by proposing a test for the baseline case where we have a pairwise model comparison over a single interval. Since we allow the models to be dynamically misspecified and/or overlapping in analogy to Vuong (1989) and Shi (2015), there are different cases for the asymptotic behaviour of our test statistic under the null of equal expected conditional coverage error. If the models are strictly non-nested, the statistic is asymptotically zero mean normal with a variance reflecting parametric and nonparametric estimation error. This is because the construction of the test statistic requires both the estimation of the parametric conditional quantile models and of the conditional coverage probability, which is

[2] On the other hand, since our approach is based on the comparison of models in terms of unconditional expected loss, our framework cannot reflect, unlike Giacomini and White (2006), the effect that estimation uncertainty has on relative forecast performance in finite samples. Moreover, even though the tests use a sequence of conditional coverage errors as inputs into the loss function, our comparison is ultimately unconditional with the drawback that the comparison remains an "average" model comparison across different states of the world.

[3] On the other hand, if the models are non-overlapping, we will say that they are strictly non-nested from now on.

estimated nonparametrically. On the other hand, when the two quantile models are overlapping, the statistic converges in probability to zero for all strictly convex loss functions, regardless of whether the models are correctly specified or not with respect to the union of the conditioning variables.

In a further step, we extend the pairwise test to the multiple intervals and/or multiple models set-up, where the null is that no competing model has a smaller expected conditional coverage error than the benchmark model for any of the intervals considered. Since this is a composite hypothesis given by the intersection of many pairwise null hypotheses, we can re-state it in terms of a finite number of weak moment inequalities as in Andrews and Soares (2010). Not surprisingly, the suggested statistic has a degenerate limiting distribution in the overlapping case when models have the same conditional coverage error almost surely.

We suggest the use of wild bootstrap critical values, along the lines of the conditional p-value approach of Hansen (1996). This has previously been extended to account for dynamic misspecification by Inoue (2001) and for parameter estimation error by Corradi and Swanson (2002). Here, we provide a further extension by showing that this procedure can properly mimic nonparametric conditional coverage estimation error as well. We show that inference based on wild bootstrap critical values is first order asymptotically valid in both the degenerate and non-degenerate case. In macroeconomic applications, such as GaR, the sample size is generally rather short and thus bootstrap based critical values are preferable to subsampling based ones, as in studies such as Escanciano and Velasco (2010).

We assess the properties of the pairwise and multiple model/interval comparison tests in Monte Carlo simulations where we explore various data generating processes (DGPs) and sample sizes. The simulations allow for different degrees of time series dependence and for different correlations among the predictors. The results indicate that our tests have good finite sample properties. In particular, the tests exhibit size control even when models are overlapping and the asymptotic distribution of the test statistic is degenerate, a likely feature of the wild bootstrap procedure whose distribution appears to collapse slower in finite samples due to the presence of various estimated expressions. The power of the test increases when we evaluate models at multiple quantiles suggesting that model comparisons over a range of quantile ranks may be more desirable.

We then provide an empirical application to GaR. According to Adrian et al. (2019), the lower quantiles of GDP growth are sensitive to the National Financial Conditions Indicator (NFCI), while the higher ones are not. The benchmark model is therefore a linear quantile model with the NFCI as a predictor. Competing models are also linear quantile models, but rely on different sets of predictors as suggested by Brownlees and Souza (2021). Additionally, we use both GDP growth and industrial production (IP) growth to proxy economic conditions, the latter of which is available on a monthly basis. When using GDP, no competing model beats the benchmark. On the other hand, when we use IP, the NFCI model is beaten by some competitors at least at a one quarter horizon. This is likely due to the higher power of the test in larger samples suggesting that the use of timelier measures of economic activity may be important for a robust comparison of different GaR models in practice.

The rest of the paper is organized as follows. Section 2 describes the general set-up of quantile models and conditional coverage testing. Section 3 introduces the test for pairwise equal expected conditional coverage error loss, provides its limiting distribution and describes how to construct valid bootstrap critical values. Section 4 provides the extension to the case of multiple intervals and/or multiple models. Section 5 shows how the tests perform in Monte Carlo simulation experiments and Section 6 provides the empirical application to GaR. Finally, Section 7 concludes the paper. The Appendix A contains the definitions of some lengthier technical expressions of the asymptotic variance and the bootstrap statistic from the main paper. Proofs of theorems can be found in the supplementary material, as well as extensions of the set-up to (nonlinear) location scale models, to two-sided intervals, to the recursive estimation scheme, and to Conditional Autoregressive Value at Risk models. The supplementary material also contains additional Monte Carlo simulations, and another empirical illustration about VaR prediction.

2. Set-up

The main objective of this paper is to compare interval predictions for future values of some target variable, y_t, from different parametric conditional quantile models. Quantile-based interval predictions differ from other types of interval forecasts in that the interval boundaries are pre-specified and given by the nominal quantile levels $[\tau_L, \tau_U]$. That is, unlike in other interval forecast settings where the forecaster aims at producing a prediction with $1-\alpha$ coverage, but the shape and the characteristics of the interval are not necessarily restricted, we deviate from such settings in that the main motivation of the paper lies in GaR and VaR type applications, where the regulator or an institution is interested in intervals at *pre-determined* quantile levels.

We will compare interval predictions of the different quantile models at these pre-specified quantile levels in terms of the coverage probability with respect to the set of conditioning variables from all models (see Section 3 for a discussion of this choice). In what follows, we will formally introduce conditional coverage, the classes of quantile models we consider, and the corresponding conditional coverage error used to construct the test statistic. Note that for notational simplicity we will only focus on one-step ahead forecasts, though the extension to general s-step ahead forecasts is immediate and is explored in the empirical section.

Consider a time series sequence $\{y_t, \mathbf{Z}_t'\}_{t=1}^T$, where y_t is a continuous target variable of interest and $\mathbf{Z}_t$ is a random vector that contains other observable predictors. We use $\mathcal{F}_t$ to denote the sigma field generated by $\{y_s, \mathbf{Z}_s'; s \leq t\}$. Given $\mathcal{F}_t$, the conditional τ-quantile of y_{t+1} is defined as:

$$q_\tau(\mathcal{F}_t) = \inf\{y : F_{y_{t+1}|\mathcal{F}_t}(y|\mathcal{F}_t) \geq \tau\},$$

where $F_{y_{t+1}|\mathcal{F}_t}(y|\mathcal{F}_t)$ is the distribution function of y_{t+1} conditional on $\mathcal{F}_t$.[4] In addition, for any $\tau_L, \tau_U \in \mathcal{T}$ with $\tau_L < \tau_U$ and $\mathcal{I}_t \subset \mathcal{F}_t$, the probability:

$$\Pr\left(q_{\tau_L}(\mathcal{I}_t) \leq y_{t+1} \leq q_{\tau_U}(\mathcal{I}_t)|\mathcal{F}_t\right) \equiv C\left([\tau_L, \tau_U]; \mathcal{F}_t\right)$$

is known as *conditional coverage* with respect to the information set $\mathcal{F}_t$. It is useful to introduce the one-sided conditional coverage where we set $\tau_L = 0$ and $\tau_U = \tau$. This is used in the case of GaR, which is defined as the lower one-sided prediction interval containing future realizations of GDP growth with τ coverage probability. In this case, the conditional coverage with respect to $\mathcal{F}_t$ just becomes a one-sided conditional probability:

$$\Pr\left(y_{t+1} \leq q_\tau(\mathcal{I}_t)|\mathcal{F}_t\right) \equiv C\left((0, \tau]; \mathcal{F}_t\right).$$

Since the purpose of this paper is to construct a comparison test for two or more candidate models used to predict the interval at levels $[\tau_L, \tau_U]$ or $(0, \tau]$, respectively, we assume that each of these models belong to one of two parametric classes of models. Specifically, let $X_{j,t}, j = 1, \ldots, J$, denote the set of conditioning variables from model j that may contain y_t and elements of $\mathbf{Z}_t$ as well as lags thereof. Since our main motivating example is the case of GaR, where conditional quantiles are typically modelled as a linear function of a macrofinancial variable (see, e.g. Prasad et al., 2019; Adrian et al., 2019), we start with the class of linear conditional quantile models of the form:

$$q_\tau(\boldsymbol{\beta}_j^\dagger; X_{j,t}) = X_{j,t}'\boldsymbol{\beta}_j^\dagger(\tau). \tag{1}$$

This model includes the quantile autoregressive (QAR) model as discussed in Koenker and Xiao (2006) and Qu (2008). In fact, letting:

$$y_{t+1} = \theta_{j,0}^\dagger(U_{t+1}) + \theta_{j,1}^\dagger(U_{t+1})y_t + \cdots + \theta_{j,p}^\dagger(U_{t+1})y_{t-p+1},$$

with U_{t+1} being a sequence of i.i.d. uniform random variables, we can write $X_{j,t} = (y_t, y_{t-1}, \ldots, y_{t-p+1})'$ and $\boldsymbol{\beta}_j(\tau) = (\theta_{j,0}(\tau), \theta_{j,1}(\tau), \ldots, \theta_{j,p}(\tau))'$ to obtain the general QAR(p) model, provided the θ_j's are monotone increasing in U_{t+1}.

We can also derive an expression like (1) for location scale models with affine transformations of the first and second conditional moments, for example:

$$y_{t+1} = X_{j,t}'\boldsymbol{\delta}_j^\dagger + (X_{j,t}'\boldsymbol{\gamma}_j^\dagger)\varepsilon_{t+1}.$$

With ε_{t+1} independent of $X_{j,t}$, we have that $q_\tau(\boldsymbol{\beta}_j^\dagger; X_{j,t}) = X_{j,t}'\boldsymbol{\beta}_j^\dagger(\tau)$ with $\boldsymbol{\beta}_j^\dagger(\tau) = \boldsymbol{\delta}_j^\dagger + \boldsymbol{\gamma}_j^\dagger q_\tau(\varepsilon_{t+1})$, where $q_\tau(\varepsilon_{t+1})$ denotes the τ unconditional quantile of the error term ε_{t+1}. On the other hand, any location scale model that does not consist of an affine transformation of the first and/or second moment(s), such as in Machado and Silva (2019), no longer gives rise to a linear conditional quantile regression model. For instance, let:

$$y_{t+1} = m\left(X_{j,t}, \boldsymbol{\theta}_{j,m}^\dagger\right) + \sigma(X_{j,t}, \boldsymbol{\theta}_{j,\sigma}^\dagger)\epsilon_{j,t+1},$$

where $\epsilon_{j,t+1} = \left(y_{t+1} - m(X_{j,t}, \boldsymbol{\theta}_{j,m}^\dagger)\right)/\sigma(X_{j,t}, \boldsymbol{\theta}_{j,\sigma}^\dagger)$, and $m(\cdot, \boldsymbol{\theta}_{j,m}^\dagger)$ as well as $\sigma(\cdot, \boldsymbol{\theta}_{j,\sigma}^\dagger)$ are some nonlinear functions indexed by finite-dimensional parameter vectors $\boldsymbol{\theta}_{j,m}^\dagger$ and $\boldsymbol{\theta}_{j,\sigma}^\dagger$. In this case, we have that the conditional quantile function $q_\tau(\boldsymbol{\theta}_j^\dagger; X_{j,t})$ with $\boldsymbol{\theta}_j^\dagger = (\boldsymbol{\theta}_{j,m}^{\dagger\prime}, \boldsymbol{\theta}_{j,\sigma}^{\dagger\prime})'$ is given by:

$$q_\tau(\boldsymbol{\theta}_j^\dagger; X_{j,t}) = m\left(X_{j,t}, \boldsymbol{\theta}_{j,m}^\dagger\right) + \sigma(X_{j,t}, \boldsymbol{\theta}_{j,\sigma}^\dagger)q_\tau(\epsilon_{j,t+1}), \tag{2}$$

which covers the various types of GARCH models used in the prevailing VaR literature (see for example Escanciano and Velasco, 2010). Moreover, since linear GARCH models can be re-written as Conditional Autoregressive Value at Risk (CAViaR) models satisfying certain parameter restrictions, note that the latter model is also covered by our set-up. The CAViaR model, first introduced by Engle and Manganelli (2004), is now a popular choice for modelling tail dynamics of financial time series data. As such, in the supplementary material, we outline how our tests can accommodate CAViaR models via the two-step quantile regression procedure suggested by Koenker and Xiao (2009).

While our set-up formally also covers the quantile models in (2), we relegate the exposition and treatment of these models including assumptions, their estimation, and their treatment in the asymptotic analysis to the supplementary material, as well as a brief empirical illustration to VaR forecasting. Thus, we will henceforth focus on linear quantile regression models, but note that all tests can also be carried out using nonlinear location scale models.

[4] In what follows, we keep the $t+1$ reference in the quantile notation $q_\tau(\mathcal{F}_t)$ implicit.

Before turning to the actual tests, we discuss estimation of the conditional quantile models introduced in (1). For $\tau \in \mathcal{T}$, define:

$$\boldsymbol{\beta}_j^\dagger(\tau) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \mathrm{E}\left[\rho_\tau\left(y_{t+1} - X_{j,t}'\boldsymbol{\beta}\right)\right], \tag{3}$$

where $\rho_\tau(u) = u(\tau - 1\{u < 0\})$ is the usual check function and $\mathcal{B}$ denotes the parameter space defined in A.3 of Section 3.2. We allow for the linear conditional quantile of model j, $q_\tau(\boldsymbol{\beta}_j^\dagger; X_{j,t}) = X_{j,t}'\boldsymbol{\beta}_j^\dagger(\tau)$, to be (dynamically) misspecified with respect to $\mathbf{X}_t = \{X_{1,t} \cup \cdots \cup X_{J,t}\}$, the union of the conditioning vectors of all candidate models under consideration.[5] This may indeed arise when the model does not use all relevant information contained in $\mathbf{X}_t$ or the functional form is misspecified.

In the forecasting literature, it is standard to split the overall sample of size T into a training sample of size R and a prediction sample of size P, and to estimate parameters in a fixed, recursive or rolling manner (e.g., West, 1996). To avoid unnecessary complication and for notational simplicity, we will state all of our main results for the case of a fixed estimation scheme, and discuss the recursive scheme in an extension (see the supplementary material). Thus, we estimate (3) using the first R observations as follows:

$$\widehat{\boldsymbol{\beta}}_{j,R}(\tau) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \frac{1}{R}\sum_{s=1}^{R-1} \rho_\tau\left(y_{s+1} - X_{j,s}'\boldsymbol{\beta}\right), \tag{4}$$

and so the estimator of the linear conditional quantile model is given by:

$$q_\tau(\widehat{\boldsymbol{\beta}}_{j,R}; X_{j,t}) = X_{j,t}'\widehat{\boldsymbol{\beta}}_{j,R}(\tau). \tag{5}$$

Since the conditional quantiles may either come from (1) or from (2), and estimation for nonlinear location scale models differs from (4) and (5), we will adopt a more generic notation hereafter. More specifically, let $\boldsymbol{\psi}_j^\dagger(\tau)$ denote a finite-dimensional parameter vector from either of the two classes of parametric models outlined above so that $\boldsymbol{\psi}_j^\dagger(\tau) = \boldsymbol{\beta}_j^\dagger(\tau)$ or $\boldsymbol{\psi}_j^\dagger(\tau) = \left(\boldsymbol{\theta}_{j,m}^\dagger, \boldsymbol{\theta}_{j,\sigma}^\dagger, \beta_j^\dagger(\tau)\right)'$. The corresponding τ-level quantile can then generically be written as $q_\tau(\boldsymbol{\psi}_j^\dagger; X_{j,t})$, while the conditional coverage of model j is defined as the (conditional) probability that y_{t+1} lies within the prediction interval of model j defined by the boundary points $q_{\tau_L}(\boldsymbol{\psi}_j^\dagger; X_{j,t})$ and $q_{\tau_U}(\boldsymbol{\psi}_j^\dagger; X_{j,t})$, respectively:

$$C_j\left([\tau_L, \tau_U]; \mathbf{X}_t\right) = \Pr\left(q_{\tau_L}(\boldsymbol{\psi}_j^\dagger; X_{j,t}) \le y_{t+1} \le q_{\tau_U}(\boldsymbol{\psi}_j^\dagger; X_{j,t}) | \mathbf{X}_t\right). \tag{6}$$

Since our test will be based on comparing coverage for different models $j = 1, \ldots, J$, given the common vector $\mathbf{X}_t$, in terms of deviations from the nominal level $\tau_U - \tau_L$ (see Section 3), we also define the following *conditional* coverage error for model j:

$$\mathcal{E}_j\left([\tau_L, \tau_U]; \mathbf{X}_t\right) = C_j\left([\tau_L, \tau_U]; \mathbf{X}_t\right) - (\tau_U - \tau_L),$$

which in the one-sided interval case becomes:

$$\mathcal{E}_j\left((0, \tau]; \mathbf{X}_t\right) = C_j\left((0, \tau]; \mathbf{X}_t\right) - \tau.$$

These errors are conditional in the sense that they are defined with respect to a specific $\mathbf{X}_t$, and a different realization of $\mathbf{X}_t$ may lead to a different sequence of errors. This reflects the fact that models may provide good predictions (in terms of coverage) in one state of the world, but not necessarily in another. For instance, while model j may have correct GaR coverage when $\mathbf{X}_t$ denotes a realization of the conditioning set in a recession, coverage may deviate from its nominal level when $\mathbf{X}_t$ represents a realization of the predictor variables in an economic expansion phase.

