Essays on Learning Theory for Time Series

Author: Jordi Llorens-Terrazas

TESI DOCTORAL UPF / Year of the thesis: 2023

THESIS SUPERVISOR
Prof. Christian Brownlees
Department of Economics and Business

To my parents, Violeta and Denis, and to my wife Emma.

Thanks

I would not be writing this thesis if it were not for my parents, who have gone far and beyond to support me in this adventure. I am forever indebted to my math teacher in high school, Pedro, who taught me how to appreciate mathematics.

Thanks to my advisor (Christian Brownlees) I never felt alone during this intellectual journey. His encouragement and support along the way were absolutely priceless. Throughout my PhD I have benefited from discussions and comments from several faculty members at UPF. The lectures by Gábor Lugosi and Piotr Zwiernik are the principal components explaining the choice of topic in this dissertation. I feel lucky to be in a department with excellent econometricians from which I got extensive feedback on my work. Special thanks to Kirill Evdokimov, Geert Mesters, Katerina Petrova and Barbara Rossi for this. The support of the Data Science Center at the Barcelona School of Economics in furthering my development as a student first, and as a teacher and consultant later has proven invaluable to me. I would like to thank Omiros Papaspiliopoulos and David Rossell for their continued trust.

Life as a PhD student gets much easier in the right company. I thank my colleagues and friends from the Data Science master and the Master in Research for being the best company one could aim for. Also, I cannot be too grateful to all those who have shared unforgettable moments with me in Amsterdam, Budapest, Colombia, Dubai, Ibiza, Malta, Oporto and Stockholm.

The last chapter of this dissertation was written during my research visit at the

University of Helsinki. Thanks to my sponsor, Mika Meitz, for making it possible. Last but not least, I thank my wife, Emma Rinneheimo (and her dog Kamu), for turning it into the warmest place on Earth for me.

Abstract

This thesis comprises three essays on statistical learning theory for time series. In the first chapter, joint with Christian Brownlees, we propose an alternative specification of a dynamic conditional correlation model based on Bregman divergences with an application to portfolio selection. The second chapter, also joint with Christian Brownlees, deals with the problem of empirical risk minimization for time series. The main result states that the performance of the empirical risk minimizer converges at a near optimal rate to the best performance attainable in a class of recursive threshold forecasts induced by the self-exciting threshold autoregressive moving average model. The third chapter derives performance guarantees for forecasting dynamic quantiles in a multivariate setup under full misspecification. The benefits of the methodology are illustrated in an application to Growth-at-Risk forecasting.

Resum

Aquesta tesi comprèn tres assajos sobre teoria de l'aprenentatge estadístic per a sèries temporals. En el primer capítol, juntament amb Christian Brownlees, proposem una especificació alternativa d'un model de correlacions condicionals dinàmiques basat en divergències de Bregman amb una aplicació a la selecció de carteres. El segon capítol, també juntament amb Christian Brownlees, tracta el problema de la minimització del risc empíric per a sèries temporals. El resultat principal estableix que la capacitat predictiva del minimitzador del risc empíric convergeix a una velocitat gairebé òptima al millor rendiment assolible en una classe d'algoritmes predictius recursius amb llindars induïts pel model autoregressiu de mitjana mòbil amb llindar autoexcitador. El tercer capítol deriva garanties per a la predicció de quantils dinàmics en una configuració multivariant sota especificació incorrecta. Els avantatges de la metodologia s'il·lustren en una aplicació a la predicció del *creixement en risc* (*Growth-at-Risk*).

Preface

The advances in computing and data availability over the last decades have paved the way for machine learning (ML) to play an influential role in virtually all areas of research. In Economics, we have seen the adoption of off-the-shelf ML methods as well as some efforts to tune and adapt those to the particularities of our discipline.[1] While machine learning is concerned about *automating* the process of learning from data, the goal of *statistical* learning theory is to *formalize* it. The vast majority of contributions in statistical learning are centered around the assumption that observations are independent of each other, although a large number of datasets in economics and finance have a temporal dimension. While econometricians are gradually incorporating more elements of statistical learning to their analyses, extensions that allow for dependent data are still relatively unexplored. My PhD dissertation is devoted to this exploration.

The first chapter exemplifies how statistical learning theory can help econometricians derive new methodologies. In joint work with Prof. Brownlees, we explored the properties of Bregman divergences and their applications to dynamic covariance modeling. We proposed a novel specification of the Dynamic Conditional Correlation (DCC) model based on an alternative normalization of the pseudo-correlation matrix called Projected DCC (Pro-DCC). Our modification consists in *projecting*, rather than *rescaling*, the pseudo-correlation matrix onto the set of correlation matrices in order to obtain a well defined conditional cor-

[1] See for instance Varian (2014); Mullainathan and Spiess (2017); Chernozhukov *et al.* (2018); Athey *et al.* (2021)

relation matrix. More specifically, we answer the question: given $\mathbf{Q}$ (positive definite but not unit diagonal), what is the closest correlation matrix to $\mathbf{Q}$? When the discrepancy between matrices is measured with a Bregman divergence, this is called a Bregman projection. An empirical application to the constituents of the S&P 100 shows that the proposed methodology performs favorably to the standard DCC in an out-of-sample asset allocation exercise.

The second and third chapters are focused on deriving guarantees for potentially misspecified time series prediction algorithms. By taking a learning theory perspective, the goal is to investigate the conditions under which these algorithms are expected to perform adequately.

The second chapter, which is joint work with Prof. Brownlees, deals with the properties of empirical risk minimization for time series. Empirical risk minimization is a standard principle for choosing algorithms in learning theory. The analysis is carried out in a general framework that covers different types of forecasting applications encountered in the literature. We are concerned with 1-step-ahead prediction of a univariate time series belonging to a class of location-scale parameter-driven processes. A class of recursive algorithms is available to forecast the time series. The algorithms are recursive in the sense that the forecast produced in a given period is a function of the lagged values of the forecast and of the time series. The relationship between the generating mechanism of the time series and the class of algorithms is not specified. Our main result establishes that the algorithm chosen by empirical risk minimization achieves asymptotically the optimal predictive performance that is attainable within the class of algorithms.

In the third chapter, I study the problem of multivariate dynamic quantile forecasting from a learning theory perspective. Despite the fact that forecasting quantiles is of obvious interest to economic agents, the theory in the dynamic quantile modeling literature focuses on *estimation under correct specification* of the quantile dynamics, and less attention is paid to *forecasting under misspecification*. I address this gap by deriving 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.

Contents

Chapter 1

PROJECTED DYNAMIC CONDITIONAL CORRELATIONS

1.1 Introduction

Estimating and forecasting the time-varying covariance matrix of asset returns is key for several applications in finance including asset allocation, risk management and systemic risk measurement. Over the years, the GARCH-DCC methodology of Engle (2002) has established itself as one of the leading paradigms in the literature due to its flexibility and ease of estimation (see also Engle and Sheppard, 2001). In a nutshell, the GARCH-DCC approach consists in modeling separately the conditional variances and the conditional correlation matrix. The conditional variances are modeled using GARCH whereas the conditional correlation matrix is modeled using the Dynamic Conditional Correlation (DCC) model. Recent re-

search in the literature that is based on GARCH-DCC includes Engle *et al.* (2019), Brownlees and Engle (2017), De Nard *et al.* (2021) and Van Os and Van Dijk (2021).

A key aspect of the DCC methodology is that the conditional correlation matrix is modeled as a function of the so called pseudo-correlation matrix. The pseudo-correlation matrix is a symmetric positive definite proxy of the conditional correlation matrix that, crucially, is not guaranteed to be a proper correlation matrix as it does not have a unit diagonal (almost surely). In order to obtain correlations, the pseudo-correlation matrix has to be appropriately normalized, and the standard strategy followed in the literature consists in rescaling this matrix (Engle, 2002; Tse and Tsui, 2002; Aielli, 2013). Engle (2009, Section 4.3) contains a discussion and a comparison of different rescaling approaches used in the literature. Despite the fact that rescaling is natural and commonly employed, it is unclear whether such an approach is in any sense optimal.

In this work we propose a modification of the standard DCC model based on an alternative normalization procedure of the pseudo-correlation matrix. Our modification consists in *projecting* the pseudo-correlation matrix onto the set of correlation matrices rather than *rescaling* it. In other words, we cast the normalization step of the pseudo-correlation matrix as a nearest-correlation matrix problem, that is the problem of finding the closest correlation matrix to a given pseudo-correlation matrix on the basis of an appropriate divergence function.

We begin this work by defining a class of projections for pseudo-correlation matrices. To do so, we first introduce the notion of Bregman divergence for sym-

metric positive definite matrices (Bregman, 1967; Banerjee *et al.*, 2005a; Dhillon and Tropp, 2007; Patton, 2020), which is used in this work to measure nearness between two symmetric positive definite matrices. This family of divergences constitutes a rich collection of divergence functions that includes many familiar losses commonly encountered in the covariance estimation literature such as the Stein and square Frobenius losses (Stein, 1986; Dey and Srinivasan, 1985; Pourahmadi, 2013). In addition, this class of loss functions has been the focus of attention in the financial econometrics literature in the context of ranking multivariate volatility models by their forecasting performance (Laurent *et al.*, 2013; Patton, 2020). In particular, the former paper establishes under mild assumptions that consistent volatility forecast rankings using conditionally unbiased proxies are obtained if and only if the loss function is of the Bregman type.

We define the projection of a pseudo-correlation matrix onto the set of correlation matrices as the correlation matrix that minimizes the Bregman matrix divergence with respect to that pseudo-correlation matrix. It is straightforward to establish that such a projection exists and is unique. Within this broad class of projections we focus in particular on the one implied by Stein's loss, which we name Stein's projection. Stein's loss is a natural loss function for covariance matrices that is related to the multivariate Wishart log-density –or equivalently, the zero mean multivariate Gaussian log-likelihood with respect to the covariance parameter–, it is widely used (Ledoit and Wolf, 2018), and it guarantees to deliver a positive definite projection. Moreover, we derive a closed form expression to compute Stein projections in the two-dimensional case and an efficient iterative

algorithm for the generic n-dimensional case.

We then introduce a novel DCC specification based on our pseudo-correlation matrix projection called Projected DCC (Pro-DCC). Simply put, the Pro-DCC corresponds to the classic DCC of Engle (2002) with the rescaling step of the pseudo-correlation matrix replaced by our proposed projection. In order to estimate the Pro-DCC we propose to follow the same multi-step procedure which is used to estimate other DCC-type models.

A simulation study is carried out to assess the performance of our projection-based methodology. We carry out two main exercises. In the first exercise we simulate i.i.d. data from a multivariate Gaussian distribution with mean zero and covariance parameter given by a correlation matrix. We then estimate the correlation matrix of the simulated data by rescaling the sample covariance matrix (i.e. the sample correlation matrix) and by projecting the sample covariance matrix onto the set of correlation matrices using Stein's projection. We find that the projection-based approach performs better than rescaling in terms of correlation estimation accuracy and that gains are larger in higher dimensional systems. In the second exercise we compare the estimation accuracy of DCC and Pro-DCC under misspecification, that is when the DGP differs from both models. In particular, we consider a dynamic equicorrelation matrix model (Engle and Kelly, 2012) in which the dynamic correlation evolves according to the cosine function, in the spirit of one of the DGPs considered in the simulation exercise of Engle (2002). We find that Pro-DCC outperforms standard DCC and that the gains increase with the dimensionality of the system and degree of cross-sectional dependence.

A Global Minimum Variance Portfolio (GMVP) exercise with the constituents of the S&P 100 is used to measure the performance of Pro-DCC. The design of the exercise is close in spirit to the one of De Nard *et al.* (2021). We compare Pro-DCC to DCC, and we consider both the standard versions of these models as well as versions that rely on nonlinear shrinkage (Ledoit and Wolf, 2020) for covariance targeting. Results show that forecasts based on the standard and nonlinear shrinkage variant of the Pro-DCC achieve the best out-of-sample performance. For completeness, we also consider GMVPs with exposure constraints in order to understand if the advantage of Pro-DCC is due to shrinkage (Jagannathan and Ma, 2003; Fan *et al.*, 2012). We find that adding 1-norm constraints substantially improves performance for both the DCC and Pro-DCC, hence suggesting that Pro-DCC performs favorably even after controlling for shrinkage.

This chapter is related to different strands of the literature. First, it is related to the literature on multivariate volatility models and the DCC. Important contributions in this area, besides the one we have already mentioned, include Bollerslev (1990) and Pakel *et al.* (2018). Classic surveys of the literature on multivariate volatility modeling are Bauwens *et al.* (2006) and Silvennoinen and Teräsvirta (2008). Second, it is related to the financial econometrics literature on large dimensional covariance estimation for asset allocation, which include the contributions of, among others, Hautsch *et al.* (2015), Hautsch and Voigt (2019), De Nard *et al.* (2021). Last, it is related to the literature on matrix projections based on Bregman divergences and the nearest-correlation matrix problem. Contributions in this area include the work of Higham (2002), Dhillon and Tropp (2007) and

Kulis *et al.* (2009).

The rest of the chapter is structured as follows. Section 1.2 introduces the methodology. Section 1.3 contains the simulation study. Section 1.4 presents the empirical application. Section 1.5 concludes the chapter. All proofs are collected in section 1.6.

1.2 Methodology

In this Section we first concisely review the DCC model of Engle (2002) and we then introduce the Pro-DCC model.

1.2.1 The DCC Model

Let $r_t = (r_{1\,t}, \ldots, r_{n\,t})'$ denote an n-dimensional vector of log returns observed at time t, for t ranging from 1 to T. The key object of interest of this work is the conditional covariance matrix of returns given past information, that is $\mathbf{\Sigma}_t = \mathbb{C}\mathrm{ov}_{t-1}(r_t)$. The GARCH-DCC framework is based on the following factorization of the conditional covariance matrix

$$\mathbf{\Sigma}_t = \mathbf{D}_t \, \mathbf{R}_t \, \mathbf{D}_t \ ,$$

where $\mathbf{D}_t$ is a $n \times n$ diagonal matrix of conditional volatilities (standard deviations) and $\mathbf{R}_t$ is the $n \times n$ conditional correlation matrix.

In the GARCH-DCC framework, the conditional volatility matrix $\mathbf{D}_t$ is typically modeled using some appropriate GARCH specification. Assuming, for in-

stance, GJR-GARCH(1,1) dynamics we have that the i-th diagonal element of $\mathbf{D}_t$, which we denote by $d_{i\,t}$, is specified as

$$d_{i\,t}^2 = \omega_i + a_i\, r_{i\,t-1}^2 + \gamma_i\, \mathbb{I}_{t-1}^{-} r_{i\,t-1}^2 + b_i\, d_{i\,t-1}^2 \,,$$

where ω_i, a_i, γ_i and b_i are the GJR-GARCH(1,1) coefficients satisfying $\omega_i > 0$, $a_i > 0$, $b_i \geqslant 0$ and $a_i + \gamma_i/2 + b_i < 1$.

The conditional correlation matrix $\mathbf{R}_t$ is modeled using the DCC specification. The DCC models the correlation process as a function of the so-called devolatilized returns that are defined as $\epsilon_t = \mathbf{D}_t^{-1} r_t$. In the DCC model the conditional correlation matrix is determined by the so-called pseudo-correlation matrix $\mathbf{Q}_t$ which evolves according to the equation

$$\mathbf{Q}_t = (1 - \alpha - \beta)\mathbf{C} + \alpha \epsilon_{t-1} \epsilon_{t-1}' + \beta \mathbf{Q}_{t-1} \,, \tag{1.1}$$

where α and β are scalar parameters that satisfy $\alpha > 0$, $\beta > 0$, $\alpha + \beta < 1$ and $\mathbf{C}$ is an $n \times n$ positive definite matrix. It is straightforward to see by recursive substitution that

$$\mathbf{Q}_t = \frac{1 - \alpha - \beta}{1 - \beta}\mathbf{C} + \alpha \sum_{i=0}^{\infty} \beta^i \epsilon_{t-1-i} \epsilon_{t-1-i}' \,. \tag{1.2}$$

A crucial aspect of the DCC model on which we build upon in the next section is that the pseudo-correlation matrix is not guaranteed to be a correlation matrix. In particular, it is clear from (1.2) that $\mathbf{Q}_t$ is symmetric positive definite but (generally) not unit diagonal. Thus, an appropriate normalization step is required to obtain a correlation matrix. The standard approach consists of *rescaling* the

pseudo-correlation matrix, that is

$$\mathbf{R}_t = \text{diag}(\mathbf{Q}_t)^{-1/2}\, \mathbf{Q}_t\, \text{diag}(\mathbf{Q}_t)^{-1/2}\,, \tag{1.3}$$

where for an $n \times n$ matrix $\mathbf{A}$, the notation $\text{diag}(\mathbf{A})$ denotes the $n \times n$ diagonal matrix with the diagonal of $\mathbf{A}$.

The GARCH-DCC family of models is estimated using a multi-step procedure motivated by a QML argument. The first step consists of estimating the conditional standard deviation matrix $\mathbf{D}_t$ by estimating n univariate GARCH models. Next, the $\mathbf{C}$ matrix is estimated by covariance targeting using the sample second moment of the estimated standardized residuals, that is

$$\widehat{\mathrm{C}} = \frac{1}{T}\sum_{t=1}^{T} \hat{\epsilon}_t \hat{\epsilon}_t'\,,$$

where $\hat{\epsilon}_{it} = r_{it}/\hat{\sigma}_{it}$ and $\hat{\sigma}_{it}$ is the estimated volatility of the first step. Last, the DCC parameters are obtained by maximizing the (Gaussian) quasi log-likelihood of the de-volatilized returns (see Engle, 2002, for details).

We remark that, albeit being intuitive, the estimation strategy put forward by Engle (2002) has some consistency issues first noted by Aielli (2013). These have motivated Aielli (2013) to introduce a "corrected" version of the model called Corrected DCC (CDCC). However, empirically, this model is found to perform similarly to the standard "uncorrected" DCC.

1.2.2 The Projected DCC Model

In this section we propose a novel DCC specification based on an alternative normalization procedure. Rather than *rescaling* the pseudo-correlation matrix as in

equation (1.3) we propose *projecting* it onto the space of correlation matrices. In other words, we cast the problem of normalizing the pseudo-correlation matrix as a nearest-correlation matrix problem, that is finding the closest correlation matrix to a given pseudo-correlation matrix. In order to introduce our projection-based model some additional machinery is required.

We begin by introducing the notion of Bregman divergence for real matrices.

Definition 1.2.1 (Bregman Divergence)**.** *Given a strictly convex and differentiable function ϕ of Legendre type,*[1] *we define the Bregman matrix divergence as*

$$d_\phi(\mathbf{M}_1, \mathbf{M}_2) = \phi(\mathbf{M}_1) - \phi(\mathbf{M}_2) - \operatorname{tr}(\nabla\phi(\mathbf{M}_2)'(\mathbf{M}_1 - \mathbf{M}_2)) \ ,$$

for any two real matrices $\mathbf{M}_1$ and $\mathbf{M}_2$.

Bregman divergences can be seen as the difference between the function ϕ evaluated at $\mathbf{M}_1$ and its first-order Taylor approximation around $\mathbf{M}_2$. Bregman divergences are a class of tractable divergences that enjoy a number of useful properties and are popular in the Machine Learning literature (Cesa-Bianchi and Lugosi, 2006). Bregman divergences are always positive, like distances, and are zero only when their arguments coincide. Unlike distances, they are not necessarily symmetric and they do not necessarily satisfy the triangle inequality. Furthermore, they are always convex with respect to their first argument and satisfy a generalized Pythagorean property (Dhillon and Tropp, 2007; Kulis *et al.*, 2009). Finally, Banerjee *et al.* (2005a) establishes the existence of a bijection between

[1] A function is of Legendre type if it is essentially smooth and essentially strictly convex (Bauschke and Borwein, 1997).

Bregman divergences and regular exponential families. In this chapter the discrepancy between a correlation matrix $\mathbf{R}$ and a pseudo-correlation matrix $\mathbf{Q}$ is measured by the divergence $d_\phi(\mathbf{R}, \mathbf{Q})$.

Depending on the choice of the function ϕ, we obtain a number of well-known loss functions for covariance matrices. If we set $\phi(\mathbf{M}) = -\ln\det(\mathbf{M})$ then we have Stein's loss,

$$d_\phi(\mathbf{M}_1, \mathbf{M}_2) = \operatorname{tr}(\mathbf{M}_1\mathbf{M}_2^{-1}) - \ln\det(\mathbf{M}_1\mathbf{M}_2^{-1}) - n\ , \tag{1.4}$$

where $\ln(\cdot)$ denotes the natural logarithm. This divergence can also be interpreted as the negative of the n-dimensional Wishart log-density (up to a constant) or, equivalently, the zero mean multivariate Gaussian log-likelihood with respect to the covariance parameter. If we set $\phi(\mathbf{M}) = \operatorname{tr}(\mathbf{M}\log\mathbf{M} - \mathbf{M})$ then we have the Von Neumann loss

$$d_\phi(\mathbf{M}_1, \mathbf{M}_2) = \operatorname{tr}(\mathbf{M}_1\log\mathbf{M}_1 - \mathbf{M}_1\log\mathbf{M}_2 - \mathbf{M}_1 + \mathbf{M}_2)\ , \tag{1.5}$$

where $\log(\cdot)$ denotes the matrix logarithm.[2] Finally, if we set $\phi(\mathbf{M}) = \|\mathbf{M}\|_F^2$ then we have the squared Frobenius loss.

Let $\mathbb{S}_+^n$ ($\mathbb{S}_{++}^n$) be the set of n-dimensional symmetric positive semidefinite (positive definite) matrices. We use Bregman divergences to introduce the following general class of projections of symmetric positive definite matrices onto

[2]For symmetric positive definite matrices, the matrix logarithm is $\log\mathbf{Q} = \mathbf{U}\log\mathbf{\Lambda}U'$, where $\boldsymbol{U}\mathbf{\Lambda}\boldsymbol{U}'$ is the eigendecomposition of $\mathbf{Q}$ and $\log\mathbf{\Lambda}$ involves taking the natural logarithm of the eigenvalues.

the set of correlation matrices –which is understood as the set of n-dimensional symmetric positive definite matrices with unit diagonal.

Lemma 1.2.1. *Let* $\mathbf{Q} \in \mathbb{S}^n_{++}$ *and let* $\mathbb{C}^n$ *denote the set of correlation matrices. Furthermore, assume that* ϕ *is a closed convex proper function of Legendre type and differentiable on* $\mathsf{int}(\mathsf{dom}(\phi)) = \mathbb{S}^n_{++}$. *Define the Bregman projection*

$$P_\phi(\mathbf{Q}) = \arg\min_{\mathbf{R}\in\mathbb{C}^n} d_\phi(\mathbf{R}, \mathbf{Q}). \tag{1.6}$$

Then we have that there exists a unique $P_\phi(\mathbf{Q}) \in \mathbb{S}^n_{++}$.

A few remarks are in order. First, we emphasize that the projection depends on the choice of the function ϕ. A natural choice that also turns out to be computationally convenient is to define a projection based on Stein's loss defined in (1.4) and the Von Neumann loss defined in (1.5). We call these projections, respectively, Stein's projection and Von Neumann projection for short.

Second, we point out that existence and uniqueness of the Bregman projection hold because (i) $\mathbb{S}^n_{++} \cap \mathbb{C}^n = \mathbb{C}^n \neq \varnothing$, (ii) ϕ is Legendre and (iii) $\mathsf{int}(\mathsf{dom}(\phi)) = \mathbb{S}^n_{++}$. The Stein and Von Neumann losses satisfy requirements (ii) and (iii), hence it follows from Bauschke and Borwein (1997) that the projection exists in $\mathbb{S}^n_{++}$ and is unique. However, the Frobenius loss does not satisfy property (iii) over the set of positive semidefinite matrices. We remark that a Frobenius projection can be uniquely defined, but it would not necessarily preserve positive definiteness.

Third, in the case of the Stein's loss we have that the projection is related to constrained maximum likelihood estimation (MLE) of the covariance of the zero-mean multivariate Gaussian with unit diagonal. In particular, the constrained

MLE would be given by $\arg\min_{\mathbf{R}\in\mathbb{C}^n} d_\phi(\mathbf{Q},\mathbf{R})$ where $\mathbf{Q}$ is the sample covariance matrix. Note that in general this differs from Pearson's sample correlation. In fact, in Example 18.3 of Kendall and Stuart (1979) it is shown that the constrained MLE for a bivariate Gaussian distribution is obtained by solving a cubic equation – which in large samples has only one real solution. In higher dimensions, finding such MLE is computationally burdensome. On the contrary, projecting the sample covariance under Stein's loss involves solving a convex problem (as opposed to the MLE) and the solution is easily obtained with a much lower computational burden.

Finally, we introduce the Pro-DCC(1,1), that is

$$\begin{aligned}\mathbf{Q}_t &= (1-\alpha-\beta)\mathbf{C} + \alpha\epsilon_{t-1}\epsilon_{t-1}' + \beta\mathbf{Q}_{t-1}\,,\\ \mathbf{R}_t &= P_\phi(\mathbf{Q}_t)\,.\end{aligned}$$

In other words, the Pro-DCC replaces the rescaling equation of the DCC (1.3) with a projection. We point out that the Pro-DCC depends on a choice of an appropriate divergence function ϕ.

A number of comments are in order. First, we remark that the projection yields the closest correlation matrix with respect to the loss induced by ϕ, which does not automatically imply any optimality properties for the purposes of forecasting or minimizing other relevant classes of losses such as the GMVP. Nevertheless, Sections 1.3 and 1.4 –which provide both simulation and empirical evidence that the projection performs favorably to standard rescaling– shed some light on this point. Second, as discussed at the end of Section 2.1, the original formulation

of the DCC model was criticized by Aielli (2013) and a corrected version of the model was introduced. We remark that we do not use the Aielli correction in the Pro-DCC methodology.

For large dimensional models, we propose estimating the model by composite likelihood as in Pakel *et al.* (2018). However, we note that for the Stein and Von Neumann cases, projecting the pseudo-correlation of any 2 assets i and j is not equal to the (i, j) entry of the projection of the entire matrix $P_\phi(\mathbf{Q}_t)$. This is because for these loss function the projection takes into account the full correlation structure and not just the correlation between assets i and j. In practice, the resulting composite likelihood estimates of the dynamic parameters do not vary substantially from their full likelihood counterparts, so this is a minor concern.

Last, we remark that, as it is widely known, the sample correlation matrix performs poorly when the concentration ratio n/T is large – see Lecture 4 in Stein (1986). For that reason, we consider using a nonlinear shrinkage estimator to rectify the in-sample bias of the sample correlation as in Ledoit and Wolf (2020).

Computing the Bregman Projection

In order to apply the Pro-DCC in practice it is key to be able to compute the projections in a computationally cheap way. We derive a closed-form expression for the projection in the 2 dimensional case for the Stein and von Neumann losses and we provide an efficient algorithm for the computation of the projection in the general n dimensional case for the Stein projection.

The following two lemmas derive the closed form of the projection.

Lemma 1.2.2. *Let* $\mathbf{Q}$ *be a* 2×2 *symmetric positive definite matrix. Consider the Bregman Projection of* $\mathbf{Q}$ *onto the set of correlation matrices under Stein's Loss. The unique minimizer of this problem is given by*

$$\hat{\rho} = \begin{cases} \frac{1-\sqrt{1+4k_s^2}}{2k_s} & k_s \neq 0 \\ 0 & k_s = 0 \end{cases}, \tag{1.7}$$

where $k_s = -\frac{q_{12}}{\det(\mathbf{Q})}$.

Lemma 1.2.3. *Let* $\mathbf{Q}$ *be a* 2×2 *symmetric positive definite matrix. Consider the Bregman Projection of* $\mathbf{Q}$ *onto the set of correlation matrices under the Von Neumann Divergence. Then, the unique minimizer of this problem is given by*

$$\hat{\rho} = \tanh(k_v) \ ,$$

where k_v *denotes the off-diagonal entry of* $\log \mathbf{Q}$.[3]

It is important to emphasize that the optimal projection in these two cases looks different from rescaling, thus implying that rescaling, at least as far as the Stein and Von Neumann divergences are concerned, is not optimal.

In the n-dimensional case we can derive an algorithm. Computing the Bregman projection $P_\phi(\mathbf{Q})$ is equivalent to solving the following optimization problem

[3]Straightforward computations show that k_v has the following analytical expression $k_v = \frac{\psi_1 \ln \lambda_1}{1+\psi_1^2} + \frac{\psi_2 \ln \lambda_2}{1+\psi_2^2}$ where

$$\lambda_i = \frac{1}{2}\left[q_{11} + q_{22} + (-1)^{i-1}\sqrt{(q_{11}-q_{22})^2 + 4q_{12}^2}\right]$$
$$\psi_i = -2q_{12}/\left(q_{11} - q_{22} + (-1)^i\sqrt{(q_{11}-q_{22})^2 + 4q_{12}^2}\right)$$

and $i = 1, 2$.

with n affine constraints (one for each diagonal element of $\mathbf{R}$):

$$\min_{\mathbf{R}\in\mathbb{S}^n_{++}} d_\phi(\mathbf{R},\mathbf{Q}) \text{ subject to } \mathbf{R}_{ii}=1 \text{ for all } i=1,\ldots,n. \tag{1.8}$$

$\mathbf{R}_{ii}$ stands for the i^{th} diagonal element of the matrix $\mathbf{R}$.

