An Oracle Inequality for Multivariate Dynamic Quantile Forecasting

Jordi Llorens-Terrazas*

First Draft: April 25, 2022
This Draft: December 3, 2022
[Click here for the latest version]

Abstract

I derive an oracle inequality for a family of possibly misspecified multivariate conditional autoregressive quantile models. The family includes standard specifications for (nonlinear) quantile prediction proposed in the literature. This inequality is used to establish that the predictor that minimizes the in-sample average check loss achieves the best out-of-sample performance within its class at a near optimal rate, even when the model is fully misspecified. An empirical application to backtesting global Growth-at-Risk shows that a combination of the generalized autoregressive conditionally heteroscedastic model and the vector autoregression for Value-at-Risk performs best out-of-sample in terms of the check loss.

Keywords: Multivariate conditional quantile, oracle inequality, time series, forecasting, Markov chain.

JEL: C14, C22, C53, C58.

*Universitat Pompeu Fabra and Barcelona School of Economics.
e-mail: jordi.llorens@upf.edu.
I would like to thank Christian Brownlees, Mika Meitz, Gabor Lugosi, Geert Mesters, Barbara Rossi, Kirill Evdokimov, Katerina Petrova, André Souza and Pierluigi Vallarino for providing numerous helpful comments. I am also indebted to participants of the 2nd International Econometrics PhD Conference at Erasmus University Rotterdam for their valuable questions and comments. I acknowledge support from the Spanish Ministry of Science and Innovation (FPI Grant PRE2019-090839).
[Click here to see the Online Appendix]

1 Introduction

Forecasting conditional quantiles of time series has a large number of applications in economics and finance. A recent popular example is the computation of Growth-at-Risk forecasts, i.e. the 5% quantile of the distribution of real gross domestic product growth given past information. Among the different methodologies proposed to forecast quantiles, the Conditional Autoregressive Value-at-Risk (CAViaR) of Engle and Manganelli (2004) stands out as one of the leading approaches in the literature due to its flexibility, parsimony and relative ease of estimation. Moreover, the CAViaR methodology is semi-parametric in the sense that it imposes mild assumptions on the data generating process (DGP) (White, Kim, and Manganelli, 2015). Despite the fact that forecasting quantiles is of obvious interest to economic agents, the theory in those papers is tailored to *estimation under correct specification* of the quantile dynamics, and less attention is paid to *forecasting under misspecification*.

This paper establishes theoretical performance guarantees for out-of-sample forecasting with a multivariate version of the CAViaR model. In practical terms, the class of forecasts is equivalent to the one-lag version of the vector autoregressive model for Value-at-Risk (VAR for VaR or VFV) of White *et al.* (2015) with a single quantile. The guarantees are obtained by deriving an *oracle inequality*, i.e. a probabilistic bound that relates the performance of an estimator to that of an ideal estimator that has best performance in the class, also known as the "oracle" (Donoho and Johnstone, 1994; Candes, 2006). The oracle inequality implies that the VFV that minimizes the in-sample average check loss achieves the oracle's out-of-sample performance in terms of the check loss at a near optimal rate, even when the model is fully misspecified. The paper allows for full misspecification in that it suffices to make nonparametric assumptions on the DGP, such as existence of a certain number of moments of the innovations and stable dynamics on the time series. This result translates into optimal out-of-sample quantile forecasting if the researcher believes that the class contains the true conditional quantile of the time series.

The theoretical framework of this paper builds upon the literature on statistical learning theory. This framework has at least three important highlights. First, the main result holds without assuming identification nor correct specification of the quantile dynamics, which are critical assumptions in the CAViaR literature (Engle and Manganelli,

2004; White *et al.*, 2015). Second, the result holds in finite samples with high probability, as opposed to being asymptotic, and it provides a specific rate of convergence for the predictive performance. Third, the theory allows to derive transparent constraints on the parameter space where the class of forecasts is stationary and ergodic. In contrast, (White *et al.*, 2015) assume the existence of some set over which the VFV is stationary and ergodic.

The proof of the main result can be broken down in three main steps. The first step is to establish existence of moments and strong mixing conditions for the loss and a "dominating process" which is similar in spirit to the domination conditions often used to obtain uniform laws of large numbers (Andrews, 1987; Pötscher and Prucha, 1989). This is accomplished through Markov chain theory (Meyn and Tweedie, 1993, Ch. 15). The novelty of the approach consists of proving that a Markov chain whose components are the DGP, the forecast, and the dominating process is V-geometrically ergodic (Liebscher, 2005; Meitz and Saikkonen, 2008a). Importantly, the strong mixing coefficients are bounded by a function with geometric decay uniformly over the parameter space, which is established using results by Roberts and Rosenthal (2004). The second step is to establish a general inequality that states that the performance of the VFV that minimizes the in-sample average check loss can be controlled by the sum of (i) the supremum of an average of differences between conditional and unconditional expected losses and (ii) the supremum of the empirical process associated with the prediction loss. In the third step, suitable bounds are derived for these two terms using, respectively, an inequality from Ibragimov (1962) and a concentration inequality for strong mixing processes (Liebscher, 1996).

The merits of the methodology are illustrated in an empirical contribution to the recent Growth-at-Risk (GaR) literature popularized by Adrian, Boyarchenko, and Giannone (2019). An out-of-sample GaR forecasting exercise shows that the past of GDP growth seems to be the key driver of the time variation in the conditional distribution of GDP growth, see also Brownlees and Souza (2021) and Catania, Luati, and Vallarino (2021). Furthermore, the results of the exercise suggest that a combination of generalized autoregressive conditionally heteroskedastic forecasts (GARCH) and VFV performs best out-of-sample. The combination exploits the dynamics on the quantiles of the standardized residuals from the AR-GARCH procedure. Although asymmetries in the conditional volatility of GDP growth do not appear to play an important role, the empirical results

of this work suggest that other types of asymmetries do still matter for the quantiles.

This paper is mainly related to three strands of the literature which share more in common than it may appear at first sight.

Dynamic Quantile Models. In a time series context, quantile regression approaches need to be adapted to account for the dependence induced by the time-ordering of the data. A natural extension is the quantile autoregressive approach developed by Koenker and Xiao (2006) and, as pointed out above, one of the most successful dynamic quantile models is the CAViaR specification by Engle and Manganelli (2004). When considering multiple quantiles of a random variable, a drawback of these approaches is the lack of an internal mechanism that avoids the quantile crossing problem. This drawback can be addressed ex-post, see Chernozhukov, Fernández-Val, and Galichon (2010), or ex-ante, see Gouriéroux and Jasiak (2008). Important contributions to the dynamic quantile literature also include White, Kim, and Manganelli (2015); Chavleishvili and Manganelli (2019); Catania and Luati (2019); Catania, Luati, and Mikkelsen (2022). Empirical illustrations as well as novel CAViaR specifications are presented in Kuester, Mittnik, and Paolella (2006); Bao, Lee, and Saltoglu (2006) for financial data and Huang, Yu, Fabozzi, and Fukushima (2009) for oil price data.

The theory in the CAViaR literature is developed under the general framework of M-estimation for dependent data. For example, the assumptions of White *et al.* (2015) – which are tailored to the goals of estimation and inference – provide an interesting benchmark to compare against the assumptions of the current paper. Overall, their assumptions can be regarded as semi-parametric in the sense that the innovation distribution may be misspecified. However, a key assumption in that paper is that there exists a unique parameter that characterizes the dynamics of the true conditional quantile of the data, i.e. identification and correct specification. In contrast, in the framework of this paper, identification and correct specification assumptions are not required.

Quasi-maximum likelihood. The oracle inequality derived in this paper can be regarded as a prediction analog of the consistency of quasi-maximum likelihood estimators. Results of this type date back to Akaike (1973) and White (1982), which studied the properties of maximum likelihood estimation for misspecified models. The main lesson from those papers is that under mild assumptions, the (quasi-) maximum likelihood

estimator (strongly) converges to the minimizer of the Kullback-Leibler Information Criterion (KLIC), which measures the discrepancy between the density of the true DGP vs the pseudo-true density (the Gaussian being the classical choice). As put by White (1982), the KLIC can be interpreted as a measure of our ignorance about the true structure of the DGP. Extensions of this type of result to M-estimators with dependent data appeared almost simultaneously in the econometrics literature (Domowitz and White, 1982; White and Domowitz, 1984).

Statistical learning theory for time series. The theory of M-estimation is able to provide useful answers to the problems of estimation and inference, but is less suitable to study the question of prediction. But seeing CAViaR as a "learning" algorithm instead of a model may prove useful. In fact, a vast literature – under the rubric of statistical learning theory – is devoted to study the prediction properties of learning algorithms. This literature is interested in a number of questions, and this paper is concerned with the following two: (*i*) to find conditions for *consistency* of learning processes, i.e. uniform convergence of a class of forecasts (Vapnik and Chervonenkis, 1971), and (*ii*) to determine the rate of convergence of the learning process (Vapnik, 1999).

An interesting feature in the learning literature is that the relationship between algorithm and data need not be specified. However, most results coming from the statistical learning literature rely on a number of assumptions that do not apply to the CAViaR models mentioned above, where data (and corresponding loss function) is non-i.i.d., unbounded, and prediction algorithms may depend on the entire past of the data. Although several efforts have been made in that literature to extend their results to time series forecasting applications, none of those provides oracle inequalities for out-of-sample forecasts based on the models cited above, nor their multivariate extensions.

The quest for forecasting performance guarantees for time series can probably be traced back to Yu (1994), which established rates of convergence for empirical processes of stationary mixing sequences – for families of predictors suitably bounded by an envelop function, which is similar in spirit to the dominating assumptions in Andrews (1987) – and Meir (2000), who provided the first generalization bounds for nonparametric time series prediction based on such results. The contributions that are probably most related to the current paper are McDonald, Shalizi, and Schervish (2017) and Kuznetsov and Mohri (2017), which provide *generalization bounds* under assumptions that allow for

mixing data and unbounded losses. It must be emphasized that generalization bounds do not imply oracle inequalities (but that the reverse implication is true); thus, the results in the current paper are not implied by the ones obtained in those papers. Kuznetsov and Mohri (2017) extend the notion of prediction performance, which is typically defined with an unconditional expectation, to *path-dependent* performance, which instead uses the expected loss conditional on past information. This is similar in spirit to the notion of conditional risk defined in the current paper.

This paper is not the first to use the framework of statistical learning theory in econometrics. Examples of this include Jiang and Tanner (2010), which studies the properties of empirical risk minimization for time series binary choice, Kock and Callot (2015), which establishes oracle inequalities for high-dimensional vector autoregressions, Brownlees and Guðmundsson (2021), which analyzes the performance of empirical risk minimization for linear regression with dependent data and Brownlees and Llorens-Terrazas (2021), which establishes similar results for a class of recursive threshold models that include as special cases the forecasts induced by ARMA(1,1) and GARCH(1,1) models. Finally, note that the framework can also be adapted to deal with policy decisions such as the allocation of treatments to individuals based on covariates (Manski, 2004; Kitagawa and Tetenov, 2018), which has recently been adapted to deal with multivariate time series (Kitagawa, Wang, and Xu, 2022).

Outline of the paper. The rest of this paper is structured as follows. Section 2 lays out the notation and presents the class of forecasts and the estimation procedure. Section 3 introduces the theoretical framework under which the main result is derived, and section 4 highlights the main steps followed to prove the claim. Section 5 contains the empirical application to Growth-at-Risk, and section 6 concludes. All proofs are relegated to the Appendix, and the more technical results and additional tables are gathered in the Online Appendix.

2 Methodology

Notation. For an $n \times 1$ real vector x, $\|x\|_r = (\sum_{i=1}^n |x_i|^r)^{1/r}$, where $r \geqslant 1$, and $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)'$, i.e. x_{-i} denotes removal of the i^{th} entry of x, $i = 1, \ldots, n$. For an $m \times n$ real matrix A, $|||A|||_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^m |a_{i\,j}|$, i.e. the maximum absolute

column sum of the matrix, and if A is square, $A^{\otimes r} = A \otimes \cdots \otimes A$, i.e. the Kronecker product taken r times. The notation $\text{vec}(A)$ represents a long vector that stacks the columns of the matrix A from left to right. For a random variable X, let $\|X\|_{L_r} = (\mathbb{E}|X|^r)^{1/r}$, where $r \geqslant 1$, and $\|X\|_{L_\infty} = \inf\{a : \Pr(|X| > a) = 0\}$ for $r = \infty$. For two real numbers a and b, denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. In this paper, $I(\cdot)$ denotes the indicator function, while $\mathbf{I}$ is used for the identity matrix. For a time series $\{X_t\}$, where t is a non-negative integer, let $\mathbb{E}_t(\cdot) = \mathbb{E}(\cdot|X_{t-1}, \ldots, X_0)$. For real x, the notation $\lfloor x \rfloor$ is used to denote the largest integer lower than or equal to x, and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

2.1 Definition of the multivariate CAViaR class.

The main goal of this paper is out-of-sample conditional quantile forecasting of a stationary, multivariate time series $\{Y_t\}$ taking values in $\mathbb{R}^N$. In the sequel, the focus is on one-step-ahead forecasting, but the results apply to h-step ahead forecasting as well (see section OA.4 in the Online Appendix). More specifically, for some $\tau_i \in [0, 1]$ and $i = 1, \ldots, N$, let $q_{it}^{\tau_i}$ denote the conditional τ_i-quantile of Y_{it} given information up to time $t - 1$. That is, q_{it} is implicitly defined as $\Pr(Y_{it} \leqslant q_{it}^{\tau_i}|Y_{t-1}, \ldots, Y_0) = \tau_i$. The following class of recursive forecasts indexed by $\theta \in \Theta_\omega \times \Theta_A \times \Theta_B \times \Theta_\lambda = \Theta \subset \mathbb{R}^p$ is available to the forecaster, and can be written in matrix notation as

$$f_{\theta t} = \omega + A s_\lambda(Y_{t-1}) + B f_{\theta t-1} \ , \tag{1}$$

where $f_{\theta t} \in \mathbb{R}^N$, $\theta = (\omega', \text{vec}(A)', \text{vec}(B)', \lambda')'$, $\omega \in \Theta_\omega \subset \mathbb{R}^{p_\omega}$, $\text{vec}(A) \in \Theta_A \subset \mathbb{R}^{p_A}$, $\text{vec}(B) \in \Theta_B \subset \mathbb{R}^{p_B}$, $\lambda \in \Theta_\lambda \subset \mathbb{R}^{p_\lambda}$, $p = p_\omega + p_A + p_B + p_\lambda$ and $s_\lambda(\cdot)$ is shorthand for $s(\cdot, \lambda)$, where $s : \mathbb{R}^N \times \mathbb{R}^{p_\lambda} \to \mathbb{R}^N$.[1] The precise assumptions on the parameters and the function s_λ are spelled out in what follows. In practice, the forecaster chooses a value $f_{\theta 0} = f_0$ (which does not depend on θ) to start the recursion.