3. Pairwise comparison

3.1. Null hypothesis and statistic

Our first objective is to test pairwise quantile forecast accuracy, measured in terms of the average distance between actual and nominal conditional coverage. We therefore introduce a loss function that will allow to penalize deviations of coverage from its nominal level for each realization of $\mathbf{X}_t$. That is, let $L(\cdot)$ denote a given loss function that satisfies Assumption A.2 below. We specify the null hypothesis as equal expected conditional coverage error between the two models, for this given loss function $L(\cdot)$. Formally, for a given pair $\tau_L, \tau_U \in \mathcal{T}$ and two models $j = 1, 2$, we test:

$$H_0 : \mathrm{E}\left(\left(L\left(\mathcal{E}_1\left([\tau_L, \tau_U]; \mathbf{X}_t\right)\right) - L\left(\mathcal{E}_2\left([\tau_L, \tau_U]; \mathbf{X}_t\right)\right)\right) 1\{\mathbf{X}_t \in \mathcal{X}\}\right) = 0 \tag{7}$$

[5] We write $\mathbf{X}_t = \{X_{1,t} \cup \cdots \cup X_{J,t}\}$ to denote the union of row vectors of possibly different dimensions. The dimension of the resulting row vector $\mathbf{X}_t$ is defined as $d = \dim(\mathbf{X}_t)$, and we assume that all variables in $\mathbf{X}_t$ are distinct predictors in the sense of exhibiting some independent variation a.s. (see A.4 in Section 3).

versus:

$$H_A : \mathrm{E}\left(\left(L\left(\mathcal{E}_1\left(\left[\tau_L, \tau_U\right]; \mathbf{X}_t\right)\right) - L\left(\mathcal{E}_2\left(\left[\tau_L, \tau_U\right]; \mathbf{X}_t\right)\right)\right) 1\{\mathbf{X}_t \in \mathcal{X}\}\right) \neq 0.$$

In the case of a one-sided interval, H_0 becomes:

$$H_0 : \mathrm{E}\left(\left(L\left(\mathcal{E}_1\left((0, \tau]; \mathbf{X}_t\right)\right) - L\left(\mathcal{E}_2\left((0, \tau]; \mathbf{X}_t\right)\right)\right) 1\{\mathbf{X}_t \in \mathcal{X}\}\right) = 0 \tag{8}$$

versus its negation. In the statement of H_0, we compute the mean over the set $\mathcal{X}$, which is a compact subset of the support of $\mathbf{X}_t$. As outlined in Section 3.2, this plays a role for the estimation of conditional coverage and could be relaxed at the cost of more complex arguments in the derivation of the asymptotic distribution of the test statistic. In addition, observe that we evaluate a model with respect to the common vector $\mathbf{X}_t$, rather than in terms of its own conditioning vector $X_{j,t}$. While the latter approach may keep the "curse of dimensionality" at bay, such a comparison will in general be misleading, especially if both conditioning vectors are strictly disjoint.[6]

In order to construct a test statistic for H_0 versus H_A, we first estimate the conditional coverage for model j in ((6)), and get a sequence of P local out-of-sample hits, where only those observations in proximity to the evaluation point $\mathbf{X}_t$ receive a positive weight. More specifically, let $q_\tau(\widehat{\boldsymbol{\psi}}_{j,R}; X_{j,t})$ denote the empirical counterpart of $q_\tau(\boldsymbol{\psi}_j^\dagger; X_{j,t})$, and let:

$$1\left\{q_{\tau_L}(\widehat{\boldsymbol{\psi}}_{j,R}; X_{j,t}) \leq y_{s+1} \leq q_{\tau_U}(\widehat{\boldsymbol{\psi}}_{j,R}; X_{j,t})\right\} \frac{1}{h^d}\mathbf{K}\left(\frac{\mathbf{X}_s - \mathbf{X}_t}{h}\right) \quad \text{for } s = R, \ldots, T-1$$

denote a sequence of estimated hits from model j, where $d = \dim(\mathbf{X}_t)$, $\mathbf{K}\left(\frac{\mathbf{u}}{h}\right) = K\left(\frac{u_1}{h}\right) \times \cdots \times K\left(\frac{u_d}{h}\right)$ is the product of d univariate kernel functions defined in Assumption A.5(iii) below, and h denotes a deterministic bandwidth sequence satisfying $h \to 0$ as $P \to \infty$. By averaging the P hits conditionally on $\mathbf{X}_t$ and weighing by the empirical density, we obtain:

$$\widehat{C}_{j,P,R}\left(\left[\tau_L, \tau_U\right]; \mathbf{X}_t\right) = \frac{1}{Ph^d}\sum_{s=R}^{T-1} \frac{1}{\widehat{f}_X\left(\mathbf{X}_t\right)} 1\left\{q_{\tau_L}(\widehat{\boldsymbol{\psi}}_{j,R}; X_{j,t}) \leq y_{s+1} \leq q_{\tau_U}(\widehat{\boldsymbol{\psi}}_{j,R}; X_{j,t})\right\} \mathbf{K}\left(\frac{\mathbf{X}_s - \mathbf{X}_t}{h}\right),$$

where:

$$\widehat{f}_X\left(\mathbf{X}_t\right) = \frac{1}{Ph^d}\sum_{s=R}^{T-1} \mathbf{K}\left(\frac{\mathbf{X}_s - \mathbf{X}_t}{h}\right) \tag{9}$$

is a standard estimator of the nonparametric density. This gives rise to the empirical conditional coverage error of model j defined as:

$$\widehat{\mathcal{E}}_{j,P,R}\left(\left[\tau_L, \tau_U\right]; \mathbf{X}_t\right) = \widehat{C}_{j,P,R}\left(\left[\tau_L, \tau_U\right]; \mathbf{X}_t\right) - \left(\tau_U - \tau_L\right),$$

which instead reads as:

$$\widehat{\mathcal{E}}_{j,P,R}\left((0, \tau]; \mathbf{X}_t\right) = \widehat{C}_{j,P,R}\left((0, \tau]; \mathbf{X}_t\right) - \tau$$

for the case of a one-sided interval.

Finally, for testing H_0 versus H_A, for the two-sided interval case, we rely on the following statistic:

$$\widehat{S}_{P,R}(\tau_L, \tau_U) \equiv \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(L\left(\widehat{\mathcal{E}}_{1,P,R}\left(\left[\tau_L, \tau_U\right]; \mathbf{X}_t\right)\right) - L\left(\widehat{\mathcal{E}}_{2,P,R}\left(\left[\tau_L, \tau_U\right]; \mathbf{X}_t\right)\right)\right) 1\{\mathbf{X}_t \in \mathcal{X}\} \tag{10}$$

while in the one-sided interval case, we use:

$$\widehat{S}_{P,R}(\tau) \equiv \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(L\left(\widehat{\mathcal{E}}_{1,P,R}\left((0, \tau]; \mathbf{X}_t\right)\right) - L\left(\widehat{\mathcal{E}}_{2,P,R}\left((0, \tau]; \mathbf{X}_t\right)\right)\right) 1\{\mathbf{X}_t \in \mathcal{X}\}. \tag{11}$$

We close this subsection with a brief discussion to acknowledge potential shortcomings of conditional coverage as a comparison measure for quantile-based interval predictions. When the nominal evaluation level of $(\tau_U - \tau_L)$ or τ lies close to the boundary of zero or one, it may be difficult to discard completely uninformative forecast models with coverage equal to zero (or one) for most realizations of $\mathbf{X}_t$. In this case, it can also be useful to assess the interval length as additional information, even if this is not the criterion required by regulators or institutions in the case of VaR or GaR. There are a few ways to do this. When the object of interest is a two-sided quantile interval $[\tau_L, \tau_U]$, a suitable loss (scoring) function

[6] For instance, suppose that the DGP is given by $q_\tau(\mathcal{F}_T) = \beta_1 X_{1,t} + \beta_2 X_{2,t} + q_\tau(e_{t+1})$, where $(\beta_1, \beta_2)' = (0, 1)'$ and all right hand side variables are independent of each other with strictly increasing marginal distribution functions everywhere. In this case, if model 1 omits $X_{2,t}$, while model 2 omits $X_{1,t}$, it still is the case that $\Pr\left(y_{t+1} \leq \beta_1^\dagger X_{1,t} + F^{-1}_{u_{1,t+1}}(\tau) \,|X_{1,t}\right) = \Pr\left(y_{t+1} \leq \beta_2^\dagger X_{2t} + F^{-1}_{u_{2,t+1}}(\tau) \,|X_{2,t}\right) = \tau$ almost surely with $u_{1,t+1} = e_{t+1} + \beta_2 X_{2,t}$ and $u_{2,t+1} = e_{t+1}$.

that also accounts for length is the interval score (e.g., Gneiting and Raftery, 2007). In the case of equal tailed, quantile bounded intervals of the form $\tau_L = \alpha/2$ and $\tau_U = (1-\alpha/2)$, the interval score takes the form:

$$(q_u - q_l) + \frac{2}{\alpha}(q_l - y_{t+1})1\{y_{t+1} < q_l\} + \frac{2}{\alpha}(y_{t+1} - q_u)1\{y_{t+1} > q_u\},$$

where q_l and q_u denote the quantile predictions of the lower and the upper bound of the interval, respectively. However, this interval score is not suitable for the case of one-sided intervals $(0, \tau]$ as it is not clear how to assess length relative to 'hits' outside the interval range in that case, and so this is not applicable in the GaR or VaR context.[7] In constructing a length statistic for the one-sided case, Brownlees and Souza (2021) suggest the following average length measure for some model j:

$$\frac{1}{P}\sum_{t=R}^{T-1} q_\tau(\widehat{\boldsymbol{\psi}}_{j,R}; X_{j,t}) - \widehat{q}_{0.01},$$

where $\widehat{q}_{0.01}$ denotes the unconditional empirical 1%-level quantile of y_{t+1}. We consider this statistic to be less informative than conditional coverage for comparing models as it simply measures the difference in their averaged conditional quantile predictions.

3.2. Limiting distribution

We now state the assumptions required to derive the limiting distribution of the test statistics in (10) and (11). Since we also consider the general case with $J > 2$ models below, we will state all conditions for a generic J. In what follows, let $\|\cdot\|$ denote the Euclidean norm and $\nabla^{(k)}g(\cdot)$ denote the kth order partial derivative of the function $g(\cdot)$ with respect to its argument.

Assumption A.1. For all $j = 1, \ldots, J$, $(y_{t+1}, X'_{j,t})'$ are strictly stationary and β -mixing with coefficients satisfying $\sum_{k=1}^{\infty} \beta(k)^{\frac{\varepsilon}{2+\varepsilon}} < \infty$, $\varepsilon > 0$.

Assumption A.1 imposes mild restrictions on the time dependence of the data. In particular, note that the condition of β-mixing (absolutely regular) data is mainly required for addressing a second order term, which becomes relevant in the overlapping case, but could otherwise be replaced by strong mixing conditions. The next assumption defines the class of loss functions we consider for the test:

Assumption A.2.

(i) The loss function $L(\cdot)$ is three times continuously differentiable on the interior of its support $\mathcal{S}_L$, where $\mathcal{S}_L \subset [-1-\epsilon, 1+\epsilon]$ for some $\epsilon > 0$, with Lipschitz continuous derivatives.

(ii) $L(u) = 0$ for $u = 0$.

(iii) $\nabla^{(1)}L(u) \leq 0$ for $u < 0$ and $\nabla^{(1)}L(u) \geq 0$ for $u > 0$.

(iv) $\nabla^{(2)}L(u) > 0$ on the interior of its support $\mathcal{S}_L$.

Assumption A.2(i)–(ii) implies that L is a generalized loss function, as defined in Granger (1999). In fact, A.2(ii) ensures that whenever u_1 is further away from zero than u_2, then $L(u_1) > L(u_2)$. A.2(i) and Assumption A.2(iv) are used in the derivation of the limiting behaviour of the statistic in the degenerate case. Note that several commonly used loss functions, such as the quadratic and linear exponential (Linex) loss, satisfy Assumption A.2. Thus, if the researcher has a preference for over- rather than under-coverage, they could choose Linex loss with $L(u) = \exp(a \cdot u) - a \cdot u - 1$, $a < 0$.

Assumption A.3.

(i) For every $j \in \{1, \ldots, J\}$ and every value x in the support of $X_{j,t}$, R_{X_j}, the conditional density function $f_{y_{t+1}|X_{j,t}}(\cdot|x)$ (abbreviated to $f_{t+1}(\cdot|x)$ in what follows) is continuous in y_{t+1} w.r.t. Lebesgue measure, uniformly bounded and bounded away from zero.

(ii) It holds that $\mathrm{E}\left(\left\|X_{j,t}\right\|^{4+\varepsilon}\right) < \infty$, where ε was defined in Assumption A.1. Moreover, for every model $j \in \{1, \ldots, J\}$ which can be written as in (1),

$$H_j(\tau) \equiv \mathrm{E}\left(f_{t+1}\left(X'_{j,t}\boldsymbol{\beta}_j^{\dagger}(\tau)\,\middle|X_{j,t}\right)X_{j,t}X'_{j,t}\right)$$

[7] Even though one may of course 'fix' a lower bound to recover a length in the one-sided interval $(0, \tau]$, this bound would be arbitrary and thus interfere with the idea of a relative forecast comparison as it is not clear how to scale the hits accordingly.

is positive definite for every $\tau \in \mathcal{T}$.

(iii) For every $j \in \{1, \ldots, J\}$ and every $\tau \in \mathcal{T}$, the parameter space of $\boldsymbol{\beta}_j^\dagger(\tau)$, denoted $\mathcal{B}$, is compact. Moreover, assume that $\boldsymbol{\beta}_j^\dagger(\tau)$ as defined in (3) is uniquely identified for all $\tau \in \mathcal{T}$ and lies in the interior of $\mathcal{B}$.

(iv) For every $j \in \{1, \ldots, J\}$ and every $\tau \in \mathcal{T}$, the parameter vector $\boldsymbol{\beta}_j^\dagger(\tau)$ defined in (3) satisfies:

$$\mathrm{E}\left(X_{j,t}(1\{y_{t+1} \leq X_{j,t}'\boldsymbol{\beta}_j^\dagger(\tau)\} - \tau)\right) = 0.$$

Assumption A.3(i)–(iv) are rather standard in the quantile regression literature and ensure consistency of the quantile regression estimator $\widehat{\boldsymbol{\beta}}_{j,R}(\tau)$ for $\boldsymbol{\beta}_j^\dagger(\tau)$ as well as weak convergence of

$$\sqrt{R}\left(\widehat{\boldsymbol{\beta}}_j(\tau) - \boldsymbol{\beta}_j^\dagger(\tau)\right)$$

to a Gaussian distribution (for example, Koenker and Xiao, 2006). Note also that A.3(i)–(ii) impose conditions on the behaviour of $f_{t+1}(\cdot|X_{j,t})$ over the support of $X_{j,t}$ as well as on its moments as we are estimating the parameter vector $\boldsymbol{\beta}_j^\dagger(\tau)$ of the quantile regression model over the whole support of $X_{j,t}$. On the other hand, A.3(iv) is a quantile version of the well-known orthogonality condition from ordinary least squares. In particular, it allows for misspecified models where $\mathrm{E}\left(1\{y_{t+1} \leq X_{j,t}'\boldsymbol{\beta}_j^\dagger(\tau)\}|X_{j,t}\right) = \tau$ does not necessarily hold with probability one, see for instance Kim and White (2003). Finally, note that when one or more nonlinear location scale model(s) are used in the comparison, an additional high-level assumption is required in that case (see supplementary material).

Finally, we impose a further regularity condition for the kernel used in the nonparametric estimation of conditional coverage and the underlying density as well as for the corresponding bandwidth.

Assumption A.4.

(i) Let $\mathcal{X}$ denote a compact, non-empty, connected set, which lies in the interior of the support of $\mathbf{X}_t = \{X_{1,t} \cup \cdots \cup X_{J,t}\}$. Assume that the density $f_X(\mathbf{X}_t)$ is continuous w.r.t. Lebesgue measure and strictly positive on $\mathcal{X}$ and that the support of $\mathbf{X}_t$ is not contained in any proper linear subspace of $\mathbb{R}^d$ almost surely.

(ii) Assume that the density of $\mathbf{X}_t$, $f_X(\cdot)$, admits r continuous partial derivatives on $\mathcal{X}$, with $r > d$, which are uniformly bounded on $\mathcal{X}$. Moreover, for every $y \in R_y$, the support of y_{t+1}, and $\mathbf{x} \in \mathcal{X}$, assume that $F_{t+1}(y|\mathbf{x})$ admits r continuous partial derivatives in y and $\mathbf{x}$ with $r > d$, which are all uniformly bounded on R_y and $\mathcal{X}$.

(iii) The univariate kernel function $K(u)$ is symmetric, has compact support, and satisfies $\int K(u)du = 1$, $\int u^l K(u)du = 0$ for $0 < l \leq (r-1)$, $\int |u^{2r}K(u)|du < \infty$.

Assumption A.5. Let $P, R \to \infty$ and $P/R \to \pi$ with $0 < \pi < \infty$. Assume that **(i)** $Ph^{2r} \to 0$ and **(ii)** $Ph^{2d} \to \infty$ for $r > d$, where r is defined in Assumption A.4.