To solve this problem, we use Bregman's *cyclic projections* method. Let $\mathbb{C}_i$ be the set of n-dimensional symmetric positive definite matrices whose i^{th} diagonal element is unity. Clearly, the set of correlation matrices $\mathbb{C}^n = \bigcap_{i=1}^n \mathbb{C}_i$. Bregman's cyclic projections method is an iterative algorithm in which one projects successively onto each basic constraint set $\mathbb{C}_i$ until the sequence of iterates converges to the Bregman projection onto the intersection $\mathbb{C}^n$. Theorem 1 establishes that this algorithm is asymptotically valid. We refer to Dhillon and Tropp (2007) and the references therein for a proof.

Theorem 1.2.1. *Suppose*

1. *ϕ is a closed convex proper function of Legendre type such that* $\mathsf{int}(\mathsf{dom}(\phi)) = \mathbb{S}^n_{++}$.

2. *$\{\mathbb{C}_i\}_{i=1}^n$ are the sets of n-dimensional symmetric positive definite matrices with unit i^{th} diagonal entry.*

3. *the control mapping $m : \mathbb{N} \to \{1,\ldots,n\}$ is a sequence that takes each output value an infinite number of times.*

For $k = 1, 2, \ldots$, define $P_{\phi,m(k)}(\mathbf{R}^{(k-1)})$ as the Bregman projection of $\mathbf{R}^{(k-1)}$ onto $\mathbb{C}_{m(k)}$. Choose $\mathbf{R}^{(0)} = \mathbf{Q} \in \mathbb{S}^n_{++}$, and form a sequence of iterates via succes-

sive Bregman projections $\mathbf{R}^{(k)} = P_{\phi,m(k)}(\mathbf{R}^{(k-1)})$. *Then the sequence of iterates* $\{\mathbf{R}^{(k)}\}$ *converges in spectral norm to* $P_\phi(\mathbf{Q})$.

Lemma 1.2.4 establishes a closed-form formula for $P_{\phi,m(k)}(\mathbf{R}^{(k-1)})$ and is a special case of the derivation in Kulis *et al.* (2009) when $\phi(\mathbf{M}) = -\ln\det(\mathbf{M})$.

Lemma 1.2.4. *Consider the setting in Theorem 1 and let* $\phi(\mathbf{M}) = -\ln\det(\mathbf{M})$. *Then, for all* $i \in \{1, \ldots, n\}$,

$$P_{\phi,i}(\mathbf{R}^{(k-1)}) = \mathbf{R}^{(k-1)} + [\mathbf{R}_{ii}^{(k-1)}]^{-2}\left(1 - \mathbf{R}_{ii}^{(k-1)}\right)\mathbf{R}^{(k-1)}e_i e_i'\mathbf{R}^{(k-1)} ,$$

where e_i *denotes the* i^{th} *canonical basis vector.*

We concisely describe this procedure in Algorithm 1. We point out that the algorithm has a complexity of $O(n^2)$ per iteration. We remark that Algorithm 1 is an iterative procedure that relies on a tolerance parameter to determine convergence. In practice, in the empirical application and simulations we have found that a tolerance value of 10^{-6} is sufficiently accurate.

1.2.3 Discussion

Projecting vs Rescaling a Pseudo-Correlation Matrix

In this section we show that the difference between rescaling and projecting can be relevant enough in many cases.

We first consider the difference between rescaling and projecting in a bivariate setting. We denote by q_{11}, q_{22} the diagonal elements of the pseudo-correlation matrix $\mathbf{Q}$ and denote by q_{12} its off-diagonal element. Simple algebra shows that when

Algorithm 1 STEIN'S PROJECTION

Compute Stein's projection of a symmetric positive definite matrix $\mathbf{Q}$ onto the set of correlation matrices.

INPUT: A symmetric positive definite matrix $\mathbf{Q}$.

INITIALIZATION
Set $\mathbf{R}^{(0)} = \mathbf{Q}$.

ITERATE UNTIL CONVERGENCE
In the k-th iteration of the algorithm choose the i-th constraint as

$$i = \arg \max_{s \in \{1,\dots,n\}} |1 - \mathbf{R}_{ss}^{(k-1)}| \,,$$

and update the projection according to the formula

$$\mathbf{R}^{(k)} = \mathbf{R}^{(k-1)} + \left[\mathbf{R}_{ii}^{(k-1)}\right]^{-2} (1 - \mathbf{R}_{ii}^{(k-1)}) \mathbf{R}^{(k-1)} e_i e_i' \mathbf{R}^{(k-1)} \,,$$

where e_i is defined as the i^{th} canonical basis vector.

COVERGENCE CRITERIA
If $\max_s |1 - \mathbf{R}_{ss}^{(k)}| <$ tolerance then stop.

OUTPUT: The projected correlation matrix $\mathbf{R}^{(k)}$.

$q_{11}q_{22} = 1$, then the expression in equation (1.7) boils down to q_{12}, which trivially coincides with rescaling q_{12} by $\sqrt{q_{11}q_{22}}$. When $q_{11}q_{22} \neq 1$, this is generally not true, as it is shown in Figure 1.1. In the top-left panel we give an example of a combination of diagonal elements whose product is one, and observe that rescaling and projecting are equivalent. If the product is greater than 1, the projected correlation is below the rescaled one, and the difference increases as the product is larger than 1. The reverse pattern occurs when the product between the diagonal elements is lower than 1. We also note that the point at which the maximum difference occurs does not correspond to the same correlation level but is a function of the product of the diagonal elements of $\mathbf{Q}$. Note that in Figure 1.1 we report only

positive correlations. An analogous pattern emerges when they are negative (if the product of diagonal elements is greater than one, then the projected correlation is above the rescaled one, and it is below otherwise).

Next, we illustrate the difference between rescaling and projecting in a large dimensional setting. Assume that the n-dimensional pseudo-correlation matrix is given by

$$\mathbf{Q} = \mathrm{Diag}(1+\xi+v)^{1/2}\mathbf{R}(\kappa_1)\,\mathrm{Diag}(1+\xi+v)^{1/2}\,,$$

where $\kappa_1 \in (0,1)$, $\xi > 0$, $\mathbf{R}(\kappa_1)$ is an $n \times n$ matrix with Toeplitz structure, v is an n-dimensional vector, and $\mathrm{Diag}(1+\xi+v)$ denotes the $n \times n$ diagonal matrix with diagonal given by $1+\xi+v$. In particular, we have that the (i,j)-th element of $\mathbf{R}(\kappa_1)$ is $\kappa_1^{|i-j|}$. The $v = (v_1, \ldots, v_n)'$ vector has its i-th element equal to $\sin(x_i)$, where $x_1 = \varepsilon, x_n = 2\pi - \varepsilon$, and $x_{i+1} = x_i + 2 \cdot \frac{\pi-\varepsilon}{n-1}$, for $\varepsilon > 0$. In Figure 1.2 we see that the difference – measured with the squared Frobenius norm divided by n – between projecting and rescaling has an inverse-U shape with respect to the magnitude of the correlations. Importantly, the figure also shows that differences become more pronounced with the matrix dimension.

Projecting as Shrinking

This sub-section explores the relationship between projecting and shrinkage. The main message is that there are some cases where the projection may be interpreted as a methodology that shrinks the eigenvalues of a given pseudo-correlation matrix compared to the rescaled version. However, this does not hold in general for all possible pseudo-correlation matrices. Hence, interpreting the projection as

shrinkage is not entirely straightforward. To see this, consider first the case where $n = 2$ with the Stein's loss. From Laurent *et al.* (2013), we have that when $\phi(\mathbf{M}) = -\ln\det(\mathbf{M})$, then

$$d_\phi(\mathbf{R}, \mathbf{Q}) = \sum_{i,j=1}^{n} \frac{\lambda_i}{\mu_j}(v_i' u_j)^2 - \sum_{i=1}^{n} \ln\frac{\lambda_i}{\mu_i} - n \,, \tag{1.9}$$

where λ_i, v_i and μ_j, u_j are the i and j-th eigenvalues/vectors of $\mathbf{R}$ and $\mathbf{Q}$, respectively. If $u_i = v_i$ for all $i = 1, 2$, and since $\lambda_1 + \lambda_2 = 2$ with $\lambda_1 > \lambda_2 > 0$, then we can re-parameterize the matrix nearness problem as a function of λ_1:

$$\min_{\mathbf{R}\in\mathbb{C}^n} d_\phi(\mathbf{R}, \mathbf{Q}) \iff \min_{\lambda_1\in(0,2)} \left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)\lambda_1 - \ln\lambda_1 - \ln(2 - \lambda_1) \,,$$

with unique solution given by

$$\lambda_1 = 1 - \frac{1}{g} + \sqrt{1 + \frac{1}{g^2}} \,, \qquad g := \frac{1}{\mu_2} - \frac{1}{\mu_1} \,,$$

which is to be compared against $2\mu_1/\text{tr}(\mathbf{Q})$, i.e. the eigenvalue obtained via rescaling. After some algebra, it can be shown that $\lambda_1 < 2\mu_1/\text{tr}(\mathbf{Q})$ whenever $\text{tr}(\mathbf{Q}) > 2$, and viceversa. Therefore, even in this simplified scenario we can see that the projection methodology may have an effect similar to shrinking the eigenvalues of the rescaled $\mathbf{Q}$, but it may as well have the opposite effect.

Clearly, the analysis becomes more complex when $n > 2$ since normalizing $\mathbf{Q}$ into a correlation matrix involves not just a change in the eigenspectrum but also rotating the eigenvectors. Under the same setup of sub-section 1.2.3, Figure 1.3 illustrates the difference between projecting $\mathbf{Q}$ versus shrinking and rescaling $\mathbf{Q}$ via the analytical nonlinear shrinkage formula of Ledoit and Wolf (2020) for different values of the concentration ratio n/T. The figure shows that the gap widens

with the degree of dependence and the dimension –as in the previous sub-section– as well as with the concentration ratio n/T, hence suggesting that projecting and shrinking are essentially different operations.

Measurement Error

In the context of the DCC model, the observed pseudo-correlation matrix $\mathbf{Q}$ is estimated. It is natural to consider the impact of measurement error on the projection methodology as opposed to rescaling. The analysis of this question is inspired by Laurent *et al.* (2013). Let $\tilde{\mathbf{Q}}$ denote the estimated value of $\mathbf{Q}$, and denote $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{P}}$ the rescaled/projected $\tilde{\mathbf{Q}}$. The structure of Bregman divergences allows us to write the average discrepancy between rescaling and projecting as

$$\mathbb{E}[d_\phi(\tilde{\mathbf{R}}, \mathbf{Q}) - d_\phi(\tilde{\mathbf{P}}, \mathbf{Q})] = \mathbb{E}[\phi(\tilde{\mathbf{R}}) - \phi(\tilde{\mathbf{P}})] - \mathbb{E}\,\mathrm{tr}(\nabla\phi(\mathbf{Q})(\tilde{\mathbf{P}} - \tilde{\mathbf{R}}))\ .$$

It follows that the measurement error does not impact the proximity rankings of $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{P}}$ provided that this discrepancy is positive. Although it seems challenging to derive general conclusions for general Legendre functions ϕ, if we consider $\phi(\mathbf{M}) = -\ln\det(\mathbf{M})$, then the proximity rankings are not affected by measurement error whenever the trace of $\mathbf{Q}$ is large enough and the eigenvalues of $\tilde{\mathbf{R}}$ are more dispersed than those of $\tilde{\mathbf{P}}$.

1.3 Simulation Study

In this section we carry out different Monte Carlo exercises to analyse the performance of Stein's projection and the Pro-DCC for correlation modeling.

1.3.1 Static Correlations

In the Monte Carlo exercise of this section we study the performance of our proposed projection in a static environment. We carry out a Monte Carlo exercise designed as follows. We simulate $T = 500$ i.i.d. random draws from an equicorrelation process given by

$$r \sim \mathcal{N}(0, \mathbf{R}), \text{ where } \mathbf{R} = (1-\rho)\mathbf{I} + \rho\iota\iota',$$

where ι is an n-dimensional vector of ones, $\rho \in (0,1)$ and $\mathbf{I}$ is the $n \times n$ identity matrix. The focus of the exercise lies in the estimation of the population correlation matrix $\mathbf{R}$. We consider two competing estimators. The first candidate estimator is Pearson's sample correlation matrix, that is

$$\hat{\mathbf{R}}^{(1)} = \operatorname{diag}(\mathbf{S})^{-1/2}\mathbf{S}\operatorname{diag}(\mathbf{S})^{-1/2} ,$$

where $\mathbf{S}$ denote the sample covariance matrix $\mathbf{S} = \frac{1}{T}\sum_{t=1}^{T} r_t r_t'$. The second estimator is the projected sample covariance matrix of the data, defined as $\hat{\mathbf{R}}^{(2)} = P_{STEIN}(\mathbf{S})$.

In each replication, we compute the loss of both estimators with respect to the true correlation matrix. The losses under consideration are the Frobenius and the MAE. These are defined as follows:

$$\mathcal{L}_{Frob}(\hat{\mathbf{R}}, \mathbf{R}) = \sqrt{\frac{1}{n}\operatorname{tr}[(\hat{\mathbf{R}} - \mathbf{R})^2]} , \tag{1.10}$$

$$\mathcal{L}_{MAE}(\hat{\mathbf{R}}, \mathbf{R}) = \frac{1}{n}\sum_{i,j} |\hat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}| . \tag{1.11}$$

Note that we divide the squared Frobenius loss and the sum of absolute errors by n to establish a fair comparison as the dimension increases. We estimate $\mathbb{E}[\mathcal{L}]$ using

the sample average of the losses obtained across 1'000 Monte Carlo replications, and repeat the same exercise for different levels of the correlation parameter ρ as well as the cross-sectional dimension n. In particular, we consider all combinations of ρ in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and n in $\{10, 23, 36, 50\}$. We remark that the concentration ratio is rather low (it only reaches a maximum of 0.1) since this exercise is based on the sample covariance, which is optimal under fixed-n large-T asymptotics but performs poorly in finite samples when the concentration ratio n/T is large. In unreported Monte Carlo exercises we have also implemented the projection of the non-linear shrinkage estimator (Ledoit and Wolf, 2020) of the covariance matrix onto the correlation set and compared them to the sample correlation. The results were also favorable to our proposed methodology.

Figure 1.4 reports the excess loss of the sample correlation matrix with respect to our proposed estimator. From this figure, it is clear that the gap between both methods is positive and increases with the dimensionality of the correlation matrix. We also observe that the gap is maximized at some intermediate value of the ρ parameter between zero and one.

To conclude this sub-section, note that the results are congruent with our findings from Section 1.2.3. The main message is that care must be taken in choosing an appropriate way of normalizing a positive definite matrix to a correlation. Rescaling is one (obvious) way of doing so, but it may not be necessarily optimal.

1.3.2 Dynamic Correlations

In the Monte Carlo exercise of this section we study the performance of our proposed projection in a dynamic environment. We carry out a Monte Carlo exercise designed as follows. We consider a dynamic equicorrelation model where the scalar correlation parameter is governed by a cosine wave, in the spirit of Engle and Kelly (2012) and one of the simulated DGPs considered in Engle (2002). More precisely, we consider the DGP given by the equations

$$\begin{aligned}
\epsilon_t &= \mathbf{R}_t^{1/2} z_t, \text{ where } z_t \sim \mathcal{N}(0, \mathbf{I}), \text{ for } t = 1, \dots, T \,, \\
\mathbf{R}_t &= (1 - \rho_t)\mathbf{I} + \rho_t \iota \iota' \,, \\
\rho_t &= \rho + 0.1 \cos(2\pi t / 200) \,.
\end{aligned}$$

In each replication, we draw $T = 1'000$ observations from the DGP described above for different levels of the ambient dimension $n = 50, 100, 200$ and $\rho = 0.4, 0.5, 0.6, 0.7, 0.8$.

To predict the sequence of matrices $\mathbf{R}_t$, we consider the DCC and Pro-DCC models. As customary, the parameters of the two specifications are estimated by maximizing the Quasi Log-Likelihood of the data with respect to the correlation matrix parameter, where the intercept matrix $\mathbf{C}$ is estimated by covariance targeting.

A few remarks are in order. In this exercise, we abstract from estimation of dynamic volatilities since the DCC and Pro-DCC methodologies employ the same univariate GARCH strategy to estimate the conditional volatilities. The exercise thus focuses on the main source of the difference between both strategies, that is

the correlation modeling step. Similar to the previous sub-section, we employ the sample covariance for the covariance targeting step, but in unreported exercises we have also implemented the non-linear shrinkage estimator and the results were also favourable to our proposed methodology.

In order to evaluate the precision of the correlation forecasts with respect to the true correlation process $\mathbf{R}_t$ we use again the Frobenius and MAE loss functions given by equations (1.10) and (1.11). For each replication of the exercise we compute the average losses across all the time periods $t = 1, \ldots, T$ and we then average again these across 1'000 replications of the Monte Carlo exercise.

The correlation plot in Figure 1.5 illustrates how the Pro-DCC correlation delivers a performance improvement over the DCC. In Figure 1.6 we report the results of the exercise for different values of n and ρ. Overall, results show that the Pro-DCC performs systematically better than rescaling and that the gains of the projections are larger when the dimension of the system is larger and when the degree of correlation is higher. These results are robust to the choice of the amplitude and period of the cosine wave. We remark that here we have focused on the positive range of the correlations to avoid issues related to the positive definiteness of the equicorrelation matrix since the corresponding lower bound for the ρ_t parameter varies with the dimension n.

1.4 Empirical Application

In this Section we carry out an out-of-sample asset allocation exercise using the constituents of the S&P 100 to assess the benefits of the Pro-DCC for forecasting. The exercise design is close in spirit to the one in De Nard *et al.* (2021). An important difference between the latter and this chapter is that we use publicly available daily data from Alpha Vantage (closing stock prices adjusted for dividends and splits) and that the focus is exclusively on the S&P 100. Hence, the investment universe is fixed rather than time-varying and smaller than the one considered in that paper.

We consider the $n = 86$ constituents of the S&P 100 that have continuously been trading between 2011-01-01 and 2019-06-30 (2'136 trading days). We transform adjusted close price data into log-returns, namely $r_{it} = 100 \times \log(P_{it}/P_{it-1})$ for $i = 1, \ldots, n$. The exercise consists in constructing the global minimum variance portfolio (GVMP) once a month on the basis of different covariance forecasts and to then measure the accuracy of the different covariance forecasts on the basis of asset allocation metrics.[4] It should be emphasized that despite the fact that portfolio rebalancing occurs on a monthly basis, all models are estimated using daily data. The GMVP is defined as $r_{\mathsf{GMVP}\,t} = w_t^{*\prime} r_t$ where

$$w_t^* = \arg\min_{w_t'\iota=1} w_t' \mathbf{\Sigma}_t w_t \ ,$$

with ι denoting an n-dimensional vector of ones. As it is well known, the mini-

[4] We follow the common convention that 21 consecutive trading days constitute one "month".

mizer of this optimization problem is given by

$$w_t^* = \frac{\mathbf{\Sigma}_t^{-1}\iota}{\iota'\mathbf{\Sigma}_t^{-1}\iota}\,. \tag{1.12}$$

The GMVP has become a fairly standard metric to evaluate covariance forecasts. In particular, the appeal of the GMVP lies in the fact that this asset allocation strategy only depends on covariance forecasts and in particular it does not rely on forecasts of the expected returns. As an additional exercise, we consider the GMVP with exposure constraints. In the literature, it is generally documented that adding exposure constraints improves the minimum-variance portfolio allocations and has an interpretation in terms of shrinkage estimation (Jagannathan and Ma, 2003; Fan *et al.*, 2012). The GMVP with 1-norm constraints is formulated as

$$w_t^* = \arg\min_{w_t'\iota=1, \|w_t\|_1 \leqslant \gamma} w_t'\mathbf{\Sigma}_t w_t, \tag{1.13}$$

for some $\gamma > 0$. We consider the well-known choices $\gamma = 1$ (the no-short sale portfolio) and $\gamma = 2$ (the 150/50 portfolio, i.e. the portfolio that allows a maximum of 50% of short positions). In order to understand whether projections represent an improvement beyond shrinkage, we compute $\gamma_t = \sum_{i=1}^n |w_{i,t}|$ for each t for the portfolio based on the Projected DCC and add the portfolio constraints of the standard DCC with exposure constraint $\gamma = \gamma_t$.[5]

To compute the GMVP from the data, we train some suitable dynamic covariance model in-sample and compute one-step ahead covariance forecasts out-of-sample on a rolling basis, keeping the parameters fixed to their in-sample estimates. The portfolio is updated every month to mitigate the costs of rebalanc-

[5]We thank one of the referees for the suggestion.

ing. We note that in De Nard *et al.* (2021) the approach to forecasting is slightly different, as they compute the average of k-step ahead forecasts for k ranging from 1 to 21. However, in order to compute those forecasts, it is assumed that $\mathbb{E}[\mathbf{R}_{t+k}|\mathcal{F}_t] = \mathbb{E}[\mathbf{Q}_{t+k}|\mathcal{F}_t]$, which is an approximation. In this work we refrain from following this strategy and we focus on one-step ahead forecasts only in order to make the comparison between DCC and Pro-DCC more transparent.

We consider different covariance forecasting strategies to construct the GMVP. We entertain: DCC, the standard version of the DCC model; Pro-DCC, the Projected DCC model based on Stein's projection; RiskMetrics, the RM2006 methodology of Zumbach (2007), and RollCov, Rolling sample covariance computed with a window of six months. For the GARCH-DCC/Pro-DCC specifications, the marginal distribution for each asset is assumed to be GARCH(1,1). For benchmarking purposes, we also report metrics for the equal-weighted portfolio ("1/N").

Moreover, since the dimensionality of the problem is fairly large, the DCC and Pro-DCC models are implemented using standard covariance targeting as well as covariance targeting based on the analytical nonlinear shrinkage (NLS) methodology from Ledoit and Wolf (2020). Also, DCC models are estimated using the composite likelihood approach with contiguous pairs for the dynamic correlation models as in Pakel *et al.* (2018), which significantly reduces the computational burden of estimation. We evaluate the performance of these different forecasting strategies using the following three out-of-sample performance measures: AV, the annualized out-of-sample average of the GMVP returns; SD, the annualized out-of-sample standard deviation of the GMVP returns, and Sharpe, the Sharpe ratio

computed as AV / SD.[6] For instance, the annualized volatility of the portfolio is given by

$$SD_{\mathsf{GMVP}} = \sqrt{252} \times \sqrt{\tau^{-1} \sum_{t=T+1}^{T+\tau} r^2_{\mathsf{GMVP}\,t}}\,,$$

where τ is the length of the out-of-sample period and T is the length of the in-sample period.

The resulting portfolio metrics are presented in Table 1.1 and results using different split dates for the in-sample and out-of-sample periods can be found in Tables 1.2 and 1.3. Results from tests of the difference in portfolio variances using HAC inference as in Ledoit and Wolf (2008) are shown in Tables 1.4, 1.5 and 1.6. The findings of the exercise can be summarised as follows. First, overall, we have that Pro-DCC-NLS and Pro-DCC-SC outperform all other candidate estimators in terms of both the standard deviation and the Sharpe ratio of the GMVP. The (analytical) nonlinear shrinkage versions of all estimators considered in the exercise present superior performance than doing no shrinkage. Second, the rel-

[6] Additionally, the following portfolio metrics are computed:

- Turnover: $\frac{1}{n(H-1)} \sum_{h=1}^{H} \|\hat{w}_{h+1} - \hat{w}_h^{hold}\|_1$, where H is the total number of months out-of-sample, and w_h^{hold} is the weight vector right before the next monthly update, that is, taking into account the price evolution of each asset during the month.
- Proportion of leverage: $\frac{1}{nH} \sum_{h=1}^{H} \sum_{i=1}^{n} \mathbf{1}\{w_{i,h} < 0\}$.
- Gamma: $\frac{1}{H} \sum_{h=1}^{H} \sum_{i=1}^{n} |w_{i,h}|$.
- Maximum weight: $\max_{i,h} w_{i,h}$.
- Minimum weight: $\min_{i,h} w_{i,h}$.

ative ranking between models does not change when introducing the nonlinear shrinkage methodology, which suggests that it uniformly improves performance irrespective of the modeling choice for the time-varying covariance matrix. Third, the empirical results show that the 1-norm constrained portfolios (denoted with γ in Tables 1.1, 1.2 and 1.3) substantially improve performance for both the DCC and Pro-DCC. In a nuthsell, these results suggest that projecting, rescaling and shrinking are essentially different operations, and that the advantage of Pro-DCC versus DCC persists even after controlling for shrinkage. Incidentally we note that the equal-weighted portfolio ("1/N") in this sample performs relatively well in comparison to other studies – despite being the worst performing portfolio in terms of standard deviation. However, it must be noted that for larger cross sections and/or periods of distress like the 2008 global financial crisis, the results from the 1/N portfolio deteriorate significantly – see for example De Nard *et al.* (2021) and Engle *et al.* (2019).

1.4.1 Alternative parameterization of the Global Minimum Variance Portfolio

We provide a new approach to analyze the weights given by the Global Minimum Variance Portfolio. We define $\mathbf{K}_t := \mathbf{\Sigma}_t^{-1}$ and $\mathbf{\Omega}_t := \mathbf{R}_t^{-1}$, and note that the $n \times n$ matrix of partial correlations can be written as

$$\boldsymbol{\varrho}_t = -\operatorname{diag}(\mathbf{\Omega}_t)^{-1/2}\mathbf{\Omega}_t \operatorname{diag}(\mathbf{\Omega}_t)^{-1/2} .$$

Defining $v_{i,t} := \frac{\sqrt{\mathbf{\Omega}_{ii,t}}}{d_{i,t}}$, after straightforward computations we have that

$$w_t^* = \frac{\mathbf{K}_t \iota}{\iota' \mathbf{K}_t \iota} = \frac{v_t \odot \boldsymbol{\varrho}_t v_t}{\iota' (v_t \odot \boldsymbol{\varrho}_t v_t)}, \tag{1.14}$$

where $v_t = (v_{1,t}, \ldots, v_{n,t})'$ and $\mathbf{\Omega}_{ii,t}$ is the i^{th} diagonal element of $\mathbf{\Omega}_t$. This means that the sign and magnitude of the GMVP weights depend on 3 factors: 1) $d_{i,t}$, the volatility of the i^{th} asset, which always shrinks the exposure to 0. 2) $s_{i,t} := \frac{1}{\sqrt{\mathbf{\Omega}_{ii,t}}}$ or in words, the "spillover" effect which is proportional to the determinant of the conditional correlation matrix $\mathbf{R}_t$ given by all assets excluding i. The higher the $s_{i,t}$, the stronger is the shrinkage towards 0. 3) $\boldsymbol{\varrho}_{ij,t}$: partial correlation between assets i and j, which measures the direct effects between assets i and j conditional on all other assets in the portfolio. Note that if the i^{th} asset is directly and strongly connected to the rest of the system, the weight is likely to be negative (i.e. increased exposure in short position).

The parameterization allows us to deepen our understanding of the GMVP weight vector. In Figure 1.7 we can visualize the alternative parameterization of the weight vector for a given date. For instance, if we focus on asset number 20, we observe that the weight implied by the DCC model is larger than the one implied by the Pro-DCC model (the actual weights on the given date are 7.5% vs 1.8%, respectively). From the second column we learn that if all assets are independent of each other, then the weight should be the lowest since this asset is the most volatile at the given date. Both models convey that partial correlation effects for this asset are on the lower end of the distribution, which makes the asset more attractive than what is implied by the volatility alone. The main difference is

thus explained by the third-to-last column (s), which captures the spillover effects from the asset in question to the rest of the system. In other words, the Pro-DCC model implies that asset 20 is indirectly connected to the remaining assets way more strongly than what is implied by the DCC model, which explains why it has a lower weight.

1.5 Conclusions

In this chapter we contribute to the DCC literature with a novel specification inspired by the literature on Bregman matrix projections and the nearest-correlation matrix problem. We demonstrate the benefits of using our proposed methodology with respect to the standard GARCH-DCC model in a simulated exercise. We also carry out a global minimum variance portfolio exercise using a set of constituents of the S&P 100. Results show that the standard and nonlinear shrinkage versions of Pro-DCC outperform all other candidate estimators of the conditional covariance matrix in terms of the standard deviation and Sharpe ratio of the GMVP.

1.6 Proofs

Proof of Lemma 1.2.1. The claim follows from Theorem 3.12(iii) in Bauschke and Borwein (1997). □

Proof of Lemma 1.2.2. Let $\mathbf{K} = \mathbf{Q}^{-1}$. Consider the $n = 2$ case:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \ \mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix} \text{ and } \mathbf{K} = \begin{bmatrix} k_1 & k_s \\ k_s & k_2 \end{bmatrix}.$$

Therefore, $\mathrm{tr}(\mathbf{RK}) = k_1 + 2k_s\rho + k_2$, $\det(\mathbf{RK}) = (k_1k_2 - k_s^2)(1-\rho^2)$, and our minimization problem can be formulated as a univariate problem:

$$\min_{\rho} f(\rho) = \min_{\rho} k_1 + 2k_s\rho + k_2 - \ln(k_1k_2 - k_s^2) - \ln(1-\rho^2) - 2.$$

Since the problem is convex and the domain of f is the open interval $(-1, 1)$, it suffices to take the first order condition and solve for ρ, which yields the result in (1.7). □

Proof of Lemma 1.2.3. Let $\mathbf{K} = \log \mathbf{Q}$. In the bivariate case, we have that

$$\mathrm{tr}(\mathbf{R}\log\mathbf{R}) - \mathrm{tr}(\mathbf{RK}) = \ln(1-\rho^2) + \rho\ln\left(\frac{1+\rho}{1-\rho}\right) - 2k_v\rho + const := f(\rho),$$

which follows since the matrix logarithm of $\mathbf{R}$ is given by

$$\log\mathbf{R} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \ln(1+\rho) & 0 \\ 0 & \ln(1-\rho) \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

Hence, the problem is equivalent to minimizing f with respect to ρ.[7] Since the problem is convex and the domain of f is the open interval $(-1, 1)$, it suffices to take the first order condition and solve for ρ:

$$\frac{-2\hat{\rho}}{1-\hat{\rho}^2} + \ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) + \hat{\rho}\frac{1-\hat{\rho}}{1+\hat{\rho}}\frac{2}{(1-\hat{\rho})^2} - 2k_v = 0.$$

[7] Note that we can ignore the terms $\mathrm{tr}(\mathbf{R})$ and $\mathrm{tr}(\mathbf{Q})$ that appear in $d_\phi(\mathbf{R}, \mathbf{Q})$ as these do not depend on ρ.