For example, a simple bivariate version of the above relates the conditional quantile

[1] To keep the theoretical analysis as simple as possible, the function s_λ is assumed to be differentiable, but the theoretical framework can accommodate arbitrarily good approximations to popularly used non-differentiable functions such as the absolute value.

forecasts of both random variables according to a vector autoregressive structure (VAR)[2]

$$f_{\theta\,1\,t} = X_t'\beta_1 + b_{11} f_{\theta\,1\,t-1} + b_{12} f_{\theta\,2\,t-1} \ ,$$
$$f_{\theta\,2\,t} = X_t'\beta_2 + b_{21} f_{\theta\,1\,t-1} + b_{22} f_{\theta\,2\,t-1} \ ,$$

where X_t represents predictors belonging to the information set up to $t-1$, which typically includes lagged values of $Y_{i\,t}$ (White *et al.*, 2015).

A number of remarks are in order. First, note that s_λ need not be differentiable as a function of λ. Second, the assumptions are general enough to accommodate multivariate versions of the symmetric and asymmetric absolute value specifications of Engle and Manganelli (2004).[3] Third, a distinguishing feature with respect to the CAViaR literature is that the relationship between Y_t and $f_{\theta\,t}$ is not specified. In particular, $q_t^\tau := (q_{1\,t}^{\tau_1}, \ldots, q_{N\,t}^{\tau_N})'$ need not be equal to $f_{\theta\,t}$. Fourth, the class can only handle a single quantile for each variable, although the quantiles may differ for each variable.[4]

2.2 Loss function.

The focus of this paper is on forecasting under the *check loss*

$$\rho_\tau(u) = u(\tau - I(u<0)) \ , \quad \tau \in [0,1] \ .$$

The check loss (also known as tick loss) can be interpreted as an asymmetric generalization of the absolute error. Setting $\tau = 1/2$ leads to the absolute error scaled by $1/2$. This allows the forecaster to incorporate the relative costs of under vs over-prediction.[5] It is well known that this loss function elicits the τ-quantile of a random variable. Technically, the forecasting problem in this paper (and in the CAViaR literature) is formulated as forecasting Y_t with respect to the check loss, even though the end goal is to forecast the unobservable q_t^τ. The question of evaluating quantile forecasts is a different and interesting problem, but it falls out of the scope of this paper. The interested reader can

[2] This example follows the terminology used in White *et al.* (2015). Arguably, the forecasting equations look more similar to the forecasts induced by a vector autoregressive moving average (VARMA).

[3] Section OA.1 in the Online Appendix provides a list of examples of data transformations allowed by Assumption A.2.

[4] The extension to multiple quantiles for each variable is possible but at the expense of more tedious proofs.

[5] Similar results to those derived in this paper also apply to asymmetric least squares Newey and Powell (1987).

refer to Engle and Manganelli (2004); Giacomini and Komunjer (2005); Komunjer (2013) for more details. It should be noted that the check loss is commonly used to assess the accuracy of quantile forecasts (Giacomini and Komunjer, 2005).

Note that standard asymptotic results for (Q)MLE require that the log-likelihood be twice differentiable, which is not the case with the check loss. Extension of the results to nonsmooth objective functions is of course feasible, and the intuition is that smoothness of the objective function can be replaced by smoothness of the limit if certain remainder terms are small. However, a proper formalization of this intuition requires proofs that are somewhat technical and lengthy (Newey and McFadden, 1994, Sec. 7.4). In contrast, the present paper does not need to deal with such technicalities since the results hold without requiring differentiability of the loss function.

2.3 Estimation.

As usual in the CAViaR literature, the parameter θ in (1) is unknown to the forecaster and needs to be estimated from the data. Let $\tau = (\tau_1, \ldots, \tau_N)' \in [0,1]^N$. The estimation problem is formulated as[6]

$$\hat{\theta}_{T,\tau} \in \arg\min_{\Theta} R_T(\theta, \tau) \ , \quad R_T(\theta, \tau) = \frac{1}{T} \sum_{t=1}^{T} l_t(\theta, \tau) \ , \tag{2}$$

and

$$l_t(\theta, \tau) = \frac{1}{N} \sum_{i=1}^{N} \rho_{\tau_i}(Y_{i\,t} - f_{\theta\, i\, t}). \tag{3}$$

Note that as in most quantile estimation problems, $\hat{\theta}_{T,\tau}$ need not be unique, and in that case one may choose $\hat{\theta}_{T,\tau}$ arbitrarily among the set of candidate minimizers of the criterion. Problem (2) is a special case of an extremum estimator, or M-estimator. While the theory of M-estimation is (obviously) focused on estimation and inference, this paper is concerned with deriving theoretical guarantees for one-step-ahead out-of-sample forecasting with $\hat{\theta}_{T,\tau}$. An important remark is that unlike in classical parametric statistics, $\theta \in \Theta$ is not indexing the family of distributions that generate $\{Y_t\}$. Instead, it only indexes the class of forecasts.

[6] In practice, the forecaster needs to choose a suitable initial value $f_{\theta\,0} = f_0$ to initiate the recursion. A typical choice is the unconditional quantiles of Y_t.

3 Theory

As it is clear from section 2, the relationship between $f_{\theta\, t}$ and Y_t is left unspecified. In particular, $f_{\theta\, t}$ need not represent the true conditional quantiles of Y_t. Nevertheless, the main result in this section states that $f_{\theta\, t}$ achieves the optimal performance within its class in the check loss sense at a near optimal rate.

3.1 Framework

Conditional risk. This section starts by formally defining the notion of performance. Let $M = \lceil \gamma T \rceil$ for some $\gamma > 0$. The *conditional risk* of $\hat{\theta}_{T,\tau}$ is defined as

$$R(\hat{\theta}_{T,\tau}, \tau) := \mathbb{E}\left[\frac{1}{M} \sum_{t=T+1}^{T+M} l_t(\hat{\theta}_{T,\tau}, \tau) \middle| Y_T, \ldots, Y_0, f_0 \right] . \tag{4}$$

It is important to remark that $R(\hat{\theta}_{T,\tau}, \tau)$ is a natural metric of *out-of-sample* performance for time series forecasting: it measures the expected average loss in one-step-ahead out-of-sample forecasting using $\hat{\theta}_{T,\tau}$ given a sample path of in-sample observations and an initial value $f_{\theta\, 0} = f_0$ chosen by the forecaster. Conditioning on the initial value allows us to analyze the properties of the *conditional* quasi-maximum likelihood rather than the *exact* maximum likelihood method (Hamilton, 1994, Ch. 5), which is typically more difficult to implement, and particularly so in misspecified settings.

Note that if the data is independent, it is simpler to define performance by taking an independent copy of the in-sample data, since the dynamics do not play any role for future forecasting, but this is not satisfactory in time series applications (Kuznetsov and Mohri, 2015). Naturally, $R(\hat{\theta}_{T,\tau}, \tau)$ is a random variable.

Dominating process. A key step in the proof of the main result is to find a process $\{d_{\theta\, t}\}$ such that $\|\theta - \dot{\theta}\|_1 \leqslant \delta$ implies that $\|f_{\theta\, t} - f_{\dot{\theta}\, t}\|_1 \leqslant \delta d_{\dot{\theta}\, t}$ for every pair $\theta, \dot{\theta} \in \Theta$ with probability 1. The dominating process in question is given by the following recursion

$$d_{\theta\, t} = 1 + C_s \left(1 + \overline{A}\right) \|Y_{t-1}\|_1 + \|f_{\theta\, t-1}\|_1 + \overline{B} d_{\theta\, t-1} + \epsilon_{d\, t} , \tag{5}$$

where $d_{\theta 0}$ is drawn from the stationary distribution[7] and C_s and $\overline{A}$ are positive finite constants, and $\{\epsilon_{dt}\}$ is an i.i.d. sequence of non-negative random variables. It follows that

$$\left| l_t(\theta, \tau) - l_t(\dot{\theta}, \tau) \right| \leqslant \frac{1}{N} \delta d_{\dot{\theta} t} \tag{6}$$

holds with probability 1. The construction of the dominating process is closely related to the smoothness conditions used to turn pointwise laws of large numbers (LLNs) into uniform LLNs over compact sets. For instance, Assumption A3 in Andrews (1987) requires that

$$\lim_{\delta \to 0} \sup_{T \geqslant 1} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \sup_{\theta \in \mathcal{B}(\dot{\theta}, \delta)} |l_t(\theta, \tau) - l_t(\dot{\theta}, \tau)| = 0 \ ,$$

where $\mathcal{B}(\dot{\theta}, \delta) = \{\theta \in \Theta : \varrho(\dot{\theta}, \theta) \leqslant \delta\}$ and ϱ can be any metric defined on Θ. It is easy to see that inequality (6) together with a suitable uniform moment requirement on $d_{\theta t}$ are enough to verify the smoothness condition A3.

Oracle inequality. An oracle inequality is a probabilistic bound that relates the performance of an estimator to that of an ideal estimator that has best performance in the class, also known as the "oracle" (Donoho and Johnstone, 1994; Candes, 2006). Following Lecué and Mendelson (2016), the M-estimator $\hat{\theta}_{T,\tau}$ satisfies an oracle inequality if the following bound

$$R(\hat{\theta}_{T,\tau}, \tau) \leqslant \inf_{\Theta} R(\theta, \tau) + r_T(N, p)$$

holds with high probability, where $r_T(N, p)$ is a term which converges to zero at a rate that depends on the sample size T, size of the cross-section N, and the complexity of the class of forecasts (quantified by p). Notice that the term does not depend on τ, suggesting that the result holds uniformly over all $\tau \in [0, 1]^N$.

The following condition is key to establish an oracle inequality for the class of multivariate CAViaR forecasts considered in this paper.

[7] Note that assumptions A.1, A.2 and A.3 are sufficient to guarantee the existence of the stationary distribution.

Condition 1 (Moments and mixing). *The following conditions are satisfied by* $\{l_t(\theta,\tau)\}$ *and* $\{d_{\theta t}\}$*, which are given by* (2) *and* (5)*:*

(i) $\theta \in \Theta \subseteq \mathbb{R}^p$*, where* Θ *is compact.*

(ii) $\{l_t(\theta,\tau)\}$ *and* $\{d_{\theta t}\}$ *are strictly stationary and* α*-mixing with* α*-mixing coefficients such that* $\alpha(m) \leqslant \exp(-C_\alpha m^{r_\alpha})$ *for some* $C_\alpha > 0$ *and* $r_\alpha > 0$ *that do not depend on* θ*.*[8]

(iii) There exists $C_L < \infty$ *such that* $\sup_\Theta \|l_t(\theta,\tau)\|_{L_k} \leqslant C_L$ *and* $\sup_\Theta \|d_{\theta t}\|_{L_k} \leqslant C_L$*, for some* $k > p + 2$*.*

(iv) The (conditional and unconditional) distribution of Y_t *is supported on* $\mathcal{Y} \subseteq \mathbb{R}^N$*, where* $\mathcal{Y}$ *has positive Lebesgue measure in* $\mathbb{R}^N$*.*

Condition 1 deserves some discussion.

The first thing to note is that Condition 1 can be verified for a large class of parameter-driven DGP's (Cox, 1981). For instance, Assumptions A.1, A.2 and A.3 imply Condition 1. This is established in this paper via a rather novel application of Markov chain theory. The novelty of the approach consists of deriving V-geometric ergodicity (Liebscher, 2005; Meitz and Saikkonen, 2008a) of the Markov chain given by the DGP, $f_{\theta t}$ and $d_{\theta t}$, which in turn implies the mixing and moment properties described in Condition 1 (Brownlees and Llorens-Terrazas, 2021). Appendix B contains a full derivation of these results.

Condition 1(i) is a standard compactness requirement on the parameter space. While compactness is typically required to guarantee the existence of a minimizer of the criterion function both in sample and in population, this paper requires compactness only to guarantee existence of a minimizer in sample. Condition 1(ii) is a strong mixing assumption (Doukhan, 1994). Although strong mixing assumptions are not the most general type of condition, they are still satisfied by a large number of models such as "stable" Markov chains with absolutely continuous innovations. An interesting example is the class of hidden Markov models given by (7) and (8). Condition 1(iii) is a moment requirement on the loss and the dominating process, which involves Y_t, $f_{\theta t}$ and $d_{\theta t}$. The requirement $k > p + 2$ follows from the choice of the proof techniques used to derive concentration inequalities for the terms on the right-hand side of (9). Condition 1(iv)

[8] See Definition 2 for a formal definition of $\alpha(m)$. $\{X_t\}$ is said to be *strongly mixing* or α*-mixing*, if $\alpha(m) \to 0$ as $m \to \infty$. While the α-mixing coefficients of $\{l_t(\theta,\tau)\}$ and $\{d_{\theta t}\}$ could be different, the condition means that they have a common upper bound.

ensures that the distribution of Y_t is sufficiently well-behaved. In particular, it rules out that Y_t might only take values in some lower-dimensional subspace of $\mathbb{R}^N$.