Assumption A.4 requires that $\mathbf{X}_t$ contains d predictors that are distinct almost surely and imposes conditions on the smoothness of the conditional distribution of y_{t+1} given $\mathbf{X}_t$ that are standard in the nonparametric estimation literature. Together with the bandwidth conditions in Assumption A.5 it allows for the use of empirical process results from Andrews and Pollard (1994) for dependent data, and also ensures that:

$$\sup_{x \in \mathcal{X}} \left|\widehat{f}_X(x) - f_X(x)\right| = o_p(1),$$

when $\widehat{f}_X(x)$ is a kernel estimator from (9). In addition, note that the condition $Ph^{2d} \to \infty$ with $r > d$ from A.5 requires the use of a fourth order kernel function in A.4 for $d = 2, 3$. This is needed for the analysis of a second order term in the asymptotic expansion of the test statistic, which becomes relevant only in the degenerate case (see below). In a similar spirit, the assumption that $0 < \pi < \infty$ is used only to simplify the convergence rates of some of the asymptotic results, and could be relaxed at the cost of more complex conditions on the relative rates at which P and R grow, see Theorem 1 and its discussion. Finally, note that part A.4(i), which restricts attention to a compact subset of the support of $\mathbf{X}_t$ can be readily relaxed at the cost of using a trimmed estimator on some slowly expanding set and more complex arguments in the proof (e.g., Andrews, 1995).

There are two distinct cases under H_0 of equal expected predictive content, namely:

CASE I:

$$\Pr\left(C_1([\tau_L, \tau_U]; \mathbf{X}_t) = C_2([\tau_L, \tau_U]; \mathbf{X}_t)\right) < 1, \tag{12}$$

CASE II:

$$\Pr\left(C_1([\tau_L, \tau_U]; \mathbf{X}_t) = C_2([\tau_L, \tau_U]; \mathbf{X}_t)\right) = 1. \tag{13}$$

To discuss the above cases in more detail, we parallel Clark and McCracken (2014), and say that models are *overlapping* under H_0 if, for a given quantile level τ, their quantile predictions are identical (in population) with probability one, i.e. $q_\tau(\boldsymbol{\psi}_1^\dagger; X_{1,t}) = q_\tau(\boldsymbol{\psi}_2^\dagger; X_{2,t})$ almost surely.[8] Otherwise, we say that the models are strictly non-nested. With this definition at hand, note that CASE I occurs under H_0 when, for a given τ_L and τ_U, the quantile models are strictly non-nested on $\mathcal{X}$. CASE II, on the other hand, could arise in a variety of scenarios (see footnote 1 below). However, given that we focus on predictions from linear quantile regression and nonlinear location scale models in this paper, CASE II arises when the two models are overlapping, i.e. $q_\tau(\boldsymbol{\psi}_1^\dagger; X_{1,t}) = q_\tau(\boldsymbol{\psi}_2^\dagger; X_{2,t})$ almost surely for the relevant τ. In fact, this latter situation can occur if both models are either correctly or incorrectly specified with respect to the common vector $\mathbf{X}_t$. As an illustrative example of CASE II with correctly specified models, consider for instance model 1, which is given by a linear quantile regression model with predictors $X_{1,t}$:

$$q_\tau(\boldsymbol{\beta}_1^\dagger; X_{1,t}) = \beta_{0,1}^\dagger(\tau) + X_{1,t}'\boldsymbol{\beta}_1^\dagger(\tau),$$

while the competitor model 2 is specified as a linear quantile regression model with different predictors $X_{2,t}$:

$$q_\tau(\boldsymbol{\beta}_2^\dagger; X_{2,t}) = \beta_{0,2}^\dagger(\tau) + X_{2,t}'\boldsymbol{\beta}_2^\dagger(\tau).$$

These models are overlapping and correctly specified when $X_{1,t}$ and $X_{2,t}$ are irrelevant regressors in the actual DGP. That is, assume that:

$$q_\tau(\mathcal{F}_t) = X_{1,t}'\boldsymbol{\beta}_1 + X_{2,t}'\boldsymbol{\beta}_2 + q_\tau(e_{t+1}), \tag{14}$$

where $q_\tau(e_{t+1})$ is the τ-level quantile for some e_{t+1} independent of $X_{1,t}$ and $X_{2,t}$. It is immediate to see that when $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \mathbf{0}$ (assuming $X_{1,t}$ and $X_{2,t}$ are of the same dimension for simplicity), for both $j = 1, 2$ in this case:

$$C_j([\tau_L, \tau_U]; \mathbf{X}_t) = \tau_U - \tau_L$$

almost surely whenever the models contain an intercept so that $\beta_{0,1}^\dagger(\tau) = \beta_{0,2}^\dagger(\tau) = q_\tau(e_{t+1})$. On the other hand, given that $\mathbf{X}_t$ is assumed to contain only distinct predictors by A.4(i), for any non-zero values of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ note that we will either be under CASE I or under the alternative hypothesis. Finally, CASE II with misspecified models (i.e., where conditional coverage is equal a.s., but not equal to the nominal level) could arise if the quantile function in (14) contained a nonlinear function of yet another predictor independent of $X_{1,t}$ and $X_{2,t}$, respectively, whose non-linearity is not picked up in the linear specification (see the additional simulations in the supplementary material for an example).

We now formally establish the limiting distribution of $\widehat{S}_{P,R}(\tau_L, \tau_U)$ and $\widehat{S}_{P,R}(\tau)$:

Theorem 1. *Let Assumptions A.1–A.5 hold. Then:*

(i) *Under H_0, in CASE I with $\Pr(C_1((0,\tau]; \mathbf{X}_t) = C_2((0,\tau]; \mathbf{X}_t) | \mathbf{X}_t \in \mathcal{X}) < 1$:*

$$\widehat{S}_{P,R}(\tau_L, \tau_U) \xrightarrow{d} G(\tau_L, \tau_U)$$

$$\widehat{S}_{P,R}(\tau) \xrightarrow{d} G(\tau),$$

where G is a Gaussian process with variance kernel $\Omega(\tau_L, \tau_U)$, or in the one-sided case, $\Omega(\tau)$, as defined in (28) in Appendix A.

(ii) *Under H_0, in CASE II:*

(a) *if either $\Pr(C_1([\tau_L, \tau_U]; \mathbf{X}_t) = C_2([\tau_L, \tau_U]; \mathbf{X}_t)) = 1$ with $C_j([\tau_L, \tau_U]; \mathbf{X}_t) = \tau_U - \tau_L$ almost surely, or $\Pr(C_1((0,\tau]; \mathbf{X}_t) = C_2((0,\tau]; \mathbf{X}_t)) = 1$ with $C_j((0,\tau]; \mathbf{X}_t) = \tau$ almost surely for $j = 1, 2$, then for any $\Delta > 0$:*

$$\lim_{P,R\to\infty} \Pr\left(|\widehat{S}_{P,R}(\tau_L, \tau_U)| \leq \Delta \frac{\sqrt{P}}{R}\right) = 1$$

$$\lim_{P,R\to\infty} \Pr\left(|\widehat{S}_{P,R}(\tau)| \leq \Delta \frac{\sqrt{P}}{R}\right) = 1.$$

(b) *if either $\Pr(C_1([\tau_L, \tau_U]; \mathbf{X}_t) = C_2([\tau_L, \tau_U]; \mathbf{X}_t)) = 1$ with $\Pr\left(C_j([\tau_L, \tau_U]; \mathbf{X}_t) = \tau_U - \tau_L\right) < 1$, or $\Pr(C_1((0,\tau]; \mathbf{X}_t) = C_2((0,\tau]; \mathbf{X}_t)) = 1$ with $\Pr\left(C_j((0,\tau]; \mathbf{X}_t) = \tau\right) < 1$ for $j = 1, 2$, then:*

$$\widehat{S}_{P,R}(\tau_L, \tau_U) \xrightarrow{p} 0$$

$$\widehat{S}_{P,R}(\tau) \xrightarrow{p} 0.$$

[8] Note that this definition comprises the strictly nested case when for instance a QAR(1) model is compared against a QAR(2) model, and both models have equal predictive content under H_0.

(iii) *Under H_A, there exists some $\varepsilon > 0$ such that:*

$$\lim_{P,R\to\infty} \Pr\left(\frac{1}{\sqrt{P}}|\widehat{S}_{P,R}(\tau_L, \tau_U)| > \varepsilon\right) = 1$$

$$\lim_{P,R\to\infty} \Pr\left(\frac{1}{\sqrt{P}}|\widehat{S}_{P,R}(\tau)| > \varepsilon\right) = 1.$$

Theorem 1(i) establishes the asymptotic distribution under H_0 in the non-overlapping CASE I. The proof of (i) reveals that, besides the average mis-coverage of both models, two distinct terms that capture the overall estimation error of models 1 and 2, respectively, also enter the asymptotic distribution of the statistic. This overall estimation error can be split into one component that stems from the estimation of the nonparametric coverage and one that derives from the estimation of the parametric quantile model.

On the other hand, Theorem 1(ii) can be further divided into subcases (a) and (b), respectively. Subcase (a) refers to a scenario where coverage errors are identical almost surely (CASE II) and models are correctly specified with respect to $\mathbf{X}_t$ so that coverage is equal to the nominal level. This can for instance arise when both models contain all relevant predictors, but also include irrelevant ones as in the aforementioned example. In this case, the statistic converges to zero in probability at the above specified rate. Technically speaking, this is because the first order derivative term in a second-order Taylor expansion of the statistics (10) or (11) around the population coverage error $\mathcal{E}_j([\tau_L, \tau_U]; \mathbf{X}_t)$ or $\mathcal{E}_j((0, \tau]; \mathbf{X}_t)$, respectively, is equal to zero almost surely, while for any strictly convex loss function satisfying Assumption A.2, the second order term is not. In fact, since $\mathcal{E}_1\left([\tau_L, \tau_U]; \mathbf{X}_j\right) = \mathcal{E}_2\left([\tau_L, \tau_U]; \mathbf{X}_j\right) = 0$ a.s., we have by Assumption A.2 that $L(0) = 0$ as well as $\nabla^{(1)}L(0) = 0$, while $\nabla^{(2)}L(0) = C$ for some positive constant C, so that the second order term becomes the lead term, and converges to zero at the rate specified in Theorem 1(ii)–(a). Note that the condition $0 < \pi < \infty$ from A.5 is not needed for this result, but only serves the purpose to simplify the convergence rates in the statement, which would otherwise depend on the relative behaviour of P and R in a more complex manner.

Similarly, subcase (b) in Theorem 1(ii) refers to a scenario where coverage errors are again identical almost surely, but unlike in subcase (a) both models are misspecified so that for instance in the one-sided interval case $\Pr\left(C_j((0, \tau]; \mathbf{X}_t) = \tau\right) < 1$, $j = 1, 2$. This practically more relevant situation for CASE II may arise when two models that overlap, i.e. $q_\tau(\boldsymbol{\psi}_1^\dagger(\tau); X_{1,t}) = q_\tau(\boldsymbol{\psi}_2^\dagger(\tau); X_{2,t})$ almost surely, are actually misspecified, for instance because both models do not capture the functional form of the DGP correctly.[9] Also in this subcase (b), although the first order derivative of the loss function is no longer zero with probability one, the statistic converges to zero in probability because of stochastic equicontinuity and because parameter estimation bias vanishes at a rate faster than $1/\sqrt{P}$.

Finally, note that the results of Theorem 1(i) and Theorem 1(ii) establish the asymptotic behaviour of the test statistic for DGPs that fall either under CASE I or under CASE II with $C_j([\tau_L, \tau_U]; \mathbf{X}_t) = \tau_U - \tau_L$ or $C_j((0, \tau]; \mathbf{X}_t) = \tau$ with probability one for $j = 1, 2$. In the next section, we will show that bootstrap critical values will provide a test of asymptotic size α in the former case, and of asymptotic size zero in the latter. In principle, one could extend the set of DGPs by also considering sequences of the form $\Pr(C_1([\tau_L, \tau_U]; \mathbf{X}_t) = C_2([\tau_L, \tau_U]; \mathbf{X}_t)) = 1 - b_P$ for some $b_P \to 0$ as $P \to \infty$. This is, however, not a trivial extension, and we leave it for future research.

3.3. Bootstrap critical values

We now suggest a wild bootstrap procedure in the spirit of the conditional p-value approach of Hansen (1996).[10] Importantly, critical values based on our procedure are first order asymptotically valid, regardless of whether we are under CASE I or II. For notational simplicity, we will focus on one-sided intervals $(0, \tau]$ for the remainder of the paper. The extension of the bootstrap statistic to two-sided intervals $[\tau_L, \tau_U]$ and to nonlinear location scale models can be found in the supplementary material.

The wild bootstrap procedure is based on an asymptotic linear representation of the test statistic. More specifically, for $j = 1, 2$, this representation depends on three distinct terms $A_{j,t}(\tau)$, $B_{j,t}(\tau)$, and $D_{j,t}(\tau)$ defined in (24), (25), and (26) in Appendix A. They respectively capture the contribution of the population coverage error, as well as the estimation error for the conditional coverage and that of the coefficients of the parametric linear quantile regression model. Corresponding estimators, which we label $\widehat{A}_{j,P,R,s}(\tau)$, $\widehat{B}_{j,P,R,s}(\tau)$, and $\widehat{D}_{j,P,R,s}(\tau)$, $j = 1, 2$, respectively, can be constructed in a straightforward manner. The exact expressions of these terms can also be found in (29), (30) and (31) of Appendix A.

[9] When quantile predictions are not model-based or the models and the data do not satisfy our assumptions, a second, rather pathological scenario that could lead to equal, but incorrect coverage is when the conditional coverage functions are themselves flat almost surely. This situation can for instance arise with extreme quantile predictions such that, using again the one sided case as an example, (conditional) coverage is either zero or one with probability one.

[10] Note that the use of subsampling based critical values, as in Angrist et al. (2006) or Escanciano and Velasco (2010) for instance, is not viable in macroeconomic applications with small sample sizes, as in the case of GaR. On the other hand, as we discuss in the context of CAViaR models in the supplementary material, when large samples of financial data are used, subsampling may be the preferred option to construct critical values.

Hence, we construct the bootstrap statistic as follows:

$$\widehat{S}^*_{P,R}(\tau) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-l_P-1} \varepsilon_t \left(\sum_{s=t}^{t+l_P} \left(\widehat{A}_{1,P,R,s}(\tau) - \widehat{A}_{2,P,R,s}(\tau) \right) + \left(\widehat{B}_{1,P,R,s}(\tau) - \widehat{B}_{2,P,R,s}(\tau) \right) \right) + \frac{\sqrt{P}}{R} \sum_{t=1}^{R-l_R-1} \eta_t \sum_{s=t}^{t+l_R} \left(\widehat{D}_{1,P,R,s}(\tau) - \widehat{D}_{2,P,R,s}(\tau) \right) \tag{15}$$

where ε_t and η_t are i.i.d. random variables independent of the data and drawn from distributions $N(0, 1/l_P)$ and $N(0, 1/l_R)$, respectively, satisfying $E(\varepsilon_t \eta_s) = 0$ for all t, s. As noted in Corradi and Swanson (2002), the lag truncation parameters l_R and l_P serve the same purpose as the block length in the block bootstrap, and account for time series dependence in the data. In addition, note that ε_t and η_t are statistically independent to capture the independence of estimation error from the parametric quantile model and of the remaining components due to the use of a fixed estimation scheme. This is no longer the case with recursive estimation, as outlined in the supplementary material.

The logic underlying the wild bootstrap statistic $\widehat{S}^*_{P,R}(\tau)$ is the following. In CASE I, conditional on the data, $\widehat{S}^*_{P,R}(\tau)$ behaves as a zero-mean random normal variable with a variance that mimics a heteroskedasticity and autocorrelation consistent (HAC) estimator of the main driving terms, namely the population coverage error component, and the terms representing nonparametric conditional coverage as well as parametric quantile model estimation error.[11] In fact, in the pairwise comparison case we could base inference in principle also on the HAC estimator more directly without relying on bootstrapped critical values. On the other hand, in the multiple comparison case in Section 4, the limiting distribution is non-standard and thus we need to rely on bootstrap critical values.

In CASE II of equal conditional coverage on the other hand, a differentiation between the case of correctly and incorrectly specified models is required: in the case of correctly specified models with coverage equal to the nominal level, the first, second, and third term of (15), respectively converge to zero in probability. Importantly, however, we show in Theorem 2(ii) below that for the case of correct conditional coverage (i.e., equal to the nominal level), the bootstrap statistic converges to zero in probability at a rate which is slower than the actual sample statistic since $l_P, l_R \to \infty$ as $T \to \infty$. This suggests that over-rejection is likely not an issue in practice. If instead l_P and l_R do not grow with the sample size, we could avoid over-rejection by introducing an infinitesimal uniformity factor as in Andrews and Shi (2013). That being said, the simulation exercise in Section 5 suggests that the distribution of the bootstrap statistic collapses slower than the distribution of the sample statistic in CASE II even when $l_P = l_R = 1$. Heuristically, this is because the wild bootstrap statistic is constructed using several additional estimated terms. Although such terms vanish at the same rate as the statistic, in finite samples they introduce additional estimation noise so that the bootstrap distribution dominates that of the sample statistic and thus prevents over rejection, at least in our extensive simulations.[12]

On the contrary, when models are misspecified and conditional coverage is not equal to the nominal level almost surely, it still holds that $\widehat{C}_{1,P,R}(\mathbf{X}_i, \tau) - \widehat{C}_{2,P,R}(\mathbf{X}_i, \tau) = o_p(1)$ and, when models are overlapping, $\| \widehat{\boldsymbol{\beta}}_{1,R}(\tau) - \widehat{\boldsymbol{\beta}}_{2,R}(\tau) \| = o_p(1)$ since $\| \boldsymbol{\beta}^{\dagger}_1(\tau) - \boldsymbol{\beta}^{\dagger}_2(\tau) \| = 0$. In this case, it can be shown that the third term of (15) still gives rise to an asymptotic distribution, while the statistic converges in probability to zero so that the test controls size asymptotically. This is because the bootstrap statistic is constructed using an asymptotic expansion for the contribution of parametric estimation error in the non-overlapping CASE I, which does not collapse whenever $X_{1,t}$ and $X_{2,t}$ exhibit independent variation.