Therefore,

$$\hat{\rho} = \frac{e^{2k_v} - 1}{e^{2k_v} + 1} = \tanh(k_v).$$

To find an analytical expression for k_v, let $\mathbf{V}\mathbf{\Lambda}\mathbf{V}'$ be the eigendecomposition of $\mathbf{Q}$, where $\mathbf{V}$ is orthonormal. It is easy to verify that the eigenvalues of $\mathbf{Q}$ are given by

$$\lambda_i = \frac{1}{2}\left[q_{11} + q_{22} + (-1)^{i-1}\sqrt{(q_{11} - q_{22})^2 + 4q_{12}^2}\right],$$

where $i = 1, 2$. Their corresponding eigenvectors are $v_i = [v_{i1}, v_{i2}]'$, where

$$v_{i1} = -2q_{12}/\left(q_{11} - q_{22} + (-1)^i\sqrt{(q_{11} - q_{22})^2 + 4q_{12}^2}\right) v_{i2} := \psi_i v_{i2}.$$

Imposing unit norm eigenvectors, we have that $v_{i2} = (1 + \psi_i^2)^{-1/2}$. Hence, it is easy to see that the (2,1) entry of the $\mathbf{K}$ matrix is given by

$$\begin{aligned} k_v &= (\ln\lambda_1)v_{11}v_{12} + (\ln\lambda_2)v_{21}v_{22} = (\ln\lambda_1)\psi_1 v_{12}^2 + (\ln\lambda_2)\psi_2 v_{22}^2 \\ &= \frac{\psi_1 \ln\lambda_1}{1 + \psi_1^2} + \frac{\psi_2 \ln\lambda_2}{1 + \psi_2^2}. \end{aligned}$$

□

Proof of Lemma 1.2.4. Let $\mathbf{R}^{(0)} = \mathbf{Q}$. Note that

$$P_{\phi,i}(\mathbf{R}^{(k-1)}) = \arg\min_{\mathbf{R}^{(k)} \in \mathbb{C}_i} d_\phi\left(\mathbf{R}^{(k)}, \mathbf{R}^{(k-1)}\right).$$

The first order condition of the Lagrangian yields the following matrix update for $\mathbf{R}^{(k)}$:

$$\begin{cases} \nabla\phi(\mathbf{R}^{(k)}) = \nabla\phi(\mathbf{R}^{(k-1)}) + \alpha e_i e_i' \\ \operatorname{tr}(\mathbf{R}^{(k)} e_i e_i') = 1 \end{cases}.$$

When $\phi(\cdot) = -\ln\det(\cdot)$, we have that $\nabla\phi(\mathbf{R}^{(k)}) = -[\mathbf{R}^{(k)}]^{-1}$, and the first equation of the system becomes

$$\mathbf{R}^{(k)} = \left([\mathbf{R}^{(k-1)}]^{-1} - \alpha e_i e_i'\right)^{-1}.$$

Using Sherman-Morrison's formula, we can re-write the first equation as

$$\mathbf{R}^{(k)} = \mathbf{R}^{(k-1)} + \frac{\alpha}{1 - \alpha e_i' \mathbf{R}^{(k-1)} e_i} \mathbf{R}^{(k-1)} e_i e_i' \mathbf{R}^{(k-1)}.$$

Note that $\mathrm{tr}(\mathbf{R}^{(k-1)} e_i e_i') = e_i' \mathbf{R}^{(k-1)} e_i = \mathbf{R}_{ii}^{(k-1)}$. It follows that $\mathrm{tr}(\mathbf{R}^{(k-1)} e_i e_i' \mathbf{R}^{(k-1)} e_i e_i') = (e_i' \mathbf{R}^{(k-1)} e_i)^2 = [\mathbf{R}_{ii}^{(k-1)}]^2$. Plugging the first equation in the second one and solving for α we get

$$\mathrm{tr}\left(\left[\mathbf{R}^{(k-1)} + \frac{\alpha}{1 - \alpha \mathbf{R}_{ii}^{(k-1)}} \mathbf{R}^{(k-1)} e_i e_i' \mathbf{R}^{(k-1)}\right] e_i e_i'\right) = 1,$$

so $\alpha = [\mathbf{R}_{ii}^{(k-1)}]^{-1} - 1$. Replacing α in the first equation of the system yields the desired result, since

$$\begin{aligned}\frac{\alpha}{1 - \alpha \mathbf{R}_{ii}^{(k-1)}} &= \frac{[\mathbf{R}_{ii}^{(k-1)}]^{-1} - 1}{1 - ([\mathbf{R}_{ii}^{(k-1)}]^{-1} - 1)\mathbf{R}_{ii}^{(k-1)}} \\ &= [\mathbf{R}_{ii}^{(k-1)}]^{-2} - [\mathbf{R}_{ii}^{(k-1)}]^{-1} = [\mathbf{R}_{ii}^{(k-1)}]^{-2}\left(1 - \mathbf{R}_{ii}^{(k-1)}\right).\end{aligned}$$

□

1.7 Tables

Split Date: 2015-12-31

Table 1.1: Portfolio selection with all constituents in S&P 100. Performance metrics for the out-of-sample period, which ranges from split date to 2019-06-

	AV	SD	Sharpe	Turnover	Leverage	Gamma	MaxWeight	MinWeight
DCC-SC	7.880	12.489	0.631	0.065	0.493	2.945	0.680	-0.125
Pro-DCC-SC	10.228	11.918	0.858	0.093	0.488	2.935	0.439	-0.125
DCC-SC($\gamma = \gamma_t$)	7.833	12.478	0.628	0.293	0.490	2.897	0.679	-0.125
DCC-SC($\gamma = 1$)	7.880	11.989	0.657	0.019	0.019	1.001	0.911	-0.004
Pro-DCC-SC($\gamma = 1$)	9.133	**11.427**	0.799	0.028	0.019	1.001	0.592	-0.004
DCC-SC($\gamma = 2$)	8.975	12.133	0.740	0.081	0.447	2.000	0.678	-0.093
Pro-DCC-SC($\gamma = 2$)	11.245	11.676	0.963	0.036	0.451	2.000	0.441	-0.116
DCC-NLS	7.745	12.362	0.626	0.071	0.492	2.756	0.685	-0.097
Pro-DCC-NLS	10.139	11.823	0.858	0.050	0.490	2.733	0.432	-0.107
RiskMetrics	8.387	12.251	0.685	0.173	0.428	4.724	0.466	-0.418
RollCov	14.931	19.799	0.754	0.559	0.463	7.787	1.017	-0.795
1/N	10.440	12.426	0.840	0.016	0.000	1.000	0.012	0.012

30. AV: annualized average portfolio return. SD: annualized volatility of portfolio returns. Sharpe: AV / SD. Turnover: $\frac{1}{n(H-1)} \sum_{h=1}^{H} \|\hat{w}_{h+1} - \hat{w}_h^{hold}\|_1$, where H is the total number of months out-of-sample, and $\hat{w}_h^{hold}$ is the weight vector right before the next monthly update, that is, taking into account the price evolution of each asset during the month. Leverage: $\frac{1}{nH} \sum_{h=1}^{H} \sum_{i=1}^{n} \mathbf{1}\{w_{i,h} < 0\}$. Gamma: $\frac{1}{H} \sum_{h=1}^{H} \sum_{i=1}^{n} |w_{i,h}|$. Maximum weight: $\max_{i,h} w_{i,h}$. Minimum weight: $\min_{i,h} w_{i,h}$. The annualized volatility of the best portfolio is highlighted in boldface.

Split Date: 2016-12-31

	AV	SD	Sharpe	Turnover	Leverage	Gamma	MaxWeight	MinWeight
DCC-SC	4.122	12.439	0.331	0.051	0.491	2.749	0.558	-0.107
Pro-DCC-SC	5.881	12.112	0.486	0.058	0.489	2.691	0.481	-0.104
DCC-SC($\gamma = \gamma_t$)	4.275	12.406	0.345	0.044	0.489	2.669	0.567	-0.107
DCC-SC($\gamma = 1$)	10.284	11.878	0.866	0.013	0.023	1.001	0.803	-0.004
Pro-DCC-SC($\gamma = 1$)	9.196	**11.454**	0.803	0.013	0.021	1.000	0.681	-0.003
DCC-SC($\gamma = 2$)	6.635	12.207	0.544	0.039	0.462	2.000	0.578	-0.087
Pro-DCC-SC($\gamma = 2$)	7.588	11.884	0.638	0.050	0.473	2.000	0.487	-0.099
DCC-NLS	4.295	12.348	0.348	0.038	0.494	2.627	0.567	-0.100
Pro-DCC-NLS	5.960	12.017	0.496	0.055	0.495	2.567	0.492	-0.092
RiskMetrics	1.229	12.100	0.102	0.125	0.416	4.329	0.388	-0.368
RollCov	1.591	15.945	0.100	0.253	0.455	7.048	0.815	-0.606
1/N	9.519	12.126	0.785	0.013	0.000	1.000	0.012	0.012

Table 1.2: Portfolio selection with all constituents in S&P 100. Performance metrics for the out-of-sample period, which ranges from split date to 2019-06-30. AV: annualized average portfolio return. SD: annualized volatility of portfolio returns. Sharpe: AV / SD. Turnover: $\frac{1}{n(H-1)} \sum_{h=1}^{H} \|\hat{w}_{h+1} - \hat{w}_h^{hold}\|_1$, where H is the total number of months out-of-sample, and $\hat{w}_h^{hold}$ is the weight vector right before the next monthly update, that is, taking into account the price evolution of each asset during the month. Leverage: $\frac{1}{nH} \sum_{h=1}^{H} \sum_{i=1}^{n} \mathbf{1}\{w_{i,h} < 0\}$. Gamma: $\frac{1}{H} \sum_{h=1}^{H} \sum_{i=1}^{n} |w_{i,h}|$. Maximum weight: $\max_{i,h} w_{i,h}$. Minimum weight: $\min_{i,h} w_{i,h}$. The annualized volatility of the best portfolio is highlighted in boldface.

Split Date: 2017-12-31

	AV	SD	Sharpe	Turnover	Leverage	Gamma	MaxWeight	MinWeight
DCC-SC	0.186	14.672	0.013	0.069	0.499	2.606	0.392	-0.136
Pro-DCC-SC	2.198	14.443	0.152	0.064	0.500	2.483	0.336	-0.096
DCC-SC($\gamma = \gamma_t$)	0.165	14.623	0.011	0.136	0.493	2.468	0.388	-0.132
DCC-SC($\gamma = 1$)	1.858	13.755	0.135	0.015	0.026	1.001	0.599	-0.004
Pro-DCC-SC($\gamma = 1$)	2.111	**13.442**	0.157	0.043	0.023	1.001	0.503	-0.002
DCC-SC($\gamma = 2$)	1.065	14.427	0.074	0.065	0.466	2.000	0.414	-0.120
Pro-DCC-SC($\gamma = 2$)	3.095	14.172	0.218	0.072	0.467	2.000	0.342	-0.085
DCC-NLS	0.105	14.581	0.007	0.066	0.503	2.511	0.395	-0.129
Pro-DCC-NLS	2.019	14.350	0.141	0.052	0.499	2.384	0.337	-0.088
RiskMetrics	-4.857	14.514	-0.335	0.159	0.426	4.776	0.386	-0.362
RollCov	-9.864	18.875	-0.523	0.232	0.462	7.715	0.814	-0.569
1/N	4.461	14.745	0.303	0.016	0.000	1.000	0.012	0.012

Table 1.3: Portfolio selection with all constituents in S&P 100. Performance metrics for the out-of-sample period, which ranges from split date to 2019-06-30. AV: annualized average portfolio return. SD: annualized volatility of portfolio returns. Sharpe: AV / SD. Turnover: $\frac{1}{n(H-1)} \sum_{h=1}^{H} \|\hat{w}_{h+1} - \hat{w}_h^{hold}\|_1$, where H is the total number of months out-of-sample, and $\hat{w}_h^{hold}$ is the weight vector right before the next monthly update, that is, taking into account the price evolution of each asset during the month. Leverage: $\frac{1}{nH} \sum_{h=1}^{H} \sum_{i=1}^{n} \mathbf{1}\{w_{i,h} < 0\}$. Gamma: $\frac{1}{H} \sum_{h=1}^{H} \sum_{i=1}^{n} |w_{i,h}|$. Maximum weight: $\max_{i,h} w_{i,h}$. Minimum weight: $\min_{i,h} w_{i,h}$. The annualized volatility of the best portfolio is highlighted in boldface.

	DCC-SC	Pro-DCC-SC	DCC-SC($\gamma = \gamma_t$)	DCC-SC($\gamma = 1$)	Pro-DCC-SC($\gamma = 1$)	DCC-SC($\gamma = 2$)	Pro-DCC-SC($\gamma = 2$)	DCC-NLS	Pro-DCC-NLS	RiskMetrics	RollCov	1/N
DCC-SC	–	-0.57*	-0.01	-0.5	-1.06***	-0.36***	-0.81***	-0.13***	-0.67**	-0.24	7.31***	-0.06
Pro-DCC-SC	0.57*	–	0.56*	0.07	-0.49	0.22	-0.24**	0.44	-0.09**	0.33	7.88***	0.51
DCC-SC($\gamma = \gamma_t$)	0.01	-0.56*	–	-0.49	-1.05***	-0.35***	-0.8***	-0.12***	-0.66**	-0.23	7.32***	-0.05
DCC-SC($\gamma = 1$)	0.5	-0.07	0.49	–	-0.56	0.14	-0.31	0.37	-0.17	0.26	7.81***	0.44
Pro-DCC-SC($\gamma = 1$)	1.06***	0.49	1.05***	0.56	–	0.71***	0.25	0.94***	0.4	0.82	8.37***	1**
DCC-SC($\gamma = 2$)	0.36***	-0.22	0.35***	-0.14	-0.71***	–	-0.46	0.23**	-0.31	0.12	7.67***	0.29
Pro-DCC-SC($\gamma = 2$)	0.81***	0.24**	0.8***	0.31	-0.25	0.46	–	0.69**	0.15**	0.58	8.12***	0.75**
DCC-NLS	0.13***	-0.44	0.12***	-0.37	-0.94***	-0.23**	-0.69**	–	-0.54*	-0.11	7.44***	0.06
Pro-DCC-NLS	0.67**	0.09**	0.66**	0.17	-0.4	0.31	-0.15**	0.54*	–	0.43	7.98***	0.6
RiskMetrics	0.24	-0.33	0.23	-0.26	-0.82	-0.12	-0.58	0.11	-0.43	–	7.55***	0.17
RollCov	-7.31***	-7.88***	-7.32***	-7.81***	-8.37***	-7.67***	-8.12***	-7.44***	-7.98***	-7.55***	–	-7.37***
1/N	0.06	-0.51	0.05	-0.44	-1**	-0.29	-0.75**	-0.06	-0.6	-0.17	7.37***	–

Table 1.4: Difference in variance test (Ledoit and Wolf, 2011) for different candidate portfolios. Split Date: 2015-12-31. Every cell computes the difference in standard deviation from the column portfolio minus the standard deviation of the row portfolio. Figures are reported in annualized terms i.e. multiplying by $\sqrt{252}$). The symbols *, **, and *** indicate that results are significant at the 10, 5 and 1%, respectively.

	DCC-SC	Pro-DCC-SC	DCC-SC($\gamma = \gamma_t$)	DCC-SC($\gamma = 1$)	Pro-DCC-SC($\gamma = 1$)	DCC-SC($\gamma = 2$)	Pro-DCC-SC($\gamma = 2$)	DCC-NLS	Pro-DCC-NLS	RiskMetrics	RollCov	1/N
DCC-SC	–	-0.33	-0.03**	-0.56*	-0.98***	-0.23***	-0.56***	-0.09***	-0.42**	-0.34	3.51***	-0.31
Pro-DCC-SC	0.33	–	0.29	-0.23	-0.66**	0.09	-0.23***	0.24	-0.09***	-0.01	3.83***	0.01
DCC-SC($\gamma = \gamma_t$)	0.03**	-0.29	–	-0.53*	-0.95***	-0.2***	-0.52***	-0.06**	-0.39*	-0.31	3.54***	-0.28
DCC-SC($\gamma = 1$)	0.56*	0.23	0.53*	–	-0.42***	0.33	0.01	0.47*	0.14	0.22	4.07***	0.25
Pro-DCC-SC($\gamma = 1$)	0.98***	0.66**	0.95***	0.42***	–	0.75***	0.43*	0.89***	0.56**	0.65	4.49***	0.67
DCC-SC($\gamma = 2$)	0.23***	-0.09	0.2***	-0.33	-0.75***	–	-0.32	0.14**	-0.19	-0.11	3.74***	-0.08
Pro-DCC-SC($\gamma = 2$)	0.56***	0.23***	0.52***	-0.01	-0.43*	0.32	–	0.46**	0.13**	0.22	4.06***	0.24
DCC-NLS	0.09***	-0.24	0.06**	-0.47*	-0.89***	-0.14**	-0.46**	–	-0.33*	-0.25	3.6***	-0.22
Pro-DCC-NLS	0.42**	0.09***	0.39*	-0.14	-0.56**	0.19	-0.13**	0.33*	–	0.08	3.93***	0.11
RiskMetrics	0.34	0.01	0.31	-0.22	-0.65	0.11	-0.22	0.25	-0.08	–	3.85***	0.03
RollCov	-3.51***	-3.83***	-3.54***	-4.07***	-4.49***	-3.74***	-4.06***	-3.6***	-3.93***	-3.85***	–	-3.82***
1/N	0.31	-0.01	0.28	-0.25	-0.67	0.08	-0.24	0.22	-0.11	-0.03	3.82***	–

Table 1.5: Difference in variance test (Ledoit and Wolf, 2011) for different candidate portfolios. Split Date: 2016-12-31. Every cell computes the difference in standard deviation from the column portfolio minus the standard deviation of the row portfolio. Figures are reported in annualized terms i.e. multiplying by $\sqrt{252}$). The symbols *, **, and *** indicate that results are significant at the 10, 5 and 1%, respectively.

	DCC-SC	Pro-DCC-SC	DCC-SC($\gamma = \gamma_t$)	DCC-SC($\gamma = 1$)	Pro-DCC-SC($\gamma = 1$)	DCC-SC($\gamma = 2$)	Pro-DCC-SC($\gamma = 2$)	DCC-NLS	Pro-DCC-NLS	RiskMetrics	RollCov	1/N
DCC-SC	–	-0.23*	-0.05**	-0.92***	-1.23***	-0.24***	-0.5***	-0.09***	-0.32***	-0.16	4.2***	0.07
Pro-DCC-SC	0.23*	–	0.18	-0.69*	-1***	-0.02	-0.27***	0.14	-0.09***	0.07	4.43***	0.3
DCC-SC($\gamma = \gamma_t$)	0.05**	-0.18	–	-0.87**	-1.18***	-0.2***	-0.45***	-0.04	-0.27**	-0.11	4.25***	0.12
DCC-SC($\gamma = 1$)	0.92***	0.69*	0.87**	–	-0.31***	0.67**	0.42	0.83**	0.59*	0.76	5.12***	0.99*
Pro-DCC-SC($\gamma = 1$)	1.23***	1***	1.18***	0.31***	–	0.98***	0.73**	1.14***	0.91***	1.07*	5.43***	1.3**
DCC-SC($\gamma = 2$)	0.24***	0.02	0.2***	-0.67**	-0.98***	–	-0.25*	0.15**	-0.08	0.09	4.45***	0.32
Pro-DCC-SC($\gamma = 2$)	0.5***	0.27***	0.45***	-0.42	-0.73**	0.25*	–	0.41***	0.18***	0.34	4.7***	0.57
DCC-NLS	0.09***	-0.14	0.04	-0.83**	-1.14***	-0.15**	-0.41***	–	-0.23*	-0.07	4.29***	0.16
Pro-DCC-NLS	0.32***	0.09***	0.27**	-0.59*	-0.91***	0.08	-0.18***	0.23*	–	0.16	4.53***	0.4
RiskMetrics	0.16	-0.07	0.11	-0.76	-1.07*	-0.09	-0.34	0.07	-0.16	–	4.36***	0.23
RollCov	-4.2***	-4.43***	-4.25***	-5.12***	-5.43***	-4.45***	-4.7***	-4.29***	-4.53***	-4.36***	–	-4.13***
1/N	-0.07	-0.3	-0.12	-0.99*	-1.3**	-0.32	-0.57	-0.16	-0.4	-0.23	4.13***	–

Table 1.6: Difference in variance test (Ledoit and Wolf, 2011) for different candidate portfolios. Split Date: 2017-12-31. Every cell computes the difference in standard deviation from the column portfolio minus the standard deviation of the row portfolio. Figures are reported in annualized terms i.e. multiplying by $\sqrt{252}$). The symbols *, **, and *** indicate that results are significant at the 10, 5 and 1%, respectively.

1.8 Figures

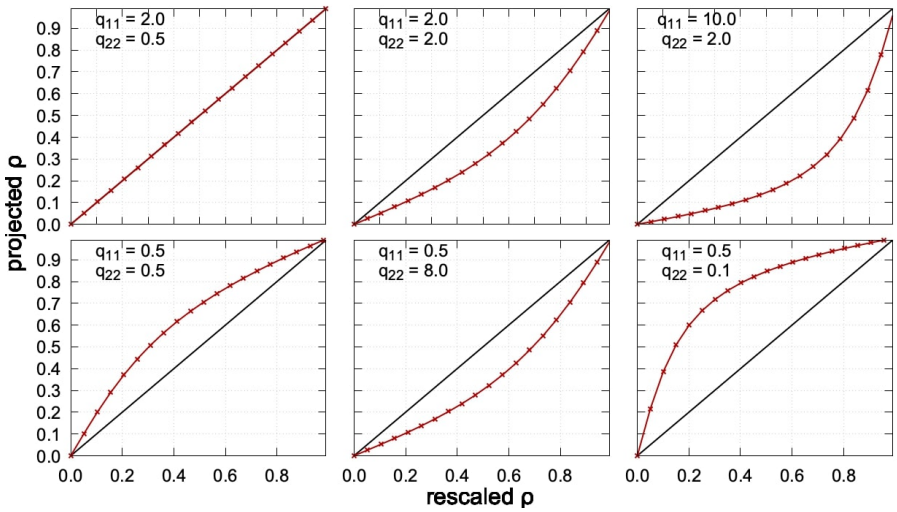


Figure 1.1: Projecting versus rescaling a pseudo-correlation matrix. Red: correlation computed using Stein's projection – see equation (1.7) – as a function of the rescaled correlation $q_{12}/\sqrt{q_{11}q_{22}}$. Black: 45 degree line.

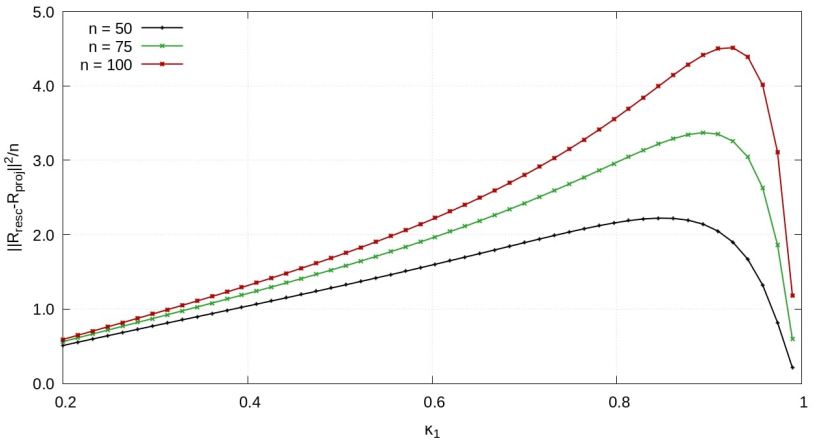


Figure 1.2: Projecting versus rescaling an n-dimensional pseudo-correlation matrix. The y-axis represents $||\mathbf{R}_{resc} - \mathbf{R}_{proj}||_2^2/n = \mathrm{tr}((\mathbf{R}_{resc} - \mathbf{R}_{proj})^2)/n$, where $\mathbf{R}_{resc} = \mathrm{diag}(\mathbf{Q})^{-1/2}\mathbf{Q}\,\mathrm{diag}(\mathbf{Q})^{-1/2}$, and $\mathbf{R}_{proj} = P_{STEIN}(\mathbf{Q})$. The pseudo-correlation matrix is $\mathbf{Q} = \mathrm{Diag}(1+\xi+v)^{1/2}\mathbf{R}(\kappa_1)\,\mathrm{Diag}(1+\xi+v)^{1/2}$, with $\kappa_1 \in (0,1)$, $\xi = 0.05$, v is an $n \times 1$ vector with i^{th} entry given by $v_i = \sin(x_i)$, where $x_{i+1} = x_i + 2(\pi-\varepsilon)/(n-1), x_1 = \varepsilon, x_n = 2\pi - \varepsilon$. The notation $\mathbf{R}(\kappa_1)$ is used to denote a Toeplitz correlation matrix with parameter κ_1, i.e. with first row/column equal to $1, \kappa_1, \kappa_1^2, \ldots, \kappa_1^{n-1}$.

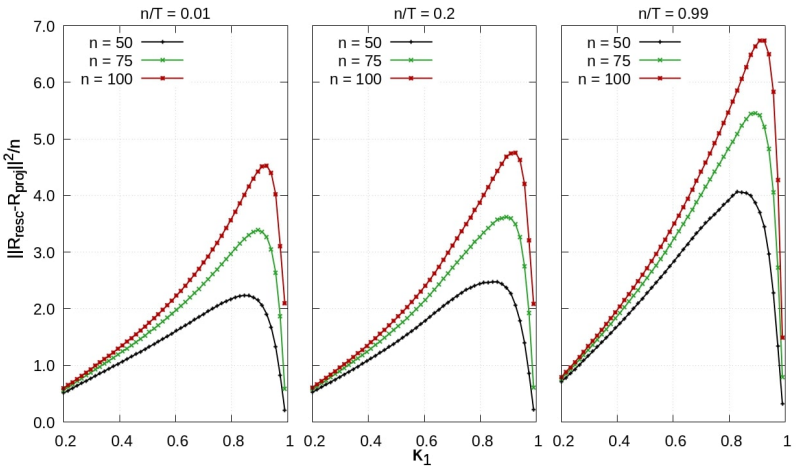


Figure 1.3: Projecting versus shrinking and rescaling an n-dimensional pseudo-correlation matrix. The y-axis represents $||\mathbf{R}_{s-r} - \mathbf{R}_{proj}||_2^2/n = \text{tr}((\mathbf{R}_{s-r} - \mathbf{R}_{proj})^2)/n$, where $\mathbf{R}_{s-r} = \text{diag}(\tilde{\mathbf{Q}})^{-1/2}\tilde{\mathbf{Q}}\,\text{diag}(\tilde{\mathbf{Q}})^{-1/2}$, $\mathbf{R}_{proj} = P_{STEIN}(\mathbf{Q})$, and $\tilde{\mathbf{Q}}$ is obtained after applying the nonlinear shrinkage estimator of Ledoit and Wolf (2020) to $\mathbf{Q}$ with bandwidth $h = T^{-1/3}$, where T is chosen to match the desired concentration ratio n/T. The pseudo-correlation matrix is $\mathbf{Q} = \text{Diag}(1+\xi+v)^{1/2}\mathbf{R}(\kappa_1)\,\text{Diag}(1+\xi+v)^{1/2}$, with $\kappa_1 \in (0,1)$, $\xi = 0.05$, v is an $n \times 1$ vector with i^{th} entry given by $v_i = \sin(x_i)$, where $x_{i+1} = x_i + 2(\pi-\varepsilon)/(n-1), x_1 = \varepsilon, x_n = 2\pi - \varepsilon$. The notation $\mathbf{R}(\kappa_1)$ is used to denote a Toeplitz correlation matrix with parameter κ_1, i.e. with first row/column equal to $1, \kappa_1, \kappa_1^2, \ldots, \kappa_1^{n-1}$.