The assumptions in Engle and Manganelli (2004) and White *et al.* (2015) provide a reasonable benchmark to establish a comparison with Condition 1. In that literature, it is assumed that there exists $\theta_{0,\tau} \in \Theta$ such that $f_{\theta_{0,\tau}\, t} = q_t^\tau$, while in this paper this is not required. The CAViaR literature assumes (inter alia) that the loss process satisfies a uniform law of large numbers (ULLN). Instead, Condition 1 can be seen as a sufficient condition to obtain the assumed ULLN from the CAViaR literature. Furthermore, Condition 1 is sufficient to establish a rate of convergence. In summary, Condition 1 is easier to verify and tailored to the goal of this paper – which is out-of-sample forecasting.

3.2 Assumptions

This sub-section gives a list of sufficient conditions under which Condition 1 holds.

Data generating process. Suppose that the data generating mechanism is given by the following hidden Markov model

$$Y_t = g_{y1}(H_t) + g_{y2}(H_t)\epsilon_{Y\,t} \tag{7}$$

$$H_t = g_{h1}(H_{t-1}) + g_{h2}(H_{t-1})\epsilon_{H\,t}\ , \tag{8}$$

where Y_t takes values in $\mathcal{Y} \subseteq \mathbb{R}^N$ and H_t takes values in $\mathcal{H} \subseteq \mathbb{R}^{p_h}$; g_{y1}, g_{y2}, g_{h1} and g_{h2} are Borel-measurable functions, and $\{\epsilon_{Y\,t}\}$ and $\{\epsilon_{H\,t}\}$ are i.i.d. sequences of random variables supported in $\mathcal{Y}$ and $\mathcal{H}$, respectively. The process is initialized at the stationary distribution, and assumption A.1 below is sufficient to guarantee its existence. To simplify notation, take $\mathcal{Y} = \mathbb{R}^N$ and $\mathcal{H} = \mathbb{R}^{p_h}$.

A.1. *The process given by equations* (7) *and* (8) *satisfies the following:*

(i) The functions g_{h1} and g_{h2} are bounded on bounded subsets of $\mathbb{R}^{p_h}$. Moreover, $\|g_{h1}(h)\|_1 \leqslant a_h\|h\|_1 + o(\|h\|_1)$ and $\|\!|g_{h2}(h)|\!\|_1 \leqslant b_h\|h\|_1 + o(\|h\|_1)$ as $\|h\|_1 \to \infty$. The matrix function $g_{h2}(h)$ is non-singular for all $h \in \mathbb{R}^{p_h}$, and $\inf_{h\in\mathbb{R}^{p_h}} |\det(g_{h2}(h))| > 0$.

(ii) The functions g_{y1} and g_{y2} are bounded on bounded subsets of $\mathbb{R}^{p_h}$. Moreover, $\|g_{y1}(h)\|_1 \leqslant C_y\|h\|_1$ and $\|\!|g_{y2}(h)|\!\|_1 \leqslant C_y\|h\|_1$ for some $C_y < \infty$. The matrix

function $g_{y2}(h)$ *is non-singular for all* $h \in \mathbb{R}^{p_h}$, *and* $\inf_{h\in\mathbb{R}^{p_h}} |\det(g_{y2}(h))| > 0$.

(iii) $\{\epsilon_{Y\,t}\}$ *and* $\{\epsilon_{H\,t}\}$ *are i.i.d. sequences of random variables with absolutely continuous distributions w.r.t. Lebesgue measure on* $\mathbb{R}^N$ *and* $\mathbb{R}^{p_h}$ *(resp.) and are supported in* $\mathbb{R}^N$ *and* $\mathbb{R}^{p_h}$ *(resp.), with densities* ϕ_Y *and* ϕ_H *that are bounded away from zero on compact subsets of* $\mathbb{R}^{p_h}$ *and* $\mathbb{R}^N$ *(resp.). The random variables* $\epsilon_{Y\,t}$ *and* $\epsilon_{H\,t}$ *satisfy* $\|\epsilon_{Y\,t}\|_{L_k} < \infty$ *and* $\|\epsilon_{H\,t}\|_{L_k} < \infty$ *(resp.) for some* $k > p + 2$.

(iv) $\mathbb{E}(a_h + b_h\|\epsilon_{H\,t}\|_1)^k < 1$.

Class of forecasts

A.2. *The class of forecasts given by* (1) *satisfies the following:*

(i) $|||B|||_1 \leqslant \overline{B} < 1$.

(ii) $\det(A) \neq 0$ *and* $|||A|||_1 \leqslant \overline{A} < \infty$.

(iii) For each $h \in \mathbb{R}^{p_h}$, *there exists some* $z \in \mathbb{R}^N$ *such that* $\det\left(\frac{\partial \tilde{s}_\lambda(h,z)}{\partial z}\right) \neq 0$, *where* $\tilde{s}_\lambda(h,z) := s_\lambda(g_{y1}(h) + g_{y2}(h)z)$.

(iv) There exists some $C_s < \infty$ *such that* $\|s_\lambda(u)\|_1 \leqslant C_s\|u\|_1$ *and* $\|s_\lambda(u) - s_{\dot\lambda}(u)\|_1 \leqslant C_s\|u\|_1\|\lambda - \dot\lambda\|_1$ *for every* u, *where* C_s *does not depend on* λ *nor* $\dot\lambda$.

(v) $\theta = (\omega', vec(A)', vec(B)', \lambda')' \in \Theta \subseteq \mathbb{R}^p$, *where* Θ *is compact.*

(vi) There exists $D_f \subseteq \mathbb{R}^N$ *such that* s_λ *is a diffeomorphism in* D_f.

Dominating process

A.3. *The dominating process given by* (5) *satisfies the following:*

(i) $\{\epsilon_{d\,t}\}$ *is an i.i.d. sequence of random variables with absolutely continuous distributions w.r.t. Lebesgue measure on* $\mathbb{R}$ *and are supported in* $[0,1]$, *with density* ϕ_d *that is bounded away from zero on compact subsets of* $[0,1]$.

Remarks. Assumption A.1 is a multivariate extension of standard assumptions used to establish geometric ergodicity of nonlinear time series models (Masry and Tjøstheim, 1995; Lu and Jiang, 2001; Lanne and Saikkonen, 2005; Meitz and Saikkonen, 2008a;

Brownlees and Llorens-Terrazas, 2021) and it allows for a fairly broad class of parameter-driven processes. Assumption A.1(i) is similar to Assumption 3.2 in Masry and Tjøstheim (1995) and it implies that (8) is dominated asymptotically by a stable linear model. As Masry and Tjøstheim (1995) emphasize, such a requirement is mild, since functions that grow everywhere faster than a stable linear model are nonstationary. Assumption A.1(ii) allows for a fair amount of flexibility in equation (7). In particular, it requires $\|Y_t\|_1$ to be bounded from above by a linear function of $\|H_t\|_1$. Assumption A.1(iii) imposes conditions on the random variables $\epsilon_{H\,t}$ and $\epsilon_{Y\,t}$ that are analogous to standard conditions used in the literature. Assumption A.1(iv) is a stability condition analogous to the one assumed in Masry and Tjøstheim (1995) or Lanne and Saikkonen (2005).

Assumption A.2(i) is a stability condition for $f_{\theta\,t}$ and $d_{\theta\,t}$. Intuitively, this assumption ensures that the forecasts have a sufficiently "fading memory" (Pötscher and Prucha, 1997). Note that A.2(i) implies that the spectral radius of B is strictly less than unity. Assumption A.2(ii) requires A to be non-singular, so Θ must avoid the region of the parameter space where $\det(A) = 0$. For instance, we may require that $|\det(A)| \geqslant \underline{A} > 0$. The upper bound $\overline{A}$ can be chosen arbitrarily by the forecaster, although higher values of $\overline{A}$ have the effect of slowing down the geometric decay rate of the strong mixing coefficients. Assumptions A.2(iii), (iv) and (vi) are relatively mild and allow for a broad class of transformations s_λ that include as special cases differentiable approximations to symmetric and asymmetric absolute values (see the Online Appendix for examples of s_λ that satisfy Assumption A.2).

Assumption A.3 is an auxiliary assumption that is useful to simplify the proof of irreducibility and aperiodicity of the "companion Markov chain" defined in (11). More specifically, the assumption permits the use of proof techniques similar in spirit to Meitz and Saikkonen (2008b, Lemma 2) and Meyn and Tweedie (1993, Ch. 7).

Condition 1 leads to an oracle inequality for the class of forecasts introduced in (1), with out-of-sample performance defined as in equation (4).

Theorem 1. *Suppose Condition 1 holds. Then, there exists a positive constant σ (uniformly over τ) such that, for all T sufficiently large, it holds that*

$$R(\hat{\theta}_T, \tau) \leqslant \inf_{\Theta} R(\theta, \tau) + 2\sigma\sqrt{\frac{p \log T}{NT}}$$

with probability at least $1 - \log^{-1} T - o(\log^{-1} T)$.

Some remarks are in order. First, if the forecaster believes that there exists $\theta_{0,\tau} \in \Theta$ such that $f_{\theta_{0,\tau} t} = q_t^\tau$, then we have the analogous result of the consistency of CAViaR for out-of-sample forecasting in finite samples and with a rate of convergence. Second, if there is no $\theta \in \Theta$ such that $f_{\theta t} = q_t^\tau$, Theorem 1 still provides finite-sample performance guarantees for out-of-sample forecasting in the check loss sense.

The constant σ^2 is application-specific and may be interpreted as an upper bound for the long run variance of the loss process. See Proposition 3 for a precise definition of σ^2. The rate of convergence $\sqrt{\log T/T}$ is sometimes referred to as the classical rate of convergence of empirical risk minimization in the learning literature for classification with i.i.d. data (Devroye *et al.*, 1996, Ch. 12). With fixed N, the theorem implies that the M-estimator is consistent with respect to the class of forecasts indexed by Θ, meaning that $|R(\hat{\theta}_{T,\tau}, \tau) - \inf_\Theta R(\theta, \tau)| \xrightarrow{\mathsf{p}} 0$ as $T \to \infty$. In other words, the M-estimator achieves asymptotically the optimal forecasting performance attainable within the class of algorithms considered.

One can interpret NT as the "effective" sample size, i.e. the number of time series multiplied by the sample size for each series. However, it should be noted that the proof techniques employed in this paper do not allow p nor N to diverge to infinity. This limits the extent to which Theorem 1 can be regarded as a "high-dimensional" result, in the sense that it cannot be used to draw conclusions about specifications for which $p \to \infty$ as $N \to \infty$. Still, it is a useful result for specifications that rely on "commonalities" on the parameters such as composite likelihood (Pakel, Shephard, and Sheppard, 2011), where p is fixed and the performance of $\hat{\theta}_{T,\tau}$ can improve by pooling information across series. An example of such a procedure is used in the empirical section.

It is important to emphasize that Theorem 1 is stronger than a consistency result for the prediction performance of the M-estimator since it is non-asymptotic (it holds for each sufficiently large T) and it provides a specific rate of convergence for the performance of the M-estimator. As will be noted in section 4, oracle inequalities can be proved with techniques similar to those used to obtain ULLNs, or "uniform convergence over a class of functions" (Vapnik and Chervonenkis, 1971). However, the oracle inequality stated in this paper is stronger than a ULLN, since it also provides information about the rate at which the performance of the forecast is approaching its optimal level (Vapnik, 1999). Lastly, we emphasize that the existence of an optimal prediction rule $\theta_{0,\tau} = \arg\min_\Theta R(\theta, \tau)$ is not required by the theorem.

3.3 Additional Discussion

This paper studies the properties of the M-estimator when the time series is generated by a parameter-driven process. Clearly, an observation-driven process may be entertained instead. In this case, the analysis of the performance of the M-estimator can be carried out using the same strategy developed in this paper. However, some of the proofs will differ and the analysis of this case is left for future research.

The theoretical framework of this paper does not require the class of algorithms to have special approximation properties or to include the optimal forecast associated with the data generating process and the loss function. What is key in the framework is that, loosely speaking, forecasts forget the past exponentially fast.

Instead of comparing the performance of the M-estimator against the optimal risk attainable in the class, one may wish to compare against the risk of the optimal 1-step-ahead forecast. For the check loss, the optimal 1-step-ahead forecast is the conditional quantile (assuming it exists) (Giacomini and Komunjer, 2005). Thus, the risk of the optimal 1-step-ahead forecast may be defined as

$$R^*(\tau) = \mathbb{E}\left[\frac{1}{M}\frac{1}{N}\sum_{t=T+1}^{T+M}\sum_{i=1}^{N}\rho_{\tau_i}(Y_{it} - q_{it}^{\tau_i})\middle| Y_T, \ldots, Y_0, f_0\right].$$

The performance of the M-estimator relative to the risk of the optimal 1-step-ahead forecast may be expressed as

$$R(\hat{\theta}_{T,\tau}, \tau) - R^*(\tau) = \left[\inf_{\Theta} R(\theta, \tau) - R^*(\tau)\right] + \left[R(\hat{\theta}_{T,\tau}, \tau) - \inf_{\Theta} R(\theta, \tau)\right].$$

The first term is called the approximation error and the second term is called the estimation error (Devroye *et al.*, 1996, Ch. 12). Notice that oracle inequalities control the estimation error. The approximation error is typically difficult to control, especially in a time series setting. There are a number of contributions that, in some sense, attempt to control the approximation error (Nelson, 1992). In general, the analysis of the approximation error requires additional assumptions. For this reason learning theory typically focuses on studying the estimation error, as it is done in this paper.

The focus of this paper is on quantile forecasting, and as such the theory is derived for the check loss function. Notwithstanding, inspection of the proof strategy reveals that similar results can be derived for other loss functions, so long as they satisfy dominance

requirements akin to (6) above. This is the case for the (asymmetric) least squares criterion proposed by Newey and Powell (1987), that is, $\varrho_{\tau_i}(u) = u^2|\tau_i - I(u < 0)|$. Note that with $l_t(\theta, \tau) = \frac{1}{N}\sum_{i=1}^{N} \varrho_{\tau_i}(Y_{it} - f_{\theta\, it})$, it holds that

$$|l_t(\theta,\tau) - l_t(\dot{\theta},\tau)| \leqslant \frac{1}{N}\left\|f_{\theta t} - f_{\dot{\theta} t}\right\|_2^2 + \frac{2}{N}\sum_{i=1}^{N}|Y_{it} - f_{\dot{\theta}\, it}||f_{\theta\, it} - f_{\dot{\theta}\, it}|\ ,$$

and it is not difficult to verify that a dominating process $d_{\theta t}$ analogous to (5) can be derived so that $\|\theta - \dot{\theta}\|_2 \leqslant \delta$ implies that $\left\|f_{\theta t} - f_{\dot{\theta} t}\right\|_2 \leqslant \delta d_{\dot{\theta} t}$ for every pair $\theta, \dot{\theta} \in \Theta$ with probability 1. However, notation and proofs do require modifications which are not pursued here.