Turning to the construction of corresponding critical values, let $c^*_{B,P,R}(\alpha/2)$ and $c^*_{B,P,R}(1-\alpha/2)$ denote the $\alpha/2$ and $1-\alpha/2$ critical values of the empirical distribution of the wild bootstrap statistic, based on B replications. The following result summarizes the above discussion about the behaviour of the test under H_0, and shows that it is consistent against fixed alternatives:

Theorem 2. *Let Assumptions A.1–A.5 hold. Also, as $P, R, B \to \infty$, $l_R, l_P \to \infty$, $l_P/\sqrt{P} \to 0$ and $l_R/\sqrt{R} \to 0$. Then:*

(i) *Under H_0, in CASE I with $\Pr(C_1((0,\tau];\mathbf{X}_t) = C_2((0,\tau];\mathbf{X}_t)|\mathbf{X}_t \in \mathcal{X}) < 1$:*

$$\lim_{P,R,B\to\infty} \Pr\left(c^*_{B,P,R}(\alpha/2) < \widehat{S}_{P,R}(\tau) < c^*_{B,P,R}(1-\alpha/2) \right) = 1-\alpha.$$

(ii) *Under H_0, in CASE II:*

$$\lim_{P,R,B\to\infty} \Pr\left(c^*_{B,P,R}(\alpha/2) < \widehat{S}_{P,R}(\tau) < c^*_{B,P,R}(1-\alpha/2) \right) = 1.$$

(iii) *Under H_A:*

$$\lim_{P,R,B\to\infty} \Pr\left(c^*_{B,P,R}(\alpha/2) < \widehat{S}_{P,R}(\tau) < c^*_{B,P,R}(1-\alpha/2) \right) = 0.$$

[11] It is noteworthy that the term used to construct $\widehat{A}_{j,P,R,s}(\tau)$, $j = 1, 2$, has been recentered to mimic the asymptotic variance also under the alternative, see (29) in Appendix A. Otherwise, under the alternative when $\mathrm{E}\left(\left(L\left(\mathcal{E}_1((0,\tau];\mathbf{X}_t) \right) - L\left(\mathcal{E}_2((0,\tau];\mathbf{X}_t) \right) \right) 1\{\mathbf{X}_t \in \mathcal{X}\} \right) \neq 0$, the bootstrap variance would diverge at rate $\sqrt{l_P}$, thus lowering the power.

[12] On the other hand, as shown in the simulation exercises of Section 5 and the supplementary material, the choice of these parameters does matter more in CASE I for the control of size at the nominal level.

Theorem 2 establishes the first order validity of the block bootstrap critical values. In particular, we have a test of size α in CASE I, and a test of size zero in CASE II.

3.4. Local power

The results from Theorems 1 and 2 suggest that the pairwise comparison test has asymptotic power against fixed alternatives. In this subsection, we analyse its local power properties by considering a drifting sequence in the coverage error differential. For simplicity, we comment on the two-sided case only in the proof and continue to focus on quantile regression models exclusively since the arguments for location scale models are similar. Thus, we define the following drifting sequence:

$$\left(L\left(\mathcal{E}_1\left((0,\tau];\mathbf{X}_t\right)\right)+\frac{\delta_1((0,\tau];\mathbf{X}_t)}{\sqrt{P}}\right)-\left(L\left(\mathcal{E}_2\left((0,\tau];\mathbf{X}_t\right)\right)+\frac{\delta_2((0,\tau];\mathbf{X}_t)}{\sqrt{P}}\right)$$

for some (measurable) continuous functions $\delta_1((0,\tau];\cdot)$ and $\delta_2((0,\tau];\cdot)$. The sequence of local alternatives $H_{A,P}$ is given by:

$$\begin{aligned}H_{A,P}:&\mathrm{E}\left(\left(\left(L\left(\mathcal{E}_1\left((0,\tau];\mathbf{X}_t\right)\right)+\frac{\delta_1((0,\tau];\mathbf{X}_t)}{\sqrt{P}}\right)-\left(L\left(\mathcal{E}_2\left((0,\tau];\mathbf{X}_t\right)\right)+\frac{\delta_2((0,\tau];\mathbf{X}_t)}{\sqrt{P}}\right)\right)1\{\mathbf{X}_t\in\mathcal{X}\}\right)\\&=\frac{\mathrm{E}\left(\left(\delta_1((0,\tau];\mathbf{X}_t)-\delta_2((0,\tau];\mathbf{X}_t)\right)1\{\mathbf{X}_t\in\mathcal{X}\}\right)}{\sqrt{P}}\equiv\frac{\zeta(\tau)}{\sqrt{P}}.\end{aligned}$$

The next result establishes that the pairwise test for a single interval $(0,\tau]$ has non-trivial power against the local alternatives as defined under $H_{A,P}$. For brevity, we state this result for CASE I only, and comment on CASE II below.

Theorem 3. *Let Assumptions A.1–A.5 hold, and $\zeta(\tau)\neq 0$. Then, under $H_{A,P}$:*

$$\begin{aligned}&\lim_{P,R,B\to\infty}Pr\left(\widehat{S}_{P,R}(\tau)>c^*_{B,P,R}(1-\alpha/2)\right)=G_{DF}\left(\zeta(\tau)-c(1-\alpha/2)\right)\\&\lim_{P,R,B\to\infty}Pr\left(\widehat{S}_{P,R}(\tau)<-c^*_{B,P,R}(1-\alpha/2)\right)=G_{DF}\left(-\zeta(\tau)-c(1-\alpha/2)\right),\end{aligned}$$

where $G_{DF}(\cdot)$ denotes the distribution function of the Gaussian process from Theorem 1(i) with variance $\Omega(\tau)$, and $c(1-\alpha/2)=\lim_{P,R,B\to\infty}c^*_{B,P,R}(1-\alpha/2)$ is the corresponding $(1-\alpha/2)$ critical value from the bootstrap empirical distribution.

Note that in CASE II with coverage equal to the nominal level on the other hand, we have that:

$$L\left(\mathcal{E}_1\left((0,\tau];\mathbf{X}_t\right)\right)-L\left(\mathcal{E}_2\left((0,\tau];\mathbf{X}_t\right)\right)=0$$

almost surely. Thus, we may define instead a sequence of 'locally overlapping' alternatives where:

$$H^{ov}_{A,P}:\left(L\left(\mathcal{E}_{1,P}\left((0,\tau];\mathbf{X}_t\right)\right)-L\left(\mathcal{E}_{2,P}\left((0,\tau];\mathbf{X}_t\right)\right)\right)1\{\mathbf{X}_t\in\mathcal{X}\}\equiv\frac{\varsigma(\tau,\mathbf{X}_t)\,1\{\mathbf{X}_t\in\mathcal{X}\}}{\sqrt{P}}$$

with $\varsigma(\tau,\mathbf{X}_t)\,1\{\mathbf{X}_t\in\mathcal{X}\}\neq 0$ almost surely. Here, provided that:

$$p\lim_{P\to\infty}\frac{1}{P}\sum_{t=R+1}^{T}\varsigma(\tau,\mathbf{X}_t)\,1\{\mathbf{X}_t\in\mathcal{X}\}\equiv\varsigma(\tau)\neq 0,$$

we obtain that $\widehat{S}_{P,R}(\tau)=\varsigma(\tau)+o_p(1)\neq 0$, while the bootstrap statistic $\widehat{S}^*_{P,R}(\tau)$ converges to zero in probability. Asymptotic power against the local alternative in $H^{ov}_{A,P}$ in the case of correct nominal coverage is therefore guaranteed, while this does not necessarily hold for CASE II when models are misspecified.

4. Multiple model & Interval comparison

So far our main results involve a model comparison of two models over a single interval. This section extends and generalizes the setting to consider the cases of pairwise model comparisons over multiple intervals and multiple models, nesting the pairwise comparison over multiple intervals or the multiple model comparison at a single interval as special cases. Specifically, the extension to multiple intervals is useful if there is no established nominal level to use in prediction intervals such as in GaR or other forecasting applications (e.g. Clements, 2014; Wang and Wu, 2012). The extension to multiple models, on the other hand, is relevant in light of the increasing number of proposed models in the empirical VaR and GaR literature. The case of testing over both multiple models and multiple intervals can also accommodate a wide range of scenarios, as in our empirical application where we want compare multiple competing GaR models at a range of possible quantile levels.

4.1. Null hypothesis and statistic

Hereafter, let model 1 be the benchmark model given by $q_\tau(\boldsymbol{\psi}_1^\dagger; X_{1,t})$, out of total set of J models with the competitor models given by $q_\tau(\boldsymbol{\psi}_j^\dagger; X_{j,t})$ for $j = 2, \ldots, J$. We compare the relative conditional coverage error of the latter models to the benchmark over M intervals, $(0, \tau_i]$, $i = 1, \ldots, M$.[13] The null hypothesis is that none of the competing models has smaller expected conditional coverage error than the benchmark at any of the intervals considered. The alternative is that at least one competitor outperforms the benchmark for at least one interval. In other words, denoting $\mathbf{X}_t^j = \{X_{1,t}, X_{j,t}\}$ as the union of the conditioning set of the benchmark and model j, the null and alternative hypothesis can be written as:

$$H_0^{RC} : \max_{j=2,\ldots,J} \max_{i=1,\ldots,M} \mathrm{E}\left(\left(L\left(\mathcal{E}_1\left((0,\tau_i];\mathbf{X}_t^j\right)\right) - L\left(\mathcal{E}_j\left((0,\tau_i];\mathbf{X}_t^j\right)\right)\right) 1\left\{\mathbf{X}_t^j \in \mathcal{X}\right\}\right) \leq 0$$

versus:

$$H_A^{RC} : \max_{j=2,\ldots,J} \max_{i=1,\ldots,M} \mathrm{E}\left(\left(L\left(\mathcal{E}_1\left((0,\tau_i];\mathbf{X}_t^j\right)\right) - L\left(\mathcal{E}_j\left((0,\tau_i];\mathbf{X}_t^j\right)\right)\right) 1\left\{\mathbf{X}_t^j \in \mathcal{X}\right\}\right) > 0.$$

Note that the above formulation of the null hypothesis in terms of weak inequalities is standard in comparison tests of multiple models (e.g., White, 2000; Hansen, 2005), and differs from the null hypothesis based on an equality in the pairwise case. Specifically, although a formulation with equalities is informative under the null hypothesis, it may not be interpretable under the alternative when the benchmark is either beaten or beats at least a competitor model, and may in fact also be rejected in a situation where some models beat the benchmark, while others do not.

The above hypothesis can be re-written as:

$$H_0^{RC} = \cap_{j=2}^{J} \cap_{i=1}^{M} H_{0,i,j}^{RC} \tag{16}$$

and:

$$H_A^{RC} = \cup_{j=2}^{J} \cup_{i=1}^{M} H_{0,i,j}^{RC,c}$$

with $H_{0,j,i}^{RC,c}$ denoting the complement of $H_{0,j,i}^{RC}$, and:

$$H_{0,i,j}^{RC} : \mathrm{E}\left(\left(L\left(\mathcal{E}_1\left((0,\tau_i];\mathbf{X}_t^j\right)\right) - L\left(\mathcal{E}_j\left((0,\tau_i];\mathbf{X}_t^j\right)\right)\right) 1\left\{\mathbf{X}_t^j \in \mathcal{X}^j\right\}\right) \leq 0.$$

As mentioned before, the general formulation of the null hypothesis with multiple models and intervals nests various other hypotheses as special cases. For instance, setting $J = 2$, we may compare the relative conditional coverage error of models 1 and 2 over M intervals, $(0, \tau_i]$, $i = 1, \ldots, M$. These intervals could arise when evaluating GaR predictions at multiple quantile levels such as 10%, 20% and 30%. On the other hand, setting $M = 1$, we may also test the null that no competing model has better predictive accuracy than the benchmark at a given interval in terms of expected conditional coverage error, for a given loss function $L(\cdot)$. This type of comparison is useful in light of the recent GaR literature (for instance Brownlees and Souza, 2021) where multiple possible models are evaluated for a single GaR quantile level and corresponds to a standard Reality Check test (White, 2000).[14]

Since our composite null hypothesis consists of a finite number of weak inequalities, we can follow the set-up of Andrews and Soares (2010). Specifically, since the number of moment weak inequalities is given by the product of the number of competing models times the number of intervals, $(J-1) \times M$, the statistic reads as:

$$\widehat{S}_{P,R}^{\max} = \sum_{j=2}^{J} \sum_{i=1}^{M} \left(\max\left\{0, \widehat{S}_{P,R}(\tau_i; j)\right\}\right)^2 \tag{17}$$

where:

$$\widehat{S}_{P,R}(\tau_i; j) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(L\left(\widehat{\mathcal{E}}_{1,P,R}\left((0,\tau_i];\mathbf{X}_t^j\right)\right) - L\left(\widehat{\mathcal{E}}_{j,P,R}\left((0,\tau_i];\mathbf{X}_t^j\right)\right)\right) 1\left\{\mathbf{X}_t^j \in \mathcal{X}^j\right\}. \tag{18}$$

Note that when using $\widehat{S}_{P,R}^{\max}$ in (17), we consider only those intervals for which $\widehat{S}_{P,R}(\tau_i; j) \geq 0$. The purpose of this trimming is to consider only intervals where violation of the null hypothesis is observed and the corresponding population moment $\mathrm{E}\left(\left(L\left(\mathcal{E}_1\left((0,\tau_i];\mathbf{X}_t\right)\right) - L\left(\mathcal{E}_2\left((0,\tau_i];\mathbf{X}_t\right)\right)\right) 1\left\{\mathbf{X}_t \in \mathcal{X}\right\}\right)$ is positive. In the next part we derive the limiting distribution of this statistic.

[13] The extension to two-sided intervals $[\tau_{i,L}, \tau_{i,U}]$ can again be found in the supplementary material.

[14] One could also consider comparisons involving model combinations from the list of competitors as well as the individual models themselves.

4.2. Limiting distribution

We now establish the limiting behaviour of this test for multiple models and multiple intervals given in (18). For simplicity, we will again outline the case where the benchmark and the competitor models are linear quantile regression models. The extension to location scale models follows by identical arguments to before. We consider the following cases:

CASE I-RC: For at least one interval $i \in \{1, \dots, M\}$ and one model $j \in \{2, \dots, J\}$:

$$\Pr\left(C_1((0,\tau_i];\mathbf{X}_t^j) = C_j((0,\tau_i];\mathbf{X}_t^j)\right) < 1.$$

CASE II-RC: For all $i = 1, \dots, M$ and $j = 2, \dots, J$:

$$\Pr\left(C_1((0,\tau_i];\mathbf{X}_t^j) = C_j((0,\tau_i];\mathbf{X}_t^j)\right) = 1.$$

Hence, in CASE I-RC there is at least one model which does not overlap with the benchmark for at least one interval, which may in fact arise if one competitor model is strictly non-nested for at least one interval. By contrast, in CASE II-RC, all competing models overlap with the benchmark model over all intervals. As before, for each model comparison pair j, the competitor and the benchmark can in principle either be misspecified or correctly specified with respect to the relevant conditioning set, $\mathbf{X}_t^j$.

Hereafter, let $\mathcal{P}$ denote the set of probability measures, $\boldsymbol{P}$ (= Pr), defined on the support of $\mathbf{X}_t$ such that Assumptions A.1, A.3 and A.4 hold, and let $\mathcal{P}_0^{RC} = \left\{\boldsymbol{P} \in \mathcal{P} : H_{0,P}^{RC} \text{ holds}\right\}$ denote the set of all DGPs under the null hypothesis. Also, denote:

$$\mathcal{P}_0^{I-RC} = \left\{\boldsymbol{P} \in \mathcal{P} : H_{0,P}^{RC} \text{ and CASE I-RC holds}\right\},$$

as the set of null DGPs such that CASE I-RC holds. Similarly:

$$\mathcal{P}_0^{IIa-RC} = \left\{\boldsymbol{P} \in \mathcal{P} : H_{0,P}^{RC} \text{ and CASE II-RC holds with } C_1((0,\tau_i];\mathbf{X}_t^j) = \tau_i \text{ a.s. for all } j = 2, \dots, J\right\}$$

and:

$$\mathcal{P}_0^{IIb-RC} = \left\{\boldsymbol{P} \in \mathcal{P} : H_{0,P}^{RC} \text{ and CASE II-RC holds with } \boldsymbol{P}\left(C_1((0,\tau_i];\mathbf{X}_t^j) = \tau_i\right) < 1 \text{ for at least one interval } i \in \{1, \dots, M\} \text{ and one model } j \in \{2, \dots, J\}\right\}$$

denote the set of null DGPs such that CASE II-RC with either $C_1((0,\tau_i];\mathbf{X}_t^j) = \tau_i$ a.s. for all $j = 2, \dots, J$ or $\boldsymbol{P}\left(C_1((0,\tau_i];\mathbf{X}_t^j) = \tau_i\right) < 1$ for at least one interval $i \in \{1, \dots, M\}$ and one model $j \in \{2, \dots, J\}$ holds. Thus, $\mathcal{P}_0^{RC}$ is formed by the union of the sets $\mathcal{P}_0^{I-RC}$, $\mathcal{P}_0^{IIa-RC}$, and $\mathcal{P}_0^{IIb-RC}$, respectively. Below we will establish the validity of bootstrap critical values uniformly over all DGPs in CASE I-RC and in CASE II-RC, respectively. Finally, denote $\mathcal{P}_A^{RC} = \left\{\boldsymbol{P} \in \mathcal{P} : H_A^{RC} \text{ holds}\right\}$ the set of DGPs under the alternative hypothesis.