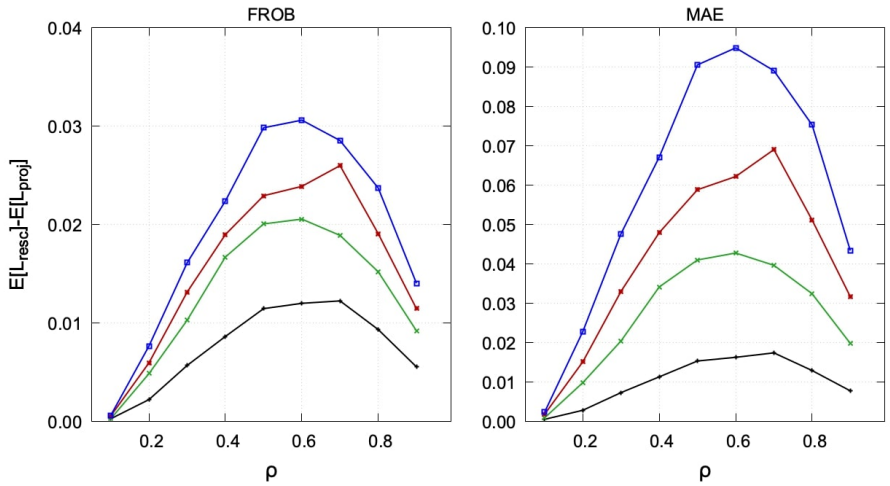


Figure 1.4: Static simulation study: Sample correlation versus Stein's projection of the sample covariance matrix onto the correlation set. The y-axis shows $\mathbb{E}[\mathcal{L}_{Frob,resc}] - \mathbb{E}[\mathcal{L}_{Frob,proj}]$ and $\mathbb{E}[\mathcal{L}_{MAE,resc}] - \mathbb{E}[\mathcal{L}_{MAE,proj}]$. For each Monte Carlo replication, we draw $T = 500$ observations from $\mathcal{N}(0, \mathbf{R})$, where $\mathbf{R} = (1-\rho)\mathbf{I} + \rho\iota\iota'$. Black, green, red and blue lines correspond to $n = 10, 23, 36, 50$, respectively.

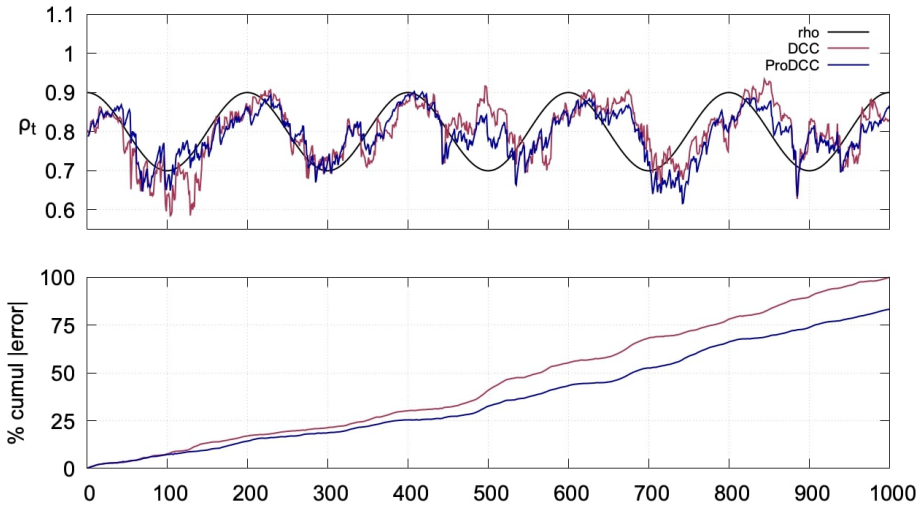


Figure 1.5: Dynamic simulation study: Top panel: Dynamic correlations for the true DGP (black), versus the corresponding estimates from DCC (red) and Pro-DCC (blue), where the true DGP is given by $\rho_t = \rho + .1\cos(2\pi t/200)$. Bottom panel: cumulative sum of absolute error for each methodology (same color legend), divided by the sum of absolute errors from the DCC model times 100.

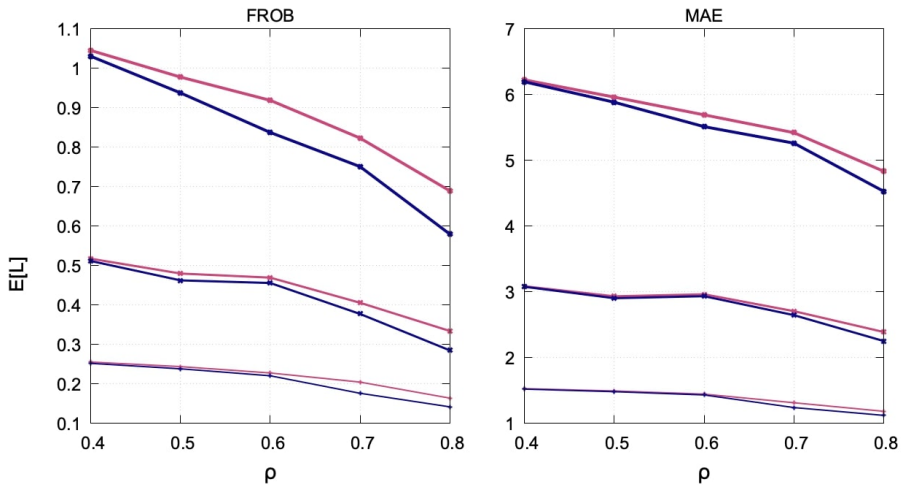


Figure 1.6: Dynamic simulation study: DCC's rescaling (red) versus Pro-DCC's projection (blue) based on Stein's loss for dimensions $n = 50$, (bottom pair), $n = 100$ (middle pair) and $n = 200$ (top pair). The y-axis shows the average Frobenius (left panel) and MAE (left panel) losses across Monte Carlo simulations, where each loss is computed as the average loss across $t = 1, \ldots, T$. The x-axis shows the magnitude of the parameter ρ, which captures the level of cross-sectional dependence of the process.

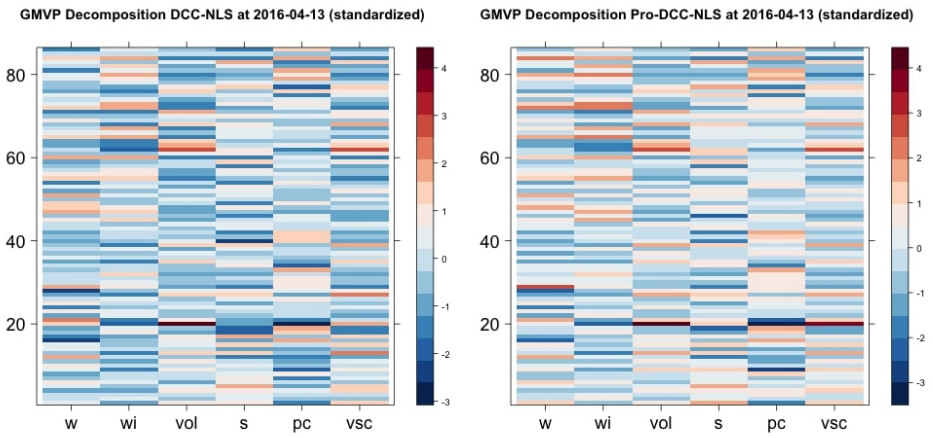


Figure 1.7: Heatmap of GMVP at 2016-04-13. Column w shows the weights implied by the model, which are contrasted to the weights under the assumption that $\mathbf{\Sigma}_t$ is diagonal (second column). The column vol shows the conditional volatilities for every asset. The column s is the vector that stacks the spillover effects $s_{i,t} := 1/\sqrt{\mathbf{\Omega}_{ii,t}}$. The column pc shows the partial correlation effects, and for every asset i they are computed as $\sum_{j \neq i} \boldsymbol{\varrho}_{ij,t}$. Finally, the column vsc contains the reciprocal of the v_t vector, so $vsc_{i,t} = v_{i,t}^{-1}$. For the purposes of visualization, all columns have been standardized.

Chapter 2

EMPIRICAL RISK MINIMIZATION FOR TIME SERIES: NONPARAMETRIC PERFORMANCE BOUNDS FOR PREDICTION

2.1 Introduction

Empirical risk minimization is a standard principle for choosing algorithms in learning theory (Vapnik and Chervonenkis, 1971; Devroye *et al.*, 1996). Simply put, empirical risk minimization consists of choosing the algorithm that minimizes

the empirical risk, that is, the average "in-sample" predictive loss. One of the goals of learning theory is to establish bounds on the expected "out-of-sample" predictive loss of the algorithm that minimizes the empirical risk relative to the optimal performance attainable in a given class of algorithms. A key feature of learning theory is its nonparametric nature, in the sense that performance bounds are typically obtained under the assumption that the generating mechanism of the data is unknown. Despite the fact that empirical risk minimization is a general principle and widely applicable, the majority of contributions in this area focus on the analysis of i.i.d. data.

In this chapter we study empirical risk minimization for time series. Our analysis is carried out in a general framework that allows to study different types of forecasting applications. We are concerned with 1-step-ahead prediction of a univariate time series belonging to a class of location-scale parameter-driven processes (Cox, 1981). The class of processes we entertain is fairly broad and it includes linear state space and stochastic volatility models. A class of recursive algorithms is available to predict the time series. The algorithms are recursive in the sense that the forecast produced in a given period is a function of the lagged values of the forecast and the time series. The class we consider is inspired by threshold models (Tong, 1990) and it includes as special cases the prediction formulae/filters of ARMA and GARCH models (as well as other specifications popular in the financial econometrics literature). The prediction accuracy of the forecasts is measured by a loss function in the Bregman class (Bregman, 1967; Banerjee *et al.*, 2005b; Laurent *et al.*, 2013; Patton, 2020), which includes the

loss functions typically used for the estimation of ARMA and GARCH models (as well as other specifications popular in the financial econometrics literature). Our analysis is nonparametric in the sense that the relationship between the data generating mechanism of the time series and the class of algorithms is not specified.

The main result of this chapter consists of establishing an oracle inequality that provides finite-sample guarantees on the predictive performance of empirical risk minimization. The oracle inequality implies that empirical risk minimization is consistent, in the sense that the algorithm chosen by empirical risk minimization achieves asymptotically the optimal predictive performance that can be attained within the class considered.

Our paper contributes to the large literature on forecasting and nonlinear time series modeling. There are two main novel contributions that are worth emphasizing. First, our main theorem implies that the algorithm chosen by empirical risk minimization achieves optimal predictive performance irrespective of whether the class of algorithms contains the optimal forecast for the target time series of interest. This is in contrast with quasi-maximum likelihood estimation theory of ARMA/GARCH models (Ling and McAleer, 2003; Francq and Zakoïan, 2004; Straumann and Mikosch, 2006) that establishes results that are analogous to the one established here but requires correct specification of the conditional mean/variance equation. Second, our main theorem provides the rate at which the performance of the algorithm chosen by empirical risk minimization converges to the optimal performance (as a function of the number of "in-sample" obser-

vations). Importantly this rate holds in finite samples as opposed to being only asymptotically valid. Moreover, this rate is optimal in the sense that our theorem recovers what is known as the classic rate of convergence of empirical risk minimization (Devroye *et al.*, 1996). These two contributions significantly weaken the conditions required for practitioners to apply optimal time series forecasting methods.

A number of illustrations showcase that our framework covers a range of applications commonly encountered in time series analysis. We consider the problem of forecasting data generated by an "AR(1) + noise" process, realized volatilities generated by a stochastic volatility process and durations generated by a stochastic volatility duration process (Ghysels *et al.*, 2004). In these applications, provided that appropriate regularity conditions are satisfied, we have that our main theorem holds, implying that the algorithm chosen by empirical risk minimization achieves optimal predictive performance.

The main result follows from five intermediate propositions. First, we establish existence of moments and strong mixing conditions of a joint process that includes the time series and the algorithm. Importantly, the strong mixing coefficients are bounded by a function with geometric decay uniformly over the class of algorithms. Second, we establish a general inequality that states that the performance of empirical risk minimization can be controlled by the sum of two quantities. The first is the supremum of an average of differences between conditional and unconditional expectations and the second is the supremum of the empirical process associated with the prediction loss of the algorithm. Finally, we

bound these two terms using, respectively, an inequality from Ibragimov and a concentration inequality for strong mixing processes.

The proposition that establishes existence of moments and strong mixing conditions of the joint system that includes the time series and the algorithm contains the main novel idea of this chapter. The result builds upon the literature on nonlinear time series models and Markov chains (Bougerol and Picard, 1992; Lanne and Saikkonen, 2005; Francq and Zakoïan, 2006; Meitz and Saikkonen, 2008a; Kristensen, 2009). The novelty with respect to the literature consists in using Markov chain theory to establish moment and dependence properties of an *algorithm*, as opposed to a *model*. More precisely, the strategy consists of embedding the time series and the algorithm in what we name a companion Markov chain. We then show that the companion Markov chain is V-geometric ergodic, which implies existence of moments and strong mixing of the time series and the algorithm (Meyn and Tweedie, 1993). The uniform bound on the strong mixing coefficients is established using results by Roberts and Rosenthal (2004). This approach is motivated by the fact that while it can be challenging to characterize the moment and dependence properties of general nonlinear processes, a number of tools are available to establish these properties for Markov nonlinear processes (Carrasco and Chen, 2002). Importantly, the result does not hinge on the approximation properties of the class of algorithms.

This chapter contributes to the literature on empirical risk minimization for dependent data. Besides a number of notable contributions, this literature is not

extensive.[1] Two closely related contributions are Jiang and Tanner (2010) and Brownlees and Guðmundsson (2021), which study empirical risk minimization for regression. The class of algorithms considered in these papers depends on a finite number of lags of the time series. In such a setting it is typically straightforward to obtain the dependence properties of the joint system composed of the times series and the algorithm. The strategy adopted in both papers consists of assuming that the time series is strong mixing and then applying standard results for functions of strong mixing processes to obtain that the joint process is also strong mixing. Such an approach is not viable in the framework of this chapter. In our setup forecasts depend on the entire past history of the time series. In this case standard results for functions of strong mixing processes do not provide useful results. Last, this chapter is related to McDonald *et al.* (2017), which studies risk properties of prediction algorithms for time series. However, that paper does not establish oracle inequalities for empirical risk minimization.

It is also important to remark that empirical risk minimization for time series is related to M-estimation for dependent data (Gallant and White, 1988; Pötscher and Prucha, 1997). There are two main differences with respect to this literature. First, the empirical risk minimization literature focuses on establishing a different set of properties (i.e. finite-sample prediction performance guarantees) in comparison to what is established in the M-estimation literature (i.e. consis-

[1] Nontrivial technical challenges arise with dependent data. Mendelson (2015) argues that some of the standard tools used in learning theory cannot be applied beyond an i.i.d. and bounded data setup.

tency and asymptotic normality). Second, the M-estimation literature for dependent data typically focuses on developing general theory on the basis of high-level assumptions. For instance, Gallant and White (1988) rely, among other requirements, on uniform NED and dominance conditions on the objective function of M-estimation. We remark that checking that these conditions hold is not always straightforward. On the contrary, in this chapter we rely on primitive assumptions to establish that conditions akin uniform NED and dominance hold.

The rest of the chapter is structured as follows. Section 2.2 introduces the framework. Section 2.3 presents empirical risk minimization and the main result. Section 2.4 discusses a number of issues related to our framework. Section 2.5 contains applications. Section 2.6 contains a simulation study. Section 2.7 outlines the proof of the main result. Concluding remarks follow in Section 2.8. Detailed proofs are in sections 2.9-2.14.

2.2 Basic Definitions and Assumptions

Data generating process. We are concerned with 1-step-ahead prediction of a time series $\{Y_t, t \geqslant 0\}$ belonging to a family of parameter-driven processes (Cox, 1981). The process $\{Y_t, t \geqslant 0\}$ takes values in $\mathcal{Y} \subseteq \mathbb{R}$ and is defined as $Y_0 = y \in \mathcal{Y}$, and

$$Y_t \quad = \quad g_{y1}(H_t) + g_{y2}(H_t)\epsilon_{Y\,t}\,, \quad t \geqslant 1\,, \tag{2.1}$$

where $\{H_t, t \geqslant 0\}$ is a hidden process, $\{\epsilon_{Y\,t}, t \geqslant 1\}$ is an i.i.d. sequence of random variables and g_{y1} and g_{y2} are Borel-measurable real functions. The process

$\{H_t, t \geqslant 0\}$ takes values in $\mathcal{H} = \text{int}(\mathcal{Y})$ and is defined as $H_0 = h \in \mathcal{H}$, and

$$H_t \;\; = \;\; g_{h1}(H_{t-1}) + g_{h2}(H_{t-1})\epsilon_{H\,t}\,, \quad t \geqslant 1\,, \tag{2.2}$$

where $\{\epsilon_{H\,t}, t \geqslant 1\}$ is an i.i.d. sequence of random variables and g_{h1} and g_{h2} are Borel-measurable real functions. We remark that in our framework, depending on the application, the target time series $\{Y_t, t \geqslant 0\}$ may denote some appropriate transformation of the data. For example, in volatility forecasting using stock returns, where interest lies in predicting the 1-step-ahead scale of stock returns, the time series $\{Y_t, t \geqslant 0\}$ may be defined as the squared return process.

The data generating process satisfies the following set of assumptions.

A.2.2.1 (Data generating process)**.** *(i) The functions g_{h1} and g_{h2} are bounded on bounded subsets of $\mathbb{R}$. There exist positive constants a and b such that $|g_{h1}(h)| \leqslant a|h| + o(|h|)$ as $|h| \to \infty$ and $|g_{h2}(h)| \leqslant b|h| + o(|h|)$ as $|h| \to \infty$. The function g_{h2} satisfies $\inf_h |g_{h2}(h)| > 0$.*

(ii) The functions g_{y1} and g_{y2} are bounded on bounded subsets of $\mathbb{R}$. There exist positive constants C_{y1} and C_{y2} such that $|g_{y1}(h)| \leqslant C_{y1}|h|$ and $|g_{y2}(h)| \leqslant C_{y2}(1 \vee |h|)$. The function g_{y1} satisfies $\inf_h g_{y1}(h) \geqslant 0$ when $\mathcal{Y} = \mathbb{R}_+$. The function g_{y2} satisfies $\inf_h g_{y2}(h) > 0$.

(iii) The random process $\{(\epsilon_{Y\,t-1}, \epsilon_{H\,t})', t \geqslant 0\}$ is i.i.d. and $(\epsilon_{Y\,t-1}, \epsilon_{H\,t})'$ has a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$ and is supported on $(\underline{\epsilon}, \infty)^2$ with $\underline{\epsilon} = -\infty$ when $\mathcal{Y} = \mathbb{R}$ and $\underline{\epsilon} = 0$ when $\mathcal{Y} = \mathbb{R}_+$. The joint density ϕ of the random vector $(\epsilon_{Y\,t-1}, \epsilon_{H\,t})'$ satisfies $\phi(\epsilon_{Y\,t-1}, \epsilon_{H\,t}) = \phi_Y(\epsilon_{Y\,t-1})\phi_H(\epsilon_{H\,t})$, where ϕ_Y and ϕ_H are densities that are

bounded away from zero on compact subsets of $(\underline{\epsilon}, \infty)$. *The random variables* $\epsilon_{Y\,t}$ *and* $\epsilon_{H\,t}$ *satisfy* $\mathbb{E}\epsilon_{H\,t}^{2r_m} < \infty$, $\mathbb{E}\epsilon_{Y\,t}^{2r_m} < \infty$ *for some* $r_m \geqslant 6$. *The random variable* $\epsilon_{Y\,t}$ *satisfies* $\mathbb{E}\left(\log \epsilon_{Y\,t}\right)^{2r_m} < \infty$ *when* $\mathcal{Y} = \mathbb{R}_+$.

(iv) The condition $\mathbb{E}(a + b|\epsilon_{H\,t}|)^{2r_m} < 1$ *holds.*

A.2.2.1 is similar to standard assumptions used to establish geometric ergodicity of nonlinear time series models (Masry and Tjøstheim, 1995; Lanne and Saikkonen, 2005; Meitz and Saikkonen, 2008a) and it allows for a fairly broad class of parameter-driven processes. Note that the $\{Y_t, t \geqslant 0\}$ process can take values on either $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = \mathbb{R}_+$ (assumptions differ slightly depending on the case). This allows us to cover different types of forecasting applications encountered in the literature. A.2.2.1(i) is similar to Assumption 3.2 in Masry and Tjøstheim (1995) and it implies that (2.2) 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. A.2.2.1(ii) allows for a fair amount of flexibility in equation (2.1). In particular, it requires $|Y_t|$ to be bounded from above by a linear function of $|H_t|$. A.2.2.1(iii) imposes conditions on the innovation processes. While this assumption does not impose these innovations to have bounded support or exponential tails (as it is customary in learning theory) this assumption requires a relatively large number of moments for the innovations to exist. In particular, this assumption rules out heavier tailed distributions that may be reasonable for financial applications. When $\mathcal{Y} = \mathbb{R}_+$ we additionally require that $2r_m$ moments of $\log \epsilon_{Y\,t}$ exist. This guarantees that the moments of some of the loss functions con-

sidered in this chapter are finite. The assumption is relatively mild as it allows, for example, for distributions with density bounded from above in a neighborhood of zero (e.g. the exponential) as well as certain distributions with unbounded density (e.g. chi-square with one degree of freedom). We remark that similar conditions are imposed in the analysis of Log-GARCH models (Francq *et al.*, 2013). Finally, the assumption states that the innovation processes are jointly i.i.d.. Again, this assumption is somehow restrictive for some financial applications as it rules out, in a stochastic volatility framework, the presence of leverage effects. A.2.2.1(iv) is a stability condition similar to the one assumed in Masry and Tjøstheim (1995) or Lanne and Saikkonen (2005).

Algorithms. A class of recursive algorithms indexed by $\theta \in \Theta \subset \mathbb{R}^p$ and denoted by $\{f_{\theta\, t}, t \geqslant 0\}$ is available to predict 1-step-ahead the time series $\{Y_t, t \geqslant 0\}$. The process $\{f_{\theta\, t}, t \geqslant 0\}$ takes values in $\mathcal{F} \subseteq \mathbb{R}$ and is defined as $f_{\theta\, 0} = f \in \mathcal{F}$ and

$$f_{\theta\, t} = \sum_{k=1}^{K} (\alpha_{0\, k} + \alpha_{1\, k} Y_{t-1} + \beta_{1\, k} f_{\theta\, t-1}) \mathbb{1}_{t-1\, k}\ , \quad t \geqslant 1\ , \tag{2.3}$$

where $\theta = (\alpha_{0\,1}, \ldots, \alpha_{0\, K}, \alpha_{1\,1}, \ldots, \alpha_{1\, K}, \beta_{1\,1}, \ldots, \beta_{1\, K})'$ with $K = p/3$, $\mathbb{1}_{t\, k} = \mathbb{1}_{\{Y_t \in \mathcal{Y}_k\}}$ and $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_K\}$ is a known partition of $\mathcal{Y}$ made of K sets referred to as regimes. The partition is of the form $\{(r_1, r_2), [r_2, r_3), \ldots, [r_K, \infty)\}$ with $-\infty = r_1 < r_2 < \cdots < r_K < \infty$ when $\mathcal{Y} = \mathbb{R}$ and $\{[r_1, r_2), [r_2, r_3), \ldots, [r_K, \infty)\}$ with $0 = r_1 < r_2 < \cdots < r_K < \infty$ when $\mathcal{Y} = \mathbb{R}_+$. The parameter vector θ is referred to as a prediction rule. We remark that the class of prediction algorithms in (2.3) corresponds to the class of 1-step-ahead prediction formulae induced by the self-exciting threshold autoregressive moving average model (SETARMA) (Tong,

1990). It is also interesting to point out that the class of prediction algorithms in (2.3) has analogies with regression trees. Like regression trees the prediction rules in (2.3) rely on partitions to capture nonlinearities in the data and possibly enhance predictive ability. As it is customary in the learning literature, we emphasize that the relationship between Y_t and $f_{\theta t}$ is not specified and (2.3) is simply an algorithm to predict Y_t. Last, in section 2.11 we show that such a class of algorithms may be interpreted as the solution of a sequential optimization problem in which the forecaster aims at minimizing a forecast tracking error accuracy measure.

The class of algorithms satisfies the following set of assumptions.

A.2.2.2 (Algorithms)**.** (i) *The set* $\Theta \subset \mathbb{R}^p$ *with* $p = 3K$ *is nonempty and such that* $\Theta \subseteq [\underline{\alpha}_0, \overline{\alpha}_0]^K \times [\underline{\alpha}_1, \overline{\alpha}_1]^K \times [0, \overline{\beta}_1]^K$ *with* $\underline{\epsilon} < \underline{\alpha}_0 < \overline{\alpha}_0 < \infty$, $0 < \underline{\alpha}_1 < \overline{\alpha}_1 < \infty$ *and* $\overline{\beta}_1 < 1$. (ii) *The number of regimes* K *satisfies* $K < (r_m - 2)/3$.

The process $\{f_{\theta t}, t \geqslant 0\}$ takes values in $\mathcal{F} = \mathbb{R}$ when $\mathcal{Y} = \mathbb{R}$ and $\mathcal{F} = [\underline{\alpha}_0, \infty)$ when $\mathcal{Y} = \mathbb{R}_+$. A.2.2.2$(i)$ is mild and imposes constraints on the class of prediction rules Θ that are analogous to standard constraints imposed in the analysis of quasi-maximum likelihood estimators of ARMA and GARCH models (Francq and Zakoïan, 2010). We remark that when $\mathcal{Y} = \mathbb{R}$ the constraint $\beta_{1\,k} \in [0, \overline{\beta}_1]$ may be relaxed to $\beta_{1\,k} \in [-\overline{\beta}_1, \overline{\beta}_1]$ at the expense of more tedious proofs. A.2.2.2(ii) states that the size of the class of prediction rules is bounded by a linear function of the number of moments of $\epsilon_{Y\,t}$ and $\epsilon_{H\,t}$. We remark that the higher is the dimensionality of Θ the higher the number of moments of the data that are required to exist in order to learn the optimal forecasting algorithm from the data,

and that in our setup the dimensionality of Θ is a linear function K. When $r_m = 6$ then we can only allow for $K = 1$ whereas when the tails of the innovations are sub-Gaussian then K can be arbitrarily large.

Loss function. The prediction accuracy of the algorithm is measured by a loss function that belongs to the Bregman class. Let $\psi : \mathcal{S} \to \mathbb{R}$ be a strictly convex and continuously differentiable function defined over a convex set $\mathcal{S} \subseteq \mathbb{R}$. Then, the Bregman loss associated with ψ for predicting Y_t with $f_{\theta\, t}$ is defined as

$$L(Y_t, f_{\theta\, t}) = \psi(Y_t) - \psi(f_{\theta\, t}) - \nabla\psi(f_{\theta\, t})(Y_t - f_{\theta\, t}) \,. \tag{2.4}$$

The Bregman class is a fairly large and tractable family of losses. In particular, the log-likelihood of random variables in the regular exponential family can be expressed as the (negative) sum of Bregman losses (up to a constant term) (Banerjee *et al.*, 2005b). Thus, the Bregman class includes the standard loss functions used for quasi-maximum likelihood estimation of time series models.

In this chapter we focus exclusively on losses that satisfy the following condition.

Condition 2.2.1 (Bregman)**.** *The loss L is such that (i) $Y_t \in \mathcal{S}$ a.s. for all $t \geqslant 0$, (ii) $\sup_{t\geqslant 1} \sup_{\theta\in\Theta} \mathbb{E}(L(Y_t, f_{\theta\, t}))^{r_m} < \infty$ and (iii) $L(f_{\theta_1\, t}, f_{\theta_2\, t}) \leqslant C_\psi(f_{\theta_1\, t} - f_{\theta_2\, t})^2$ a.s. for all $t \geqslant 1$, for any $\theta_1, \theta_2 \in \Theta$ and for some positive constant C_ψ.*

Table 2.2 contains a number of Bregman losses that satisfy Condition 2.2.1 given A.2.2.1 and A.2.2.2. We remark that when $\mathcal{Y} = \mathbb{R}$ only the first two losses are admissible whereas when $\mathcal{Y} = \mathbb{R}_+$ all the losses in the table are allowed.

The table contains both well known and lesser known loss functions.[2] The table includes the loss that corresponds to the log-likelihood of the Gaussian (with known variance) with respect to the mean parameter, which is the classic square loss. This loss function is typically used for maximum likelihood estimation of ARMA models. The table also contains the loss associated with the log-likelihood of the NEF-GHS (with known number of convolutions) with respect to the natural parameter (Morris, 1982), which to the best of our knowledge is not extensively used in the time series literature. Next, the table includes the loss associated with the log-likelihood of the gamma (with known shape) with respect to the mean parameter, which in the volatility forecasting literature is known as the QLIKE loss (Patton, 2011).[3] We recall that by appropriately constraining the shape parameter, the gamma distribution nests the exponential and chi-square distributions. This loss function is typically used for maximum likelihood estimation of MEM (Engle and Gallo, 2006), ACD models (Engle and Russell, 1998) and GARCH models. Finally, the table includes the losses associated with the log-likelihoods of the Poisson and negative binomial (with known number of failures) with respect to the mean parameter.[4] These loss functions are typically used for maximum likelihood estimation of dynamic models for count data (Agosto *et al.*, 2016). We remark that our framework does not allow for $\{Y_t, t \geqslant 0\}$ to take values on a

[2]The random variables listed in Table 2.2 are all the random variables in the natural exponential family with quadratic variance function and unbounded support (Morris, 1982).

[3]The standard definition of the QLIKE is $L(Y_t, f_{\theta\,t}) = Y_t/f_{\theta\,t} + \log f_{\theta\,t}$. This is equivalent to our definition for optimization purposes with respect to θ.

[4]We follow the convention that $0 \log 0 = 0$, hence $\text{dom}(\psi) = \mathbb{R}_+$ in both cases.

countable set. That said, these losses satisfy our regularity conditions and may be used for empirical risk minimization. The analysis of empirical risk minimization in this case can be carried out using the same strategy developed in this chapter, but some of the proofs would differ.