4 Sketch of proof of Theorem 1

This section explains the main steps to derive the proof of Theorem 1, which are broken down in four different propositions. Proofs can be found in Appendix A.

Step 1: Basic inequality. The first step consists of noting that the discrepancy between $R(\hat{\theta}_T, \tau)$ and $\inf_\Theta R(\theta, \tau)$ – also known as "regret" in the learning literature – can be upper bounded by two key terms.

Proposition 1. *Let* $\overline{R}(\theta, \tau) = \mathbb{E}l_t(\theta, \tau)$. *Then,*

$$R(\hat{\theta}_{T,\tau}, \tau) - \inf_\Theta R(\theta, \tau) \leqslant 2\sup_\Theta |R_T(\theta, \tau) - \overline{R}(\theta, \tau)| + 2\sup_\Theta |R(\theta, \tau) - \overline{R}(\theta, \tau)|\ . \quad (9)$$

It is important to emphasize that Proposition 1 is a general result that only requires the loss process to be stationary.[9] Note that when the data is i.i.d., $R(\theta, \tau) = \overline{R}(\theta, \tau)$ and the inequality in Proposition 1 corresponds to the classic inequality derived in Vapnik and Chervonenkis (1974) (Devroye *et al.*, 1996), which is routinely used to derive bounds on the performance of empirical risk minimization.

The first term on the right hand side of (9) is the supremum of the empirical process associated with the prediction loss $l_t(\theta, \tau)$. The second term is the supremum of the average difference between conditional and unconditional expectations of the prediction

[9] To clarify, $\hat{\theta}_{T,\tau}$ is obtained by fixing an initial value f_0, but the analysis can be carried out with the process $f_{\theta t}$ initialized at the stationary distribution because Proposition 1 only involves the conditional expectation defined in (4), which is already conditioned on f_0.

loss over the out-of-sample period.

Step 2: Covering. The second step is summarized in the following.

Proposition 2. *Suppose Condition 1 is satisfied. Then, for any $\varepsilon > 0$ it holds that*

$$\begin{aligned}&\Pr\left(\sup_{\Theta}|R_T(\theta,\tau)-\overline{R}(\theta,\tau)|>\frac{\varepsilon}{2}\right)\\&\quad\leqslant\left(1+\frac{24C_\Theta C_d}{N\varepsilon}\right)^p\sup_{\Theta}\left[P_1^T\left(l_t(\theta,\tau),\frac{\varepsilon}{4}\right)+P_1^T\left(d_{\theta\,t},C_d\right)\right]\ ,\end{aligned}$$

where $P_a^b(U_t,\varepsilon)=\Pr\left(\left|\frac{1}{b-a+1}\sum_{t=a}^b U_t-\mathbb{E}U_t\right|>\varepsilon\right)$, and

$$\begin{aligned}&\Pr\left(\sup_{\Theta}|R(\theta,\tau)-\overline{R}(\theta,\tau)|>\frac{\varepsilon}{2}\right)\\&\quad\leqslant\left(1+\frac{24C_\Theta C_d}{N\varepsilon}\right)^p\sup_{\Theta}\left[P_{T+1}^{T+M}\left(\mathbb{E}_T l_t(\theta,\tau),\frac{\varepsilon}{4}\right)+P_{T+1}^{T+M}\left(\mathbb{E}_T d_{\theta\,t},C_d\right)\right]\ ,\end{aligned}$$

where $C_\Theta=\sup_\Theta\|\theta\|_1$ and $C_d=\sup_\Theta\|d_{\theta\,t}\|_{L_1}$.

Proposition 2 relies on a "covering argument" which has appeared in the literature to establish uniform laws of large numbers (Amemiya, 1985; Davidson, 1994) and in empirical risk minimization for time series (Jiang and Tanner, 2010).

Step 3: Concentration inequality (part I). The third step uses a slight modification of a well known concentration inequality for sums of α-mixing processes (Liebscher, 1996). Proposition 3 formalizes the result.

Proposition 3. *Suppose Condition 1 is satisfied. Then, for all T sufficiently large and for $\varepsilon_T=\sigma\sqrt{\frac{p\log T}{NT}}$, it holds that*

$$\begin{aligned}&\left(1+\frac{24C_\Theta C_d}{N\varepsilon_T}\right)^p\sup_{\Theta}\Pr\left(\left|\frac{1}{T}\sum_{t=1}^T l_t(\theta,\tau)-\mathbb{E}l_t(\theta,\tau)\right|>\frac{\varepsilon_T}{4}\right)\leqslant\frac{1}{\log T}\ \textit{and}\\&\left(1+\frac{24C_\Theta C_d}{N\varepsilon_T}\right)^p\sup_{\Theta}\Pr\left(\left|\frac{1}{T}\sum_{t=1}^T d_{\theta\,t}-\mathbb{E}d_{\theta\,t}\right|>C_d\right)\leqslant o\left(\frac{1}{\log T}\right)\ \textit{as}\ T\to\infty\ ,\end{aligned}$$

where $\sigma^2=8(2^{1-1/k}+1)C_L^2\left(\frac{1}{4^{1-2/k}}+2\sum_{m=1}^\infty\exp\left(-C_\alpha m^{r_\alpha}\right)^{1-\frac{2}{k}}\right)$.

Step 4: Concentration inequality (part II). The fourth step – summarized in Proposition 4 – uses a well known result by Ibragimov (1962) that establishes a bound

on the L_p-norm of the discrepancy between conditional and unconditional expectations of α-mixing processes.

Proposition 4. *Suppose Condition 1 is satisfied. Then, for all T sufficiently large and for $\varepsilon_T = \sigma\sqrt{\frac{p\log T}{NT}}$, it holds that*

$$\left(1+\frac{24C_\Theta C_d}{N\varepsilon_T}\right)^p \sup_\Theta \Pr\left(\left|\frac{1}{M}\sum_{t=T+1}^{T+M}\mathbb{E}_T l_t(\theta,\tau)-\mathbb{E}l_t(\theta,\tau)\right|>\frac{\varepsilon_T}{4}\right)\leqslant\frac{1}{\log T} \text{ and}$$
$$\left(1+\frac{24C_\Theta C_d}{N\varepsilon_T}\right)^p \sup_\Theta \Pr\left(\left|\frac{1}{M}\sum_{t=T+1}^{T+M}\mathbb{E}_T d_{\theta\,t}-\mathbb{E}d_{\theta\,t}\right|>C_d\right)\leqslant o\left(\frac{1}{\log T}\right) \text{ as } T\to\infty\ ,$$

where σ^2 is defined in Proposition 3.

It follows from Propositions 2, 3 and 4 that, for all T sufficiently large,

$$2\sup_\Theta|R(\theta,\tau)-\overline{R}(\theta,\tau)|+2\sup_\Theta|R_T(\theta,\tau)-\overline{R}(\theta,\tau)|\leqslant 2\sigma\sqrt{\frac{p\log T}{NT}}$$

holds with high probability. This fact and Proposition 1 imply Theorem 1.

Proof of Theorem 1. Follows by Condition 1 and Propositions 1, 2, 3 and 4. □

5 Application to backtesting global Growth-at-Risk

The International Monetary Fund (IMF) has recently popularized a risk measure for GDP growth called Growth-at-Risk (GaR), which is the worst-case scenario GDP growth at a given coverage level and is the analog of the classic Value-at-Risk (VaR) used in risk management. Several institutions such as the IMF or the European Central Bank publish GaR for major world economies on a routine basis. One of the appealing features of quantile regression is that it allows direct linkage of downside risk predictors to the quantiles of GDP growth.

This application explores the use of the multivariate CAViaR class defined in the theoretical framework of this paper. The CAViaR class is closely related to the quantile regression techniques put forward by Adrian *et al.* (2019). A key difference is the recursive nature of the CAViaR forecasts, which rely on the entire past of GDP growth – similarly to GARCH models. In fact, GARCH forecasts that use no information other than the past of GDP growth exhibit better performance than quantile regressions that use

external information such as the national financial conditions index (NFCI) (Brownlees and Souza, 2021). This suggests that – quite remarkably – the (entire) past of GDP growth seems to be the key driver of the time variation in the conditional distribution of GDP growth. The present paper also investigates the "synergies" between GARCH and CAViaR.

Description of the exercise. The data consists of a balanced panel of GDP growth rates for 24 OECD countries that spans from 1961Q1 to 2019Q1. The sample comprises all countries for which GDP data are available since at least 1973Q1 to match some of the predictors used in the quantile regression analysis. GDP growth rates are defined as the quarterly percentage change in seasonally adjusted real GDP and are obtained from the OECD database.

The specifications considered in the exercise can be classified in three broad types. First, a class of GARCH(1,1) models is entertained, estimated via the pooled GARCH procedure proposed by (Pakel, Shephard, and Sheppard, 2011). The pooled GARCH procedure relies on a specification where the dynamic parameters of the GARCH recursion are common for all countries and are estimated via composite (quasi) maximum likelihood, while the intercept parameter is country-specific and estimated via variance targeting. This is done because in relatively short time series such as GDP growth, it is challenging to obtain stable parameter estimates (Brownlees *et al.*, 2011). Results are reported for both GARCH models estimated on GDP growth – labeled as GARCH in Table 5 – and on the residuals of an AR(1) – labeled as AR-GARCH.

Second, a number of quantile regression models (QR) are implemented following Adrian *et al.* (2019). Quantile regression requires specifying a set of downside risk predictors. The list of variables includes country-specific variables such as the national financial conditions index (NFCI), credit-to-GDP gap and growth (CG and CR), term spread (TS), housing prices (HP), the World Uncertainty Index (WUI), and economic policy uncertainty (EPU), as well as global predictors such as the global real activity factor (GF), stock variance (SV), credit spread (CS), and the geopolitical risk index (GPR). The details on the data availability, construction and imputation can be found in Brownlees and Souza (2021).

Third, a number of special cases of (1) are implemented, labeled as pooled VFV in

Table 5. All pooled VFV specifications take the form

$$f_{\theta\, i\, t} = \omega_i + \alpha s_\lambda(Y_{i\, t-1}) + \beta f_{\theta\, i\, t-1}, \quad i = 1, \ldots, N,$$

where $s_\lambda(u) = b(\sqrt{1 + (u/b)^2} - 1)\ |\tau - I(u < 0)|$, $\tau \in [0, 1]$, $b \in [\underline{b}, \bar{b}]$, $\underline{b} > 0$ and $\lambda = (\tau, b)'$. Note that $s_\lambda(u)$ is an arbitrarily good approximation of $|u||\tau - I(u < 0)|$ as $b \to 0^+$, which corresponds to the symmetric absolute value (Sym) and asymmetric slope (Asym) specifications introduced by Engle and Manganelli (2004) if $\tau = 1/2$ and $\tau \neq 1/2$, respectively. In the asymmetric specification, τ is set to 0.05. See Example 3 in the Online Appendix for a verification of Assumption A.2 for this choice of the function s_λ. The restrictions on the parameters α, β and $\omega_1, \ldots, \omega_N$ are naturally deduced from Assumption A.2.

The pooled VFV specifications impose that the dynamic parameters α and β are common for all countries and are estimated with the procedure described in equation (2). This procedure is analogous to the composite likelihood approach mentioned above – but with the check loss instead of the Gaussian (quasi) likelihood. Results are reported for VFV specifications estimated on (*i*) GDP growth, labeled as VFV in Table 5; (*ii*) the residuals of an AR(1) (VFV-AR); and (*iii*) on the standardized residuals of the pooled GARCH, both on GDP growth and on the AR(1) residuals (GARCH-VFV and AR-GARCH-VFV, respectively).

Recursive estimation is carried out for all specifications under consideration for each quarter from 1973Q1 to 2016Q4 and out-of-sample forecasts are computed starting from 1983Q4. Starting the forecasting exercise from 1983Q4 implies that the out-of-sample period is based on approximately 75% of the available data.

Marginal GaR forecasts are evaluated using the check loss over the out-of-sample period, that is,

$$\text{CL} = \frac{1}{M} \sum_{t=T+1}^{T+M} \sum_{i=1}^{N} \rho_\tau(Y_{i\, t} - f_{\theta\, i\, t}). \tag{10}$$

For completeness, Table 5 also reports Coverage and Length, which are defined as

$$\text{Cov} = \frac{1}{MN} \sum_{t=T+1}^{T+M} \sum_{i=1}^{N} I\left(Y_{i\, t} > f_{\hat{\theta}_{T,\tau}\, i\, t}\right), \quad \text{Len} = \frac{1}{MN} \sum_{t=T+1}^{T+M} \sum_{i=1}^{N} \left(\hat{Q}_{0.99}(Y_i) - f_{\hat{\theta}_{T,\tau}\, i\, t}\right),$$

where $\hat{Q}_{0.99}(Y_i)$ denotes the unconditional 99% empirical quantile of the i^{th} series estimated on the entire sample. All else being equal, GaR forecasts with a smaller length are typically preferred.