To establish the limiting distribution of $\widehat{S}_{P,R}^{\max}$, denote by $\boldsymbol{V}$ the asymptotic $M(J-1) \times M(J-1)$ dimensional variance–covariance matrix for CASE I-RC, whose principal diagonal elements are v_{kk} with $k = (j-2)M + i$ with $i = 1, \dots, M$ and $j = 2, \dots, J$, and whose off-diagonal elements are given by $v_{kk'}$, with $k' = (j'-2)M + i'$ and $i' \neq i$ and/or $j' \neq j$ (see (32) and (33) in the Appendix A for an exact definition of v_{kk} and $v_{kk'}$).

We are now ready to study the limiting behaviour of the test statistic. That is, recalling the definition $\widehat{S}_{P,R}(\tau_i; j)$ in (18), we denote:

$$\begin{pmatrix} \widehat{S}_{P,R,1} \\ \vdots \\ \widehat{S}_{P,R,M(J-1)} \end{pmatrix} = \begin{pmatrix} \widehat{S}_{P,R}(\tau_1; 2) \\ \vdots \\ \widehat{S}_{P,R}(\tau_M; J-1) \end{pmatrix}$$

so that $\widehat{S}_{P,R}^{\max}$ can be written as:

$$\widehat{S}_{P,R}^{\max} = \sum_{k=1}^{(J-1)M} \left(\max\left\{0, \widehat{S}_{P,R,k}\right\}\right)^2.$$

Also, for each $k = (j-2)M + i$, $i = 1, \dots, M$ and $j = 2, \dots, J$, let $\varsigma_k = \lim_{P\to\infty} \mu_{k,P}$ with:

$$\mu_{k,P} = \sqrt{P}\mathrm{E}_{\boldsymbol{P}}\left(\left(L\left(\mathcal{E}_1\left((0,\tau_i];\mathbf{X}_t^j\right)\right) - L\left(\mathcal{E}_j\left((0,\tau_i];\mathbf{X}_t^j\right)\right)\right) 1\left\{\mathbf{X}_t^j \in \mathcal{X}^j\right\}\right) \tag{19}$$

and note that, for a fixed $\boldsymbol{P} \in \mathcal{P}_0^{RC}$, we have either $\varsigma_k = 0$ or $\varsigma_k = -\infty$, depending on whether $\mu_{k,P} = 0$ or $\mu_{k,P} < 0$, respectively. The limiting behaviour of $\widehat{S}_{P,R}^{\max}$ is given by the following theorem:

Theorem 4. *Let Assumptions A.1–A.5 hold and let Z_k denote the kth element of a $M(J-1)$-dimensional zero mean normal random vector with variance–covariance matrix equal to $\boldsymbol{V}$ as defined in (32) and (33) in Appendix A. Assume that $\boldsymbol{V}$ is positive semidefinite. Then:*

(i) *Under H_0^{RC}, for a given $\boldsymbol{P} \in \mathcal{P}_0^{I-RC}$ with $\Pr\left(C_1((0,\tau_i];\mathbf{X}_t^j) = C_j((0,\tau_i];\mathbf{X}_t^j)|\mathbf{X}_t^j \in \mathcal{X}\right) < 1$ for at least one $j \in \{2,\dots,J\}$ and one $i \in \{1,\dots,M\}$:*

$$\widehat{S}_{P,R}^{\max} \xrightarrow{d} \sum_{k=1}^{M(J-1)} \left(\max\{0, Z_k + \varsigma_k\}\right)^2,$$

where $Z_k + \varsigma_k = -\infty$ when $\varsigma_k = -\infty$.

(ii) *Under H_0^{RC}:*

(a) For a given $\boldsymbol{P} \in \mathcal{P}_0^{IIa-RC}$ and any $\Delta > 0$:

$$\lim_{P,R\to\infty} \boldsymbol{P}\left(\widehat{S}_{P,R}^{\max} \leq \Delta \frac{\sqrt{P}}{R}\right) = 1.$$

(b) For a given $\boldsymbol{P} \in \mathcal{P}_0^{IIb-RC}$ and any $\Delta > 0$:

$$\widehat{S}_{P,R}^{\max} \xrightarrow{\boldsymbol{P}} 0.$$

(iii) Under H_A^{RC}, there exists $\varepsilon > 0$, such that for a given $\boldsymbol{P} \in \mathcal{P}_A^{RC}$

$$\lim_{P,R\to 1} \boldsymbol{P}\left(\frac{1}{\sqrt{P}}\widehat{S}_{P,R}^{\max} > \varepsilon\right) = 1.$$

Theorem 4 provides the analogous results for the multiple intervals and models case as those in Theorem 1. Specifically, under CASE I-RC the asymptotic distribution is driven by all non-slack models, and also reflects the two layers of estimation error across all models. Notice, however, that the statements in Theorem 4 only hold for a fixed distribution $\boldsymbol{P} \in \mathcal{P}$, while our goal is to make inference uniformly over all DGPs within the sets $\mathcal{P}_0^{I-RC}$, $\mathcal{P}_0^{IIa-RC}$, and $\mathcal{P}_0^{IIb-RC}$, respectively.

Indeed, as we cannot consistently estimate ς_k uniformly over $\mathcal{P}_0^{I-RC}$, we now introduce the wild bootstrap statistic which, conditional on the sample, properly mimics the limiting distribution of $\widehat{S}_{P,R}^{\max}$ uniformly over all DGPs in the respective null set. We then show that inference based on wild bootstrap critical values is asymptotically correct and non-conservative whenever at least one moment condition holds with equality. Thus, recall the different components of the bootstrap statistic $\widehat{A}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)$, $\widehat{B}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)$, and $\widehat{D}_{j,t}\left(\tau_i\right)$ from Section 3.3, where we make the dependence in $\widehat{A}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)$ and $\widehat{B}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)$ on a specific conditioning set $\mathbf{X}_t^j$ now explicit, and suppress the dependence on P and R to avoid further notational clutter. That is, given $\widehat{A}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)$, $\widehat{B}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)$, and $\widehat{D}_{j,t}\left(\tau_i\right)$ for each competitor model, j, and interval, i, we define the bootstrap statistic for the pairwise comparison with the benchmark as:

$$\begin{aligned}\widehat{S}_{P,R}^{*}(\tau_i;j) = \frac{1}{\sqrt{P}}\sum_{t=R}^{T-l_P-1}\varepsilon_t\left(\sum_{s=t}^{t+l_P}\left(\widehat{A}_{1,s}\left(\tau_i,\mathbf{X}_s^j\right) - \widehat{A}_{j,s}\left(\tau_i,\mathbf{X}_s^j\right) + \widehat{B}_{1,s}\left(\tau_i,\mathbf{X}_s^j\right) - \widehat{B}_{j,s}\left(\tau_i,\mathbf{X}_s^j\right)\right)\right)\\ + \frac{\sqrt{P}}{R}\sum_{t=1}^{R-l_R-1}\eta_t\left(\sum_{s=t}^{t+l_R}\left(\widehat{D}_{1,s}\left(\tau_i\right) - \widehat{D}_{j,s}\left(\tau_i\right)\right)\right)\end{aligned}$$

where ε_t and η_t are as defined in Section 3.3. The bootstrap statistics for each pairwise comparison become:

$$\begin{pmatrix}\widehat{S}_{P,R,1}^{*}\\ \vdots\\ \widehat{S}_{P,R,M(J-1)}^{*}\end{pmatrix} = \begin{pmatrix}\widehat{S}_{P,R}^{*}(\tau_1;2)\\ \vdots\\ \widehat{S}_{P,R}^{*}(\tau_M;J-1)\end{pmatrix}$$

Finally, for the moment selection, we also need to construct a HAC estimator of the diagonal elements v_{kk} of the asymptotic variance. We denote this estimator by $\widehat{v}_{kk,P,R}$, and refer the reader again to (34) for its exact definition. The final bootstrap statistic for the test now reads as:

$$\widehat{S}_{P,R}^{*\max} = \sum_{k=1}^{M(J-1)}\left(\max\left\{0,\widehat{S}_{P,R,k}^{*}1\left\{\widehat{S}_{P,R,k} \geq -\sqrt{\widehat{v}_{kk,P,R}}\cdot\kappa_P\right\}\right\}\right)^2$$

with $\kappa_P \to \infty$ as $P \to \infty$. Thus, following Andrews and Soares (2010), the bootstrap statistic discards "poor" model alternatives k which are clearly dominated by the benchmark. Specifically, whenever the event $\widehat{S}_{P,R,k} < -\sqrt{\widehat{v}_{kk,P,R}}\cdot\kappa_P$

occurs, it holds with probability one that:

$$\max\left\{0, \widehat{S}^{*}_{P,R,1} 1\left\{\widehat{S}_{P,R,k} \geq -\sqrt{\widehat{v}_{kk,P,R}} \cdot \kappa_P\right\}\right\} = 0.$$

As a result, sufficiently negative empirical moment conditions do not contribute to the bootstrap statistic, and the wild bootstrap statistic asymptotically eliminates inequalities that are too slack. This trimming procedure is common in the superior predictive ability testing literature (e.g., Hansen, 2005), and helps to enhance the power properties of these tests in finite samples. In this context, the fact that $\widehat{A}_{1j,t}\left(\tau_i, \mathbf{X}^j_t\right)$, which comprises of $\widehat{A}_{1,t}\left(\tau_i, \mathbf{X}^j_t\right)$ and $\widehat{A}_{j,t}\left(\tau_i, \mathbf{X}^j_t\right)$ as defined in (29), is centred around its sample mean plays an additional role. In fact, it ensures that $\widehat{v}_{kk,P,R}$ converges to the "true" variance of the kth moment condition. Without recentering, $\widehat{v}_{kk,P,R}$ is an estimator of the second moment, and it would diverge to infinity at rate $\sqrt{l_P}$, whenever $\mu_k < 0$. As a consequence, we could fail to eliminate (a subset of) slack moment conditions.

Let $c^{*\max}_{B,R,P,1-\alpha}$ be the $(1-\alpha)$-th critical values of the empirical distribution of $\widehat{S}^{*\max}_{P,R}$ based on B replications. We have:

Theorem 5. *Let Assumptions A.1–A.5 hold and assume that $\boldsymbol{V}$ from Theorem 4 is positive semidefinite. Also, as $P, R, B \to \infty$, $l_R, l_P \to \infty$, $l_P/\sqrt{P} \to 0$ and $l_R/\sqrt{R} \to 0$, $\frac{\kappa_P}{\log\log P} \to \infty$ and $\kappa_P/\sqrt{P} \to 0$. Then:*

(i) *Under H^{RC}_0, in CASE I-RC with $\Pr\left(C_1((0,\tau_i];\mathbf{X}^j_t) = C_j((0,\tau_i];\mathbf{X}^j_t)|\mathbf{X}^j_t \in \mathcal{X}\right) < 1$ for at least one $j \in \{2,\ldots,J\}$,*

$$\lim_{B,P,R\to\infty} \sup_{\boldsymbol{P}\in\mathcal{P}^{I-RC}_0} \boldsymbol{P}\left(\widehat{S}^{\max}_{P,R} \geq c^{*\max}_{B,R,P,1-\alpha}\right) \leq \alpha$$

and if for some k, $\mu_{k,P} = 0$, with $\mu_{k,P}$ as in (19),

$$\lim_{B,P,R\to\infty} \sup_{\boldsymbol{P}\in\mathcal{P}^{I-RC}_0} \boldsymbol{P}\left(\widehat{S}^{\max}_{P,R} \geq c^{*\max}_{B,R,P,1-\alpha}\right) = \alpha.$$

(ii) *Under H^{RC}_0, in CASE II-RC for $k = a, b$:*

$$\lim_{B,P,R\to\infty} \sup_{\boldsymbol{P}\in\mathcal{P}^{IIk-RC}_0} \boldsymbol{P}\left(\widehat{S}^{\max}_{P,R} \geq c^{*\max}_{B,R,P,1-\alpha}\right) = 0.$$

(iii) *Under H^{RC}_A and a given $\boldsymbol{P} \in \mathcal{P}^{RC}_A$:*

$$\lim_{B,R,P\to\infty} \boldsymbol{P}\left(\widehat{S}^{\max}_{P,R} \geq c^{*\max}_{B,R,P,1-\alpha}\right) = 1.$$

It is immediate to see from the statement in Theorem 5(i)–(ii) that wild bootstrap based critical values are asymptotically valid uniformly over all DGPs within each case under H^{RC}_0. In other words, Theorem 5(i) establishes that the test controls size at level α uniformly over $\boldsymbol{P} \in \mathcal{P}^{I-RC}_0$ as $P \to \infty$ when at least one moment is binding, and at level at most α otherwise. On the other hand, Theorem 5(ii) shows that, regardless of whether $\boldsymbol{P} \in \mathcal{P}^{IIa-RC}_0$ or $\boldsymbol{P} \in \mathcal{P}^{IIb-RC}_0$, we obtain a test that is of asymptotic size zero. Moreover, we note that in CASE I-RC, whenever at least one model has the same coverage error as the benchmark over at least one interval, then inference is no longer asymptotically conservative, and wild bootstrap critical values provide a test with correct asymptotic size.[15]

Finally, note that comparisons across different forecasting horizons have become increasingly important in recent years (e.g., Clark et al., 2020; Quaedvlieg, 2021). The bootstrap procedure developed in this section may indeed be straightforwardly adapted to accommodate such a multi-horizon set-up, where the test is carried out simultaneously across horizons $s = 1,\ldots,S$. Formally, we may be interested in testing that none of the competing models has smaller expected conditional coverage error at any of the intervals across all forecasting horizons against the alternative that there exists at least one competitor model that outperforms the benchmark for at least one interval and one horizon. That is:

$$\max_{j=2,\ldots,J}\max_{i=1,\ldots,M}\max_{s=1,\ldots,S} \mathrm{E}\left(\left(L\left(\mathcal{E}_{1,s}\left((0,\tau_i];\mathbf{X}^j_t\right)\right) - L\left(\mathcal{E}_{j,s}\left((0,\tau_i];\mathbf{X}^j_t\right)\right)\right) 1\left\{\mathbf{X}^j_t \in \mathcal{X}\right\}\right) \leq 0$$

versus:

$$\max_{j=2,\ldots,J}\max_{i=1,\ldots,M}\max_{s=1,\ldots,S} \mathrm{E}\left(\left(L\left(\mathcal{E}_{1,s}\left((0,\tau_i];\mathbf{X}^j_t\right)\right) - L\left(\mathcal{E}_{j,s}\left((0,\tau_i];\mathbf{X}^j_t\right)\right)\right) 1\left\{\mathbf{X}^j_t \in \mathcal{X}\right\}\right) > 0,$$

where $\mathcal{E}_{j,s}\left((0,\tau_i];\mathbf{X}^j_t\right), j = 1,\ldots,J$, denotes the coverage error of model j for the interval $(0,\tau_i]$ at horizon s. The key point is to use the same draws for ε_t and η_t across horizons $s = 1,\ldots,S$ in the construction of the wild bootstrap statistic. In this way, the dependence among moment inequalities across forecasting horizons is properly captured and results akin to Theorem 5 hold.

[15] Note that based on the results of Theorems 4 and 5, a local power statement could be formulated in analogy to Theorem 3, though we omit this for brevity reasons.

5. Monte Carlo simulation

In this section we run Monte Carlo simulations to explore the size and power of our tests in finite samples. We exploit a simple DGP for the target variable y_{t+1} which allows us to explore the results from the theory above for CASE I and II under the null, and the alternative. The DGP is a linear model involving only two predictors $X_{1,t}$ and $X_{2,t}$:

$$y_{t+1} = \beta_1 X_{1,t} + \beta_2 X_{2,t} + e_{t+1}. \tag{20}$$

In order to make the simulations as realistic as possible, we allow y_{t+1} to have serial correlation as well as the possibility that $X_{1,t}$ and $X_{2,t}$ are correlated, as is the case in most economic and financial applications. We therefore let $X_{1,t}$ and $X_{2,t}$ follow autoregressive processes $X_{j,t} = \rho X_{j,t-1} + v_{j,t}$ for $j = 1, 2$, so that y_{t+1} also has serial dependence through the parameter ρ. We let the errors $v_{j,t}$ follow a multivariate normal distribution with variance equal to $1-\rho^2$ and covariance equal to ϕ, so that they are allowed to be correlated by varying the parameter ϕ and both have an unconditional variance equal to unity. In the baseline results the time dependence is set to $\rho = 0.5$ and there is no correlation among the $X_{j,t}$ variables ($\phi = 0$) but we will also discuss the results with higher serial correlation ($\rho = 0.7, 0.9$) and correlated $X_{j,t}$ variables ($\phi = 0.25$). The error term e_{t+1} in (20) is drawn from a standard normal distribution and we will check the robustness of the results to this error distribution.