$\mathcal{S}$	$\psi(u)$	$L(u,v)$	Log-likelihood
$\mathbb{R}$	u^2	$(u-v)^2$	Gaussian
$\mathbb{R}$	$u\tan^{-1}(u) - \frac{1}{2}\log(1+u^2)$	$u\left[\tan^{-1}(u) - \tan^{-1}(v)\right] + \frac{1}{2}\log\frac{1+v^2}{1+u^2}$	NEF-GHS
$\mathbb{R}_{++}$	$-\log u$	$\frac{u}{v} - \log\frac{u}{v} - 1$	Gamma
$\mathbb{R}_{+}$	$u\log u - u$	$u\log\frac{u}{v} - (u-v)$	Poisson
$\mathbb{R}_{+}$	$u\log\frac{u}{1+u} - \log(1+u)$	$u\log\frac{u}{v} + (1+u)\log\frac{1+v}{1+u}$	Negative Binomial

Table 2.1: Regular Bregman losses that satisfy Condition 2.2.1 given A.2.2.1 and A.2.2.2. We recall that $\mathbb{R}_+$ is the set $\{x \in \mathbb{R} : x \geqslant 0\}$ and $\mathbb{R}_{++}$ is the set $\{x \in \mathbb{R} : x > 0\}$.

Dominating process. We introduce a dominating process $\{d_{\theta\, t}, t \geqslant 0\}$ that plays a key role in the theoretical analysis of this chapter. This process bounds the absolute difference between the forecast processes associated with two different prediction rules. The process $\{d_{\theta\, t}, t \geqslant 0\}$ takes values in $\mathcal{D} = [1, \infty)$ and is defined as $d_{\theta\, 0} = d \in \mathcal{D}$ and

$$d_{\theta\, t} = 1 + |Y_{t-1}| + |f_{\theta\, t-1}| + \overline{\beta}_1 d_{\theta\, t-1}\,, \quad t \geqslant 1\,. \tag{2.5}$$

As it is established in one of the intermediate results of this chapter, this process has the property that for any $\delta \in (0, 1]$ and for any $\theta, \dot{\theta} \in \Theta$ such that $\|\theta - \dot{\theta}\|_2 \leqslant \delta$ it holds that $|f_{\theta\, t} - f_{\dot{\theta}\, t}| \leqslant \delta d_{\dot{\theta}\, t}$ for all $t \geqslant 0$. This property and

the generalized triangular equality for Bregman losses imply that $L(Y_t, f_{\theta\, t}) \leqslant L(Y_t, f_{\dot{\theta}\, t}) + \delta C_\psi(d^2_{\dot{\theta}\, t} + 2|Y_t - f_{\dot{\theta}\, t}|d_{\dot{\theta}\, t})$ for $t \geqslant 1$.

2.3 Empirical Risk Minimization

We are interested in choosing a prediction rule θ from a sequence of "in-sample" observations $\{Y_1, \ldots, Y_T\}$ to forecast 1-step-ahead a sequence of "out-of-sample" observations $\{Y_{T+1}, \ldots .Y_{T+M}\}$. The number of out-of-sample observations is $M = \lceil \gamma T \rceil$ for some $\gamma > 0$. The accuracy of a prediction rule θ is measured by the out-of-sample 1-step-ahead conditional risk, which is defined as

$$R(\theta) = \mathbb{E}\left[\frac{1}{M} \sum_{t=T+1}^{T+M} L(Y_t, f_{\theta\, t}) \middle| Y_T, \ldots, Y_1 \right] . \tag{2.6}$$

A natural strategy for choosing a prediction rule θ consists of picking the one that minimizes the in-sample 1-step-ahead empirical risk. The empirical risk minimizer (ERM) is defined as

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} R_T(\theta) \ , \text{ where } R_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} L(Y_t, f_{\theta\, t}) \ . \tag{2.7}$$

If more than one prediction rule achieves the minimum we may pick one arbitrarily. In (2.6) and (2.7) we remark that $f_{\theta\, 1}$ is computed using $Y_0 = y$ and $f_{\theta\, 0} = f$ that are fixed, known and that do not depend on θ.[5]

One of the goals of learning theory is to establish a bound on the performance of the ERM relative to the optimal risk that can be achieved within the class of

[5]The initial value Y_0 can be a pre-sample observation assumed to be fixed or a fixed value set at the outset of the analysis. Note that when Y_0 is a pre-sample observation then the empirical risk in (2.7) can be thought of as the analog of the conditional log-likelihood of θ given Y_0.

prediction rules considered. We measure the accuracy of the ERM on the basis of the conditional out-of-sample risk, which is defined as

$$R(\hat{\theta}) = \mathbb{E}\left[\frac{1}{M} \sum_{t=T+1}^{T+M} L(Y_t, \hat{f}_t) \middle| Y_T, \ldots, Y_1 \right] , \tag{2.8}$$

where $\hat{f}_t = f_{\hat{\theta}\, t}$ and $f_0 \in \mathcal{F}$. The performance measure in (2.8) can be interpreted as the out-of-sample conditional risk of the ERM obtained from the in-sample observations. The following theorem establishes such a bound and is our main result.

Theorem 2.3.1. *Suppose A.2.2.1 and A.2.2.2 are satisfied. Then there exists a constant σ^2 such that, for all T sufficiently large, we have that*

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

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

The inequality in Theorem 2.3.1 is commonly referred to as an *oracle inequality*, and it provides finite-sample guarantees on the performance of the ERM.[6] The constant σ^2 is application-specific and may be interpreted as an upper bound for the long run variance of the loss process. We define the constant precisely in Proposition 2.7.5. The rate of convergence $\sqrt{\log T/T}$ is sometimes referred to

[6]We remark that it is possible to obtain explicit bounds for the minimum T and for the probability of the oracle inequality. Moreover, it is straightforward to see from the intermediate results of this chapter that it is possible to sharpen the rate of the probability upper bound of the oracle inequality as well as the absolute constant. However, we have not pursued this and we have solely focused on recovering the "classic" rate of convergence $\sqrt{\log T/T}$.

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). The theorem implies that in our framework the ERM is consistent with respect to the class of prediction rules Θ, meaning that $|R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta)| \xrightarrow{p} 0$. In other words, the ERM achieves asymptotically the optimal performance attainable within the class of algorithms considered. It is important to emphasize that Theorem 2.3.1 is stronger than a consistency result for the prediction performance of the ERM 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 ERM. Last, we emphasize that the theorem does not require the existence of an optimal prediction rule $\theta^* = \arg\min_{\theta \in \Theta} R(\theta)$.

2.4 Additional Discussion

Our analysis studies the properties of the ERM 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 ERM can be carried out using the same strategy developed in this chapter. However, some of the proofs will differ and we leave the analysis of this case for future research. In particular, we acknowledge that establishing some of the intermediate results required to establish Theorem 2.3.1 is more challenging in an observation-driven setup.

The class of recursive algorithms we entertain is fairly flexible and builds upon

the class of threshold models that have a well established tradition. We remark that our results may be extended to alternative classes of recursive algorithms and do not inherently depend on the functional form of the algorithmic class we consider in this chapter. In particular, our analysis does not require the class of algorithms to have special approximation properties or to include the optimal 1-step-ahead forecast. What is key in our framework is that, loosely speaking, the algorithms "forget the past sufficiently fast".

In many applications h-step-ahead predictions for $h > 1$ are of interest as well. A natural strategy for h-step-ahead "direct forecasting" is to modify the loss in (2.7) to be $L(Y_{t+h-1}, f_{\theta\, t})$. We conjecture that empirical risk minimization in this case is consistent and that this may be established using the proof strategy put forward here.

Instead of comparing the performance of the ERM against the optimal risk attainable in the class, one may wish to compare against the risk of the optimal 1-step-ahead forecast. For loss functions in the Bregman class the optimal 1-step-ahead forecast is the conditional mean (assuming it exists) (Banerjee *et al.*, 2005b). Thus, the risk of the optimal 1-step-ahead forecast may be defined as

$$R^* = \mathbb{E}\left[\frac{1}{M}\sum_{t=T+1}^{T+M} L(Y_t, \mu_t)\middle| Y_T, \ldots, Y_1\right],$$

where $\mu_t = \mathbb{E}(Y_t|Y_{t-1}, \ldots, Y_1)$ for $t > 1$. The performance of the ERM relative to the risk of the optimal 1-step-ahead foreast may be expressed as

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

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 data dependent 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 we do in this chapter.

Last, our main theorem establishes prediction performance guarantees for empirical risk minimization trained using the in-sample observations. In practice it is natural to update the empirical risk minimizer on the basis of newly available out-of-sample observations on the basis of either a rolling or a recursive training scheme. We remark that our theorem cannot be straightforwardly extended to establish bounds for these alternative training schemes and that we intend to explore this issue in future research.

2.5 Applications

Forecasting an AR(1) plus noise. Consider forecasting the AR(1) plus noise process

$$\begin{aligned} Y_t &= H_t + \epsilon_{Y\,t}\,, \\ H_t &= \mu_H + \varrho(H_{t-1} - \mu_H) + \epsilon_{H\,t}\,, \end{aligned}$$

for $t \geqslant 1$, where $\{\epsilon_{Y\,t}, t \geqslant 1\}$ and $\{\epsilon_{H\,t}, t \geqslant 1\}$ are a joint i.i.d. sequence of Gaussian random variables and $\varrho \in [0, 1)$. The class of algorithms defined in

(2.3) may be used for forecasting setting $K = 1$, which corresponds to the 1-step-ahead prediction formula of the ARMA(1,1). Prediction accuracy is measured by the square loss. Then, A.2.2.1 and A.2.2.2 are satisfied and Theorem 2.3.1 holds.

Note that in this application the approximation error converges to zero when T is large (when Θ is suitably chosen). In fact, the class of algorithms includes the steady state Kalman filter, implying that the class includes a forecast process that converges to the conditional mean of Y_t given its past as t grows large.

Forecasting Volatility. Consider forecasting realized volatilities (Andersen *et al.*, 2003) generated by the nonlinear stochastic volatility process[7]

$$\begin{aligned} RV_t &= \sigma_t^2 \epsilon_{RV\,t}\,, \\ \sigma_t^2 &= g_{h1}(\sigma_{t-1}^2) + g_{h2}(\sigma_{t-1}^2)\epsilon_{\sigma^2\,t}\,, \end{aligned}$$

for $t \geqslant 1$, where $\{\epsilon_{RV\,t}, t \geqslant 1\}$ and $\{\epsilon_{\sigma^2\,t}, t \geqslant 1\}$ are a joint i.i.d. sequence of gamma random variables and g_{y1} and g_{y2} are Borel-measurable real functions that satisfy A.2.2.1.(i). We assume that $\epsilon_{RV\,t}$ is unit mean, which implies that RV_t is a conditionally unbiased proxy for the volatility σ_t^2.[8] The class of algorithms defined in (2.3) may be used for forecasting setting $K = 1$, which corresponds to

[7]We remark that the model considered in this application should be interpreted as a reduced form approximation. The results obtained by Meddahi (2003) imply that a more flexible framework than the one considered here is required to allow for the continuous-time stochastic volatility models entertained in the realized volatility literature and we have not pursued this. Our framework is consistent with the continuous-time stochastic volatility model with constant intra-daily volatility used in Patton (2011).

[8]The realized volatility measurement error in the model is multiplicative. The analysis of this section can also be carried out in the case of an additive measurement error.

the 1-step-ahead prediction formula of the MEM(1,1) or ARMA(1,1). Prediction accuracy is measured by the QLIKE or the square loss. Then, A.2.2.1 and A.2.2.2 are satisfied and Theorem 2.3.1 holds.

Theorem 2.3.1 implies that the ERM achieves the optimal performance for realized volatility prediction. However, interest typically lies in forecasting the latent volatility process $\{\sigma_t^2, t \geqslant 0\}$ rather than its noisy measurement. Building upon (Hansen and Lunde, 2006; Patton, 2011) we establish further properties of the ERM. We measure the accuracy of a prediction rule θ for predicting the volatility process $\{\sigma_t^2, t \geqslant 0\}$ using the out-of-sample 1-step-ahead conditional risk

$$R_{\text{Vol}}(\theta) = \mathbb{E}\left[\frac{1}{M}\sum_{t=T+1}^{T+M} L(\sigma_t^2, f_{\theta\, t})\middle| RV_T, \ldots, RV_1\right] .$$

The loss in predicting the volatility process $\{\sigma_t^2, t \geqslant 0\}$ satisfies

$$L(\sigma_t^2, f_{\theta\, t}) = L(RV_t, f_{\theta\, t}) + L(\sigma_t^2, RV_t) - (\sigma_t^2 - RV_t)\nabla\psi(f_{\theta\, t}) + (\sigma_t^2 - RV_t)\nabla\psi(RV_t), \tag{2.9}$$

which follows from the generalized triangular equality for Bregman losses. In (2.9) we have that the second and fourth terms do not depend on the algorithm and the third term has a conditional expectation of zero given the past, because of the independence between of $\epsilon_{RV\, t}$ and $\epsilon_{\sigma^2\, t}$. The decomposition in (2.9) and Theorem 2.3.1 imply that

$$\left|R_{\text{Vol}}(\hat{\theta}) - \inf_{\theta\in\Theta} R_{\text{Vol}}(\theta)\right| = \left|R(\hat{\theta}) - \inf_{\theta\in\Theta} R(\theta)\right| \xrightarrow{p} 0 .$$

Thus, the ERM based on the noisy realized volatility measure chooses an algorithm with optimal performance for volatility forecasting within the class of algo-

rithms.

Forecasting Durations. Consider forecasting durations generated by a modified version of the stochastic volatility duration process of Ghysels *et al.* (2004)

$$d_t = -\frac{\log[1-\Phi(\eta_{1\,t})]}{c + aG(b, \Phi(F_t))}$$
$$F_t = \psi F_{t-1} + \sqrt{1-\psi^2}\,\eta_{2\,t}\,,$$

where $\Phi(\cdot)$ is the distribution of a standard normal and $G(b,\cdot)$ is the quantile function of the Gamma(b,b) distribution and $(a,c,\psi) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times [0,1)$ are parameters. We remark that the difference between the process defined above and the model of Ghysels *et al.* (2004) lies in the c parameter, which is equal to zero in that paper. It is straightforward to verify that this process can be cast as the data generating process in (2.1)–(2.2).[9] The class of algorithms defined in (2.3) may be used for forecasting setting $K = 1$, which corresponds to the 1-step-ahead prediction formula of the ACD(1,1). Prediction accuracy is measured by the Bregman loss associated with the gamma random variable. Then, A.2.2.1 and A.2.2.2 are satisfied and Theorem 2.3.1 holds.

2.6 Simulation Study

We carry out a simulation study to assess numerically the performance of empirical risk minimization. Consider the time series generated by the stochastic

[9]To see this simply note that $\epsilon_{Y\,t} = -\log[1-\Phi(\eta_{1\,t})]$, $\epsilon_{H\,t} = \exp(\sqrt{1-\psi^2}\;\eta_{2\,t})$, $H_t = \exp(F_{2\,t})$, $g_{h1}(h) = 0$, $g_{h2}(h) = \exp(\psi \log h) = h^{\psi}$, $g_{y1}(h) = 0$, and $g_{y2}(h) = 1/[c + aG(b, \Phi(\log(h)))]$.

volatility process

$$r_t = \sqrt{\sigma_t^2} z_t$$

$$\log \sigma_t^2 = -1.0 + 0.99(\log \sigma_{t-1}^2 + 1.0) + \eta_t \ ,$$

where $z_t \sim \mathcal{N}(0,1)$, $\eta_t \sim \mathcal{N}(0, 0.1)$ and $\log \sigma_0^2 = -1.0$. In section 2.14 it is verified that this data generating process satisfies A.2.2.1 for $Y_t = r_t^2$ and $H_t = \sigma_t^2$. We predict $Y_t = r_t^2$ using the 1-step-ahead GARCH(1,1) prediction rule

$$f_{\theta\, t} = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 f_{\theta\, t-1} \ ,$$

where $f_{\theta\, 0}$ is set at the sample variance. The QLIKE loss measures prediction accuracy.

We use simulations to compute a number of quantities related with this setup. First, we estimate the $(1 - 1/\log T)$-th quantile of the distribution of $R(\hat{\theta})$, which we denote RB_{sim}. This is obtained by simulating S paths $\{Y_1^{(s)}, \ldots, Y_{T+M}^{(s)}\}$ and then computing for each path

$$R^{(s)}(\hat{\theta}) = \frac{1}{M} \sum_{t=T+1}^{T+M} L(Y_t^{(s)}, f_{\hat{\theta}^{(s)}\, t}^{(s)}) \ ,$$

where $\hat{\theta}^{(s)}$ is the emprirical risk minimizer computed from the first T observations of the s-th path. RB_{sim} is then obtained as the $(1 - 1/\log T)$-th quantile of the $R^{(s)}(\hat{\theta})$ draws. Second, we use simulations to estimate $\inf_{\theta \in \Theta} \overline{R}(\theta)$. This is obtained by simulating S paths and then minimizing the average predictive loss across all paths, that is

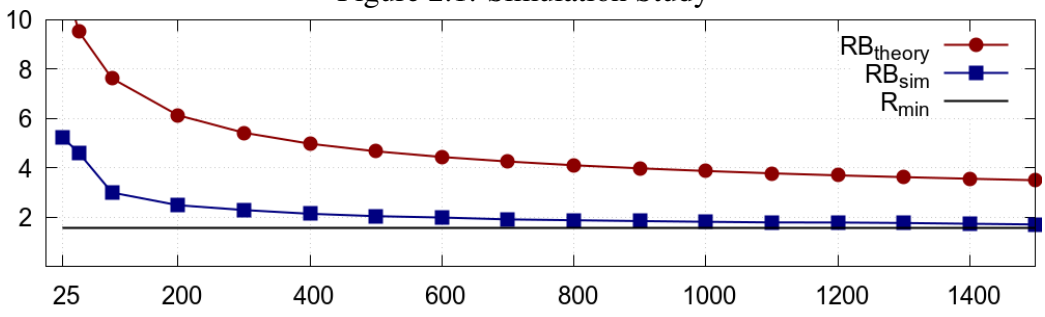
$$R_{min} = \min_{\theta \in \Theta} \left\{ \frac{1}{SM} \sum_{s=1}^{S} \sum_{t=T+1}^{T+M} L(Y_t^{(s)}, f_{\theta\, t}^{(s)}) \right\} \ . \tag{2.10}$$

Third, we compute a heuristic estimate of the risk bound implied by Theorem 2.3.1, which is given by

$$RB_{theory} = R_{min} + \hat{\sigma}\sqrt{\frac{3\log T}{T}}\,,$$

where $\hat{\sigma}$ is approximated by the long run variance of the loss process associated with the θ that minimizes the expression in (2.10). It is important to emphasize that this is a heuristic conservative choice of the value of the constant σ. We carry out this simulation study for different values of T ranging from 25 to 1500 and setting the value of M to 1000. The total number of replications is $S = 10000$. Figure 2.1 reports the results of the simulation exercise. The figure highlights a

Figure 2.1: Simulation Study



The figure reports (from top to bottom) the heuristic estimate of the risk bound of Theorem 2.3.1 (RB_{theory}), the estimate of the $(1 - 1/\log T)$-th quantile of the distribution of $R(\hat{\theta})$ (RB_{sim}) and the estimate of the $\inf_{\theta\in\Theta} \overline{R}(\theta)$ (R_{min}).

number of facts. First, the predictive performance of the empirical risk minimizer converges to the optimal predictive performance attainable in the class. In particular, for $T \geqslant 200$ the performance gap of the empirical risk minimizer is less than 20% relative to the optimal predictive performance attainable. Second, the heuristic risk bound based on Theorem 2.3.1 holds uniformly over all T considered,

albeit being fairly conservative.

2.7 Proof of Theorem 2.3.1

2.7.1 Companion Markov Chain

We begin by introducing a Markov chain associated with the process $\{(Y_t, f_{\theta\,t})', t \geqslant 0\}$. We recall a number of notions from Markov chain theory. Notation and definitions are based on Meyn and Tweedie (1993). The discrete-time process $\{X_t, t \geqslant 0\}$ is a time-homogeneous Markov chain with state space $\mathcal{X} \subseteq \mathbb{R}^d$ 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 $P_X^n(x, A) = \mathbb{P}(X_{t+n} \in A | X_t = x)$ for all $t \in \mathbb{Z}_+$. As customary, $P_X^1(x, A)$ is denoted by $P_X(x, A)$. We use $\pi_X : \mathcal{B}(\mathcal{X}) \to [0,1]$ to denote the invariant measure of the Markov chain (assuming it exists), that is, the probability measure such that for each $A \in \mathcal{B}(\mathcal{X})$ it holds that $\pi_X(A) = \int_{\mathcal{X}} \pi_X(dx) P_X(x, A)$.

Define the companion Markov chain $\{X_{\theta\,t}, t \geqslant 0\}$ that takes values in $\mathcal{X} = \mathcal{H} \times \mathcal{F} \times \mathcal{D}$ and is given by $X_{\theta\,0} = x = (h, f, d)' \in \mathcal{H} \times \mathcal{F} \times \mathcal{D}$ and $X_{\theta\,t} = (H_t, f_{\theta\,t}, d_{\theta\,t})'$, with

$$\begin{bmatrix} H_t \\ f_{\theta\,t} \\ d_{\theta\,t} \end{bmatrix} = \begin{bmatrix} g_{h1}(H_{t-1}) + g_{h2}(H_{t-1}) Z_{1\,t} \\ \sum_{k=1}^{K} \{\alpha_{0\,k} + \alpha_{1\,k} \left[g_{y1}(H_{t-1}) + g_{y2}(H_{t-1}) Z_{2\,t} \right] + \beta_{1\,k} f_{\theta\,t-1} \} \mathbb{1}_{t-1\,k} \\ 1 + |g_{y1}(H_{t-1}) + g_{y2}(H_{t-1}) Z_{2\,t}| + |f_{\theta\,t-1}| + \overline{\beta}_1 d_{\theta\,t-1} \end{bmatrix} \tag{2.11}$$

for $t \geqslant 1$, where $\mathbb{1}_{t-1\,k} = \mathbb{1}_{\{g_{y1}(H_{t-1})+g_{y2}(H_{t-1})Z_{2\,t}\in\mathcal{Y}_k\}}$, $Z_{1\,t} = \epsilon_{H\,t}$ and $Z_{2\,t} = \epsilon_{Y\,t-1}$. We are interested in establishing that the companion Markov chain $\{X_{\theta\,t}, t \geqslant 0\}$ is V_X-geometrically ergodic (Meyn and Tweedie, 1993; Meitz and Saikkonen, 2008a).

Definition 2.7.1 (V_X-geometric ergodicity)**.** *A Markov chain $\{X_t, t \geqslant 0\}$ 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*

$$\sup_{v:|v|\leqslant V_X} \left| \int_{\mathcal{X}} P_X^n(x, dx_n)v(x_n) - \int_{\mathcal{X}} \pi_X(dx_n)v(x_n) \right| \leqslant \rho^n M_x \,, \tag{2.12}$$

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

A number of remarks are in order. First, the definition implicitly assumes that the expectation of the function V_X with respect to the measure π_X exists. Second, a Markov chain that is V_X-geometric ergodic has convenient moment and dependence properties. If we choose $V_X = 1$ then we have that (2.12) coincides with the definition of geometric ergodicity, which implies β- and α-mixing. Last, V_X-geometric ergodicity implies that the unconditional expectation of $v(X)$ exists for any function v such that $|v| \leqslant V_X$.

The following lemma established V_X-geometric ergodicity of $\{X_{\theta\,t}, t \geqslant 0\}$.

Lemma 2.7.1. *Suppose A.2.2.1 and A.2.2.2 are satisfied. Then $\{X_{\theta\,t}, t \geqslant 0\}$ is V_X-geometrically ergodic with $V_X(x) = 1 + \|x\|_1^{2r_m}$.*

The proof of this lemma and all subsequent results is postponed to sections 2.9 and 2.10. The proof of Lemma 2.7.1 is based on establishing that the Markov

chain is irreducible, aperiodic and satisfies the so-called drift criterion. The claim then follows from Theorem 15.0.1 of Meyn and Tweedie (1993), which is a classic result that is routinely employed to establish stability of nonlinear time series models.

The following lemma establishes that the constants ρ and M_x in Definition 2.7.1 can be chosen independently of θ in the case of geometric ergodicity (that is, when $V_X = 1$).[10]

Lemma 2.7.2. *Suppose A.2.2.1 and A.2.2.2 are satisfied. Then, there exist positive constants* $\rho \in (0,1)$ *and* $R < \infty$ *that do not depend on* θ *such that* $\{X_{\theta\, t}, t \geqslant 0\}$ *satisfies*

$$\sup_{v:|v|\leqslant 1} \left| \int_{\mathcal{X}} P_X^n(x, dx_n) v(x_n) - \int_{\mathcal{X}} \pi_X(dx_n) v(x_n) \right| \leqslant R\tilde{V}_X(x)\rho^n \,, \tag{2.13}$$

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

The proof of Lemma 2.7.2 consists of an application of Theorem 12 of Roberts and Rosenthal (2004). We remark that 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 2.7.2 is that the dependence properties of the companion Markov chain $\{X_{\theta\, t}, t \geqslant 0\}$ can be characterized independently of θ.

We use the properties of the companion Markov chain $\{X_{\theta\, t}, t \geqslant 0\}$ to establish the properties of the joint process $\{(Y_t, X_{\theta\, t})', t \geqslant 0\}$. The following lemma

[10] We omit the subscript θ from x to simplify the notation, but the dependence on θ is understood.

establishes the connection between the transition kernels of $\{X_{\theta t}, t \geqslant 0\}$ and $\{(Y_t, X_{\theta t})', t \geqslant 0\}$.

Lemma 2.7.3. *Consider the Markov chain $\{(Y_t, X_{\theta t})', t \geqslant 0\}$ defined above. Let $\pi_{Y|X}(dy|x_t)$ denote the (invariant) conditional distribution of Y_t given $X_{\theta t} = x_t$. Then, its n-step transition kernel is given by*

$$P^n_{Y,X}((y,x), d(y_n, x_n)) = \pi_{Y|X}(dy_n|x_n) \int_{\mathcal{H}} P^{n-1}_X(\tilde{x}, dx_n) P_H(h, dh_1), \quad n \geqslant 2, \tag{2.14}$$

where P_H is the transition kernel of $\{H_t, t \geqslant 0\}$, and

$\tilde{x} = \tilde{x}(y, x, h_1) = (h_1, \sum_{k=1}^{K} (\alpha_{0\,k} + \alpha_{1\,k} y + \beta_{1\,k} f) \mathbb{1}_{\{y \in \mathcal{Y}_k\}}, 1 + |y| + |f| + \overline{\beta}_1 d)'$.

The proof of the lemma builds upon the analysis of GARCH models of Meitz and Saikkonen (2008a). The structure given by equations (2.1), (2.2), (2.3) and (2.5) allows us to cast $\{(Y_t, X_{\theta t})', t \geqslant 0\}$ as a Markov chain with Dirac measure as the initial distribution. We remark that the analysis of $\{X_{\theta t}, t \geqslant 0\}$ differs depending on whether the process is studied in isolation or jointly with the process $\{Y_t, t \geqslant 0\}$. The random vector $X_{\theta t}$ depends on Y_{t-1}. When the process $\{X_{\theta t}, t \geqslant 0\}$ is analyzed in the joint system $\{(Y_t, X_{\theta t})', t \geqslant 0\}$ we have that the 1-step-ahead transition kernel of the process conditions on Y_{t-1}. However, when $\{X_{\theta t}, t \geqslant 0\}$ is analyzed in isolation we have that the 1-step-ahead transition kernel of the process does not condition on Y_{t-1}.

The following lemma establishes that $\{(Y_t, X_{\theta t})', t \geqslant 0\}$ inherits the moment and dependence properties of the companion Markov chain $\{X_{\theta t}, t \geqslant 0\}$.

Lemma 2.7.4. *Suppose A.2.2.1 and A.2.2.2 are satisfied. Then* (i) $\{(Y_t, X_{\theta\, t})', t \geqslant 0\}$ *is* $V_{Y,X}$*-geometrically ergodic with* $V_{Y,X}(y,x) = 1 + |y|^{2r_m} + \|x\|_1^{2r_m}$*; and* (ii) *there exist positive constants* $\rho \in (0,1)$ *and* $R < \infty$ *that do not depend on* θ *such that* $\{(Y_t, X_{\theta\, t})', t \geqslant 0\}$ *satisfies*

$$\sup_{v:|v|\leqslant 1} \left| \int_{\mathcal{Y}\times\mathcal{X}} [P_{Y,X}^n((y,x), d(y_n,x_n)) - \pi_{Y,X}(d(y_n,x_n))]v(y_n,x_n) \right| \leqslant R\tilde{V}_X(\check{x})\rho^n,$$

for all $(y,x)' \in \mathcal{Y}\times\mathcal{X}$ *and for all* $n \geqslant 2$*, and* $\check{x} = (h, \overline{\overline{\alpha_0}} + \overline{\alpha}_1|y| + \overline{\beta}_1|f|, 1 + |y| + |f| + \overline{\beta}_1 d)'$*, where* $\overline{\overline{\alpha_0}} = |\underline{\alpha}_0| \vee |\overline{\alpha}_0|$.