Table 1: 95% GaR Marginal Forecast Evaluation

Method	Specification	Cov	Len	Check
Benchmark	Historical	94.41	5.42	0.14
Pooled GARCH	GARCH	93.28	5.17	4.56
Pooled GARCH	AR-GARCH	93.12	5.07	11.66
Pooled VFV	Sym	93.59	5.28	−0.22
Pooled VFV	Asym	94.82	5.37	8.45
Pooled VFV	AR-VFV Sym	93.94	5.20	11.52
Pooled VFV	AR-VFV Asym	95.36	5.40	2.87
Pooled GARCH-VFV	GARCH-VFV Sym	93.09	5.18	4.55
Pooled GARCH-VFV	GARCH-VFV Asym	94.07	5.24	12.60
Pooled GARCH-VFV	AR-GARCH-VFV Sym	93.06	5.09	12.87
Pooled GARCH-VFV	AR-GARCH-VFV Asym	93.53	5.12	12.51
QR	NFCI	92.77	5.17	3.85
QR	NFCI + TS	91.13	5.08	−0.13
QR	NFCI + TS + GF	90.72	5.09	−1.23
QR	Full	89.39	5.15	−19.18

Cov: Average empirical coverage; Len: average empirical length; Check: first row: average check loss of the historical benchmark; remaining rows: percentage improvement in average check loss relative to historical benchmark.

The results of the exercise can be summarized as follows. First, the VFV specifications on the standardized residuals of (AR-) GARCH perform best out-of-sample. The approach exploits non-obvious dynamics of the standardized residuals of the GARCH procedure. The dynamics are not obvious in the sense that they are not captured by inspection of the autocorrelation function of the standardized residuals nor their absolute values or squares. In addition, empirical support in favor of AR-GARCH-CAViaR methodologies has been documented in Kuester *et al.* (2006), which use more than 30 years of daily return data on the NASDAQ Composite Index. Panel Diebold-Mariano tests statistics of superior predictive ability based on the check loss are reported in Table OA.6.

Second, a comparison between GARCH versus the VFV reveals that the GARCH specification outperforms the VFV in terms of the check loss, whereas the VFV specifica-

tion provides better out-of-sample coverage. This is perhaps surprising in the sense that the VFV specification is designed to minimize the check loss function. This suggests that the approach to GaR forecasting using conditional volatility is particularly useful in this dataset. Another possible explanation may be that the GARCH specification benefits more from exploiting "commonalities" in conditional variance with respect to the VFV specification, which exploits commonalities in conditional quantiles.

Third, asymmetries in conditional volatility of GDP growth do not play an important role, but they still matter for the quantiles. The results from the specifications Pooled VFV (Sym) vs Pooled VFV (Asym) in Table 5 suggest that negative growth rates have more predictive power for conditional quantiles than positive ones in the check loss sense. However, the narrative changes when the VFV specifications are run on the residuals of AR or AR-GARCH. This is perhaps not surprising since the relevant asymmetries are found at the zero growth level.

To sum up, these forecasting results suggest that using the entire past of GDP growth provides a benchmark that is easy to implement and hard to beat even by cross-sectional quantile regression approaches based on external information such as the NFCI.

6 Concluding Remarks

This paper establishes theoretical guarantees for out-of-sample multivariate dynamic quantile forecasts. A key feature of the analysis is that the relationship between the data generating process and the class of algorithms is unspecified. The main result implies that the predictor that minimizes the in-sample average check loss achieves asymptotically the optimal predictive performance that is attainable within the class, even when it is fully misspecified.

To put it differently, this paper shows that the conditional quasi-maximum likelihood estimator achieves the oracle's out-of-sample predictive performance within the class of VAR for VaR specifications considered here. A crucial condition to obtain this type of result is that the data and the forecast forget their past sufficiently fast and that enough moments exist. The paper also gives a set of primitive assumptions that are sufficient to validate this condition.

This work exemplifies how to combine the tools of statistical learning theory and nonlinear time series to obtain performance guarantees for time series forecasting. Following

the "algorithmic modeling" culture fostered by Breiman (2001), this paper hopefully paves the way for the development of new forecasting strategies for time series applications with minimal assumptions on how the data is generated.

A Proofs

A.1 Main Result

Proof of Proposition 1. Let $\{l_t^G(\theta,\tau)\}$ be an independent copy of $\{l_t(\theta,\tau)\}$ (initialized at the stationary distribution). Define $\overline{R}(\hat{\theta}_{T,\tau},\tau) = \mathbb{E}l_t^G(\hat{\theta}_{T,\tau},\tau)$. By the properties of infimum and supremum and the definition of empirical risk minimizer (i.e. $R_T(\theta,\tau) \geqslant R_T(\hat{\theta}_{T,\tau},\tau)$ for all $\theta \in \Theta$), we have that

$$
\begin{aligned}
&R(\hat{\theta}_{T,\tau},\tau) - \inf_{\Theta} R(\theta,\tau) \\
&\quad = R(\hat{\theta}_{T,\tau},\tau) - \overline{R}(\hat{\theta}_{T,\tau},\tau) + \overline{R}(\hat{\theta}_{T,\tau},\tau) - \inf_{\Theta}\left[\overline{R}(\theta,\tau) + R(\theta,\tau) - \overline{R}(\theta,\tau)\right] \\
&\quad \leqslant \left[R(\hat{\theta}_{T,\tau},\tau) - \overline{R}(\hat{\theta}_{T,\tau},\tau)\right] + \left[\overline{R}(\hat{\theta}_{T,\tau},\tau) - \inf_{\Theta}\overline{R}(\theta,\tau)\right] - \inf_{\Theta}\left[R(\theta,\tau) - \overline{R}(\theta,\tau)\right] \\
&\quad \leqslant 2\sup_{\Theta}|R(\theta,\tau) - \overline{R}(\theta,\tau)| + \left[\overline{R}(\hat{\theta}_{T,\tau},\tau) - \inf_{\Theta}\overline{R}(\theta,\tau)\right] \\
&\quad \leqslant 2\sup_{\Theta}|R(\theta,\tau) - \overline{R}(\theta,\tau)| + 2\sup_{\Theta}|R_T(\theta,\tau) - \overline{R}(\theta,\tau)| \,,
\end{aligned}
$$

where the last inequality follows by Lemma 8.2 in Devroye, Györfi, and Lugosi (1996). □

Proof of Proposition 2. The proof is based on a covering argument similar in spirit to Jiang and Tanner (2010, Prop. 2). Let $\{\Theta_j\}_{j=1}^{N_\delta}$, where $\Theta_j = \{\theta : \|\theta - \theta_j\|_1 \leqslant \delta, \theta_j \in \Theta\}$ be a δ-covering of Θ and N_δ is the covering number. The choice of $\delta > 0$ will be determined in what follows. By the union bound it follows that

$$
\begin{aligned}
&\Pr\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{T} l_t(\theta,\tau) - \mathbb{E}l_t(\theta,\tau)\right| > \frac{\varepsilon}{2}\right) \\
&\qquad \leqslant \sum_{j=1}^{N_\delta}\Pr\left(\sup_{\theta\in\Theta_j}\left|\frac{1}{T}\sum_{t=1}^{T} l_t(\theta,\tau) - \mathbb{E}l_t(\theta,\tau)\right| > \frac{\varepsilon}{2}\right).
\end{aligned}
$$

Add and subtract $l_t(\theta_j,\tau) - \mathbb{E}l_t(\theta_j,\tau)$, use the fact that if $|a+b| > \varepsilon$, then either $|a| > \varepsilon/2$ or $|b| > \varepsilon/2$, and again by the union bound we can write

$$
\begin{aligned}
&\Pr\left(\sup_{\theta\in\Theta_j}\left|\frac{1}{T}\sum_{t=1}^{T} l_t(\theta,\tau) - \mathbb{E}l_t(\theta,\tau)\right| > \frac{\varepsilon}{2}\right) \\
&\leqslant \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} l_t(\theta_j,\tau) - \mathbb{E}l_t(\theta_j,\tau)\right| > \frac{\varepsilon}{4}\right) \\
&\quad + \Pr\left(\sup_{\theta\in\Theta_j}\left|\frac{1}{T}\sum_{t=1}^{T}[l_t(\theta,\tau) - l_t(\theta_j,\tau)] - \mathbb{E}[l_t(\theta,\tau) - l_t(\theta_j,\tau)]\right| > \frac{\varepsilon}{4}\right) \,.
\end{aligned}
$$

Now, by (6) we have that $|l_t(\theta,\tau) - l_t(\theta_j,\tau)| \leqslant \frac{\delta}{N} d_{\theta_j t}$ with probability 1, which is proven in Lemma OA.2 (see Online Appendix). By the triangular inequality, the second term

is bounded above by

$$\Pr\left(\sup_{\theta\in\Theta_j}\frac{1}{T}\sum_{t=1}^{T}|l_t(\theta,\tau)-l_t(\theta_j,\tau)|+|\mathbb{E}[l_t(\theta,\tau)-l_t(\theta_j,\tau)]|>\frac{\varepsilon}{4}\right)$$
$$\leqslant \Pr\left(\frac{1}{T}\sum_{t=1}^{T}d_{\theta_j t}+\mathbb{E}d_{\theta_j t}>\frac{N\varepsilon}{4\delta}\right)\ .$$

Furthermore, since $\sup_\Theta \mathbb{E}(d_{\theta_j t})\leqslant C_d$ for some $C_d<\infty$, by choosing $\delta=N\varepsilon/(12C_d)$ it follows that

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}d_{\theta_j t}+\mathbb{E}d_{\theta_j t}>3C_d\right)=\Pr\left(\frac{1}{T}\sum_{t=1}^{T}d_{\theta_j t}-\mathbb{E}d_{\theta_j t}>3C_d-2\mathbb{E}d_{\theta_j t}\right)$$
$$\leqslant \Pr\left(\frac{1}{T}\sum_{t=1}^{T}d_{\theta_j t}-\mathbb{E}d_{\theta_j t}>C_d\right).$$

Finally, the claim follows by noting that

$$N_\delta\leqslant\left(1+\frac{2C_\Theta}{\delta}\right)^p=\left(1+\frac{24C_\Theta C_d}{N\varepsilon}\right)^p.$$

The same covering argument applies to the second part of the claim with $l_t(\theta,\tau)$ and $d_{\theta t}$ replaced by $\mathbb{E}_T l_t(\theta,\tau)$ and $\mathbb{E}_T d_{\theta t}$, respectively. This is because $|\mathbb{E}_T l_t(\theta,\tau)-\mathbb{E}_T l_t(\theta_j,\tau)|\leqslant \mathbb{E}_T|l_t(\theta,\tau)-l_t(\theta_j,\tau)|\leqslant\frac{2}{N}\delta\mathbb{E}_T d_{\theta_j t}$ by Jensen's inequality and the order-preserving property of the conditional expectation. □

Proof of Proposition 3. Let $\widetilde{U}_{\theta t}=l_t(\theta,\tau)-\mathbb{E}l_t(\theta,\tau)$ and $\widetilde{V}_{\theta t}=d_{\theta t}-\mathbb{E}d_{\theta t}$. To simplify notation, the subscript θ in $\{\widetilde{U}_{\theta t}\}$ is omitted. Define $M_T=\lfloor T^{\frac{1}{2}-\frac{p+1}{2(k-1)}}\log^{-\frac{1}{2}}T\rfloor$ and $b_T=C_b T^{\frac{p+1}{2(k-1)}}(\log T)^{-\frac{p-1}{2(k-1)}}$ where C_b is a positive constant to be chosen in what follows. Let $\widetilde{U}_t=U'_t+U''_t$ where $U'_t=l_t(\theta,\tau)I(l_t(\theta,\tau)\leqslant b_T)-\mathbb{E}\left(l_t(\theta,\tau)I(l_t(\theta,\tau)\leqslant b_T)\right)$ and $U''_t=l_t(\theta,\tau)I(l_t(\theta,\tau)>b_T)-\mathbb{E}\left(l_t(\theta,\tau)I(l_t(\theta,\tau)>b_T)\right)$. Then,

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\widetilde{U}_t\right|>\frac{\varepsilon_T}{4}\right)\leqslant\Pr\left(\left|\sum_{t=1}^{T}U'_t\right|>\frac{T\varepsilon_T}{8}\right)+\Pr\left(\left|\sum_{t=1}^{T}U''_t\right|>\frac{T\varepsilon_T}{8}\right)\ .$$

The sequence $\{U'_t\}$ has the same mixing properties as $\{\widetilde{U}_t\}$ and $\|U'_t\|_{L_\infty}<b_T$ since $l_t(\theta,\tau)\geqslant 0$. Then for all T sufficiently large and $p<k-2$ the conditions of Theorem 2.1 in Liebscher (1996) are satisfied since $M_T\in\{1,\ldots,T\}$ and $T\varepsilon_T/8>4M_T b_T$. By application of that theorem and noting that $\{l_t(\theta,\tau)\}$ is stationary and non-negative,

$$\Pr\left(\left|\sum_{t=1}^{T}U'_t\right|>\frac{T\varepsilon_T}{8}\right)\leqslant 4\exp\left(-\frac{T\varepsilon_T^2}{\frac{4096}{M_T}\mathbb{E}\left(\sum_{t=1}^{M_T}U'_t\right)^2+\frac{64}{3}M_T b_T\varepsilon_T}\right)$$
$$+4\frac{T}{M_T}\exp\left(-C_\alpha M_T^{r_\alpha}\right)\ .$$

Let $\gamma(m) = |\mathrm{Cov}(U'_t, U'_{t+m})|$ for $m = 0, \ldots, T-1$. Then, $\mathbb{E}\left(\sum_{t=1}^{M_T} U'_t\right)^2 \leqslant M_T(\gamma(0) + 2\sum_{m=1}^{\infty}\gamma(m))$. Noting that $l_t(\theta,\tau) \geqslant 0$ and $k \geqslant 2$, Davydov's inequality (Davidson, 1994, Corollary 14.3) implies

$$\begin{aligned}\gamma(m) &\leqslant 2(2^{1-1/k}+1)\alpha(m)^{1-2/k}\|U'_t\|_{L_k}\|U'_{t+m}\|_{L_k} \\ &= 2(2^{1-1/k}+1)\alpha(m)^{1-2/k}\|U'_t\|^2_{L_k} \quad \text{(by stationarity)}\end{aligned}$$

for $m = 0, \ldots, T-1$. Also note that for any $k > 1$ we have

$$\|U'_t\|_{L_k} \leqslant 2\|l_t(\theta,\tau)\|_{L_k} \leqslant 2C_L$$

by Jensen's inequality, and the last inequality holds by Condition 1(*iii*). Thus,

$$\mathbb{E}\left(\sum_{t=1}^{M_T} U'_t\right)^2 \leqslant M_T 8(2^{1-1/k}+1)C_L^2\left(\frac{1}{4^{1-2/k}} + 2\sum_{m=1}^{\infty}\exp\left(-C_\alpha m^{r_\alpha}\right)^{1-\frac{2}{k}}\right) := M_T\sigma^2.$$