We will first evaluate the performance of the pairwise model comparison test by making quantile predictions using two quantile regression models (later we will introduce an additional model to assess the performance of the test with multiple models). The first model is the benchmark and uses only $X_{1,t}$, whereas the second model uses only $X_{2,t}$. For models $j = 1, 2$ we write this as:

$$q_\tau(\boldsymbol{\beta}_j^\dagger; X_{j,t}) = \beta_{0j}^\dagger(\tau) + \beta_{1j}^\dagger(\tau) X_{j,t}, \tag{21}$$

where $\boldsymbol{\beta}_j^\dagger = \left(\beta_{0j}^\dagger(\tau), \beta_{1j}^\dagger(\tau)\right)'$. This corresponds to the linear conditional quantile model (1) above. We estimate the model parameters using the quantile regression estimator described in (4), which aligns this simulation set-up with our theoretical results.

There will be three different specifications for the parameters (β_1, β_2) in the DGP in (20) which will allow us to assess different properties of the tests:

DGP1: $(\beta_1, \beta_2) = (1, 1)$

DGP2: $(\beta_1, \beta_2) = (0, 0)$

DGP3: $(\beta_1, \beta_2) = (0, 1)$

In DGP1, both $X_{1,t}$ and $X_{2,t}$ are present in the conditional quantile function. This means that both of the single-variable models are equally misspecified so we are in the non-degenerate CASE I under the null. On the other hand, in DGP2 neither $X_{1,t}$ nor $X_{2,t}$ are present in the conditional quantile function so both models are correctly specified and we are in the overlapping CASE II under the null.[16] Finally, in DGP3 only $X_{2,t}$ features in the conditional quantile function, so the second model is correctly specified and the first is not. This final DGP allows us to assess the power of the test.

In the calculation of the test statistic, we will use the quadratic loss function $L(x) = x^2$ for evaluation. We will trim the variables according to the indicator $1\{\mathbf{X}_t \in \mathcal{X}\}$ in (11) using a 1% trimming rule for the lower and upper tail of each variable. Regarding the sample sizes of the study, we set $T \in \{240, 480, 960\}$ and use the fixed estimation scheme to estimate the quantile regression coefficients $\widehat{\boldsymbol{\beta}}_{j,R}(\tau)$. We set $\pi = 1$ where $\pi = \lim_{T\to\infty} P/R$ as in Assumption A.5, so that the estimation and evaluation windows are equal to $P = \{120, 240, 480\}$. For the bootstrap, we generate one bootstrap draw over $B = 1999$ simulations using the warp speed method of Giacomini et al. (2013) and we will display results for block lengths $l \in \{1, 2, 5\}$. For calculating the conditional coverage we use the Epanechnikov kernel and a rule-of-thumb bandwidth.[17] In computing the density term $\widehat{f}_{t+1}(\cdot)$ in the bootstrap we use the Hall and Sheather (1988) bandwidth used in a similar context by Qu (2008).[18]

We will also run a variety of robustness checks relative to this baseline set-up, in addition to the modifications of the time dependence and correlation in the $X_{j,t}$ variables mentioned above. Specifically, we will also look into the results to compare: trimming versus no trimming in $1\{\mathbf{X}_t \in \mathcal{X}\}$; the fourth-order versus second-order kernel; and Gaussian versus t-distributed errors.

[16] To explore the other subcase of Case II where we have overlapping models which are both incorrectly specified, we provide an additional DGP and results which can be found in the supplementary material.

[17] This rule-of-thumb uses the rate conditions depending on the order of the kernel as imposed by Assumption A.4. For the fourth-order kernel we use a bandwidth of $h_P = KP^{-1/6}$, setting the constant term K to allow roughly one third of observations to receive a non-zero weight. In practice we tend to prefer the use of the second-order kernel, which uses a bandwidth rule $h_P = KP^{-1/3}$, as it is more widely used in practice and gives very similar results to the fourth-order kernel. We will present results for both cases.

[18] Specifically, this bandwidth is $h_P = 0.5P^{-1/3}z_\tau^{2/3}\left[1.5\phi^2(\Phi^{-1}(\tau))/((2\Phi^{-1}(\tau))^2+1)\right]^{1/3}$ where z_τ satisfies $\Phi(z_\tau) = 1-\tau/2$ and the scaling by 0.5 ensures that $h \approx 0.02$ when $\tau = 0.1$ and $P = 120$.

Table 1
Rejection rates: Pairwise - single quantile level.

$T = 240$		DGP1	DGP2	DGP3
$l = 1$	$\tau = 0.1$	0.0575	0.0150	0.0960
	$\tau = 0.2$	0.0815	0.0145	0.2726
	$\tau = 0.3$	0.0520	0.0150	0.5373
$l = 2$	$\tau = 0.1$	0.0550	0.0150	0.0895
	$\tau = 0.2$	0.0605	0.0130	0.2901
	$\tau = 0.3$	0.0525	0.0140	0.5408
$l = 5$	$\tau = 0.1$	0.0495	0.0140	0.0850
	$\tau = 0.2$	0.0465	0.0135	0.2551
	$\tau = 0.3$	0.0420	0.0145	0.4937
$T = 480$		DGP1	DGP2	DGP3
$l = 1$	$\tau = 0.1$	0.0755	0.0085	0.2766
	$\tau = 0.2$	0.0825	0.0105	0.7484
	$\tau = 0.3$	0.0855	0.0170	0.9730
$l = 2$	$\tau = 0.1$	0.0615	0.0105	0.2626
	$\tau = 0.2$	0.0890	0.0110	0.7249
	$\tau = 0.3$	0.0925	0.0175	0.9710
$l = 5$	$\tau = 0.1$	0.0540	0.0115	0.2496
	$\tau = 0.2$	0.0720	0.0135	0.7834
	$\tau = 0.3$	0.0765	0.0170	0.9695
$T = 960$		DGP1	DGP2	DGP3
$l = 1$	$\tau = 0.1$	0.0690	0.0095	0.6473
	$\tau = 0.2$	0.1011	0.0125	0.9895
	$\tau = 0.3$	0.1036	0.0155	1.0000
$l = 2$	$\tau = 0.1$	0.0640	0.0100	0.5533
	$\tau = 0.2$	0.0895	0.0130	0.9890
	$\tau = 0.3$	0.0905	0.0165	1.0000
$l = 5$	$\tau = 0.1$	0.0630	0.0110	0.5638
	$\tau = 0.2$	0.0770	0.0165	0.9865
	$\tau = 0.3$	0.0760	0.0195	1.0000

Notes: The cases of DGP1 through DGP3 correspond to (β_1, β_2) equal to $(1, 1)$, $(0, 0)$, and $(0, 1)$ in (20). In this pairwise model set-up we have the non-degenerate CASE I under DGP1, the overlapping CASE II under DGP2 and we are under the alternative for DGP3.

5.1. Pairwise comparison - Single quantile level

We firstly explore the results of Theorems 1 and 2 by assessing the performance of the $\widehat{S}_{P,R}(\tau)$ statistic in (11) for the pairwise model comparison context for a single τ level. In all simulations, our focus is on the one-sided interval case $(0, \tau]$. We will first discuss the performance of $\widehat{S}_{P,R}(\tau)$ for individual values from $\tau \in \{0.1, 0.2, 0.3\}$, while in the next section we will turn to the test which assesses models across multiple τ levels. The use of τ values smaller than the median is in light of the GaR context where the primary focus is on the left tail of the distribution.

Table 1 displays the two-sided rejection rates for the pairwise test at the 10% significance level. DGP1 corresponds to the non-degenerate CASE I under the null and Table 1 shows that we see rejection rates close to the nominal size, as expected. The results are best with a truncation lag length of $l = 1$ which is due to the relatively low time series dependence in the simulated data. The rejection rates decrease with l which mirrors the finding of Inoue (2001). Although the test appears to be slightly undersized for the smallest sample size ($T = 240$, $P = 120$), with rejection rates around 5%–8% for $l = 1$, this improves as the sample size increases ($T = 960$, $P = 480$). Here we see rejection rates very close to the nominal size especially for quantile levels $\tau = 0.2$ and $\tau = 0.3$.

Moving to DGP2, which corresponds to the overlapping models scenario, we observe rejection rates close to zero and far below nominal size for all τ levels and truncation lag lengths. This is in line with Theorems 1 and 2 which show the degeneracy of $\widehat{S}_{P,R}(\tau)$ in CASE II where the size of the test is expected to be zero.

Regarding the power of the test, the results for DGP3 in Table 1 show that the rejection rate increases towards unity with the sample size. Especially in the case of quantile level $\tau = 0.3$, which is furthest from the tail of the distribution, we see good power properties of the test in medium sample sizes. The power rises from around 55% for the smallest sample size ($T = 240$, $P = 120$) to above 95% when the sample size increases ($T = 480$, $P = 240$). When the sample size is very low ($T = 240$, $P = 120$), the test has low power in the cases of $\tau = 0.1$ and $\tau = 0.2$, though this situation improves as T increases.

We also performed several robustness checks relative to the baseline set-up, the results of which are displayed in the supplementary material as Tables S1 to S9. In Tables S1 to S3 we repeat the analysis of Tables 1 to 3 but with a small sample size of $T = 120$ implying a very small in-sample and out-of-sample window length of $R = P = 60$. Although size control is still good, the power of the test obviously drops, falling to below 20% in the multiple models set-up. We would advise caution when testing on sample sizes this small; a situation which might arise when quarterly GDP is used in countries

Table 2
Rejection rates: Pairwise - multiple quantile levels.

$T = 240$	DGP1	DGP2	DGP3
$l = 1$	0.0640	0.0115	0.5853
$l = 2$	0.0640	0.0100	0.6088
$l = 5$	0.0555	0.0125	0.5633
$T = 480$	DGP1	DGP2	DGP3
$l = 1$	0.0965	0.0105	0.9785
$l = 2$	0.0905	0.0100	0.9745
$l = 5$	0.0790	0.0110	0.9745
$T = 960$	DGP1	DGP2	DGP3
$l = 1$	0.1016	0.0160	1.0000
$l = 2$	0.0770	0.0155	1.0000
$l = 5$	0.0790	0.0210	1.0000

Notes: Same as for Table 1.

with a short data history. On the other hand, our main simulations above show that samples with $T = 240$ observations may already deliver reasonable power. In Tables S4 and S5 we increase the time series dependence parameter to $\rho = 0.7$ and $\rho = 0.9$. As expected, we find that larger truncation lag lengths like $l = 10$ or $l = 20$ are required for reasonable size control. In Table S6 we allow correlation in the $X_{j,t}$ variables with very similar results, finding that rejection rates typically reduce slightly towards zero as compared to the baseline set-up. In Table S7 the trimming is set to zero instead of 1%, and the results are very similar to the main results. Table S8 reports the results when the fourth-order Epanechnikov kernel is used instead of the second-order. The results are broadly similar to the second-order kernel results in Table 1 above, despite a slightly worse performance due to the presence of negative kernel weights. Finally, in Table S9 we use a Student's-t distribution for the error term (rescaled to have unit variance) in place of the standard normal assumption. These results are also in line with the main set-up, with size control being ensured in DGP1 with a larger truncation lag length of $l = 2$.

5.2. Pairwise comparison - Multiple quantile levels

We now turn to the performance of the $\widehat{S}_{P,R}^{\max}$ test in (17) when it is applied in the context of a pairwise comparison across multiple quantiles. Here the null hypothesis is that model 1 has equal or smaller coverage error than model 2 across all quantile levels $\tau \in \{0.1, 0.2, 0.3\}$. Under DGP1 corresponding to CASE I-RC, the two models are equally misspecified across all τ levels, so in this instance we are in the least favourable case under the null. The other results for DGP2 and DGP3 will assess the overlapping CASE II-RC under the null and the power respectively. Results are again presented for a significance level of 10%. We repeat all of the robustness checks from above which are available from the authors on request.

The results in Table 2 show promising size properties in the least favourable case under the null (DGP1). Focussing on the $l = 1$ results, we see rejection rates of 6.4%, 9.6% and 10.1% as we increase the sample size. These results show improvement over the single quantile tests from the previous section. Most importantly, we find that the power of the test (DGP3) is much greater than in the single quantile test. Notably, the power is roughly 60% for $l = 1$ at $T = 240$ and this rises swiftly to unity with the sample size. This is in contrast to the previous section where power was as low as 10% for $\tau = 0.1$ in the single quantile level test in Table 1. For the overlapping case (DGP2) we see similar findings to the previous section. In particular, the rejection rate is close to zero which is in line with Theorems 4 and 5.

5.3. Multiple model comparison - Multiple quantile levels

In addition to the two single-variable models used for the pairwise comparison results, we now introduce a third model. This will enable us to assess the performance of the $\widehat{S}_{P,R}^{\max}$ test in (17) for comparing multiple models ($J = 3$) at multiple τ levels. Specifically, we will make predictions using a conditional quantile model which uses both predictors $\mathbf{X}_t = \left(X_{1,t}, X_{2,t}\right)'$:

$$q_\tau(\boldsymbol{\beta}^\dagger; \mathbf{X}_t) = \beta_0^\dagger(\tau) + \beta_1^\dagger(\tau) X_{1,t} + \beta_2^\dagger(\tau) X_{2,t}. \tag{22}$$

As the benchmark model, we retain the first model in (21) which uses only $X_{1,t}$ as a predictor. On introducing the additional model, this set-up will assess different properties of the test under DGP1. Specifically, under DGP1 the benchmark model is now dominated by the new model in (22) so we are under the alternative and not the null as in the previous sections. Under DGP2, we have a similar case to the previous section as all three models overlapping and we are in the degenerate CASE II-RC under the null. Under DGP3, the benchmark is dominated by both of the other models and we are under the alternative.

Table 3
Rejection rates: Multiple models - multiple quantile levels - benchmark $X_{1,t}$.

$T = 240$	DGP1	DGP2	DGP3
$l = 1$	0.2436	0.0030	0.5923
$l = 2$	0.2281	0.0035	0.5763
$l = 5$	0.1846	0.0030	0.5583
$T = 480$	DGP1	DGP2	DGP3
$l = 1$	0.7589	0.0050	0.9740
$l = 2$	0.7059	0.0040	0.9780
$l = 5$	0.6963	0.0050	0.9710
$T = 960$	DGP1	DGP2	DGP3
$l = 1$	0.9965	0.0055	1.0000
$l = 2$	0.9960	0.0045	1.0000
$l = 5$	0.9945	0.0025	1.0000

Notes: The cases of DGP1 through DGP3 correspond to (β_1, β_2) equal to (1, 1), (0, 0), and (0, 1) in (20). In this multiple model set-up with $X_{1,t}$ in the benchmark model and $X_{2,t}$ and $\mathbf{X}_t$ being used in the others, the benchmark is worse than the $\mathbf{X}_t$ model under DGP1, worse than both models under DGP3, and overlapping under DGP2.

Table 3 displays the rejection rates for this multiple model test. The rejection rates are found to be very close to zero under DGP2 which is to be expected in light of the theory and the results of the previous sections. Under DGP1 and DGP3 we see that the power of the test improves to unity with the sample size. The rejection rate is lower under DGP1 than DGP3 which reflects that under DGP1 the benchmark is only beaten by a single model, whereas under DGP3 the benchmark is beaten by both of the other competing models.

In summary of all of the above simulation results, we find that the various versions of our test have good size and power properties under different DGPs and samples sizes. While the power of the test can be somewhat low in small samples when interest is in the pairwise model comparison at a single quantile level near the tail, such as $\tau = 0.1$, this situation is greatly improved when performing the pairwise test across different quantiles, even in small samples. The tests also perform well when multiple models are under consideration.

6. Empirical application

In this section we apply our test to evaluate the out-of-sample specification of the recent GaR framework of Adrian et al. (2019). The idea behind their approach is that economic conditions vary with financial conditions in the lower part of the distribution, but not in the upper part. Their two-step method uses a quantile regression approach in which economic activity, measured by real GDP growth, is regressed on a National Financial Conditions Index to obtain out-of-sample forecasts. In this empirical application we aim to explore the robustness of this specification in terms of conditional coverage, especially (i) by seeing whether models using other candidate predictors suggested by Brownlees and Souza (2021) can provide more accurate predictions in terms of conditional coverage and (ii) assessing whether the quarterly results are robust when economic activity is instead proxied by monthly industrial production growth, which is highly correlated with real GDP growth but has many more available observations. This latter exercise is important as it is difficult to perform robust inference in the traditional GaR context which uses short quarterly data spans.