Finally, we establish the moment and dependence properties of $\{(Y_t, X_{\theta\, t})', t \geqslant 0\}$. We define the L_r norm of a random variable X as $\|X\|_{L_r} = (\mathbb{E}|X|^r)^{1/r}$ for any $r \in [1,\infty)$. The α-mixing coefficients of the process $\{(Y_t, X'_{\theta\, t})', t \geqslant 0\}$ are defined as

$$\alpha(l) = \sup_{s\geqslant 1} \sup_{A\in\mathcal{F}_0^s, B\in\mathcal{F}_{s+l}^\infty} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\,\mathbb{P}(B)| \ ,$$

where $\mathcal{F}_j^k = \sigma(\{(Y_t, X'_{\theta\, t})' : j \leqslant t \leqslant k\})$.

Proposition 2.7.1. *Suppose A.2.2.1 and A.2.2.2 are satisfied. Then,* $\{(Y_t, X'_{\theta\, t})', t \geqslant 0\}$ (i) *satisfies* $\sup_{t\geqslant 1}\|Y_t\|_{L_{2r_m}} < \infty$, $\sup_{t\geqslant 1}\|H_t\|_{L_{2r_m}} < \infty$, $\sup_{t\geqslant 1}\sup_{\theta\in\Theta}\|f_{\theta\, t}\|_{L_{2r_m}} < \infty$, $\sup_{t\geqslant 1}\sup_{\theta\in\Theta}\|d_{\theta\, t}\|_{L_{2r_m}} < \infty$*;* (ii) *has* α*-mixing coefficients that satisfy* $\alpha(l) \leqslant \exp(-C_\alpha l^{r_\alpha})$ *for some* $C_\alpha > 0$ *and* $r_\alpha > 0$ *that do not depend on* θ*; and* (iii) *its distribution converges to the invariant measure* $\pi_{Y,X}$*, which admits* $2r_m$ *moments.*

2.7.2 Establishing Performance Bounds for the ERM

We introduce a general inequality to bound the performance of the ERM.

Proposition 2.7.2. *Let* $\overline{R}(\theta) = \mathbb{E}L(Y_1^G, f_{\theta 1}^G)$*, where* $\{(Y_t^G, f_{\theta t}^G)', t \geqslant 0\}$ *is an independent copy of* $\{(Y_t, f_{\theta t})', t \geqslant 0\}$ *initialized at the stationary distribution. Then, it holds that*

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \leqslant 2 \sup_{\theta \in \Theta} |R(\theta) - \overline{R}(\theta)| + 2 \sup_{\theta \in \Theta} |R_T(\theta) - \overline{R}(\theta)| \,. \tag{2.15}$$

It is important to emphasize that Proposition 2.7.2 is a general result that only requires the "ghost" stationary loss process $\{L(Y_t^G, f_{\theta t}^G), t \geqslant 0\}$ to exist. Note that the existence of the process is established in Proposition 2.7.1 and is a consequence of the $V_{Y,X}$-geometric ergodicity of the process $\{(Y_t, X_{\theta t}')', t \geqslant 0\}$ (Lemma 2.7.4). We note that when the data is i.i.d. we have that $R(\theta) = \overline{R}(\theta)$ and the inequality in Proposition 2.7.2 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 the ERM.

In what follows we establish upper bounds for the two terms on the right hand side of (2.15) that hold uniformly over Θ. Proposition 2.7.3 is based on a covering argument similar to Jiang and Tanner (2010). Importantly, the proof of Proposition 2.7.3 relies on the properties of the dominating process $\{d_{\theta t}, t \geqslant 0\}$.

Proposition 2.7.3. *Suppose A.2.2.1 and A.2.2.2 are satisfied.*

Let $\mathbb{E}_T(\cdot) = \mathbb{E}(\cdot | Y_T, \ldots, Y_1)$*. Then, for any* $\varepsilon \in (0, 24C_d]$*, any* $T \geqslant 4C_U/\varepsilon$ *and*

any $M \geqslant 4C_U/\varepsilon$*, we have*

$$\begin{aligned}&\mathbb{P}\left(\sup_{\theta\in\Theta}|R_T(\theta)-\overline{R}(\theta)|>\frac{\varepsilon}{2}\right)\\&\leqslant\left(1+\frac{48C_\Theta C_d}{\varepsilon}\right)^p\sup_{\theta\in\Theta}\left[P_1^T\left(U_{\theta\,t},\frac{\varepsilon}{8}\right)+P_1^T\left(V_{\theta\,t},C_d\right)\right],\end{aligned}$$

where $P_{t_1}^{t_2}(U_t,\varepsilon)=\mathbb{P}\left(\left|\frac{1}{t_2-t_1+1}\sum_{t=t_1}^{t_2}U_t-\mathbb{E}\,U_t\right|>\varepsilon\right)$*, and*

$$\begin{aligned}&\mathbb{P}\left(\sup_{\theta\in\Theta}|R(\theta)-\overline{R}(\theta)|>\frac{\varepsilon}{2}\right)\\&\leqslant\left(1+\frac{48C_\Theta C_d}{\varepsilon}\right)^p\sup_{\theta\in\Theta}\left[P_{T+1}^{T+M}\left(\mathbb{E}_T\,U_{\theta\,t},\frac{\varepsilon}{8}\right)+P_{T+1}^{T+M}\left(\mathbb{E}_T\,V_{\theta\,t},C_d\right)\right],\end{aligned}$$

where $C_\Theta=\sup_{\theta\in\Theta}\|\theta\|_2$*,* $C_d=C_\psi\sup_\theta\|d_{\theta\,t}^2+2|Y_t-f_{\theta\,t}|d_{\theta\,t}\|_{L_1}$*,* $U_{\theta\,t}=L(Y_t,f_{\theta\,t})$*,* $U_{\theta\,t}^G=L(Y_t^G,f_{\theta\,t}^G)$*,* $V_{\theta\,t}=C_\psi\left(d_{\theta\,t}^2+2|Y_t-f_{\theta\,t}|d_{\theta\,t}\right)$*, and* $C_U=6\sup_{\theta\in\Theta}\|U_{\theta\,1}^G\|_{L_2}\sum_{l=1}^\infty\exp(-C_\alpha l^{r_\alpha})^{1/2}$.

The first term of the inequality in (2.15) is the supremum of a difference between an average of conditional and unconditional expectations. Proposition 2.7.4 bounds this term using Proposition 2.7.3 and Ibragimov's inequality (Davidson, 1994, Theorem 14.2).

Proposition 2.7.4. *Suppose A.2.2.1 and A.2.2.2 are satisfied. Then, for all* T *sufficiently large and for* $\varepsilon_T=\sigma\sqrt{p\log T/T}$*, it holds that*

$$\begin{aligned}&\left(1+\frac{48C_\Theta C_d}{\varepsilon_T}\right)^p\sup_{\theta\in\Theta}\mathbb{P}\left(\left|\frac{1}{M}\sum_{t=T+1}^{T+M}\mathbb{E}_T\,U_{\theta\,t}-\mathbb{E}\,U_{\theta\,t}\right|>\frac{\varepsilon_T}{8}\right)\leqslant\frac{1}{\log T}\ \textit{and}\\&\left(1+\frac{48C_\Theta C_d}{\varepsilon_T}\right)^p\sup_{\theta\in\Theta}\mathbb{P}\left(\left|\frac{1}{M}\sum_{t=T+1}^{T+M}\mathbb{E}_T\,V_{\theta\,t}-\mathbb{E}\,V_{\theta\,t}\right|>C_d\right)\leqslant o\left(\frac{1}{\log T}\right),\end{aligned}$$

as $T\to\infty$*, where* $\sigma^2=16\frac{r_m}{r_m-2}C_m^2(1+2\sum_{l=1}^\infty\exp\left(-C_\alpha l^{r_\alpha}\right)^{1-\frac{2}{r_m}})$ *and* $C_m=\sup_{t\geqslant1}\sup_{\theta\in\Theta}(\|U_{\theta\,t}\|_{L_{r_m}}\vee\|V_{\theta\,t}\|_{L_{r_m}})$.

The second term of the inequality in (2.15) is the supremum of the empirical process associated with the prediction loss process. We bound this using Proposition 2.7.3 and a concentration inequality for α-mixing processes. Proposition 2.7.5 is based on a Bernstein-type inequality for α-mixing sequences (Liebscher, 1996). The constant σ^2 that appears in Proposition 2.7.4 is the proportionality constant in Theorem 2.3.1. This may be interpreted as an upper bound on long run variance associated with the loss process. It depends on the r_m moment of the loss process and the α-mixing coefficients. It is worth noting that in the i.i.d. case the constant would be smaller and would reduce to $\sigma^2 = 16\frac{r_m}{r_m-2}C_m^2$.

Proposition 2.7.5. *Suppose A.2.2.1 and A.2.2.2 are satisfied. Then, for all T sufficiently large and for $\varepsilon_T = \sigma\sqrt{p\log T/T}$, it holds that*

$$\left(1+\frac{48C_\Theta C_d}{\varepsilon_T}\right)^p \sup_{\theta\in\Theta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}U_{\theta t}-\mathbb{E}\,U_{\theta t}\right|>\frac{\varepsilon_T}{8}\right)\leqslant\frac{1}{\log T}\ \textit{and}$$
$$\left(1+\frac{48C_\Theta C_d}{\varepsilon_T}\right)^p \sup_{\theta\in\Theta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}V_{\theta t}-\mathbb{E}\,V_{\theta t}\right|>C_d\right)\leqslant o\left(\frac{1}{\log T}\right),$$

as $T\to\infty$, where σ is defined in Proposition 2.7.4.

It follows from Propositions 2.7.1, 2.7.3, 2.7.4 and 2.7.5 that, for all T sufficiently large,

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

holds with high probability. This fact and Proposition 2.7.2 imply Theorem 2.3.1.

2.8 Conclusions

Leo Breiman forcefully argued that there are two main philosophies to analyze data (Breiman, 2001), the data modeling and the algorithmic modeling cultures. The data modeling culture is based on assuming that the data is generated by a (partially) known model whereas the algorithmic modeling culture pursues to be agnostic about the data generating mechanism. It is fair to say that the majority of research in the time series literature is typically carried out through the lens of the data modeling culture, whereas the fraction of contributions from the algorithmic modeling perspective is meager. In this work we take the algorithmic standpoint and study the performance of empirical risk minimization to choose an algorithm to forecast 1-step-ahead a time series. A key feature of the analysis is that the relationship between the time series and the class of algorithms is not specified. Our main result implies that the algorithm chosen by empirical risk minimization achieves asymptotically the optimal predictive performance that is attainable within the class. The algorithmic modeling culture paves the way for the development of new forecasting strategies for time series applications. Using the tools introduced in the nonlinear time series literature it is possible to develop general nonparametric theory to study algorithmic forecasting strategies from primitive assumptions.

2.9 Proofs of Sections 2.3 and 2.7

To simplify the analysis and without loss of generality we assume that $\Theta = [\underline{\alpha}_0, \overline{\alpha}_0]^K \times [\underline{\alpha}_1, \overline{\alpha}_1]^K \times [0, \overline{\beta}_1]^K$. To simplify notation we write $W_{\theta t} = (Y_t, X_{\theta t})'$.

Proof of Theorem 2.3.1. The claim follows from Propositions 2.7.1, 2.7.2, 2.7.3, 2.7.4 and 2.7.5. □

Proof of Lemma 2.7.1. We apply Lemmas 2.10.1, 2.10.2 and 2.10.3 together with Theorem 15.0.1 of Meyn and Tweedie (1993) to obtain that $\{X_{\theta t}, t \geqslant 0\}$ is q_X-geometrically ergodic with $q_X(x) = 1 + (\kappa' \dot{x})^{2r_m}$, where $\dot{x} = (|h|, |f|, |d|)'$, and the vector $\kappa \in (0,1)^3$ is defined in Lemma 2.10.1. Moreover, it is easy to see that Lemma 2.10.3 still holds with $q_X(x)$ replaced with $q_X(x)/\underline{\kappa}^{2r_m}$, where $\underline{\kappa}$ is the minimum of the components of κ. The claim follows by noting that $V_X(x) = 1 + \|x\|_1^{2r_m} \leqslant q_X(x)/\underline{\kappa}^{2r_m}$. □

Proof of Lemma 2.7.2. The claim of the Lemma follows from an application of Theorem 12 by Roberts and Rosenthal (2004). Define $\tilde{q}_X(x) = 1 + \tilde{\kappa}_h|h| + \tilde{\kappa}_f|f| + \tilde{\kappa}_d|d| = 1 + \tilde{\kappa}'\dot{x}$ where $\tilde{\kappa} \in (0,1)^3$ as well as the set $\tilde{S}_{2\epsilon} = \left\{x \in \mathcal{X} : \tilde{\kappa}'\dot{x} \leqslant \tilde{M}\right\}$. By arguments analogous to those used to claim that $S_{2\epsilon}$ defined in Lemma 2.10.1 is small we can show that we can choose a $\tilde{\kappa}$ such that for any $x \in \tilde{S}_{2\epsilon}$ and any $A \in \mathcal{B}(\mathcal{X})$ it holds that $P_X^2(x, A) \geqslant \tilde{c}_* \tilde{\varphi}(A)$, where $\tilde{c}_* \in (0,1)$ and $\tilde{\varphi}(A) = \mu_{Leb}(A \cap \tilde{D})$ is Lebesgue measure restricted to an open rectangular region $\tilde{D}$, which is the analogue of D defined in Lemma 2.10.1. As we remark in the proof of Lemma 2.10.1 $\tilde{c}_*$ and $\tilde{D}$ do not depend on θ.

It is easily verified that $\tilde{q}_X(x)$ satisfies the drift criterion by the same arguments as in Lemma 2.10.3, and that $\tilde{q}_X(x) \leqslant \tilde{V}_X(x)$. Define $\lambda^{-1} = 1 - \tilde{\gamma}_1 + \tilde{\gamma}_2/(2 + \tilde{M})$, where $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are analogous to γ_1 and γ_2 in Lemma 2.10.3 and do not depend on θ, and $\tilde{M} = \inf_{x \in S_{2\epsilon}^c} \tilde{q}_X(x)$. The proof strategy of Theorem 12 by Roberts and Rosenthal (2004) is based on a coupling argument. To this end we use $\{X_{\theta t}^G, t \geqslant 0\}$ to denote an independent copy of the Markov chain $\{X_{\theta t}, t \geqslant 0\}$ started at the stationary distribution, namely $X_{\theta 0}^G \sim \pi_X$. We define $B = \max\{1, \lambda^2(1 - \tilde{c}_*)\overline{R}\}$, where the constant $\overline{R}$ is computed in Lemma 2.13.2. We distinguish two cases. Note that in both cases we are applying Lemma 2.13.2.

(i) Suppose that $\lambda^{-1} < 1$. Then the assumptions of Proposition 11 and Theorem 12 by Roberts and Rosenthal (2004) are satisfied, thus applying the theorem we have that for any $j \in \{1, \ldots, n\}$,

$$\begin{aligned} &\sup_{v:|v| \leqslant 1} \left| \int_{\mathcal{X}} \left[P_X^n(x, dx_1) - \pi_X(dx_1) \right] v(x_1) \right| \\ &\qquad \leqslant (1 - \tilde{c}_*)^j + \lambda^{-n} \frac{B^{j-1}}{2} \left(\tilde{q}_X(x) + \mathbb{E}\tilde{q}_X(X_{\theta t}^G) \right) \end{aligned}$$

holds for all $x \in \mathcal{X}$ and all $n \geqslant 1$. Let $\overline{V} = 1 + \|H_t^G\|_{L_1} + \sup_{\theta \in \Theta} \|f_{\theta t}^G\|_{L_1} + \sup_{\theta \in \Theta} \|d_{\theta t}^G\|_{L_1}$. Obviously, $\overline{V} \geqslant \mathbb{E}\tilde{q}_X(X_{\theta t}^G)$. Furthermore, $\overline{V} < \infty$ by Lemma 2.7.1 and Proposition 2.7.1(i). Set $j = \lfloor rn \rfloor$ for sufficiently small $r > 0$ so that the bound converges to zero at a geometric rate. We now have that (2.13) holds with $\rho = (1 - \tilde{c}_*)^r \vee (\lambda^{-1} B^r) < 1$ and $R = 2\overline{V}$ (note that $\tilde{q}_X \geqslant 1$). The result follows since ρ and R do not depend on θ.

(ii) In the case $\lambda^{-1} \geqslant 1$, we can find an enlargement of $\tilde{S}_{2\epsilon}$ for which the result in (i) still holds (Roberts and Rosenthal, 2004). We choose M' such that

$2 + M' > \tilde{\gamma}_2/\tilde{\gamma}_1$. Note that $\tilde{S}'_{2\,\epsilon} = \{x \in \mathcal{X} : \tilde{\kappa}'\dot{x} \leqslant M'\}$ is still a small set by the same arguments used in the proof of Part II of Lemma 2.10.1. Consequently, $P_X^2(x, A) \geqslant \tilde{c}'_* \tilde{\varphi}'(A)$ for all $x \in \tilde{S}'_{2\,\epsilon}$, where $\tilde{c}'_*$ is possibly smaller than $\tilde{c}_*$ but strictly positive (and independent of θ), $\tilde{\varphi}'(A) = \mu_{Leb}(A \cap \tilde{D}')$, and $\tilde{D}'$ is analogous to $\tilde{D}$ in part (i). Clearly, $\lambda'^{-1} = 1 - \tilde{\gamma}_1 + \frac{\tilde{\gamma}_2}{2+M'} < 1$. The claim holds by the same arguments as in (i) with λ and $\tilde{c}_*$ replaced by λ' and $\tilde{c}'_*$. □

Proof of Lemma 2.7.3. For all $n \geqslant 2$ we write

$$\begin{aligned} P_W^n(w, dw_n) &= \pi_{Y|X}(dy_n|x_n)\mathbb{P}(dx_n|w) \\ &= \pi_{Y|X}(dy_n|x_n) \int_{\mathcal{H}} \mathbb{P}(dx_n|w, h_1) P_H(h, dh_1) \,, \end{aligned}$$

where the last equality follows because the H_t component of $W_{\theta\,t}$ is a Markov chain of its own. Define $\tilde{f}_\theta = f_{\theta\,1} = \sum_k \{\alpha_{0k} + \alpha_{1k} y + \beta_1 f\} \mathbb{1}_{\{y \in \mathcal{Y}_k\}}$, and $\tilde{d} = d_1 = 1 + |y| + |f| + \overline{\beta}_1 d$. Note that by the i.i.d. assumption on the innovations $Z_{1\,t}$ and $Z_{2\,t}$ we have that the $X_{\theta\,t}$ component of $\{W_{\theta\,t}, t \geqslant 0\}$ has a 2-step transition mechanism which is entirely similar to the 1-step transition mechanism of the companion Markov chain defined in (2.11) with initial value given by $\tilde{x}(w, h_1)$. We denote $\tilde{P}_X^n((w, h_1), dx_n) = \mathbb{P}(dx_n|w, h_1)$. Note that $\tilde{P}_X^2((w, h_1), dx_2) = P_X^1(\tilde{x}, dx_2)$ where $\tilde{x} = \tilde{x}(w, h_1) = (h_1, \tilde{f}_\theta, \tilde{d})$. We have $\tilde{P}_X^3((w, h_1), dx_3) = \int_{\mathcal{X}} P_X(\tilde{x}, dx_2) P_X(x_2, dx_3) = P_X^2(\tilde{x}, dx_3)$. By induction, $\tilde{P}_X^n((w, h_1), dx_n) = P_X^{n-1}(\tilde{x}, dx_n)$, and the result follows. □

Proof of Lemma 2.7.4. (i) First, $\{X_{\theta\,t}, t \geqslant 0\}$ viewed as a separate Markov chain

is V_X-geometrically ergodic by Lemma 2.7.1. We begin by showing that

$$\mathbb{E}_{Y|X}\left(V_{Y,X}(W_{\theta\,t})|X_{\theta\,t}=x\right)\equiv\int_{\mathcal{Y}}V_{Y,X}(y,x)\pi_{Y|X}(dy|x)<C\cdot V_X(x)\,. \quad (2.16)$$

For all $x\in\mathcal{X}$, by Assumption A.2.2.1(ii) we have that

$$\begin{aligned}
&\mathbb{E}_{Y|X}\left(V_{Y,X}(W_{\theta\,t})|X_{\theta\,t}=x\right)\\
&=V_X(x)+\mathbb{E}_{Y|X}\left(|g_{y1}(h)+g_{y2}(h)\epsilon_{Y\,t}|^{2r_m}\right)\\
&\leqslant V_X(x)+2^{2r_m-1}|g_{y1}(h)|^{2r_m}+2^{2r_m-1}|g_{y2}(h)|^{2r_m}\mathbb{E}_{Y|X}(|\epsilon_{Y\,t}|^{2r_m})\\
&\leqslant V_X(x)+C\|x\|_1^{2r_m}\leqslant C\cdot V_X(x)\,,
\end{aligned}$$

where the constant $0<C<\infty$ may change from line to line. To satisfy the definition of $V_{Y,X}$-geometric ergodicity, we must have that $\mathbb{E}\left(V_{Y,X}(Y_t,X_{\theta\,t})\right)<\infty$, where the expectation is taken with respect to the invariant measure $\pi_{Y,X}$. By (2.16) we have that

$$\begin{aligned}
\mathbb{E}\left(V_{Y,X}(W_{\theta\,t})\right)&=\int_{\mathcal{X}}\pi_X(dx)\int_{\mathcal{Y}}V_{Y,X}(y,x)\pi_{Y|X}(dy|x)\\
&\leqslant\int_{\mathcal{X}}\pi_X(dx)C\cdot V_X(x)<\infty,
\end{aligned}$$

as expected. For any $w=(y,x)'\in\mathcal{Y}\times\mathcal{X}$ and all $n\geqslant 2$ we have that

$$\begin{aligned}
&\sup_{v:|v|\leqslant V_{Y,X}}\left|\int_{\mathcal{Y}\times\mathcal{X}}\left[P_{Y,X}^n(w,dw_n)-\pi_{Y,X}(dw_n)\right]v(w_n)\right|\\
&=\sup_{v:|v|\leqslant V_{Y,X}}\left|\int_{\mathcal{X}}\left(\int_{\mathcal{H}}P_X^{n-1}(\tilde{x},dx_n)P_H(h,dh_1)-\pi_X(dx_n)\right)\int_{\mathcal{Y}}\pi_{Y|X}(dy_n|x_n)v(y_n,x_n)\right|\\
&\quad\leqslant C\sup_{v':|v'|\leqslant V_X}\left|\int_{\mathcal{H}}\left(\int_{\mathcal{X}}\left[P_X^{n-1}(\tilde{x},dx_n)-\pi_X(dx_n)\right]v'(x_n)\right)P_H(h,dh_1)\right|\\
&\quad\leqslant C\int_{\mathcal{H}}\sup_{v':|v'|\leqslant V_X}\left|\int_{\mathcal{X}}\left[P_X^{n-1}(\tilde{x},dx_n)-\pi_X(dx_n)\right]v'(x_n)\right|P_H(h,dh_1)\\
&\quad\leqslant CR_\theta\rho_\theta^{n-1}\mathbb{E}\left(V_X(\tilde{x})|H_0=h\right)\,, \quad (2.17)
\end{aligned}$$

where $R_\theta < \infty$, $\rho_\theta < 1$. The equality follows by Lemma 2.7.3, the first inequality is a consequence of (2.16) and the last inequality is implied by the drift criterion that we have used in the proof of Lemma 2.7.1. Furthermore, A.2.2.1(i), (iii), (iv) and (2.19) imply

$$\begin{aligned}\mathbb{E}\left(V_X(\tilde{x})|H_0 = h\right) &\leqslant 1 + 3^{2r_m-1}\left(\mathbb{E}\left(|H_1|^{2r_m}|H_0 = h\right) + |\tilde{f}_\theta|^{2r_m} + |\tilde{d}|^{2r_m}\right) \\ &< 1 + C \cdot \begin{cases} |\tilde{f}_\theta|^{2r_m} + |\tilde{d}|^{2r_m} + |h|^{2r_m}\,, & |h| > M_\epsilon \\ 1 + |\tilde{f}_\theta|^{2r_m} + |\tilde{d}|^{2r_m}\,, & |h| \leqslant M_\epsilon \end{cases} \\ &< CV_X(\check{x})\,,\end{aligned}$$

where $1 < C < \infty$ may change from line to line and the choice of ϵ is such that $\mathbb{E}(a + b^\epsilon|\epsilon_{H\,t}| + \epsilon)^{2r_m} < 1$ (A.2.2.1(iv)).

(ii) Repeating the same arguments as in (i) with $2r_m = 1$ and with $\sup_{v:|v|\leqslant 1}$ instead of $\sup_{v:|v|\leqslant V_{Y,X}}$, we can use Lemma 2.7.2 in the last inequality of (2.17) instead of the standard drift criterion to obtain constants $\rho \in (0,1)$ and $R < \infty$ that do not depend on θ. The proof is completed by noting that we can redefine R to absorb $C\rho^{-1}$. □

Proof of Proposition 2.7.1. (i) By the same arguments used to arrive at (2.19), which rely on A.2.2.1, take the L_{2r_m}-norm on both sides of (2.19) and apply Minkowski's inequality to get

$$\begin{aligned}\|H_t\|_{L_{2r_m}} &\leqslant \underbrace{\|a + b^\epsilon|\epsilon_{H\,t}| + \epsilon\|_{L_{2r_m}}}_{\varrho_\epsilon < 1}\|H_{t-1}\|_{L_{2r_m}} + \|\overline{g}_h^\epsilon(1 + |\epsilon_{H\,t}|)\mathbb{1}_{\{|H_{t-1}|\leqslant M_\epsilon\}}\|_{L_{2r_m}} \\ &\leqslant \varrho_\epsilon\|H_{t-1}\|_{L_{2r_m}} + C_\epsilon,\end{aligned}$$

where $C_\epsilon < \infty$, and we have used that $|H_{t-1}|^{2r_m} \mathbb{1}_{\{|H_{t-1}|>M_\epsilon\}} \leqslant |H_{t-1}|^{2r_m}$. Thus, $\sup_{t\geqslant 1} \|H_t\|_{L_{2r_m}} \leqslant |h| + \frac{C_\epsilon}{1-\varrho_\epsilon} < \infty$. A.2.2.1 implies that there exists $C_{\epsilon_Y} < \infty$ such that

$$\begin{aligned}\sup_{t\geqslant 1} \|Y_t\|_{L_{2r_m}} &\leqslant \sup_{t\geqslant 1} \left[\|g_{y1}(H_t)\|_{L_{2r_m}} + \|g_{y2}(H_t)\|_{L_{2r_m}} \|\epsilon_{Y\,t}\|_{L_{2r_m}}\right] \\ &\leqslant C_{y1} \sup_{t\geqslant 1} \|H_t\|_{L_{2r_m}} + C_{y2} \sup_{t\geqslant 1} \|H_t\|_{L_{2r_m}} C_{\epsilon_Y} < \infty.\end{aligned}$$

Furthermore, since $\alpha_{1\,k} > 0$, we have

$$\|f_{\theta\,t}\|_{L_{2r_m}} \leqslant \overline{\overline{\alpha_0}} + \overline{\alpha}_1 \|Y_{t-1}\|_{L_{2r_m}} + \overline{\beta}_1 \|f_{\theta\,t-1}\|_{L_{2r_m}},$$

where $\overline{\overline{\alpha_0}} = |\underline{\alpha}_0| \vee |\overline{\alpha}_0|$. Therefore,

$$\sup_{t\geqslant 1} \sup_{\theta\in\Theta} \|f_{\theta\,t}\|_{L_{2r_m}} \leqslant \sup_{\theta\in\Theta} |f_{\theta\,0}| + \frac{\overline{\overline{\alpha_0}} + \overline{\alpha}_1 \sup_{t\geqslant 1} \|Y_{t-1}\|_{L_{2r_m}}}{1-\overline{\beta}_1} < \infty.$$

Finally, we have

$$\sup_{t\geqslant 1} \sup_{\theta\in\Theta} \|d_{\theta\,t}\|_{L_{2r_m}} \leqslant 1 + \frac{\sup_{t\geqslant 1} \|Y_{t-1}\|_{L_{2r_m}} + \sup_{t\geqslant 1} \sup_{\theta\in\Theta} \|f_{\theta\,t-1}\|_{L_{2r_m}}}{1-\overline{\beta}_1} < \infty.$$

(ii) It is enough to show that $\{W_{\theta\,t}, t \geqslant 0\}$ is geometrically β-mixing, since $\alpha(l) \leqslant \beta(l)$, where $\beta(l) = \sup_{t\in\mathbb{Z}} \frac{1}{2} \sup \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|$, and the supremum is taken over all pairs of finite partitions $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_J\}$ of Ω such that $A_i \in \sigma\{W_{\theta\,s} : s \leqslant t\}$, $i = 1, \ldots, I$, and $B_j \in \sigma\{W_{\theta\,s} : s \geqslant t+l\}$, $j = 1, \ldots, J$. Let $\delta_w(A) = \mathbb{1}\{w \in A\}$ for any $A \in \mathcal{B}(\mathcal{Y}\times\mathcal{X})$. By Proposition 4 in Liebscher (2005), $\{W_{\theta\,t}, t \geqslant 0\}$ is β-mixing with geometrically decaying mixing numbers if (a) $\int_{\mathcal{Y}\times\mathcal{X}} V_X(x_0)\delta_{y,x}(dw_0) = V_X(x) < \infty$, and (b) $\{W_{\theta\,t}, t \geqslant 0\}$ is Q-geometrically ergodic in the sense of Liebscher (2005) with

$Q(w) = V_X(x)$. Condition (a) holds for all $w \in \mathcal{Y} \times \mathcal{X}$. For condition (b), we first need to show that $\int_{\mathcal{Y}\times\mathcal{X}} V_X(x_n)\pi_{Y,X}(dw_n) < \infty$. This follows from

$$\int_{\mathcal{Y}\times\mathcal{X}} V_X(x_n)\pi_{Y,X}(dw_n) = \int_{\mathcal{X}} V_X(x_n)\pi_X(dx_n) \int_{\mathcal{Y}} \pi_{Y|X}(dy_n|x_n) < \infty,$$

where the last inequality follows from the V_X-geometric ergodicity of $\{X_{\theta t}, t \geqslant 0\}$. As for the remaining part of condition (b), notice that from Lemma 2.7.4(ii) we have that $\left\|P^l_{Y,X}(w,\cdot) - \pi_{Y,X}\right\|_{TV} \leqslant R\tilde{V}_X(\check{x})\rho^l \wedge 1$, where

$$\left\|P^l_{Y,X}(w,\cdot) - \pi_{Y,X}\right\|_{TV} = \sup_{v:|v|\leqslant 1} \left| \int_{\mathcal{Y}\times\mathcal{X}} \left[P^l_{Y,X}(w, dw_l) - \pi_{Y,X}(dw_l)\right] v(w_l) \right| ,$$

which completes the proof of condition (b). It remains to be shown that the rate of decay does not depend on θ. For any probability measure τ on $\mathcal{Y}\times\mathcal{X}$, define $\xi_l(\tau) = \int_{\mathcal{Y}\times\mathcal{X}} \left\|P^l_{Y,X}(w,\cdot) - \pi_{Y,X}\right\|_{TV} \cdot \tau(dw)$. By virtue of part (ii) of Lemma 2.7.4 we compute that $\xi_l(\pi_{Y,X}) \leqslant R\check{V}\rho^l$, where

$$\check{V} = 2 + \|H^G_1\|_{L_1} + (1+\overline{\alpha}_1)\|Y^G_1\|_{L_1} + (1+\overline{\beta}_1)\sup_{\theta\in\Theta}\|f^G_{\theta 1}\|_{L_1} + \overline{\beta}_1 \sup_{\theta\in\Theta}\|d^G_{\theta 1}\|_{L_1} ,$$

and $\xi_l(\delta_{y,x}) = \left\|P^l_{Y,X}(w,\cdot) - \pi_{Y,X}\right\|_{TV} \leqslant R\tilde{V}_X(\check{x})\rho^l \wedge 1$. Now, by Proposition 3 in Liebscher (2005) we have that for any $w \in \mathcal{Y}\times\mathcal{X}$, and $m = \lfloor l/2 \rfloor$, $\beta(l) \leqslant 3\xi_m(\delta_{y,x}) + \xi_m(\pi_{Y,X}) \leqslant R\left(\check{V} + 3\tilde{V}_X(\check{x})\right)\rho^m \wedge 1$. It is not difficult to verify that $\alpha(l) \leqslant \beta(l) \leqslant \exp\left(-C_\alpha l^{r_\alpha}\right) \wedge 1$ for all $l \geqslant 1$. The choice of C_α and r_α depends on $R, \check{V}, \rho$ and $\tilde{V}_X(\check{x})$. Note that the rate of decay of the uniform bound for the α-mixing coefficients does not depend on θ (Lemma 2.7.2). The claim follows by redefining R and noting that $\tilde{V}_X \geqslant 1$.