Then, for all T sufficiently large, since $p, k > 1$ are such that $p < k-2$, it holds that

$$\left(1 + \frac{24C_\Theta C_d}{N\varepsilon_T}\right)^p \Pr\left(\left|\sum_{t=1}^{T} U'_t\right| > \frac{T\varepsilon_T}{8}\right) = o(\log^{-1} T)\ .$$

Furthermore,

$$\begin{aligned}&\left(1 + \frac{24C_\Theta C_d}{N\varepsilon_T}\right)^p \Pr\left(\left|\sum_{t=1}^{T} U''_t\right| > \frac{T\varepsilon_T}{8}\right) \overset{(a)}{\leqslant} \left(1 + \frac{24C_\Theta C_d}{N\varepsilon_T}\right)^p \frac{8}{T\varepsilon_T}\mathbb{E}\left|\sum_{t=1}^{T} U''_t\right| \\ &\leqslant \left(1 + \frac{24C_\Theta C_d}{N\varepsilon_T}\right)^p \frac{16}{\varepsilon_T}\mathbb{E}\left[l_t(\theta,\tau)I(l_t(\theta,\tau) > b_T)\right] \overset{(b)}{\leqslant} \left(1 + \frac{24C_\Theta C_d}{N\varepsilon_T}\right)^p \frac{16}{\varepsilon_T}\frac{C_L^k}{b_T^{k-1}} \\ &\overset{(c)}{\leqslant} \log^{-1} T\ ,\end{aligned}$$

where (a) follows from Markov's inequality (b) because $\mathbb{E}(|X|I(|X| > b)) \leqslant \mathbb{E}(|X|^r)/b^{r-1}$ for any random variable X with finite r-th moment and positive constant b and Condition 1(*iii*), and (c) from a sufficiently large choice of the constant C_b, for sufficiently large T and noting that N, p and k are fixed. The sequence $\{\tilde{V}_{\theta t}\}$ can be analysed using the same strategy (using the exact same choice of M_T and b_T used for $\tilde{U}_t$). □

Proof of Proposition 4. By Proposition 2, we have that

$$\begin{aligned}&\Pr\left(\sup_\Theta\left|R(\theta,\tau) - \overline{R}(\theta,\tau)\right| > \frac{\varepsilon}{2}\right) \\ &\quad\leqslant \left(1 + \frac{24C_\Theta C_d}{\varepsilon}\right)^p \sup_\Theta\left\{P_{T+1}^{T+M}\left(\mathbb{E}_T l_t(\theta,\tau), \frac{\varepsilon}{4}\right) + P_{T+1}^{T+M}\left(\mathbb{E}_T d_{\theta t}, C_d\right)\right\}\ .\end{aligned}$$

By Markov's inequality,

$$\sup_{\Theta} \Pr\left(\left|\frac{1}{M}\sum_{t=T+1}^{T+M} \mathbb{E}_T l_t(\theta,\tau) - \mathbb{E} l_t(\theta,\tau)\right| > \varepsilon\right) \leqslant \frac{\sup_{\Theta} \mathbb{E}\left|\frac{1}{M}\sum_{t=T+1}^{T+M} \mathbb{E}_T l_t(\theta,\tau) - \mathbb{E} l_t(\theta,\tau)\right|^p}{\varepsilon^p} .$$

By Ibragimov's inequality (Davidson, 1994, Theorem 14.2), we have that for $k > p \geqslant 1$,

$$\sup_{\Theta} \left\|\mathbb{E}_T l_t(\theta,\tau) - \mathbb{E} l_t(\theta,\tau)\right\|_{L_p} \leqslant 2(2^{1/p}+1)\alpha(m)^{1/p-1/k} \sup_{\Theta} \left\|l_t(\theta,\tau)\right\|_{L_k} , \quad m = t - T ,$$

where $\sup_{\Theta} \|l_t(\theta,\tau)\|_{L_k} < C_L < \infty$ by Condition 1. Consequently, and because of the exponential decay of the α-mixing coefficients,

$$\sup_{\Theta} \mathbb{E}\left|\frac{1}{M}\sum_{t=T+1}^{T+M} \mathbb{E}_T l_t(\theta,\tau) - \mathbb{E} l_t(\theta,\tau)\right|^p \leqslant \frac{\sigma^{2p}}{\gamma^p T^p} ,$$

where we have used that $M = \lceil \gamma T \rceil$. Let $\varepsilon_T = \sigma\sqrt{\frac{p \log T}{NT}}$. It follows that

$$\left(1 + \frac{24 C_\Theta C_d}{N\varepsilon_T}\right)^p \sup_{\Theta} P_{T+1}^{T+M}\left(\mathbb{E}_T l_t(\theta,\tau), \frac{\varepsilon_T}{4}\right) \leqslant \frac{C}{\gamma^p N^p T^p \varepsilon_T^{2p}} = O\left(\log^{-p} T\right) .$$

for some $C < \infty$. By Condition 1, $d_{\theta\, t}$ is also α-mixing with exponentially decaying coefficients and $\sup_{\Theta} \|d_{\theta\, t}\|_{L_k} < \infty$. The same arguments as above lead to the bound

$$\left(1 + \frac{24 C_\Theta C_d}{N\varepsilon_T}\right)^p \sup_{\Theta} P_{T+1}^{T+M}\left(\mathbb{E}_T d_{\theta\, t}, C_d\right) \leqslant \frac{C}{\gamma^p N^p T^p \varepsilon_T^{p}} = o(\log^{-p} T)$$

for all T sufficiently large. □

B Verification of Condition 1

This section starts by recalling a number of notions from Markov chain theory. Notation and definitions are based on Meyn and Tweedie (1993). The discrete-time process $\{X_t\}$ is a time-homogeneous Markov chain with state space $\mathcal{X} \subseteq \mathbb{R}^{p_x}$ and equipped with a Borel σ-algebra $\mathcal{B}(\mathcal{X})$ if for each $n \in \mathbb{N}$ there exists an n-step transition probability kernel $P_X^n : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ such that $P_X^n(x,\mathcal{A}) = \Pr(X_{t+n} \in \mathcal{A} | X_t = x)$ for all $t \in \mathbb{Z}_+$. As customary, $P_X^1(x,\mathcal{A})$ is denoted by $P_X(x,\mathcal{A})$. Let $\pi_X : \mathcal{B}(\mathcal{X}) \to [0,1]$ denote the invariant measure of the Markov chain (assuming it exists), that is, the probability measure such that for each $\mathcal{A} \in \mathcal{B}(\mathcal{X})$ it holds that $\pi_X(\mathcal{A}) = \int_{\mathcal{X}} \pi_X(dx) P_X(x,\mathcal{A})$.

B.1 Companion Markov chain

Let $X_t = (X_{1\,t}', X_{2\,t}', X_{3\,t})'$ be defined as

$$\begin{bmatrix} X_{1\,t} \\ X_{2\,t} \\ X_{3\,t} \end{bmatrix} = \begin{bmatrix} g_{h1}(X_{1\,t-1}) + g_{h2}(X_{1\,t-1})Z_{1\,t} \\ \omega + A\tilde{s}_\lambda(X_{1\,t-1}, Z_{2\,t}) + BX_{2\,t-1} \\ 1 + C_s\left(1 + \overline{A}\right)\|\tilde{Y}(X_{1\,t-1}, Z_{2\,t})\|_1 + \|X_{2\,t-1}\|_1 + \overline{B}X_{3\,t-1} + Z_{3\,t} \end{bmatrix}, \quad (11)$$

where

$$\begin{aligned} \tilde{s}_\lambda(X_{1\,t-1}, Z_{2\,t}) &= s_\lambda(\tilde{Y}(X_{1\,t-1}, Z_{2\,t})) \\ \tilde{Y}(X_{1\,t-1}, Z_{2\,t}) &= g_{y1}(X_{1\,t-1}) + g_{y2}(X_{1\,t-1})Z_{2\,t}, \end{aligned}$$

and $Z_{1\,t} = \epsilon_{H\,t}$, $Z_{2\,t} = \epsilon_{Y\,t-1}$, and $Z_{3\,t} = \epsilon_{d\,t}$. The state space of the companion Markov chain is $\mathcal{X} := \mathbb{R}^{p_h} \times \mathbb{R}^N \times [1, \infty) \subset \mathbb{R}^{p_x}$, where $p_x = p_h + N + 1$.

B.2 V-geometric ergodicity

The concept of V-geometric ergodicity used in this paper is the same as in Meitz and Saikkonen (2008a). Note that this is stronger than $\mathcal{Q}$-geometric ergodicity (Liebscher, 2005).

Definition 1 (V_X-geometric ergodicity)**.** *A Markov chain $\{X_t\}$ is V_X-geometrically ergodic if there exists a real valued function $V_X : \mathcal{X} \to [1, \infty)$, a probability measure π_X on $\mathcal{B}(\mathcal{X})$, and constants $\rho < 1$ and $M_x < \infty$ (depending on x) such that*

$$\sup_{v:|v| \leqslant V_X} \left| \int_{\mathcal{X}} P_X^n(x, dx_n)v(x_n) - \int_{\mathcal{X}} \pi_X(dx_n)v(x_n) \right| \leqslant \rho^n M_x \ , \quad (12)$$

for all $x \in \mathcal{X}$ and all $n \geqslant 1$.

Verification of Condition 1 begins by establishing the V-geometric ergodicity of the companion Markov chain $\{X_t\}$. The proof follows by Lemmas B.1 and B.2 (Meyn and Tweedie, 1993).

Lemma B.1 (Irreducibility and Aperiodicity of X_t.)**.** *Let X_t be the Markov chain defined in* (11)*. Then, X_t is irreducible and aperiodic.*

Proof. Start by noting that X_t in (11) can be cast as a nonlinear state space model NSS(F) (Meyn and Tweedie, 1993), i.e. $X_t = F(X_{t-1}, (Z_{1\,t}', Z_{2\,t}', Z_{3\,t})')$ with F defined in an obvious way.[10] For the chain to be irreducible we first need that the *controllability matrix* has full rank. More specifically, the *rank condition* states that for each initial value $x \in \mathcal{X} \subseteq \mathbb{R}^{p_x}$, there exists some $n \in \mathbb{Z}_+$ and a sequence $Z^* = (Z_1^*, \ldots, Z_n^*) \in$

[10] Note that in our derivation it is only required that F be differentiable with respect to Z and not the states or the parameters.

$\times_{i=1}^{n}(\mathbb{R}^{p_h} \times \mathbb{R}^N \times \mathbb{R}_+)$ such that $\text{rank}C_x^n(Z^*) = p_x$ (Meyn and Tweedie, 1993, Eq. 7.13). The controllability matrix for $n = 1$ is defined as the derivative of the transition function with respect to the vector of innovations, i.e.

$$C_x^1(Z^*) = \frac{\partial F}{\partial Z'} = \begin{bmatrix} g_{h2}(x_1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A\frac{\partial \tilde{s}_\lambda(x_1, Z_2)}{\partial Z_2} & \mathbf{0} \\ \mathbf{0} & \bullet & 1 \end{bmatrix}.$$

By Assumptions A.1(i) and A.2(ii)-(iii), we have that for every $x \in \mathcal{X}$ we can find a $Z^* \in \mathbb{R}^{p_h} \times \mathbb{R}^N \times \mathbb{R}_+$ such that

$$\det(C_x^1(Z^*)) = \det(g_{h2}(x_1)) \det(A) \det\left(\frac{\partial \tilde{s}_\lambda(x_1, Z_2)}{\partial Z_2}\right) \neq 0 \ .$$

The claim follows after finding a globally attracting state (Meyn and Tweedie, 1993; Meitz and Saikkonen, 2008b). To do this, the first step is to find a fixed point of the map. It is enough to do this for a choice of Z. Let $Z_1^* = g_{h2}(x_1^*)^{-1}[x_1^* - g_{h1}(x_1^*)]$, for an arbitrary $x_1^* \in \mathbb{R}^{p_h}$. Note that Z_1 exists by Assumption A.1(i). Choose $Z = Z^* = (Z_1^{*\prime}, 0', 0)$. Then, x_1^* is a fixed point for the first component of the map (F_1).

$$x_2^* = (\mathbf{I} - B)^{-1}\left[\omega + As_\lambda(g_{y1}(x_1^*))\right]$$

is a fixed point for the second component of the map (F_2), and by Assumption A.2(i) it is clear that $x_2^* \in \mathbb{R}^N$. Finally, given x_1^* and x_2^*, we have that

$$x_3^* = \frac{1 + C_s(1 + \overline{A})\|\tilde{Y}(x_1^*, 0)\|_1 + \|x_2^*\|_1}{1 - \overline{B}}$$

is a fixed point for the third component of the map (F_3), where $x_3^* \in [1, \infty)$. It follows that $x^* = (x_1^{*\prime}, x_2^{*\prime}, x_3^*)'$ is a fixed point of the map F. Next, one needs to show that the fixed point is attainable for a choice of shock sequence. But this is also accomplished by setting the shocks to zero and noting that $X_{1\,t} \to x_1^*$ as $t \to \infty$, and the same conclusion holds for $X_{2\,t}$ and $X_{3\,t}$. It follows that the companion Markov chain is both irreducible and aperiodic. □

Lemma B.2 (Drift Criterion for X_t)**.** *Let X_t be the Markov chain defined in* (11)*. Then,*

$$\mathbb{E}(V_X(X_t)|X_{t-1} = x) \leqslant (1 - \gamma_1)V_X(x) + \gamma_2 I(x \in \mathcal{S}),$$

where $V_X(x) = 1 + \|x\|_1^k$, $\gamma_1 > 0$, $\gamma_2 < \infty$ and $\mathcal{S}$ is a compact set.