6.1. Set-up and data

The models used in our application are an augmented version of the quantile autoregressive model of Koenker and Xiao (2006) which has a linear quantile regression form matching our set-up and simulation sections. Specifically, we use quantile regressions of future s-step ahead economic conditions[19] (y_{t+s}) on an autoregressive term (y_t) and a single predictor (X_{jt}) which is different across models $j = 1, \ldots, J$:

$$q_\tau(\boldsymbol{\beta}_j^\dagger; y_t, X_{jt}) = \beta_{0j}^\dagger(\tau) + \beta_{1j}^\dagger(\tau) y_t + \beta_{2j}^\dagger(\tau) X_{jt}. \tag{23}$$

The economic conditions variable y_t is either real GDP growth in the quarterly case or IP growth in the monthly case. The benchmark model (which we refer to as "QAR(1)+NFCI") uses the Chicago Fed NFCI as in Adrian et al. (2019). We will begin our analysis with a pairwise comparison ($J = 2$) where the competitor model ("QAR(1)+SV") uses stock volatility (SV), suggested by Brownlees and Souza (2021), which is proxied by the squared daily returns on the S&P500, averaged to the quarterly or monthly frequency. In particular, note that if we view NFCI and SV as noisy predictors of some common

[19] We use the forward growth rate for y_{t+s} as in Brownlees and Souza (2021) though we also checked the results when using the cumulative s-period growth as in Adrian et al. (2019). The results are quite similar, though the cumulative growth rate transformation has the drawback of imparting a high degree of serial correlation onto the series.

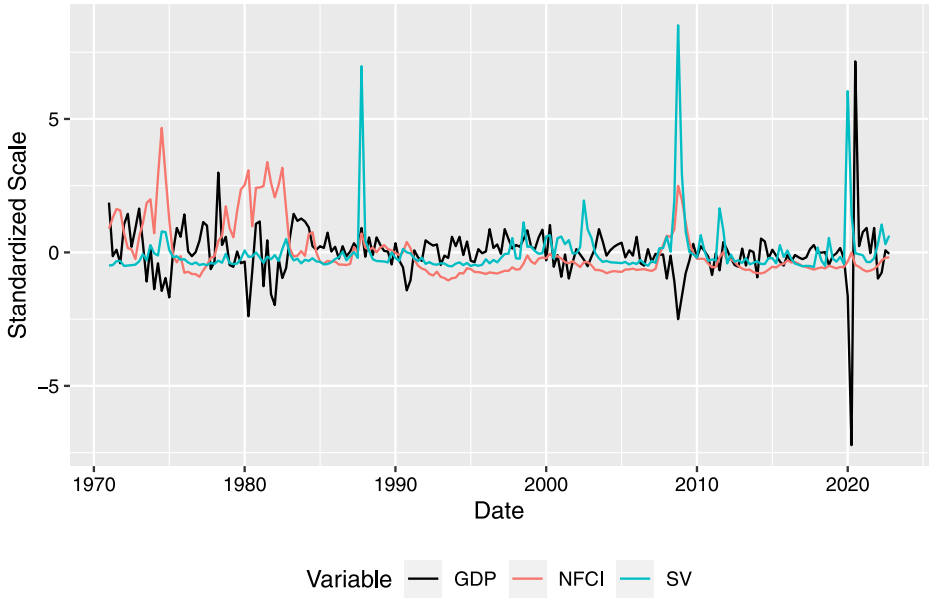

Fig. 1. Real GDP growth vs. Chicago Fed NFCI.

latent factor, say $X_{jt} = F_t + u_{jt}$ with $j = 1, 2$, where F_t denotes the common factor, while X_{1t} and X_{2t} are the observed NFCI and SV measures, respectively, a variety of scenarios might arise in principle. Specifically, provided that (i) each factor j is relevant in the sense that $\Pr\left(\beta_{0j}^{\dagger}(\tau) + \beta_{1j}^{\dagger}(\tau)y_t + \beta_{2j}^{\dagger}(\tau)X_{jt} = \widetilde{\beta}_{0j}^{\dagger}(\tau) + \widetilde{\beta}_{1j}^{\dagger}(\tau)y_t\right) < 1$, (ii) $\mathrm{Var}\left(u_{jt}\right) > 0$ for $j = 1, 2$, and (iii) $\Pr\left(u_{1t} = u_{2t}\right) < 1$, we have that:

$$\Pr\left(X_{1,t}'\boldsymbol{\beta}_1^{\dagger}(\tau) = X_{2,t}'\boldsymbol{\beta}_2^{\dagger}(\tau)\right) < 1$$

and thus models 1 and 2 do not overlap. Whether we are in CASE I under H_0 or under the alternative hypothesis then depends on the underlying DGPs. Similarly, if at least one of the above conditions (i)–(iii) fails to hold, then we are in CASE II under H_0.

The data for GDP, IP and NFCI are taken from the FRED Economic Data[20] service whereas the S&P500 data for SV are taken from Yahoo! Finance.[21] The dataset runs from January 1971 to February 2023, which yields a quarterly sample size of $T_Q = 208$ for the GDP growth model and a monthly sample size of $T_M = 626$ in the IP growth model. We will focus on the one quarter ahead and one year ahead forecast horizons, as in Adrian et al. (2019), in other words $s_Q \in \{1, 4\}$ with quarterly data or $s_M \in \{3, 12\}$ for the monthly case. Fig. 1 plots the quarterly real GDP growth rate against the NFCI and SV series, which confirms that the negative correlation between economic and financial conditions is more pronounced during lower-tail events in real GDP growth. It also highlights the pronounced drop and rebound of GDP growth in the second and third quarters of 2020 when the Covid-19 pandemic reached the U.S. This was mirrored by a single period increase in SV in 2020Q2 though there was little significant impact on the NFCI.

Regarding the out-of-sample set-up, we use the fixed estimation scheme as outlined in the rest of the paper and we set the quarterly evaluation window to be $P_Q = 120$ so as to match the set-up used in the simulation section as closely as possible. The equivalent monthly evaluation window is $P_M = 360$. We therefore make quantile forecasts for 30 years spanning from the early 1990s to the start of 2023. The rest of the set-up is exactly as in the simulation section, namely: quantile levels $\tau \in \{0.1, 0.2, 0.3\}$ will be used (first to assess the results of the single-quantile test across different quantiles and then for the multiple quantile test), truncation lag lengths $l \in \{1, 2, 5\}$ will be used with $B = 1999$ bootstrap draws, the trimming fraction is set to 1%, the second-order Epanechnikov kernel is used and all of the bandwidth rules are as described above.[22]

[20] See: https://fred.stlouisfed.org/ [Last Accessed: 12/04/2023]; The FRED codes for GDP, IP and NFCI are A191RL1Q225SBEA, INDPRO and NFCI, respectively.

[21] See: https://finance.yahoo.com/quote/%5EGSPC/ [Last accessed: 12/04/2023].

[22] In principle, it is possible to adopt a data-driven choice of the truncation lag length as in Inoue (2001), which is based on a simulation driven response surface method (see p.170, Inoue, 2001). However, we do not pursue the approach here as the methodology is based on stylized simulations. We prefer instead to present the results for a range of lag choices and check the sensitivity of the results in this regard.

Table 4
QAR(1)+NFCI vs. QAR(1)+SV - Pairwise comparison - single quantile level.

		Real GDP growth				IP growth			
		$s_Q = 1$		$s_Q = 4$		$s_M = 3$		$s_M = 12$	
		Stat	p-value	Stat	p-value	Stat	p-value	Stat	p-value
$l = 1$	$\tau = 0.1$	0.4181	0.3132	−0.0407	0.1831	0.3813	0.3102	−0.0422	0.3162
	$\tau = 0.2$	0.5151	0.3742	0.0339	0.9235	0.5335	0.1841	0.2170	0.8904
	$\tau = 0.3$	0.9344	0.2861	0.1663	0.7524	0.8262	0.0760	0.3338	0.4662
$l = 2$	$\tau = 0.1$	–	0.3302	–	0.2341	–	0.3592	–	0.3922
	$\tau = 0.2$	–	0.4142	–	0.9715	–	0.1781	–	0.8884
	$\tau = 0.3$	–	0.3392	–	0.7814	–	0.0750	–	0.4812
$l = 5$	$\tau = 0.1$	–	0.3112	–	0.2931	–	0.3232	–	0.4602
	$\tau = 0.2$	–	0.4592	–	0.9365	–	0.2171	–	0.8894
	$\tau = 0.3$	–	0.3542	–	0.7664	–	0.1181	–	0.4922

Table 5
QAR(1)+NFCI versus QAR(1)+SV - Pairwise multiple quantile test.

	Real GDP growth				IP growth			
	$s_Q = 1$		$s_Q = 4$		$s_M = 3$		$s_M = 12$	
	Stat	p-value	Stat	p-value	Stat	p-value	Stat	p-value
$l = 1$	1.3133	0.1621	0.0288	0.5058	1.1127	0.0525	0.1585	0.5858
$l = 2$	–	0.1806	–	0.5098	–	0.0600	–	0.5823
$l = 5$	–	0.2001	–	0.4897	–	0.0815	–	0.5968

6.2. Results: Pairwise comparison - Single quantile level

We first present the results for the pairwise test at a single quantile level, corresponding to the $\widehat{S}_{P,R}(\tau)$ statistic in (11). The results are displayed in Table 4 which shows the test statistic ("Stat") at the various quantile levels as well as the two-sided p-values calculated using different block lengths. We firstly note that, in terms of the bootstrap procedure, the p-values appear to be very stable across l. When quarterly real GDP growth is used to proxy economic conditions there are no rejections of the null, showing that the QAR(1)+SV model does not improve over the benchmark QAR(1)+NFCI model at any quantile level, neither for quarter-ahead nor year-ahead predictions.

However, when we change the economic activity variable to IP growth which allows us to greatly increase the sample size, the results change somewhat. In the right panel of Table 4 we see that for one quarter ahead prediction ($s_M = 3$) there is some evidence at the 10% significance level (with a p-value of 7.6%) that the QAR(1)+SV model improves over the QAR(1)+NFCI model at quantile level $\tau = 0.3$ though not at the other quantile levels. This indicates that the NFCI model may be reasonable in the far left tail of economic activity but not when interest is in a less extreme definition of downside risk. There is still no evidence of improvement for the one year ahead ($s_M = 12$) horizon at any quantile level.

Relative to the existing literature, our findings seem to align with the evidence of Brownlees and Souza (2021) who find little statistically significant evidence of the QAR(1)+NFCI model being outperformed for real GDP growth. However, the fact that the results change slightly when the sample size is increased and the dependent variable is IP growth can potentially be explained by the power gains from using larger samples which we document in our simulations. It is possible that existing GaR studies are unable to uncover robust evidence in favour of new models solely due to the choice of quarterly data.

6.3. Results: Pairwise test - Multiple quantiles

We now extend the results from the previous section by performing the test which operates across multiple quantile levels. This is a useful exercise as it is possible that rejections of the null do not occur when taking multiple quantile levels into consideration, even if a rejection does occur in isolation at some specific quantile level. Table 5 displays the test statistic in (17) as well as the one-sided p-values calculated across the different block lengths we consider.

The results are in line with the previous findings. In the case of real GDP growth there is no evidence that the QAR(1)+NFCI model is outperformed when looking jointly over $\tau \in \{0.1, 0.2, 0.3\}$. On the other hand, the results for IP growth in the right panel of Table 5 show that the QAR(1)+NFCI model is outperformed by the QAR(1)+SV model for the one quarter ahead horizon ($s_M = 3$), almost with rejection at the 5% level. As in the previous section, this result could be driven by the higher power of the test in the larger sample size that we have with monthly IP growth. Overall, there is some suggestion that the baseline specification of Adrian et al. (2019) might be improved upon when the definition of downside risk is extended to a wider set of quantiles in the lower tail of the distribution.

Table 6
Multiple model, multiple quantile test.

	Real GDP Growth				IP Growth			
	$s_Q = 1$		$s_Q = 4$		$s_M = 3$		$s_M = 12$	
	Stat	p-value	Stat	p-value	Stat	p-value	Stat	p-value
$l = 1$	1.8015	0.4132	0.0288	0.9585	2.3935	0.1076	0.1874	0.8264
$l = 2$	–	0.4012	–	0.9665	–	0.1246	–	0.8354
$l = 5$	–	0.4142	–	0.9470	–	0.1471	–	0.8489

6.4. Results: Multiple models - Multiple quantiles

We now check whether there are other candidate predictors which deliver further gains over the QAR(1)+SV method. This will enable us to apply our test for multiple models and multiple quantiles using the most general version of the $\widehat{S}_{P,R}^{\max}$ statistic in (17). To do so, we consider a further set of three variables in addition to the NFCI and SV predictors, which are those deemed most successful by the study of Brownlees and Souza (2021). This brings the total number of models to $J = 5$. The additional variables we use are: the global real economic activity factor (GF) of the Federal Reserve Bank of Dallas; a term spread (TS) equal to the 10-year treasury rate minus the 1-year rate; and house prices (HP) which we proxy by the CPI of housing in order to allow for monthly analysis. All variables are taken from FRED and are available at both the monthly and quarterly frequency.[23] The test is carried out jointly across all models and all quantile levels $\tau \in \{0.1, 0.2, 0.3\}$ as above.

The results, displayed in Table 6, are generally in line with the previous findings. In the case of real GDP growth, the test statistic increases at $s_Q = 1$ which is driven mostly by the QAR(1)+GF model, although the largest contribution is still from QAR(1)+SV.[24] Nevertheless, there is still no rejection of the null for real GDP growth, so there is no evidence that any model improves over the QAR(1)+NFCI model at any quantile level.[25] Similarly, the test statistic for IP growth also rises most for the quarter ahead horizon $s_M = 3$ due to the contribution of the QAR(1)+GF model while the effect of the SV variable still dominates. The fact that the GF variable increases the test statistic makes sense given that it is cited as one of the key GaR predictor variables in the study of Brownlees and Souza (2021). As in the previous set of results, the most significant result comes for IP at $s_M = 3$, though in this case there is not quite enough evidence to reject the null at the 10% level with a p-value of 10.8% for $l = 1$. The overall conclusion would therefore have been slightly different had we checked only the results with multiple models and not looked at the pairwise results.

7. Conclusion

This paper introduces tests for evaluating interval predictions from competing quantile models over different quantile levels. We start with the baseline case of a pairwise comparison for a single quantile level, where we compare the expected conditional coverage errors for some given loss function. We focus on coverage as this is the functional of interest in the case of VaR and GaR, as defined by regulators or institutional bodies. Our test is capable of comparing possibly (dynamically) misspecified and overlapping models, and is extended to cover cases which are of empirical relevance: the comparison of multiple intervals, multiple models, and both multiple models and multiple intervals. An extension to multiple forecast horizons is also outlined. The asymptotic properties of the tests are derived and a wild bootstrap procedure developed, which provides first order valid inference even in cases where degeneracy occurs due to the presence of nested or overlapping models.

The finite sample properties of the tests are explored in Monte Carlo simulations, which show overall good size and power properties. We apply our method in backtesting the Adrian et al. (2019) quantile regression specification for GaR prediction by exploring the possibility of financial predictors other than the NFCI in predicting risks to economic activity. While the test does not reject the null of equal expected conditional coverage error loss for the GaR of real GDP growth, when we use higher frequency monthly industrial production data, we find some evidence that the baseline NFCI model can be beaten by other financial predictors at some quantile levels. Our results suggest that future studies should consider using timelier measures of economic activity than real GDP in constructing GaR measures, as it is difficult to provide robust model evaluation tests for predicting tail events in real output growth using small quarterly data spans. We also suggest that GaR measures are evaluated over a range of quantile ranks as these tests are found to have better power properties in simulations than those based on a single quantile level.

[23] See: https://fred.stlouisfed.org/ [Last Accessed: 12/04/2023]. The FRED codes for GF, TS and HP are IGREA, DGS10 minus DGS1 and CPIHOSSL respectively.

[24] The individual test statistics for each $j = 2, \ldots, 5$ are not presented but available on request.

[25] In principle, in situations with no rejection of the null one might consider ranking models according to other statistics like the "length" of the prediction as in Brownlees and Souza (2021). We find that this does not have a particularly meaningful interpretation in the one-sided interval case we explore here. Additionally, when we computed such metrics for the various models, horizons and quantile levels under consideration, there does not appear to be any model with uniformly 'better' length than another model.