(iii) The existence of the stationary distribution with $2r_m$ moments of $\{(Y_t, X'_{\theta t})', t \geqslant$

$0\}$ follows by its $V_{Y,X}$-geometric ergodicity, which is established in Lemma 2.7.4.

□

Proof of Proposition 2.7.2. We begin by noting that $\{L(Y_t^G, f_{\theta t}^G)', t \geqslant 0\}$ is a stationary process. Define $\overline{R}(\hat{\theta}) = \mathbb{E}\, L(Y_1^G, f_{\hat{\theta}1}^G)$. The properties of infimum and supremum and the definition of empirical risk minimizer (i.e. $R_T(\theta) \geqslant R_T(\hat{\theta})$ for all $\theta \in \Theta$) imply

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

where the last inequality follows from Lemma 8.2 in Devroye *et al.* (1996). □

Proof of Proposition 2.7.3. Let $U_{\theta t} = L(Y_t, f_{\theta t})$ and $U_{\theta t}^G = L(Y_t^G, f_{\theta t}^G)$. Adding and subtracting $\mathbb{E}\, U_{\theta t}$, we have

$$\begin{aligned} \mathbb{P}\left(\sup_{\theta\in\Theta} \left|R_T(\theta) - \overline{R}(\theta)\right| > \frac{\varepsilon}{2}\right) &\leqslant \mathbb{P}\left(\sup_{\theta\in\Theta} \left|\frac{1}{T}\sum_{t=1}^{T} (U_{\theta t} - \mathbb{E}\, U_{\theta t})\right| > \frac{\varepsilon}{4}\right) \\ &\quad + \mathbb{P}\left(\sup_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} |\,\mathbb{E}\, U_{\theta t} - \mathbb{E}\, U_{\theta 1}^G| > \frac{\varepsilon}{4}\right). \end{aligned}$$

Recall that $\{W_{\theta t}, t \geqslant 0\}$ is initialized at the Dirac measure. Furthermore, $\{W_{\theta t}, t \geqslant 0\}$ and $\{W_{\theta t}^G, t \geqslant 0\}$ have the same transition kernel, so $\mathbb{E}\, U_{\theta t} = \mathbb{E}(U_{\theta t}|W_{\theta 0} = w) = \mathbb{E}(U_{\theta t}^G|W_{\theta 0}^G = w)$. Since $\mathbb{E}\, U_{\theta t} - \mathbb{E}\, U_{\theta t}^G$ is not random,

$$|\,\mathbb{E}\, U_{\theta t} - \mathbb{E}\, U_{\theta t}^G| = |\,\mathbb{E}(U_{\theta t}^G|W_{\theta 0}^G = w) - \mathbb{E}\, U_{\theta t}^G| \leqslant 6\exp(-C_\alpha t^{r_\alpha})^{1/2} \sup_{\theta\in\Theta} \|U_{\theta t}^G\|_{L_2},$$

which follows from Ibragimov's inequality and Proposition $1(ii)$. We have $\mathbb{P}\left(\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}|\,\mathbb{E}\,U_{\theta\,t}-\mathbb{E}\,U_{\theta\,1}^{G}|>\frac{\varepsilon}{4}\right)\leqslant\mathbb{P}\left(\frac{C_U}{T}>\frac{\varepsilon}{4}\right)$, where $C_U<\infty$ by Proposition $1(ii)$ and (iii).[11] This implies that

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}|\,\mathbb{E}\,U_{\theta\,t}-\mathbb{E}\,U_{\theta\,1}^{G}|>\frac{\varepsilon}{4}\right)=0$$

for all $T\geqslant 4C_U/\varepsilon$.

Let $\Theta_i=\{\theta\in\mathbb{R}^p:\|\theta-\theta_i\|_2\leqslant\delta\}$ with $\theta_i\in\Theta$ for $i=1,\ldots,N_\delta$ denote a δ-covering of Θ for some $\delta\in(0,1]$. Then, we have that for all $T\geqslant 4C_U/\varepsilon$,

$$\begin{aligned}\mathbb{P}\left(\sup_{\theta\in\Theta}\left|R_T(\theta)-\overline{R}(\theta)\right|>\frac{\varepsilon}{2}\right)&\leqslant\sum_{i=1}^{N_\delta}\mathbb{P}\left(\sup_{\theta\in\Theta_i}\left|\frac{1}{T}\sum_{t=1}^{T}(U_{\theta\,t}-\mathbb{E}U_{\theta\,t})\right|>\frac{\varepsilon}{4}\right)\\&\leqslant\sum_{i=1}^{N_\delta}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}(U_{i\,t}-\mathbb{E}U_{i\,t})\right|>\frac{\varepsilon}{8}\right)\\&+\sum_{i=1}^{N_\delta}\mathbb{P}\left(\sup_{\theta\in\Theta_i}\left|\frac{1}{T}\sum_{t=1}^{T}(U_{\theta\,t}-\mathbb{E}U_{\theta\,t})-\left(\frac{1}{T}\sum_{t=1}^{T}U_{i\,t}-\mathbb{E}U_{i\,t}\right)\right|>\frac{\varepsilon}{8}\right),\end{aligned}$$

where $U_{i\,t}=U_{\theta_i\,t}$. Lemma 2.13.3 establishes that (i) for each $\theta\in\Theta_i$ we have $|U_{\theta\,t}-U_{i\,t}|\leqslant\delta V_{i\,t}=\delta C_\psi(d_{\theta_i\,t}^2+2|Y_t-f_{\theta_i\,t}|d_{\theta_i\,t})$ and (ii) there exists a positive constant C_d (that does not depend on i and t) such that for all $\delta\in(0,1]$ we have

[11] By Condition 2.2.1(iii), $\sup_{\theta\in\Theta}\|U_{\theta\,1}^G\|_{L_2}\leqslant 2C_\psi\left(\|Y_1^G\|_{L_4}+\sup_{\theta\in\Theta}\|f_{\theta\,1}^G\|_{L_4}\right)$. By Proposition 2.7.1$(iii)$, $\|Y_t^G\|_{L_4}<\infty$ and by the same arguments as in the proof of Proposition $1(i)$, we have $\|f_{\theta\,t}^G\|_{L_4}\leqslant\overline{\overline{\alpha_0}}+\overline{\alpha}_1\|Y_{t-1}^G\|_{L_4}+\overline{\beta}_1\|f_{\theta\,t-1}^G\|_{L_4}$, but by stationarity, we have $\sup_{\theta\in\Theta}\|f_{\theta\,1}^G\|_{L_4}\leqslant\frac{\overline{\overline{\alpha_0}}+\overline{\alpha}_1\|Y_1^G\|_{L_2}}{1-\overline{\beta}_1}<\infty$.

that $\sup_{t\geqslant 1}\mathbb{E}V_{i\,t}\leqslant C_d$. Set $\delta=\varepsilon/(24C_d)$. Then, for all $\varepsilon<24C_d$, we get

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\theta\in\Theta_i}\left|\frac{1}{T}\sum_{t=1}^{T}(U_{\theta\,t}-U_{i\,t})-\mathbb{E}(U_{\theta\,t}-U_{i\,t})\right|>\frac{\varepsilon}{8}\right)\\
&\leqslant\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(V_{i\,t}+\mathbb{E}V_{i\,t})>\frac{\varepsilon}{8\delta}\right)\\
&=\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(V_{i\,t}-\mathbb{E}V_{i\,t})>\frac{\varepsilon}{8\delta}-\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\,V_{i\,t}\right)\\
&\leqslant\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(V_{i\,t}-\mathbb{E}V_{i\,t})>C_d\right)
\end{aligned}$$

holds. The claim follows after noting that $N_\delta\leqslant(1+2C_\Theta/\delta)^p=(1+48C_\Theta C_d/\varepsilon)^p$.[12] It is straightforward to check that the same covering argument applies to the second part of the claim with $U_{\theta\,t}$ and $V_{\theta\,t}$ replaced by $\mathbb{E}_T\,U_{\theta\,t}$ and $\mathbb{E}_T\,V_{\theta\,t}$, respectively. This is because $|\,\mathbb{E}_T\,U_{\theta\,t}-\mathbb{E}_T\,U_{\theta_i\,t}|\leqslant\mathbb{E}_T\,|U_{\theta\,t}-U_{\theta_i\,t}|\leqslant\delta\,\mathbb{E}_T\,V_{\theta_i\,t}$ by Jensen's inequality and the order-preserving property of the conditional expectation. □

Proof of Proposition 2.7.4. By Markov's inequality, for any $\varepsilon>0$, we have

$$\sup_{\theta\in\Theta}P_{T+1}^{T+M}(\mathbb{E}_T\,U_{\theta\,t},\varepsilon)\leqslant\frac{\sup_{\theta\in\Theta}\mathbb{E}\left|\frac{1}{M}\sum_{t=T+1}^{T+M}(\mathbb{E}_T\,U_{\theta\,t}-\mathbb{E}\,U_{\theta\,t})\right|^p}{\varepsilon^p}\,. \tag{2.18}$$

By Ibragimov's inequality, we have that for $r_m>p\geqslant 1$,

$$\begin{aligned}
&\sup_{\theta\in\Theta}\|\mathbb{E}_T\,U_{\theta\,t}-\mathbb{E}\,U_{\theta\,t}\|_{L_p}\\
&\leqslant 2(2^{1/p}+1)\alpha(l)^{1/p-1/r_m}\sup_{t\geqslant 1}\sup_{\theta\in\Theta}\|U_{\theta\,t}\|_{L_{r_m}}\,,\quad l=t-T\,,
\end{aligned}$$

where $\sup_{t\geqslant 1}\sup_\Theta\|U_{\theta\,t}\|_{L_{r_m}}<\infty$, which exists by Proposition 2.7.1. Consequently, and because of the exponential decay of the strong mixing coefficients,

[12] The detailed computations of this inequality are shown in 2.12.

the numerator in (2.18) is bounded above by C/T^p for some $C < \infty$, where we have used that $M = \lceil \gamma T \rceil$. Let $\varepsilon_T = \sigma\sqrt{\frac{p \log T}{T}}$. It follows that

$$\left(1 + \frac{48 C_\Theta C_d}{\varepsilon_T}\right)^p \sup_{\theta \in \Theta} P_{T+1}^{T+M}\left(\mathbb{E}_T\, U_{\theta t}, \frac{\varepsilon_T}{8}\right) \leqslant \frac{C}{T^p \varepsilon_T^{2p}} = O\left(\log^{-p} T\right) .$$

Finally, since $V_{\theta t}$ is also strong mixing with exponentially decaying coefficients and $\sup_{t \geqslant 1} \sup_{\theta \in \Theta} \|V_{\theta t}\|_{L_{r_m}} < \infty$, the result follows by repeating the same arguments. □

Proof of Proposition 2.7.5. Let $\widetilde{U}_{\theta t} = U_{\theta t} - \mathbb{E}\, U_{\theta t}$ and $\widetilde{V}_{\theta t} = V_{\theta t} - \mathbb{E}\, V_{\theta t}$. The analysis of the sequences $\{\widetilde{U}_{\theta t}, t \geqslant 0\}$ and $\{\widetilde{V}_{\theta t}, t \geqslant 0\}$ is analogous. Here we focus on $\{\widetilde{U}_{\theta t}, t \geqslant 0\}$. To simplify notation we omit the subscript θ in the notation of the sequence $\{\widetilde{U}_{\theta t}, t \geqslant 0\}$.

Let $\sum_{t=1}^T \widetilde{U}_t = \sum_{t=1}^T U_t' + \sum_{t=1}^T U_t''$ where $U_t' = U_t \mathbb{1}_{\{U_t \leqslant b_T\}} - \mathbb{E}\left(U_t \mathbb{1}_{\{U_t \leqslant b_T\}}\right)$ and $U_t'' = U_t \mathbb{1}_{\{U_t > b_T\}} - \mathbb{E}\left(U_t \mathbb{1}_{\{U_t > b_T\}}\right)$. We then have that

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T \widetilde{U}_t\right| > \frac{\varepsilon_T}{8}\right) \leqslant \mathbb{P}\left(\left|\sum_{t=1}^T U_t'\right| > \frac{T\varepsilon_T}{16}\right) + \mathbb{P}\left(\left|\sum_{t=1}^T U_t''\right| > \frac{T\varepsilon_T}{16}\right) .$$

Define $M_T = \lfloor T^{\frac{1}{2} - \frac{p+1}{2(r_m - 1)}} \log^{-\frac{1}{2}} T \rfloor$ and $b_T = C_b T^{\frac{p+1}{2(r_m-1)}} (p \log T)^{-\frac{p-1}{2(r_m-1)}}$ where C_b is a positive constant to be chosen in what follows. The sequence $\{U_t'\}_{t=1}^T$ has the same mixing properties of $\{\widetilde{U}_t\}_{t=1}^T$ and $\|U_t'\|_{L_\infty} \leqslant b_T$. Then for all T sufficiently large and $p < r_m - 2$ the conditions of Theorem 2.1 in Liebscher (1996) are

satisfied since $M_T \in \{1, \ldots, T\}$ and $4M_T b_T \leqslant T\varepsilon_T/16$. Then, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{t=1}^{T} U'_t\right| > \frac{T\varepsilon_T}{16}\right) \\
&\leqslant 4\exp\left(-\frac{T\varepsilon_T^2}{\frac{16384}{M_T}\sup_{0\leqslant t\leqslant T-1}\mathbb{E}(\sum_{s=t+1}^{(t+M_T)\wedge T} U'_s)^2 + \frac{128}{3}M_T b_T \varepsilon_T}\right) \\
&\quad + 4\frac{T}{M_T}\exp\left(-C_\alpha M_T^{r_\alpha}\right) .
\end{aligned}$$

Let $\gamma_t(l) = |\mathrm{Cov}(U'_t, U'_{t+l})|$ for $l = 0, \ldots, T-1$. Let $C_m = \sup_{t\geqslant 1}\sup_{\theta\in\Theta}\|U_t\|_{L_{r_m}}$. Noting that $L(Y_t, f_{\theta\, t}) \geqslant 0$, Davydov's inequality implies that

$$\gamma_t(l) \;\leqslant\; 2\frac{r_m}{r_m-2}2^{1-\frac{2}{r_m}}\alpha(l)^{1-\frac{2}{r_m}}\|U'_t\|_{L_{r_m}}\|U'_{t+l}\|_{L_{r_m}} \leqslant 16\frac{r_m}{r_m-2}C_m^2\alpha(l)^{1-\frac{2}{r_m}} ,$$

for $l = 0, \ldots, T-1$, and we use the fact that for any r we have $\|U'_t\|_{L_r} \leqslant 2\|U_t\|_{L_r}$. Thus, we have $\sup_{0\leqslant t\leqslant T-1}\mathbb{E}(\sum_{s=t+1}^{(t+M_T)\wedge T} U'_t)^2 \leqslant M_T 16\frac{r_m}{r_m-2}C_m^2(1 + 2\sum_{l=1}^{\infty}\exp\left(-C_\alpha l^{r_\alpha}\right)^{1-\frac{2}{r_m}}) = M_T\sigma^2$. Then, for all T sufficiently large, after some algebra it holds that

$$\left(1 + \frac{48C_\Theta C_d}{\varepsilon_T}\right)^p \mathbb{P}\left(\left|\sum_{t=1}^{T} U'_t\right| > \frac{T\varepsilon_T}{16}\right) = o(\log^{-1} T) .$$

Let $C_m = \sup_{t\geqslant 1}\sup_{\theta\in\Theta}(\|U_{\theta\, t}\|_{L_{r_m}} \vee \|V_{\theta\, t}\|_{L_{r_m}}) < \infty$ by Proposition 2.7.1. We note that for all T sufficiently large,

$$\begin{aligned}
&\left(1 + \frac{48C_\Theta C_d}{\varepsilon_T}\right)^p \mathbb{P}\left(\left|\sum_{t=1}^{T} U''_t\right| > \frac{T\varepsilon_T}{16}\right) \overset{(a)}{\leqslant} \left(1 + \frac{48C_\Theta C_d}{\varepsilon_T}\right)^p \frac{16}{T\varepsilon_T}\mathbb{E}\left|\sum_{t=1}^{T} U''_t\right| \\
&\leqslant \left(1 + \frac{48C_\Theta C_d}{\varepsilon_T}\right)^p \frac{32}{\varepsilon_T}\sup_{t\geqslant 1}\mathbb{E}\left|U_t \mathbb{1}_{\{U_t > b_T\}}\right| \overset{(b)}{\leqslant} \left(1 + \frac{48C_\Theta C_d}{\varepsilon_T}\right)^p \frac{32}{\varepsilon_T}\frac{C_m^{r_m}}{b_T^{r_m-1}} \\
&\leqslant (1 + 48C_\Theta C_d)^p \frac{32}{\varepsilon_T^{p+1}}\frac{C_m^{r_m}}{b_T^{r_m-1}} \overset{(c)}{\leqslant} \log^{-1} T ,
\end{aligned}$$

where (a) follows from Markov's inequality (b) from the inequality $\mathbb{E}(|X\mathbb{1}_{\{|X|>b\}}|) \leqslant \mathbb{E}(|X|^r)/b^{r-1}$ for any random variable X wih finite r-th moment and positive constant b, and (c) from a sufficiently large choice of the constant C_b. The sequence $\tilde{V}_{\theta\,t}$ can be analysed using the same strategy (using the same choice of M_T and b_T used for $\tilde{U}_T$). □

2.10 Irreducibility, Aperiodicity, Drift Criterion

Before we proceed, we establish upper bounds on $|H_t|$, $|f_{\theta\,t}|$ and $d_{\theta\,t}$. By A.2.2.1(i) we have that for any $\epsilon > 0$ there exists some $1 < M_\epsilon < \infty$ such that $|H_t| \leqslant (a + b^\epsilon|Z_{1\,t}| + \epsilon)|H_{t-1}|$ for all $|H_{t-1}| > M_\epsilon$, where $b^\epsilon := b + \epsilon$. The same assumption also implies that when $|H_{t-1}| \leqslant M_\epsilon$ we have $|H_t| \leqslant |g_{h1}(H_{t-1})| + |g_{h2}(H_{t-1})|\,|Z_{1\,t}| \leqslant \overline{g}_h^\epsilon(1 + |Z_{1\,t}|)$, where $\overline{g}_h^\epsilon = \sup|g_{h1}(h)| \vee \sup|g_{h2}(h)| < \infty$ and the supremums are taken with respect to $h \in [-M_\epsilon, M_\epsilon] \cap \mathcal{H}$. Hence we have that

$$|H_t| \leqslant (a + b^\epsilon|Z_{1\,t}| + \epsilon)|H_{t-1}|\mathbb{1}_{\{|H_{t-1}|>M_\epsilon\}} + \overline{g}_h^\epsilon(1 + |Z_{1\,t}|)\mathbb{1}_{\{|H_{t-1}|\leqslant M_\epsilon\}}\,. \tag{2.19}$$

By A.2.2.2(i) we have $|f_{\theta\,t}| \leqslant \overline{\overline{\alpha_0}} + \overline{\alpha}_1|g_{y1}(H_{t-1})| + \overline{\alpha}_1|g_{y2}(H_{t-1})|\,|Z_{2\,t}| + \overline{\beta}_1|f_{\theta\,t-1}|$, where $\overline{\overline{\alpha_0}} = |\underline{\alpha}_0| \vee |\overline{\alpha}_0|$. Furthermore, it follows from A.2.2.1(i) and (ii) that

$$|f_{\theta\,t}| \leqslant \begin{cases} \overline{\alpha}_1 C_y(1 + \epsilon + |Z_{2\,t}|)\,|H_{t-1}| + \overline{\beta}_1|f_{\theta\,t-1}| & |H_{t-1}| > M_\epsilon \\ \overline{\overline{\alpha_0}} + \overline{\alpha}_1\overline{g}_y^\epsilon(1 + |Z_{2\,t}|) + \overline{\beta}_1|f_{\theta\,t-1}| & |H_{t-1}| \leqslant M_\epsilon \end{cases}, \tag{2.20}$$

where $C_y = C_{y1} \vee C_{y2}$, $\overline{g}_y^\epsilon = \sup |g_{y1}(h)| \vee \sup |g_{y2}(h)| < \infty$ and the supremums are taken with respect to $h \in [-M_\epsilon, M_\epsilon]$. We note that M_ϵ can be redefined if necessary in order to "remove" the constant $\overline{\overline{\alpha_0}}$ from the bound when $|H_{t-1}| > M_\epsilon$. Lastly, using analogous arguments we have

$$d_{\theta\, t} \leqslant \begin{cases} C_y(1 + \epsilon + |Z_{2\,t}|)\, |H_{t-1}| + |f_{\theta\, t-1}| + \overline{\beta}_1 |d_{\theta\, t-1}| \,, & |H_{t-1}| > M_\epsilon \\ 1 + \overline{g}_y^\epsilon (1 + |Z_{2\,t}|) + |f_{\theta\, t-1}| + \overline{\beta}_1 |d_{\theta\, t-1}| \,, & |H_{t-1}| \leqslant M_\epsilon \end{cases} \tag{2.21}$$

where again M_ϵ may be redefined if necessary in order to "remove" the constant 1 from the bound when $|H_{t-1}| > M_\epsilon$.

Second, we introduce a partition of the state space $\mathcal{X}$. Let $\kappa = (\kappa_h, \kappa_f, \kappa_d)' \in (0,1)^3$ where the specific choice of this vector will be determined in what follows. We define the sets

$$S_{2\,\epsilon} = \{(h, f, d) \in \mathcal{X} : \kappa_h |h| + \kappa_f |f| + \kappa_d |d| \leqslant M\} \quad \text{and} \quad S_{1\,\epsilon} = \mathcal{X} \backslash S_{2\,\epsilon}, \tag{2.22}$$

where M is a positive constant (note that in general $M \neq M_\epsilon$).

Third, let $\rho_{z\,\epsilon} = a + b^\epsilon |Z_{1\,t}| + \epsilon$, $C_{y,z}^\epsilon = C_y(1 + \epsilon + |Z_{2\,t}|)$ and define the matrix

$$\mathbf{C}_\epsilon(Z_t) = \begin{bmatrix} \rho_{z\,\epsilon} & 0 & 0 \\ \overline{\alpha}_1 C_{y,z}^\epsilon & \overline{\beta}_1 + \epsilon & 0 \\ C_{y,z}^\epsilon & 1 & \overline{\beta}_1 + \epsilon \end{bmatrix}.$$

A.2.2.1(iii) implies that $\mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m}\right)$ exists.

Finally, we set $\epsilon > 0$ in a way such that $\mathbb{E}(a + b^\epsilon |Z_{1\,t}| + \epsilon)^{2r_m} < 1$ and $\overline{\beta}_1 + \epsilon < 1$. A.2.2.1$(iv)$ and A.2.2.2 imply that an ϵ that satisfies these constraints exists. For

this particular choice of ϵ, we have that the spectral radius of $\mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m}\right)$ is less than one. Such a choice of ϵ will be assumed throughout.

Lemma 2.10.1 (Irreducibility)**.** *Consider the setting of Proposition 2.7.1. There exists an open rectangular region $D \subset \mathcal{X}$ that does not depend on θ or x such that the Markov chain $\{X_{\theta\, t}, t \geqslant 0\}$ is φ-irreducible with $\varphi(A) = \mu_{Leb}(A \cap D)$ for any $A \in \mathcal{B}(\mathcal{X})$.*

Proof. We follow the strategy of Lanne and Saikkonen (2005) and Meitz and Saikkonen (2008a). It suffices to show the following three intermediate results.

I. For any $x \in S_{1\,\epsilon}$ there exists an $n \in \mathbb{Z}_+$ such that $P_X^n(x, S_{2\,\epsilon}) > 0$.

II. For any $A \in \mathcal{B}(\mathcal{X})$ it holds that $\inf_{x \in S_{2\,\epsilon}} P_X^2(x, A \cap D) \geqslant c_* \mu_{Leb}(A \cap D)$, where D is an open rectangular region to be specified in what follows and c_* is a positive scalar. Both D and c_* do not depend on θ or x.

III. For any $x \in S_{1\,\epsilon}$ there exists an $n \in \mathbb{Z}_+$ such that for any $A \in \mathcal{B}(\mathcal{X})$ it holds that $P_X^{n+2}(x, A \cap D) > 0$ whenever $\mu_{Leb}(A \cap D) > 0$.

I. Define the event $\Omega_n = \{\omega \in \Omega : |Z_{1\,t}| \leqslant \mathbb{E}|Z_{1\,t}| \text{ and } |Z_{2\,t}| \leqslant \mathbb{E}|Z_{2\,t}|\,, t = 1, \ldots, n\}$ for an arbitrary n and note that $\mathbb{P}(\Omega_n) > 0$. Define the auxiliary vector $\dot{X}_{\theta\, t} = (|H_t|, |f_{\theta\, t}|, d_{\theta\, t})'$. To establish part I we show that for any $\kappa \in (0,1)^3$ and for each $t = 1, \ldots, n$ we have that when $X_{\theta\, t-1} \in S_{1\,\epsilon}$ the inequality

$$\left(\kappa' \dot{X}_{\theta\, t}\right)^{2r_m} \leqslant \left(\kappa^{\otimes 2r_m}\right)' \mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m}\right) \dot{X}_{\theta\, t-1}^{\otimes 2r_m} \tag{2.23}$$

holds given Ω_n. We distinguish the cases (i) $|H_{t-1}| > M_\epsilon$ and (ii) $|H_{t-1}| \leqslant M_\epsilon$.

(i) From (2.19), (2.20) and (2.21) we have that

$$\left(\kappa' \dot{X}_{\theta\, t}\right)^{2r_m} \leqslant \left(\kappa' \mathbf{C}_\epsilon(Z_t) \dot{X}_{\theta\, t-1}\right)^{2r_m} = \left(\kappa^{\otimes 2r_m}\right)' \mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m} \dot{X}_{\theta\, t-1}^{\otimes 2r_m}, \quad (2.24)$$

where the last equality follows from properties of Kronecker products. By adding and subtracting $\mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m}\right)$ in (2.24) we obtain

$$\begin{aligned}\left(\kappa' \dot{X}_{\theta\, t}\right)^{2r_m} &\leqslant \left(\kappa^{\otimes 2r_m}\right)' \mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m}\right) \dot{X}_{\theta\, t-1}^{\otimes 2r_m} \\ &\quad + \left(\kappa^{\otimes 2r_m}\right)' \left\{\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m} - \mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m}\right)\right\} \dot{X}_{\theta\, t-1}^{\otimes 2r_m}. \quad (2.25)\end{aligned}$$

The random elements of the matrix $\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m}$ are of the form $C(a + b^\epsilon |Z_{1\,t}| + \epsilon)^{j_1}(1 + \epsilon + |Z_{2\,t}|)^{j_2}$ for $j_1, j_2 \geqslant 0$ such that $j_1 + j_2 \leqslant 2r_m$, where C denotes some positive constant (that depends on ϵ). Conditionally on Ω_n, it follows from the independence between $Z_{1\,t}$ and $Z_{2\,t}$ as well as Jensen's inequality that the random elements are bounded from above by their expectations. This establishes that the bound in (2.23) holds in case (i).