Proof. First, since X_t is a T-chain, it follows that every compact set is small (Meyn and Tweedie, 1993). Let $q_X(x) = 1 + (\kappa'\dot{x})^k$ where $\kappa = (\kappa_1, \kappa_2, \kappa_3)' \in \times_{i=1}^{3}(0, 1)$ and $\dot{x} = (\|x_1\|_1, \|x_2\|_1, |x_3|)'$. Note that $V_X(x) \leqslant q_X(x)/\underline{\kappa}^k$, where $\underline{\kappa}$ denotes the minimum of the components of κ. Thus, it suffices to show that the drift criterion holds with $q_X(x)$

with the compact set $S_{2\epsilon}$ defined below (Lanne and Saikkonen, 2005, Appendix A). By Assumption A.1(i), for every $\epsilon > 0$ there exists $M'_\epsilon < \infty$ such that

$$\|g_{h1}(x_1) + g_{h2}(x_1)Z_{1t}\|_1 \leqslant (a_h + b_h^\epsilon \|Z_{1t}\|_1 + \epsilon)\,\|x_1\|_1$$

holds for all $\|x_1\|_1 > M'_\epsilon$, where $b_h^\epsilon = b_h + \epsilon$. In particular, $\epsilon > 0$ is chosen small enough such that $\mathbb{E}(a_h + b_h^\epsilon \|Z_{1t}\|_1 + \epsilon)^k < 1$ and $\overline{B} + \epsilon < 1$. Such a choice is possible by Assumptions A.1(iv) and A.2(i), respectively.

Now, let $S_{2\epsilon} = \{x \in \mathcal{X} : \kappa'\dot{x} \leqslant M_\epsilon\}$, which is compact, and $S_{1\epsilon} = \mathcal{X}\backslash S_{2\epsilon}$.[11] The proof proceeds by analyzing the cases $\|x_1\|_1 > M'_\epsilon$ and $\|x_1\|_1 \leqslant M'_\epsilon$ separately.[12]

Case $\|x_1\|_1 > M'_\epsilon$. By Assumptions A.2(iv) and A.1(ii),

$$\|\omega + A\tilde{s}(x_1, Z_{2t}) + Bx_2\|_1 \leqslant \|\omega\|_1 + |||A|||_1 C_s C_y (1 + \|Z_{2t}\|_1)\|x_1\|_1 + |||B|||_1 \|x_2\|_1 \ .$$

Note that M'_ϵ may be enlarged if necessary so that

$$\begin{aligned} &\|\omega\|_1 + |||A|||_1 C_s C_y (1 + \|Z_{2t}\|_1)\|x_1\|_1 + |||B|||_1 \|x_2\|_1 \\ &\quad \leqslant \overline{A} C_s C_y (1 + \epsilon + \|Z_{2t}\|_1)\|x_1\|_1 + \overline{B}\|x_2\|_1 \ , \end{aligned}$$

where $\overline{A} < \infty$ is a uniform upper bound for $|||A|||_1$ over Θ by Assumption A.2(v). Also note that $|||B|||_1 \leqslant \overline{B} < 1$ by Assumption A.2(i). Similarly,

$$\begin{aligned} &1 + C_s\left(1 + \overline{A}\right)\|g_{y1}(x_1) + g_{y2}(x_1)Z_{2t}\|_1 + \|x_2\|_1 + \overline{B}|x_3| + Z_{3t} \\ &\quad \leqslant 2 + C_s\left(1 + \overline{A}\right)\|g_{y1}(x_1) + g_{y2}(x_1)Z_{2t}\|_1 + \|x_2\|_1 + \overline{B}|x_3| \\ &\quad \leqslant C_s\left(1 + \overline{A}\right) C_y (1 + \epsilon + \|Z_{2t}\|_1)\|x_1\|_1 + \|x_2\|_1 + \overline{B}|x_3| \ , \end{aligned}$$

where the first inequality uses Assumption A.3(i). Let $\rho_{Z\epsilon} = a_h + b_h^\epsilon \|Z_{1t}\|_1 + \epsilon$, and $C_{yZ}^\epsilon = C_y(1 + \epsilon + \|Z_{2t}\|_1)$. It follows that[13]

$$1 + (\kappa'\dot{X}_t)^k \leqslant (\kappa'\mathbf{C}_\epsilon(Z_t)\dot{X}_{t-1})^k \ ,$$

where the 3×3 matrix $\mathbf{C}_\epsilon(Z_t)$ is defined as

$$\mathbf{C}_\epsilon(Z_t) = \begin{bmatrix} \rho_{Z\epsilon} & 0 & 0 \\ \overline{A} C_s C_{yZ}^\epsilon & \overline{B} + \epsilon & 0 \\ C_s\left(1 + \overline{A}\right) C_{yZ}^\epsilon & 1 & \overline{B} + \epsilon \end{bmatrix}.$$

[11] Note that M_ϵ is larger than M'_ϵ. In particular, $M_\epsilon = \|\overline{C}_{z\epsilon}\|_{L_k}/\epsilon + M'_\epsilon$ with $\overline{C}_{z\epsilon}$ defined below.
[12] Note that the conclusions in both cases hold for any choice of $\kappa \in \times_{i=1}^3 (0, 1)$.
[13] Note that M'_ϵ can be enlarged if necessary to absorb the constant 1.

Note that for the chosen ϵ, the spectral radius of $\mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes k}\right)$ is strictly less than one. By properties of Kronecker products, it holds that

$$\mathbb{E}(q_X(X_t)|X_{t-1}=x) \leqslant (\kappa^{\otimes k})'\mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes k}\right)\dot{x}^{\otimes k} \ . \tag{13}$$

Case $\|x_1\|_1 \leqslant M'_\epsilon$. Note that by Assumption A.1(i),

$$\|g_{h1}(x_1) + g_{h2}(x_1)Z_{1\,t}\|_1 \leqslant \underbrace{\overline{g}_h^\epsilon\,(1+\|Z_{1\,t}\|_1)}_{C_1} \ ,$$

where $\overline{g}_h^\epsilon := \sup_{M'_\epsilon}\|g_{h1}(x_1)\|_1 \vee \sup_{M'_\epsilon}\|\!|g_{h2}(x_1)|\!\|_1$ and $\sup_{M'_\epsilon}$ is the supremum over the set $\{x_1 \in \mathbb{R}^{p_h} : \|x_1\|_1 \leqslant M'_\epsilon\}$. Moreover,

$$\|\omega + A\tilde{s}(x_1, Z_{2\,t}) + Bx_2\|_1 \leqslant \underbrace{\|\omega\|_1 + \overline{A}\ \overline{g}_y^\epsilon\,(1+\|Z_{2\,t}\|_1)}_{C_2} + \overline{B}\|x_2\|_1$$

and

$$\begin{aligned} &1 + C_s\left(1+\overline{A}\right)\|g_{y1}(x_1) + g_{y2}(x_1)Z_{2\,t}\|_1 + \|x_2\|_1 + \overline{B}|x_3| + Z_{3\,t} \\ &\quad\leqslant \underbrace{2 + C_s\left(1+\overline{A}\right)\overline{g}_y^\epsilon\,(1+\|Z_{2\,t}\|_1) + \|x_2\|_1}_{C_3} + \overline{B}|x_3| \end{aligned}$$

where $\overline{g}_y^\epsilon = \sup_{M'_\epsilon}\|g_{y1}(x_1)\|_1 \vee \sup_{M'_\epsilon}\|\!|g_{y2}(x_1)|\!\|_1$. From the previous inequalities one obtains

$$\begin{aligned} \mathbb{E}\left(q_X(X_t)|X_{t-1}=x\right) &\leqslant \mathbb{E}\left(\overline{C}_{z\,\epsilon} + \kappa_2\overline{B}\|x_2\|_1 + \kappa_3\|x_2\|_1 + \kappa_3\overline{B}\|x_3\|_1\right)^k \\ &\leqslant \left(\|\overline{C}_{z\,\epsilon}\|_{L_k} + \kappa'_{-1}\mathbf{B}\dot{x}_{-1}\right)^k \ , \end{aligned}$$

where $\overline{C}_{z\,\epsilon} = C_1 + C_2 + C_3 + C_4$, where $C_4 < \infty$ is a constant that absorbs the 1 in q_X and $\mathbf{B}$ is a 2×2 lower triangular matrix with diagonal entries $\mathbf{B}_{11} = \mathbf{B}_{22} = \overline{B}$ and off-diagonal entry $\mathbf{B}_{21} = 1$. The first inequality uses the fact that $\kappa \in \times_{i=1}^3(0,1)$, and the second uses Minkowski's inequality.

Note that $\kappa_1\|x_1\|_1 + \kappa'_{-1}\dot{x}_{-1} > M_\epsilon$ is true whenever $x \in S_{1\,\epsilon}$. Choose $M_\epsilon = \frac{\|\overline{C}_{z\,\epsilon}\|_{L_k}}{\epsilon} + M'_\epsilon$. Since, $\kappa_1\|x_1\|_1 < \|x_1\|_1 \leqslant M'_\epsilon$, it follows that

$$M'_\epsilon + \kappa'_{-1}\dot{x}_{-1} > \kappa_1\|x_1\|_1 + \kappa'_{-1}\dot{x}_{-1} > \frac{\|\overline{C}_{z\,\epsilon}\|_{L_k}}{\epsilon} + M'_\epsilon \ ,$$

so $\epsilon\kappa'_{-1}\dot{x}_{-1} > \|\overline{C}_{z\,\epsilon}\|_{L_k}$. Thus, $\|\overline{C}_{z\,\epsilon}\|_{L_k} + \kappa'_{-1}\mathbf{B}\dot{x}_{-1} < \kappa'_{-1}(\mathbf{B}+\epsilon\mathbf{I})\dot{x}_{-1} := \kappa'_{-1}\mathbf{B}_\epsilon\dot{x}_{-1}$. Notice that $\mathbf{B}_\epsilon$ is the 2×2 lower diagonal block of $\mathbf{C}_\epsilon(Z_t)$, so one can write

$$\kappa'_{-1}\mathbf{B}_\epsilon\dot{x}_{-1} \leqslant \kappa'\mathbf{C}_\epsilon(Z_t)\dot{x} \ .$$

Again by properties of the Kronecker product it follows that the bound in (13) also holds in this case. Therefore, in both cases $\|x_1\|_1 > M'_\epsilon$ and $\|x_1\|_1 \leqslant M'_\epsilon$ we obtain the same bound for any $\kappa \in \times_{i=1}^n (0,1)$ whenever $x \in S_{1\,\epsilon}$. Thus, by Lemma A.2. of Ling and McAleer (2003) it follows that we can choose $\kappa \in \times_{i=1}^n (0,1)$ such that $v = (\mathbf{I} - \mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes k}\right))'\kappa^{\otimes k}$ has positive components.[14] One can now conclude that for all $x \in S_{1\,\epsilon}$, it holds that

$$\mathbb{E}\left(q_X(X_t)|X_{t-1} = x\right) \leqslant (1-\gamma_1)(\kappa^{\otimes k})'\dot{x}^{\otimes k} \ ,$$

where $\gamma_1 \in (0,1)$ is the minimum of the components of v.

On the other hand, it follows from Assumptions A.1, A.2 and A.3 that

$$\sup_{\substack{x \in S_{2\,\epsilon} \\ \theta \in \Theta}} \mathbb{E}(q_X(X_t)|X_{t-1} = x) \leqslant \gamma_2 < \infty, \quad x \in S_{2\,\epsilon} \ ,$$

where the expectation exists and it is bounded over Θ for every $x \in S_{2\,\epsilon}$ provided that $\|Z_{1\,t}\|_1$ and $\|Z_{2\,t}\|_1$ have k moments. Since $(1-\gamma_1)q_X(x)$ is positive, the claim holds when $x \in S_{2\,\epsilon}$, which completes the proof. □

Lemmas B.3, B.4, B.5 and Proposition 5 below are slight modifications of Lemmas 2, 3, 4 and Proposition 1 of Brownlees and Llorens-Terrazas (2021). For completeness, full derivations of the proofs are available in the Online Appendix. The following lemma establishes that the constants ρ and M_x in Definition 1 in the case of geometric ergodicity (that is, when $V_X = 1$) can be chosen so that they do not depend on θ.

Lemma B.3. *Suppose Assumptions A.1, A.2 and A.3 are satisfied. Then, there exist positive constants $\rho \in (0,1)$ and $R < \infty$ that do not depend on θ such that $\{X_t\}$ satisfies*

$$\sup_{v:|v| \leqslant 1} \left| \int_{\mathcal{X}} P_X^n(x, dx_n) v(x_n) - \int_{\mathcal{X}} \pi_X(dx_n) v(x_n) \right| \leqslant R\tilde{V}_X(x)\rho^n \ ,$$

for all $x \in \mathcal{X}$ and all $n \geqslant 1$, and $\tilde{V}_X(x) = 1 + \|x\|_1$.

The proof of Lemma B.3 is based on an application of Theorem 12 of Roberts and Rosenthal (2004). The MCMC literature has developed a number of results that allow to establish explicit geometric ergodicity convergence rates (Rosenthal, 1995). The important implication of Lemma B.3 is that the dependence properties of the companion Markov chain $\{X_t\}$ can be characterized independently of θ.

The next step of the analysis consists of using the properties of the companion Markov chain $\{X_t\}$ to establish the properties of the joint process $W_t = \{(Y_t', S_t')'\} =$

[14] Recall that $\mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes k}\right)$ has a spectral radius strictly less than 1. As noted by Lanne and Saikkonen (2005), inspection of the proof of Lemma A.2. in Ling and McAleer (2003) reveals that it means no loss of generality to assume that the components of κ are bounded by unity.

$\{(Y_t', H_t', f_{\theta t}', d_{\theta t})'\}$.[15] The following lemma establishes the connection between the transition kernels of $\{X_t\}$ and $\{W_t\}$.