Appendix A. Asymptotic variance

For $j = 1, 2$, define the quantities:

$$A_{j,t}(\tau) = 1\{\mathbf{X}_t \in \mathcal{X}\} L\left(C_j\left(\boldsymbol{\psi}_j^{\dagger}(\tau); \mathbf{X}_t\right) - \tau\right), \tag{24}$$

$$B_{j,t}(\tau) = 1\{\mathbf{X}_t \in \mathcal{X}\} \nabla^{(1)} L\left(C_j\left(\boldsymbol{\psi}_j^{\dagger}(\tau); \mathbf{X}_t\right) - \tau\right)\left(1\{y_{t+1} \leq q_\tau(\boldsymbol{\psi}_j^{\dagger}(\tau); X_{j,t})\} - F_{t+1}(q_\tau(\boldsymbol{\psi}_j^{\dagger}(\tau); X_{j,t})|\mathbf{X}_t)\right), \tag{25}$$

and:

$$D_{j,t}(\tau) = \varphi(\boldsymbol{\psi}_1^{\dagger}(\tau); y_{t+1}, X_{1,t}), \tag{26}$$

with:

$$\varphi(\boldsymbol{\psi}_j^{\dagger}(\tau); y_{t+1}, X_{j,t}) = \Lambda_j(\tau)\left(H_j(\tau)^{-1} X_{j,t}\left(1\left\{y_{t+1} - X_{j,t}\boldsymbol{\beta}_j^{\dagger}(\tau) \leq 0\right\} - \tau\right)\right), \tag{27}$$

$\Lambda_j(\tau) = \mathrm{E}\left(1\{\mathbf{X}_t \in \mathcal{X}\}\left(\nabla^{(1)} L\left(C_j((0,\tau]; \mathbf{X}_t) - \tau\right) f_{t+1}\left(X_{j,t}'\boldsymbol{\beta}_j^{\dagger}(\tau)|\mathbf{X}_t\right) X_{j,t}'\right)\right)$ (see supplementary material for the corresponding definition in the location scale case), and $H_j(\tau)$ defined in A.3(ii). The variance–covariance kernel in Theorem 1(i) for the one-sided case is given by:

$$\begin{aligned}\Omega(\tau) = {} & \mathrm{Avar}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(A_{1,t}(\tau) - A_{2,t}(\tau)\right)\right) + \mathrm{Avar}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(B_{1,t}(\tau) - B_{2,t}(\tau)\right)\right) \\ & + \mathrm{Avar}\left(\frac{\sqrt{P}}{R}\sum_{t=1}^{R-1}\left(D_{1,t}(\tau) - D_{2,t}(\tau)\right)\right) \\ & + 2\mathrm{Acov}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(A_{1,t}(\tau) - A_{2,t}(\tau)\right), \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(B_{1,t}(\tau) - B_{2,t}(\tau)\right)\right),\end{aligned} \tag{28}$$

where $\Omega(\tau)$ follows since:

$$\begin{aligned}& \mathrm{Acov}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(A_{1,t}(\tau) - A_{2,t}(\tau)\right), \frac{\sqrt{P}}{R}\sum_{t=1}^{R-1}\left(D_{1,t}(\tau) - D_{2,t}(\tau)\right)\right) \\ = {} & \mathrm{Acov}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(B_{1,t}(\tau) - B_{2,t}(\tau)\right), \frac{\sqrt{P}}{R}\sum_{t=1}^{R-1}\left(D_{1,t}(\tau) - D_{2,t}(\tau)\right)\right) \\ = {} & 0\end{aligned}$$

due to the use of the fixed estimation scheme.

Appendix B. Bootstrap statistic

Pairwise Comparison: For $j = 1, 2$, the expressions $\widehat{A}_{j,P,R,t}(\tau)$, $\widehat{B}_{j,P,R,t}(\tau)$, and $\widehat{D}_{j,P,R,t}(\tau)$ of the bootstrap statistic are defined as follows:

$$\widehat{A}_{j,P,R,t}(\tau) = 1\{\mathbf{X}_t \in \mathcal{X}\}\left(L\left(\widehat{C}_{j,P,R}((0,\tau]; \mathbf{X}_t) - \tau\right) - \frac{1}{P}\sum_{s=R}^{T-1} L\left(\widehat{C}_{j,P,R}((0,\tau]; \mathbf{X}_s) - \tau\right)\right), \tag{29}$$

$$\widehat{B}_{j,P,R,t}(\tau) = 1\{\mathbf{X}_t \in \mathcal{X}\} \nabla^{(1)} L\left(\widehat{C}_{j,P,R}((0,\tau]; \mathbf{X}_t) - \tau\right)\left(1\left\{y_{t+1} \leq X_{j,t}'\widehat{\boldsymbol{\beta}}_{j,R}(\tau)\right\} - \widehat{F}_{t+1,P}\left(X_{j,t}'\widehat{\boldsymbol{\beta}}_{j,R}(\tau)|\mathbf{X}_t\right)\right), \tag{30}$$

where:

$$\widehat{F}_{t+1,P}\left(X_{j,t}'\widehat{\boldsymbol{\beta}}_{j,R}(\tau)|\mathbf{X}_t\right) = \widehat{C}_{j,P,R}((0,\tau]; \mathbf{X}_t) = \frac{\sum_{s=R}^{T-1} 1\left\{y_{s+1} \leq X_{j,t}'\widehat{\boldsymbol{\beta}}_{1,R}(\tau)\right\}\mathbf{K}\left(\frac{\mathbf{X}_s - \mathbf{X}_t}{h}\right)}{\sum_{s=R}^{T-1}\mathbf{K}\left(\frac{\mathbf{X}_s - \mathbf{X}_t}{h}\right)}$$

and:

$$\widehat{D}_{j,P,R,t}(\tau) = \widehat{\Lambda}_{j,P,R}(\tau)\left(\widehat{H}_{j,R}^{-1}(\tau) X_{j,t}\left(1\left\{y_{t+1} \leq X_{j,t}'\widehat{\boldsymbol{\beta}}_{j,R}(\tau)\right\}\right) - \tau\right), \tag{31}$$

with:

$$\widehat{H}_{j,R}(\tau) = \frac{1}{R}\sum_{t=1}^{R-1}\left(\widehat{f}_{t+1}\left(X_{j,t}'\widehat{\boldsymbol{\beta}}_{j,R}(\tau)|X_{j,t}\right) X_{j,t}X_{jt}'\right)$$

$$\widehat{\Lambda}_{j,P,R}(\tau) = \frac{1}{P}\sum_{t=R}^{T-1} 1\{\mathbf{X}_t \in \mathcal{X}\}\left(\nabla^{(1)} L\left(\widehat{C}_{j,P,R}\left((0,\tau];\mathbf{X}_t\right) - \tau\right)\widehat{f}_{t+1,P}\left(X'_{j,t}\widehat{\boldsymbol{\beta}}_{j,R}(\tau)|\mathbf{X}_t\right)X'_{j,t}\right)$$

where:

$$\widehat{f}_{t+1,P}\left(X'_{j,t}\widehat{\boldsymbol{\beta}}_{j,R}(\tau)|\mathbf{X}_t\right) = \frac{\widehat{C}_{j,P,R}\left((0,\tau_k];\mathbf{X}_t\right) - \widehat{C}_{j,P,R}\left((0,\tau_{k-1}];\mathbf{X}_i\right)}{\left(X'_{j,t}\widehat{\boldsymbol{\beta}}_{j,R}(\tau_k) - X'_{j,t}\widehat{\boldsymbol{\beta}}_{j,R}(\tau_{k-1})\right)}$$

for $\tau_{k-1} < \tau < \tau_k$ with $\tau_k - \tau_{k-1} \to 0$ as $P \to \infty$.

Multiple Models & Intervals: In the multiple model and interval case of Section 4, recall that $\boldsymbol{V}$ is an $M(J-1)\times M(J-1)$ dimensional matrix, whose principal diagonal elements are given by v_{kk}, with $k = (j-2)M + i$ with $i = 1,\dots,M$ and $j = 2,\dots,J$:

$$\begin{aligned} v_{kk} = \text{Avar}\Bigg(&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{A}_{1,t}\left(\tau_i,\mathbf{X}_t^j\right) - \widehat{A}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right) + \widehat{B}_{1,t}\left(\tau_i,\mathbf{X}_t^j\right) - \widehat{B}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)\right) \\ &+\frac{\sqrt{P}}{R}\sum_{t=1}^{R-1}\left(\widehat{D}_{1,t}\left(\tau_i\right) - \widehat{D}_{j,t}\left(\tau_i\right)\right)\Bigg) \end{aligned} \tag{32}$$

and whose off-diagonal elements are given by $v_{kk'}$, with $k' = (j'-2)M + i'$ and $i' \neq i$ and/or $j' \neq j$:

$$\begin{aligned} v_{kk'} = \text{Acov}\Bigg(&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{A}_{1,t}\left(\tau_i,\mathbf{X}_t^j\right) - \widehat{A}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right) + \widehat{B}_{1,t}\left(\tau_i,\mathbf{X}_t^j\right) - \widehat{B}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)\right) \\ &+\frac{\sqrt{P}}{R}\sum_{t=1}^{R-1}\left(\widehat{D}_{1,t}\left(\tau_i\right) - \widehat{D}_{j,t}\left(\tau_i\right)\right), \\ &\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{A}_{1,t}\left(\tau_{i'},\mathbf{X}_t^{j'}\right) - \widehat{A}_{j',t}\left(\tau_{i'},\mathbf{X}_t^{j'}\right) + \widehat{B}_{1,t}\left(\tau_{i'},\mathbf{X}_t^{j'}\right) - \widehat{B}_{j',t}\left(\tau_{i'},\mathbf{X}_t^{j'}\right)\right) \\ &+\frac{\sqrt{P}}{R}\sum_{t=1}^{R-1}\left(\widehat{D}_{1,t}\left(\tau_{i'}\right) - \widehat{D}_{j',t}\left(\tau_{i'}\right)\right)\Bigg). \end{aligned} \tag{33}$$

Finally, we also need to construct a HAC estimator of the diagonal elements v_{kk} of the asymptotic variance. Thus, let $\pi_s = 1 - \frac{s}{l_P+1}$ where $s = 1,\dots,l_P$ is the lag truncation parameter, where we assume $l_P = l_R$ for simplicity. In addition, let $\widehat{A}_{1j,t}\left(\tau_i,\mathbf{X}_t^j\right) = \left(\widehat{A}_{1,t}\left(\tau_i,\mathbf{X}_t^j\right) - \widehat{A}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)\right)$, $\widehat{B}_{1j,t}\left(\tau_i,\mathbf{X}_t^j\right) = \left(\widehat{B}_{1,t}\left(\tau_i,\mathbf{X}_t^j\right) - \widehat{B}_{j,t}\left(\tau_i,\mathbf{X}_t^j\right)\right)$ and also $\widehat{D}_{1j,t}\left(\tau_i\right) = \left(\widehat{D}_{1,t}\left(\tau_i\right) - \widehat{D}_{j,t}\left(\tau_i\right)\right)$. Then, for $k = (j-2)M + i$, $j = 2,\dots,J$, $i = 1,\dots,M$ we may construct the HAC estimator of v_{kk} as:

$$\begin{aligned} \widehat{v}_{kk,P,R} = &\frac{1}{P}\sum_{t=R+l_P}^{T-l_P-1}\sum_{s=-l_P}^{l_P}\pi_s\widehat{A}_{1j,t}\left(\tau_i,\mathbf{X}_t^j\right)\widehat{A}_{1j,t-s}\left(\tau_i,\mathbf{X}_{t-s}^j\right) + \frac{1}{P}\sum_{t=R+l_P}^{T-l_P-1}\sum_{s=-l_P}^{l_P}\pi_s\widehat{B}_{1j,t}\left(\tau_i,\mathbf{X}_t^j\right)\widehat{B}_{1j,t-s}\left(\tau_i,\mathbf{X}_{t-s}^j\right) \\ &+\frac{1}{R}\sum_{t=1+l_R}^{R-l_R-1}\sum_{s=-l_R}^{l_R}\pi_s\widehat{D}_{1j,t}\left(\tau_i\right)\widehat{D}_{1j,t-s}\left(\tau_i\right) + \frac{2}{P}\sum_{t=R+l_P}^{T-l_P}\sum_{s=-l_P}^{l_P}\pi_s\widehat{A}_{1j,t}\left(\tau_i,\mathbf{X}_t^j\right)\widehat{B}_{1j,t-s}\left(\tau_i,\mathbf{X}_{t-s}^j\right), \end{aligned} \tag{34}$$

since in the fixed estimation scheme, the asymptotic covariance between the parametric estimation error and the other components is zero.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2023.105490.

References

Adrian, T., Boyarchenko, N., Giannone, D., 2019. Vulnerable growth. Amer. Econ. Rev. 109 (4), 1263–1289.
Andrews, D., 1995. Nonparametric kernel estimation for semiparametric models. Econom. Theory 11, 560–596.
Andrews, D., Pollard, D., 1994. An introduction to functional central limit theorems for dependent stochastic processes. Internat. Statist. Rev. 62 (1), 119–132.
Andrews, D., Shi, X., 2013. Inference based on conditional moment inequalities. Econometrica 81 (2), 609–666.
Andrews, D., Soares, G., 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78, 119–157.

Angrist, J., Chernozhukov, V., Fernandez-Val, I., 2006. Quantile regression under misspecification, with an application to the U.S. wage structure. Econometrica 74 (2), 539–563.
Brownlees, C., Souza, A.B., 2021. Backtesting global growth-at-risk. J. Monetary Econ. 118, 312–330.
Christoffersen, P., 1998. Evaluating interval forecasts. Internat. Econom. Rev. 39 (4), 841–862.
Clark, T., McCracken, M., 2001. Tests of equal forecast accuracy and encompassing forNested models. J. Econometrics 105 (1), 85–110.
Clark, T., McCracken, M., 2014. Tests of equal forecast accuracy for overlapping models. J. Appl. Econometrics 29, 415–430.
Clark, T., McCracken, M., Mertens, E., 2020. Modeling time-varying uncertainty of multiple-horizon forecast errors. Rev. Econ. Stat. 102 (1), 17–33.
Clements, M.P., 2014. Forecast uncertainty - ex ante and ex post: U.S. inflation and output growth. J. Bus. Econom. Statist. 32 (2), 206–216.
Corradi, V., Swanson, N., 2002. A consistent test for nonlinear out of sample predictive accuracy. J. Econometrics 110, 353–381.
Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econom. Statist. 13 (3), 253–263.
Engle, R., Manganelli, S., 2004. CAViaR: Conditional autoregressive value at risk by regression quantiles. J. Bus. Econom. Statist. 22 (4), 367–381.
Escanciano, J.C., Olmo, J., 2010. Backtesting parametric value-at-risk with estimation risk. J. Bus. Econom. Statist. 28 (1), 36–51.
Escanciano, J., Velasco, C., 2010. Specification tests of parametric dynamic conditional quantiles. J. Econometrics 159, 209–221.
Giacomini, R., Politis, D.N., White, H., 2013. A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. Econom. Theory 29 (3), 567–589.
Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74 (6), 1545–1578.
Gneiting, T., 2011. Quantiles as optimal point predictors. Int. J. Forecast. 27, 197–207.
Gneiting, T., Raftery, A., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378.
Gonzalez-Rivera, G., Maldonado, J., Ruiz, E., 2019. Growth in stress. Int. J. Forecast. 35, 948–966.
Granger, C., 1999. Outline of forecast theory using generalized cost functions. Span. Econom. Rev. 1, 161–173.
Hall, P., Sheather, S.J., 1988. On the distribution of a studentized quantile. J. R. Stat. Soc. Ser. B Stat. Methodol. 50 (3), 381–391.
Hansen, B., 1996. Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64, 413–430.
Hansen, P., 2005. A test for superior predictive ability. J. Bus. Econom. Statist. 23 (4), 365–380.
Horvath, P., Li, J., Liao, Z., Patton, A., 2022. A consistent specification test for dynamic quantile models. Quant. Econ. 13, 125–151.
Inoue, A., 2001. Testing for distributional changes in time series. Econom. Theory 17, 156–187.
Kim, T.-H., White, H., 2003. Estimation, inference, and specification testing for possibly misspecified quantile regression. In: Fomby, T., Carter Hill, R. (Eds.), Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later. In: Advances in Econometrics, vol. 17, Emerald Group Publishing Limited, Bingley, pp. 107–132.
Koenker, R., Xiao, Z., 2006. Quantile autoregression. J. Amer. Statist. Assoc. 101 (475), 980–990.
Koenker, R., Xiao, Z., 2009. Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models. J. Amer. Statist. Assoc. 109 (488), 1696–1712.
Li, J., Liao, Z., Quaedvlieg, R., 2022. Conditional superior predictive ability. Rev. Econom. Stud. 89, 843–875.
Machado, J.A., Silva, J.S., 2019. Quantiles via moments. J. Econometrics 213 (1), 145–173.
Manzan, S., 2015. Forecasting the distribution of economic variables in a data-rich environment. J. Bus. Econom. Statist. 33 (1), 144–164.
McCracken, M., 2020. Diverging tests of equal predictive ability. Econometrica 88, 1753–1754.
Plagborg-Møller, M., Reichlin, L., Ricco, G., Hasenzagl, T., 2020. When is growth at risk? Technical report, BPEA Conference Draft, Spring.
Prasad, A., Elekdag, S., Jeasakul, P., Lafarguette, R., Alter, A., Feng, A., Wang, C., 2019. Growth at Risk: Concept and Application in IMF Country Surveillance. IMF Working Paper 19/36, International Monetary Fund.
Qu, Z., 2008. Testing for structural change in regression quantiles. J. Econometrics 146, 170–184.
Quaedvlieg, R., 2021. Multi-horizon forecast comparison. J. Bus. Econom. Statist. 39 (1), 40–53.
Reichlin, L., Ricco, G., Hasenzagl, T., 2020. Financial Variables as Predictors of Real Growth Vulnerability. Technical Report 05/2020, Deutsche Bundesbank Working Paper.
Shi, X., 2015. A nondegenerate Vuong test. Quant. Econ. 6, 85–121.
Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333.
Wang, J., Wu, J., 2012. The taylor rule and forecast intervals for exchange rates. J. Money Credit Bank. 44 (1), 103–144.
West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64 (5), 1067–1084.
White, H., 2000. A reality check for data snooping. Econometrica 68 (5), 1097–1126.
Zhu, Y., Timmermann, A., 2020. Can Two Forecasts Have the Same Conditional Expected Accuracy?. arXiv working paper, arXiv:2006.03238v1.