Lemma B.4. *Consider the Markov chain $\{W_t\}$. Let $\pi_{Y|S}(dy|S)$ denote the (invariant) conditional distribution of Y_t given $S_t = s_t$. Then, its n-step transition kernel is given by*

$$P_W^n(w, dw_n) = \pi_{Y|S}(dy_n|s_n) \int_0^1 \int_{\mathcal{H}} P_X^{n-1}(\tilde{x}, ds_n) P_H(h, dh_1) \Pr(d\epsilon_{d1}), \quad n \geqslant 2,$$

where P_H is the transition kernel of $\{H_t\}$, and
$\tilde{x} = \tilde{x}(w, h_1, \epsilon_{d1}) = (h_1, \omega + As_\lambda(y) + Bf, 1 + C_s\left(1 + \overline{A}\right)\|y\|_1 + \|f\|_1 + \overline{B}d + \epsilon_{d1})'$.

The proof of the lemma builds upon the analysis of GARCH models of Meitz and Saikkonen (2008a). The structure given by equations (1), (5), (7) and (8) admits casting $\{W_t\}$ as a Markov chain.

The following lemma establishes that $\{W_t\}$ inherits the moment and dependence properties of the companion Markov chain $\{X_t\}$.

Lemma B.5. *Suppose Assumptions A.1, A.2 and A.3 are satisfied. Then (i) $\{W_t\}$ is V_W-geometrically ergodic with $V_W(w) = 1 + \|y\|_1^k + \|s\|_1^k$; and (ii) there exist positive constants $\rho \in (0,1)$ and $R < \infty$ that do not depend on θ such that $\{W_t\}$ satisfies*

$$\sup_{v:|v|\leqslant 1} \left| \int_{\mathcal{Y}\times\mathcal{X}} [P_W^n(w, dw_n) - \pi_W(dw_n)] v(w_n) \right| \leqslant R\tilde{V}_X(\check{s})\rho^n,$$

for all $w \in \mathcal{Y} \times \mathcal{X}$ and for all $n \geqslant 2$, and

$$\check{s} = (h, \overline{\omega} + \overline{A}C_s\|y\|_1 + \overline{B}\|f\|_1, 2 + C_s\left(1 + \overline{A}\right)\|y\|_1 + \|f\|_1 + \overline{B}d)'.$$

Finally, the moment and dependence properties of $\{W_t\}$ are established.

Definition 2. *For a stationary process $\{X_t\}$, its α-mixing coefficients are defined by*

$$\alpha(m) = \begin{cases} 1/4 & m = 0 \\ \sup_{\mathcal{A}\in\mathcal{F}_0^s, \mathcal{B}\in\mathcal{F}_{s+m}^\infty} |\Pr(\mathcal{A} \cap \mathcal{B}) - \Pr(\mathcal{A})\Pr(\mathcal{B})| & m \geqslant 1 \end{cases}$$

where $s \in \mathbb{Z}$, and $\mathcal{F}_0^s$ and $\mathcal{F}_{s+m}^\infty$ denote the σ-algebras generated by $\{X_t : 0 \leqslant t \leqslant s\}$ and $\{X_t : s + m \leqslant t \leqslant \infty\}$ respectively.

Proposition 5. *Suppose Assumptions A.1, A.2 and A.3 are satisfied. Then, the process $\{W_t\}$ (i) satisfies $\|\|Y_t\|_1\|_{L_k} < \infty$, $\|\|H_t\|_1\|_{L_k} < \infty$, $\sup_\Theta \|\|f_{\theta t}\|_1\|_{L_k} < \infty$ and $\sup_\Theta \|d_{\theta t}\|_{L_k} < \infty$; and (ii) if $W_0 \sim \pi_W$, it is strictly stationary and α-mixing with*

[15] The subscript θ is omitted from S_t and W_t to simplify the notation, but the dependence on θ is understood.

α-mixing coefficients that satisfy $\alpha(m) \leqslant \exp(-C_\alpha m^{r_\alpha})$ for some $C_\alpha > 0$ and $r_\alpha > 0$ that do not depend on θ.

The verification of Condition 1 concludes with the following result.

Lemma B.6. *Suppose Proposition 5 holds. Then, Condition 1 holds.*

Proof. Condition 1(i) is verified by finding a suitable compact set $\Theta \subset \mathbb{R}^p$ compatible with Assumptions A.2(i) and (ii). For example, let $\Theta = \Theta_\omega \times \Theta_A \times \Theta_B \times \Theta_\lambda$, where

$$\begin{aligned}
\Theta_\omega &= \{\omega \in \mathbb{R}^{p_\omega} : \|\omega\|_1 \leqslant \overline{\omega} < \infty\} , \\
\Theta_A &= \{\mathrm{vec}(A) \in \mathbb{R}^{p_A} : 0 < \underline{A} \leqslant |\det(A)|, \|\!|A|\!\|_1 \leqslant \overline{A}\} , \\
\Theta_B &= \{\mathrm{vec}(B) \in \mathbb{R}^{p_B} : \|\!|B|\!\|_1 \leqslant \overline{B}\} , \\
\Theta_\lambda &= \{\lambda \in \mathbb{R}^{p_\lambda} : \|\lambda\|_1 \leqslant \overline{\lambda} < \infty\} ,
\end{aligned}$$

and $p = p_\omega + p_A + p_B + p_\lambda$. Note that Θ_ω, Θ_A, Θ_B, and Θ_λ are compact and nonempty.[16] Condition 1(ii) holds because $l_t(\theta, \tau) = \frac{1}{N}\sum_{i=1}^N \rho_{\tau_i}(Y_{i\,t} - f_{\theta\,i\,t})$ and $d_{\theta\,t}$ are both measurable functions of W_t, which is strictly stationary and α-mixing with coefficients that satisfy $\alpha(m) \leqslant \exp(-C_\alpha m^{r_\alpha})$ for some C_α and $r_\alpha > 0$ that do not depend on θ by Proposition 5. To verify Condition 1(iii), note that by (3), one can write

$$l_t(\theta, \tau) \leqslant \frac{1}{N}\sum_{i=1}^N |Y_{i\,t}| + \frac{1}{N}\sum_{i=1}^N |f_{\theta\,i\,t}| = \frac{1}{N}\|Y_t\|_1 + \frac{1}{N}\|f_{\theta\,t}\|_1 .$$

Thus,

$$\|l_t(\theta, \tau)\|_{L_k} \leqslant \frac{1}{N}\left\|\|Y_t\|_1 + \|f_{\theta\,t}\|_1\right\|_{L_k} \leqslant \frac{1}{N}\left\|\|Y_t\|_1\right\|_{L_k} + \frac{1}{N}\left\|\|f_{\theta\,t}\|_1\right\|_{L_k} ,$$

but by Proposition 5, $\left\|\|Y_t\|_1\right\|_{L_k} < \infty$, $\sup_\Theta \left\|\|f_{\theta\,t}\|_1\right\|_{L_k} < \infty$ and $\sup_\Theta \|d_{\theta\,t}\|_{L_k} < \infty$, which completes the proof. □

[16] The fact that Θ_A is compact and nonempty is verified in the Online Appendix.

References

Adrian, T., Boyarchenko, N., and Giannone, D. (2019). Vulnerable growth. *American Economic Review*, **109**(4), 1263–89.

Akaike, H. (1973). Information theory and an extension of the likelihood principle. In *Proceedings of the Second International Symposium of Information Theory*.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Andrews, D. W. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica: Journal of the Econometric Society*, pages 1465–1471.

Bao, Y., Lee, T.-H., and Saltoglu, B. (2006). Evaluating predictive performance of value-at-risk models in emerging markets: a reality check. *Journal of Forecasting*, **25**(2), 101–128.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, **16**, 199 – 231.

Brownlees, C. and Guðmundsson, G. S. (2021). Performance of empirical risk minimization for linear regression with dependent data. *arXiv preprint arXiv:2104.12127*.

Brownlees, C. and Llorens-Terrazas, J. (2021). Empirical Risk Minimization for Time Series: Nonparametric Performance Bounds for Prediction. *Available at SSRN 3900432*.

Brownlees, C. and Souza, A. B. (2021). Backtesting global growth-at-risk. *Journal of Monetary Economics*, **118**, 312–330.

Brownlees, C., Engle, R., and Kelly, B. (2011). A practical guide to volatility forecasting through calm and storm. *Journal of Risk*, **14**(2), 3–22.

Candes, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta numerica*, **15**, 257–325.

Catania, L. and Luati, A. (2019). Semiparametric modeling of multiple quantiles. *Available at SSRN 3494995*.

Catania, L., Luati, A., and Vallarino, P. (2021). Economic vulnerability is state dependent. CREATES Research Papers 2021-09, Department of Economics and Business Economics, Aarhus University.

Catania, L., Luati, A., and Mikkelsen, E. B. (2022). Dynamic Multiple Quantile Models. *Available at SSRN 3727513*.

Chavleishvili, S. and Manganelli, S. (2019). Forecasting and stress testing with quantile vector autoregression. Working Paper Series 2330, European Central Bank.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, **78**(3), 1093–1125.

Cox, D. R. (1981). Statistical Analysis of Time Series: Some Recent Developments. *Scandinavian Journal of Statistics*, **8**, 93–115.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. OUP Oxford.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Domowitz, I. and White, H. (1982). Misspecified models with dependent observations. *Journal of Econometrics*, **20**(1), 35–58.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**(3), 425–455.

Doukhan, P. (1994). *Mixing*. Springer-Verlag, New York.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of business & economic statistics*, **22**(4), 367–381.

Giacomini, R. and Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics*, **23**(4), 416–431.

Gouriéroux, C. and Jasiak, J. (2008). Dynamic quantile models. *Journal of econometrics*, **147**(1), 198–205.

Hamilton, J. (1994). *Time series analysis*. Princeton Univ. Press, Princeton, NJ.

Huang, D., Yu, B., Fabozzi, F. J., and Fukushima, M. (2009). Caviar-based forecast for oil price risk. *Energy Economics*, **31**(4), 511–518.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability & Its Applications*, **7**(4), 349–382.

Jiang, W. and Tanner, M. A. (2010). Risk minimization for time series binary choice with variable selection. *Econometric Theory*, **26**(5), 1437–1452.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, **86**(2), 591–616.

Kitagawa, T., Wang, W., and Xu, M. (2022). Policy Choice in Time Series by Empirical Welfare Maximization. *arXiv*.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, **186**(2), 325–344.

Koenker, R. and Xiao, Z. (2006). Quantile autoregression. *Journal of the American statistical association*, **101**(475), 980–990.

Komunjer, I. (2013). Quantile prediction. In *Handbook of economic forecasting*, volume 2, pages 961–994. Elsevier.

Kuester, K., Mittnik, S., and Paolella, M. S. (2006). Value-at-Risk Prediction: A Comparison of Alternative Strategies. *Journal of financial econometrics*, **4**(1), 53–89.

Kuznetsov, V. and Mohri, M. (2015). Learning theory and algorithms for forecasting non-stationary time series. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Kuznetsov, V. and Mohri, M. (2017). Generalization bounds for non-stationary mixing processes. *Machine Learning*, **106**(1), 93–117.

Lanne, M. and Saikkonen, P. (2005). Non-linear garch models for highly persistent volatility. *The Econometrics Journal*, **8**(2), 251–276.

Lecué, G. and Mendelson, S. (2016). Performance of empirical risk minimization in linear aggregation. *Bernoulli*, **22**(3), 1520–1534.

Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stochastic Processes and Their Applications*, **65**(1), 69–80.

Liebscher, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, **26**(5), 669–689.

Ling, S. and McAleer, M. (2003). Asymptotic theory for a vector ARMA-GARCH model. *Econometric theory*, **19**(2), 280–310.

Lu, Z. and Jiang, Z. (2001). L_1 geometric ergodicity of a multivariate nonlinear AR model with an ARCH term. *Statistics & Probability Letters*, **51**, 121–130.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, **72**(4), 1221–1246.

Masry, E. and Tjøstheim, D. (1995). Nonparametric Estimation and Identification of Nonlinear ARCH Time Series: Strong Convergence and Asymptotic Normality. *Econometric Theory*, **11**, 258–289.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2017). Nonparametric risk bounds for time-series forecasting. *The Journal of Machine Learning Research*, **18**(1), 1044–1083.

Meir, R. (2000). Nonparametric time series prediction through adaptive model selection. *Machine learning*, **39**(1), 5–34.

Meitz, M. and Saikkonen, P. (2008a). Ergodicity, mixing, and existence of moments of a class of Markov models with applications to GARCH and ACD models. *Econometric Theory*, **24**(5), 1291–1320.

Meitz, M. and Saikkonen, P. (2008b). Stability of nonlinear AR-GARCH models. *Journal of Time Series Analysis*, **29**(3), 453–475.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Springer Science & Business Media.

Nelson, D. B. (1992). Filtering and forecasting with misspecified ARCH models I: Getting the right variance with the wrong model. *Journal of econometrics*, **52**(1-2), 61–90.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, **4**, 2111–2245.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

Pakel, C., Shephard, N., and Sheppard, K. (2011). Nuisance parameters, composite likelihoods and a panel of GARCH models. *Statistica Sinica*, pages 307–329.

Pötscher, B. M. and Prucha, I. (1997). *Dynamic nonlinear econometric models: Asymptotic theory*. Springer Science & Business Media.

Pötscher, B. M. and Prucha, I. R. (1989). A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica: Journal of the Econometric Society*,

pages 675–683.

Roberts, G. and Rosenthal, J. (2004). General State Space Markov Chains and MCMC algorithms. *Probability Surveys*, **1**, 20–71.

Rosenthal, J. (1995). Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *Journal of the Americal Statistical Association*, **90**, 558–566.

Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **197**, 264–280.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, **10**(5), 988–999.

Vapnik, V. N. and Chervonenkis, A. Y. (1974). *Theory of Pattern Recognition [in Russian]*. Nauka.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25.

White, H. and Domowitz, I. (1984). Nonlinear regression with dependent observations. *Econometrica: Journal of the Econometric Society*, pages 143–161.

White, H., Kim, T.-H., and Manganelli, S. (2015). VAR for VaR: Measuring tail dependence using multivariate regression quantiles. *Journal of Econometrics*, **187**(1), 169–188.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116.