(ii) From (2.19), (2.20) and (2.21) we can write

$$\left(\kappa' \dot{X}_{\theta\, t}\right)^{2r_m} \leqslant \left(\overline{C}_{z\,\epsilon} + \kappa_{\mathbf{f}}' \overline{B} \mathbf{f}_{\theta\, t-1}\right)^{2r_m},$$

where $\overline{C}_{z\,\epsilon} = \kappa_h \overline{H}_{z\,\epsilon} + \kappa_f(\overline{\overline{\alpha_0}} + \overline{\alpha}_1 \overline{Y}_{z\,\epsilon}) + \kappa_d(1 + \overline{Y}_{z\,\epsilon})$, $\overline{H}_{z\,\epsilon} = \overline{g}_h^\epsilon(1 + |Z_{1\,t}|)$, $\overline{Y}_{z\,\epsilon} = \overline{g}_y^\epsilon(1 + |Z_{2\,t}|)$, $\kappa_{\mathbf{f}} = (\kappa_f, \kappa_d)'$, $\mathbf{f}_{\theta\, t-1} = (|f_{\theta\, t-1}|, |d_{\theta\, t-1}|)'$ and $\overline{B}$ is a 2×2 lower triangular matrix with $\overline{B}_{11} = \overline{B}_{22} = \overline{\beta}_1$ and $\overline{B}_{21} = 1$. On the event Ω_n, it is straightforward to verify that $\overline{C}_{z\,\epsilon} \leqslant \left\|\overline{C}_{z\,\epsilon}\right\|_{L_1}$. Combining conditions $X_{\theta\, t-1} \in S_{1\,\epsilon}$ and $|H_{t-1}| \leqslant M_\epsilon$ we have $\kappa_{\mathbf{f}}' \mathbf{f}_{\theta\, t-1} > M - \kappa_h |H_{t-1}| \geqslant M - \kappa_h M_\epsilon$. Thus, we can choose M large enough such that $\epsilon\,(M - \kappa_h M_\epsilon) > \left\|\overline{C}_{z\,\epsilon}\right\|_{L_{2r_m}} \geqslant \left\|\overline{C}_{z\,\epsilon}\right\|_{L_1}$.

Note that this choice of M is independent of t. Such a choice of M is kept fixed throughtout our derivations.[13] Then, conditionally on Ω_n and whenever $X_{\theta\, t-1} \in S_{1\,\epsilon}$, we have

$$\left(\overline{C}_{z\,\epsilon} + \kappa_{\mathbf{f}}'\overline{B}\mathbf{f}_{\theta\, t-1}\right)^{2r_m} \leqslant \left(\left\|\overline{C}_{z\,\epsilon}\right\|_{L_{2r_m}} + \kappa_{\mathbf{f}}'\overline{B}\mathbf{f}_{\theta\, t-1}\right)^{2r_m} \leqslant \left(\kappa_{\mathbf{f}}'\overline{B}_{\epsilon}\mathbf{f}_{\theta\, t-1}\right)^{2r_m} ,$$

where $\overline{B}_{\epsilon} = \overline{B} + \epsilon I$. We note that $\kappa_{\mathbf{f}}'\overline{B}_{\epsilon}\mathbf{f}_{\theta\, t-1} = \kappa'\mathbf{C}_{\epsilon}(Z_t)\ddot{X}_{\theta\, t-1}$, where $\ddot{X}_{\theta\, t-1} = \left(0, \mathbf{f}_{\theta\, t-1}'\right)'$, and we have that

$$\left(\kappa'\mathbf{C}_{\epsilon}(Z_t)\ddot{X}_{\theta\, t-1}\right)^{2r_m} \leqslant (\kappa^{\otimes 2r_m})'\mathbb{E}\left(\mathbf{C}_{\epsilon}(Z_t)^{\otimes 2r_m}\right)\dot{X}_{\theta\, t-1}^{\otimes 2r_m} , \tag{2.26}$$

where we use the definition of $\ddot{X}_{\theta\, t-1}$. This establishes that the bound in (2.23) holds in case (ii). By Lemma A.2. of Ling and McAleer (2003) we can choose $\kappa \in (0,1)^3$ such that the vector $v = \left(I - \mathbb{E}\left(\mathbf{C}_{\epsilon}(Z_t)^{\otimes 2r_m}\right)\right)'\kappa^{\otimes 2r_m}$ has positive components.[14] In particular, we remark that the vector v does not depend on θ. We use $\underline{v}$ to denote the minimum of the components of v. Thus from (2.23) it follows that

$$\begin{aligned}
&\left(\kappa'\dot{X}_{\theta\, t}\right)^{2r_m} \\
&\leqslant \left(\kappa'\dot{X}_{\theta\, t-1}\right)^{2r_m} - v'\dot{X}_{t-1}^{\otimes 2r_m} = \left(\kappa'\dot{X}_{\theta\, t-1}\right)^{2r_m}\left(1 - \frac{v'\dot{X}_{\theta\, t-1}^{\otimes 2r_m}}{\left(\kappa^{\otimes 2r_m}\right)'\dot{X}_{\theta\, t-1}^{\otimes 2r_m}}\right) \\
&\leqslant (1-\underline{v})\left(\kappa'\dot{X}_{\theta\, t-1}\right)^{2r_m} ,
\end{aligned} \tag{2.27}$$

where $\underline{v} \in (0,1)$. By repeated application of (2.27) starting from $X_{\theta\, 0} = x \in S_{1\,\epsilon}$ we have $\left(\kappa'\dot{X}_{\theta\, n}\right)^{2r_m} \leqslant (1-\underline{v})^n\left(\kappa'\dot{x}\right)^{2r_m}$. Since $\underline{v} \in (0,1)$ we have that for

[13] We remark that A.2.2.1 implies that $\left\|\overline{C}_{z\,\epsilon}\right\|_{L_{2r_m}}$ exists.

[14] Recall that the matrix $\mathbb{E}\left(\mathbf{C}_{\epsilon}(Z_t)^{\otimes 2r_m}\right)$ has a spectral radius that is strictly less than unity. As noted by Lanne and Saikkonen (2005), the given proof makes clear that it means no loss of generality to assume that the components of κ are bounded by unity.

any $x \in S_{1\epsilon}$ there exists a sufficiently large n such that the right hand side of the inequality is smaller than M^{2r_m}. Thus, we have that $X_{\theta n} \in S_{2\epsilon}$ with positive probability.

II. First we write $P_X^2(x, A \cap D) = \mathbb{E}(\ \mathbb{E}(\mathbb{1}_{\{X_{\theta 2} \in A \cap D\}} | H_1, X_{\theta 0})\ |\ X_{\theta 0} = x\ )$, for any $x \in S_{2\epsilon}$, $A \in \mathcal{B}(\mathcal{X})$, $D \subset \mathcal{X}$ such that D is an open rectangular region (to be specified in what follows). Let $\underline{h}_1 \geqslant \sup_{|h|<M/\kappa_h} g_{h1}(h)$. The result is obtained by showing the following intermediate results. (i) $\inf_{h_1 \in [\underline{h}_1, \underline{h}_1+1]} \mathbb{E}(\mathbb{1}_{\{X_{\theta 2} \in A \cap D\}} | H_1 = h_1, X_{\theta 0} = x) \geqslant c' \mu_{Leb}(A \cap D)$, where c' is a positive scalar that does not depend on θ or x. (ii) $P_X^2(x, A \cap D) \geqslant c'' \inf_{h_1 \in [\underline{h}_1, \underline{h}_1+1]} \mathbb{E}(\mathbb{1}_{\{X_{\theta 2} \in A \cap D\}} | H_1 = h_1, X_{\theta 0} = x)$, where c'' is a positive scalar that does not depend on θ or x.

(i) Set $\underline{Z}_{21} = \sup_{|h|<M/\kappa_h} (R - g_{y1}(h))/g_{y2}(h)$ and

$\underline{Z}_{22} = \sup_{h_1 \in [\underline{h}_1, \underline{h}_1+1]} (R - g_{y1}(h_1))/g_{y2}(h_1)$, where $R > \frac{\overline{\beta}_1 M/\kappa_f - \underline{\alpha}_0}{\underline{\alpha}_1} \vee r_K \vee \sup_{|h|<M/\kappa_h} g_{y1}(h) \vee \sup_{h_1 \in [\underline{h}_1, \underline{h}_1+1]} g_{y1}(h_1)$, and note that $\underline{Z}_{21}$ and $\underline{Z}_{22}$ do not depend on x, h_1 or θ. Then it holds that

$$\begin{aligned} &\inf_{h_1 \in [\underline{h}_1, \underline{h}_1+1]} \mathbb{E}(\mathbb{1}_{\{X_{\theta 2} \in A \cap D\}} | H_1 = h_1, X_{\theta 0} = x) \\ &\quad \geqslant \inf_{h_1 \in [\underline{h}_1, \underline{h}_1+1]} \int_0^\infty \int_{\underline{Z}_{21}}^\infty \int_{\underline{Z}_{22}}^\infty \mathbb{1}_{\{X_{\theta 2} \in A \cap D\}} \phi_H(Z_{12}) \phi_Y(Z_{21}) \phi_Y(Z_{22}) dZ_{12} dZ_{21} dZ_{22}\ . \end{aligned}$$

Over the integration range of the right hand side a number of properties hold. First, we have that $g_{y1}(h) + g_{y2}(h) Z_{21} > R$ and $g_{y1}(h_1) + g_{y2}(h_1) Z_{22} > R$, and $f_{\theta 1} > 0$. Furthermore, we have that the map between $X_{\theta 2}$ and $(Z_{12}, Z_{22}, Z_{21})'$ is linear and is given by

$$X_{\theta 2} = c + G \begin{bmatrix} Z_{12} & Z_{22} & Z_{21} \end{bmatrix}'\ , \tag{2.28}$$

where the expression for the vector $c = (c_h, c_f, c_d)'$ and the 3×3 block-diagonal matrix G are given in (2.34) and (2.35). Furthermore, G is invertible uniformly over $S_{2\,\epsilon}$, Θ and $[\underline{h}_1, \underline{h}_1 + 1]$. In fact, we have that $\det G \in [G_l, G_u]$, and it is clear that $G_l > 0$ by A.2.2.1 and A.2.2.2.[15] Define $Z_{1\,2}(X_{\theta\,2}) = \frac{H_2 - c_h}{g_{h2}(h_1)}$, and

$$\begin{aligned} Z_{2\,2}(X_{\theta\,2}) &= \frac{g_{h2}(h_1) g_{y2}(h) [\alpha_{1\,K} + \overline{\beta}_1)(f_{\theta\,2} - c_f) - \alpha_{1\,K}\beta_{1\,K}(d_{\theta\,2} - c_d)]}{\det G} \\ Z_{2\,1}(X_{\theta\,2}) &= \frac{g_{h2}(h_1) g_{y2}(h_1) [\alpha_{1\,K}(d_{\theta\,2} - c_d) - (f_{\theta\,2} - c_f)]}{\det G} \,. \end{aligned}$$

We observe that the constraints $Z_{2\,1}(X_{\theta\,2}) > \underline{Z}_{1\,2}$ and $Z_{2\,2}(X_{\theta\,2}) > \underline{Z}_{2\,2}$ impose upper and lower bounds on $d_{\theta\,2}$ which are linear functions of $f_{\theta\,2}$ with positive slopes. In fact, from A.2.2.2 we have that the minimum discrepancy between slopes is given by $\inf_{\theta \in \Theta} \frac{\alpha_{1\,K} + \overline{\beta}_1}{\alpha_{1\,K}\beta_{1\,K}} - \frac{1}{\alpha_{1\,K}} = \inf_{\theta \in \Theta} \frac{\alpha_{1\,K} + \overline{\beta}_1 - \beta_{1\,K}}{\alpha_{1\,K}\beta_{1\,K}} = \frac{1}{\overline{\beta}_1} > 0$. It follows that the intersection of images of the map defined in (2.28) with respect to $\theta \in \Theta$, $x \in S_{2\,\epsilon}$ and $h_1 \in [\underline{h}_1, \underline{h}_1 + 1]$ contains the following set[16]

$$\{X_{\theta\,2} \in \mathcal{X} : H_2 > \underline{H_2}, f_{\theta\,2} > \underline{f}_{2}, \underline{d}_2(f_{\theta\,2}) < d_{\theta\,2} < \overline{d}_2(f_{\theta\,2})\} \,. \tag{2.29}$$

We remark that A.2.2.2 implies that such a set is non-empty and it contains sets of positive Lebesgue measure. Thus, we can pick D as an open rectangular region in the intersection of (2.29) and $S_{1\,\epsilon}$. Clearly, D does not depend on θ, x or h_1. Next, by the change of variable theorem we obtain that

$$\begin{aligned} &\inf_{h_1 \in [\underline{h}_1, \underline{h}_1 + 1]} \int_0^\infty \int_{\underline{Z}_{2\,1}}^\infty \int_{\underline{Z}_{2\,2}}^\infty \mathbb{1}_{\{X_{\theta\,2} \in A \cap D\}} \phi_H(Z_{1\,2}) \phi_Y(Z_{2\,1}) \phi_Y(Z_{2\,2}) dZ_{1\,2} dZ_{2\,1} dZ_{2\,2} \\ &\quad \geqslant \inf_{\substack{x \in S_{2\,\epsilon} \\ h_1 \in [\underline{h}_1, \underline{h}_1 + 1] \\ \theta \in \Theta \\ X_{\theta\,2} \in A \cap D}} \det G^{-1} \phi_H(Z_{1\,2}(X_{\theta\,2})) \phi_Y(Z_{2\,1}(X_{\theta\,2})) \phi_Y(Z_{2\,2}(X_{\theta\,2})) \int_{A \cap D} dX_{\theta\,2}. \end{aligned}$$

[15] See (2.36) and (2.37) for the expression for G_l and G_u.

[16] See (2.38), (2.39),(2.40) and (2.41) for the expressions for $\underline{H}_2$, $\underline{f}_2$, $\underline{d}_2(f_{\theta\,2})$ and $\overline{d}_2(f_{\theta\,2})$.

The boundedness conditions on $g_{h2}, g_{y1}, g_{y2}, \phi_H$ and ϕ_Y imply that

$$\inf_{\substack{x\in S_{2\,\epsilon}\\ h_1\in[\underline{h}_1,\underline{h}_1+1]\\ \theta\in\Theta\\ X_{\theta\,2}\in A\cap D}} \det G^{-1}\phi_H(Z_{1\,2}(X_{\theta\,2}))\phi_Y(Z_{2\,1}(X_{\theta\,2}))\phi_Y(Z_{2\,2}(X_{\theta\,2})) \geqslant c' > 0\ ,$$

where c' does not depend on θ, x or h_1. The claim of part (i) then follows.

(ii) We have that

$$\begin{aligned} P_X^2(x, A\cap D) &\geqslant \int_0^\infty \mathbb{E}(\mathbb{1}_{\{X_{\theta\,2}\in A\cap D\}}|H_1 = h_1, X_{\theta\,0} = x)\phi_H(Z_{1\,1})dZ_{1\,1} \\ &\geqslant \inf_{h_1\in[\underline{h}_1,\underline{h}_1+1]} \mathbb{E}(\mathbb{1}_{\{X_{\theta\,2}\in A\cap D\}}|H_1 = h_1, X_{\theta\,0} = x) \\ &\quad\times \int_{\underline{h}_1}^{\underline{h}_1+1} \frac{1}{g_{h2}(h)}\phi_H\left(\frac{h_1 - g_{h1}(h)}{g_{h2}(h)}\right) dh_1\ , \end{aligned}$$

where the last inequality follows by the choice of $\underline{h}_1$, and $g_{h2}(h)$ is strictly positive by assumption. Moreover, the boundedness conditions on g_{h1}, g_{h2}, and ϕ_H imply that $\inf_{|h|<M/\kappa_h} g_{h2}(h)^{-1}\phi_H\left(g_{h2}(h)^{-1}(h_1 - g_{h1}(h))\right) \geqslant c' > 0$, where we emphasize that c' does not depend on x or θ. This concludes the second part. Combining parts (i) and (ii), the result in II holds with $c_* = c'c'' > 0$.

III. The Chapman-Kolmogorov equations imply that for any $x \in S_{1\,\epsilon}$ it holds that

$$\begin{aligned} P_X^{n+2}(x, A\cap D) &\geqslant \textstyle\int_{S_{2\,\epsilon}} P_X^n(x, dx_n)P_X^2(x_n, A\cap D) \\ &\geqslant c_*\mu_{Leb}(A\cap D)P_X^n(x, S_{2\,\epsilon}) > 0, \end{aligned}$$

where the last two inequalities follow from Parts I and II (for a sufficiently large n). □

Lemma 2.10.2 (Aperiodicity)**.** *Consider the setting of Lemma 2.7.1. Then, the Markov chain $\{X_{\theta\,t}, t \geqslant 0\}$ is aperiodic.*

Proof of Lemma 2.10.2. It follows from Proposition A1.1 in Chan (1990), that to establish aperiodicity it suffices to show that that for each $x \in D$ there exists an $n \in \mathbb{Z}_+$ such that $P_X^{n+2}(x, D) > 0$ and $P_X^{n+3}(x, D) > 0$, where D is a small set. We divide the proof in three parts. In part (i) we show that the set D defined in Lemma 2.10.1 is a small set. In part (ii) we show that for each $x \in D$ there exists an n such that $P_X^{n+2}(x, D) > 0$. In part (iii) we show that for the same x and same n defined in part (ii) it holds that $P_X^{n+3}(x, D) > 0$.

(i) We note that by repeating the arguments in Part II of Lemma 2.10.1 with $S_{2\epsilon}$ replaced by D we have that for any $A \in \mathcal{B}(\mathcal{X})$ there exist $c'_* > 0$ and an open rectangular region D' such that $\inf_{x \in D} P_X^2(x, A \cap D') \geqslant c'_* \mu_{Leb}(A \cap D')$.

(ii) It follows from Parts I and II of Lemma 2.10.1 that for any $x \in D$ there exists an n such that $P_X^n(x, S_{2\epsilon}) > 0$ and for any $x \in S_{2\epsilon}$ we have that $P_X^2(x, D) > 0$. The claim is implied by the Chapman-Kolmogorov equation.

(iii) Note that in the proof of Lemma 2.10.1 we can choose an M and M' with $M > M'$ such that $P_X^n(x, S_{2\epsilon}) > 0$ and $P_X^{n+1}(x, S'_{2\epsilon}) > 0$ where $S'_{2\epsilon} = \{(h, f, d)' \in \mathcal{X} : \kappa_h|h| + \kappa_f|f| + \kappa_d|d| \leqslant M'\}$. It is straightforward to see in the proof of Lemma 2.10.1 that M can be chosen as any sufficiently large constant. Furthermore, we have that $\inf_{x \in S'_{2\epsilon}} P_X^2(x, D) \geqslant \inf_{x \in S_{2\epsilon}} P_X^2(x, D) \geqslant c_* \mu_{Leb}(D) > 0$. The Chapman-Kolmogorov equation gives the claim as

$$P_X^{n+3}(x, D) \geqslant \int_{S'_{2\epsilon}} P_X^{n+1}(x, dx_{n+1}) P_X^2(x_{n+1}, D) \geqslant c_* \mu_{Leb}(D) P_X^{n+1}(x, S'_{2\epsilon}) > 0.$$

□

Lemma 2.10.3. *Consider the setting of Lemma 2.7.1. Then, the Markov chain*

$\{X_{\theta\, t}, t \geqslant 0\}$ *satisfies* $\mathbb{E}\left(q_X(X_{\theta\, t})|X_{\theta\, t-1} = x\right) \leqslant (1-\gamma_1)q_X(x) + \gamma_2 \mathbb{1}_{\{x\in C\}}$ *for some* $\gamma_1 > 0$ *and* $\gamma_2 < \infty$ *where* C *is a small set. Furthermore,* γ_1*,* γ_2 *and* C *do not depend on* θ*.*

Proof of Lemma 2.10.3. Set C equal to $S_{2\,\epsilon}$ and note that Part II in the proof of Lemma 2.10.1 establishes that $S_{2\,\epsilon}$ is a small set that does not depend on θ. When $x \in S_{1\,\epsilon}$, we distinguish two cases: (i) $|h| > M_\epsilon$ or (ii) $|h| \leqslant M_\epsilon$.

Case (i). From (2.24) we have that

$$\mathbb{E}_x(q_X(X_{\theta\, t})) - 1 \leqslant \mathbb{E}_x\left(\kappa' \mathbf{C}_\epsilon(Z_t)\dot{x}\right)^{2r_m} = \left(\kappa^{\otimes 2r_m}\right)' \mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m}\right) \dot{x}^{\otimes 2r_m}\,. \tag{2.30}$$

Following steps analogous to the ones used to go from (2.23) to (2.27) we have that

$$\mathbb{E}_x\left(q_X(X_{\theta\, t})|X_{\theta\, t-1} = x\right) \leqslant 1 + \left(\kappa'\dot{x}\right)^{2r_m} - v'\dot{x}^{\otimes 2r_m} \leqslant (1-\gamma_1)q_X(x)\,,$$

where $\gamma_1 \in (0,1)$ and does not depend θ.

Case (ii). From Part I of Lemma 2.10.1 (case (ii)) it follows that $\mathbb{E}_x\left(q_X(X_{\theta\, t})\right) - 1 \leqslant \mathbb{E}_x\left(\overline{C}_{z\,\epsilon} + \kappa_{\mathbf{f}}'\overline{B}\mathbf{f}\right)^{2r_m}$. We observe that

$$\begin{aligned}\mathbb{E}_x\left(\overline{C}_{z\,\epsilon} + \kappa_{\mathbf{f}}'\overline{B}\mathbf{f}\right)^{2r_m} &= \left(\left(\mathbb{E}_x\left(\overline{C}_{z\,\epsilon} + \kappa_{\mathbf{f}}'\overline{B}\mathbf{f}\right)^{2r_m}\right)^{\frac{1}{2r_m}}\right)^{2r_m} \\ &\leqslant \left(\left\|\overline{C}_{z\,\epsilon}\right\|_{L_{2r_m}} + \kappa_{\mathbf{f}}'\overline{B}\mathbf{f}\right)^{2r_m}\,,\end{aligned}$$

and note that the assumptions on the innovations imply that $\left\|\overline{C}_{z\,\epsilon}\right\|_{L_{2r_m}}$ exists. Using steps analogous to those used to get to (2.26) we have that

$$\begin{aligned}\left(\left\|\overline{C}_{z\,\epsilon}\right\|_{L_{2r_m}} + \kappa_{\mathbf{f}}'\overline{B}\mathbf{f}\right)^{2r_m} &\leqslant \mathbb{E}_x\left(\kappa' \mathbf{C}_\epsilon(Z_t)\ddot{x}\right)^{2r_m} \\ &\leqslant (\kappa^{\otimes 2r_m})'\mathbb{E}\left(\mathbf{C}_\epsilon(Z_t)^{\otimes 2r_m}\right)\dot{x}^{\otimes 2r_m}\,.\end{aligned}$$

The claim of case (ii) then follows using the same steps of case (i) after equation (2.30). When $x \in S_{2\,\epsilon}$ it follows from A.2.2.1 and the definition of $S_{2\,\epsilon}$ that $\sup_{\substack{x \in S_{2\,\epsilon} \\ \theta \in \Theta}} \mathbb{E}\left(q_X(X_{\theta\,t}) | X_{\theta\,t-1} = x\right) \leqslant \gamma_2 < \infty$, where we have used the fact the expectation exists and it is bounded over Θ for every $x \in S_{2\,\epsilon}$ provided that $Z_{1\,t}$ and $Z_{2\,t}$ have $2r_m$ moments. Since $(1-\gamma_1)q_X(x)$ is positive the claim holds when $x \in S_{2\,\epsilon}$. □

2.11 Recursive prediction as a solution of a sequential optimization problem.

The class of algorithms defined in (2.3) was introduced without any justification other than its close connection to standard models used in the literature. In this section we show that this class of algorithms may be motivated as the solution of a sequential optimization problem. The analysis is inspired by the research by Creal *et al.* (2013) and Harvey (2013) on GAS/DCS models and by Gijbels *et al.* (1999); Harvey and Chakravarty (2008) on the relation between nonparametric estimators and time series models.[17]

Let $\{f_t, t \geqslant 0\}$ be defined as $f_0 = f \in \mathcal{F}$ and

$$f_t = \arg \min_{f \in \text{int}(\mathcal{S})} Q_t(f) \,, \tag{2.31}$$

[17]The 1-step-ahead prediction formula implied by GAS/DCS models is sometimes motivated as the approximate solution of a local estimation problem based on a generic (and sufficiently regular) likelihood function. The class of algorithms we introduce for 1-step-ahead prediction can be interpreted as the exact solution of a local estimation problem based on a Bregman loss.

where Q_t is the tracking error function defined as

$$Q_t(f) = w_1 L(\bar{f}, f) + w_2 L(Y_{t-1}, f) + w_3 L(f_{t-1}, f) \, ,$$

where L denotes a loss in the Bregman family and $(w_1, w_2, w_3) = w \in \Delta^3$ with Δ^3 denoting the 3-dimensional simplex. The tracking error is a convex combination of the divergences with respect to the constant $\bar{f}$, the previous observation and the previous forecast. If $f_0 = \bar{f}$,[18] it is straightforward to verify that

$$Q_t(f) \propto \sum_{i=0}^{t-1} k\left(\frac{x_t - x_{t-i}}{h}\right) L\left(Y_{t-i-1}, f\right) + \lambda L(\bar{f}, f) \, , \tag{2.32}$$

where $\{x_t, t \geqslant 0\}$ is a deterministic sequence defined as $x_t = t$ for each $t \geqslant 0$, $k(u) = \exp(u)\mathbb{1}_{\{u \leqslant 0\}}$, $h = 1/\ln(w_3)$ and $\lambda = w_2^{-1} - \sum_{i=1}^{t} w_3^{i-1}$. Thus, the tracking error can equivalently be thought of as the objective function of a local constant regression plus a regularization term that penalizes deviations from the constant $\bar{f}$. The solution of this optimization problem is

$$f_t = w_1 \bar{f} + w_2 Y_{t-1} + w_3 f_{t-1} \, , \tag{2.33}$$

which coincides with the class of algorithms in (2.3) provided that $\theta = (\alpha_0, \alpha_1, \beta_1)'$ with $\alpha_0 = w_1 \bar{f}$, $\alpha_1 = w_2$ and $\beta_1 = w_3$. Note that empirical risk minimization may be interpreted as choosing the set of weights w and the constant $\bar{f}$ in the objective function Q_t that minimize the in-sample empirical prediction loss.

[18] We remark that the choice $f_0 = \bar{f}$ is made only for expository purposes, as it simplifies the notation in (2.32). This would imply that the initial value for the forecast process is determined by empirical risk minimization, which we do not cover in our framework.

2.12 Detailed Computations of Lemma B.1/Part II

Over the integration range described in Part II of Lemma B.1 and conditionally on H_1, we have that $X_{\theta 2} = c + G[Z_{12}, Z_{22}, Z_{21}]'$, where $c = (c_h, c_f, c_d)'$,

$$c = \begin{bmatrix} g_{h1}(h_1) \\ \alpha_{0K}(1+\beta_{1K}) + \beta_{1K}^2 f + \alpha_{1K} g_{y1}(h_1) + \alpha_{1K}\beta_{1K} g_{y1}(h) \\ (1 + \alpha_{0K} + \beta_{1K} f + \overline{\beta}_1 + \overline{\beta}_1 f) + \overline{\beta}_1^2 d + g_{y1}(h_1) + (\alpha_{1K} + \overline{\beta}_1) g_{y1}(h) \end{bmatrix}, \tag{2.34}$$

$$G = \begin{bmatrix} g_{h2}(h_1) & 0 & 0 \\ 0 & \alpha_{1K} g_{y2}(h_1) & \alpha_{1K}\beta_{1K} g_{y2}(h) \\ 0 & g_{y2}(h_1) & (\alpha_{1K} + \overline{\beta}_1) g_{y2}(h) \end{bmatrix}. \tag{2.35}$$

Moreover, we have that $\det G \in [G_l, G_u]$, where

$$G_l = \underline{\alpha}_1^2 \inf_{\substack{x \in S_{2\epsilon} \\ h_1 \in [\underline{h}_1, \underline{h}_1 + 1]}} g_{h2}(h_1) g_{y2}(h) g_{y2}(h_1) \tag{2.36}$$

$$G_u = \overline{\alpha}_1(\overline{\alpha}_1 + \overline{\beta}_1) \sup_{\substack{x \in S_{2\epsilon} \\ h_1 \in [\underline{h}_1, \underline{h}_1 + 1]}} g_{h2}(h_1) g_{y2}(h) g_{y2}(h_1)\,. \tag{2.37}$$

The set described in (2.29) is determined by the following quantities:

$$\underline{H}_2 = \sup c_h \tag{2.38}$$

$$\underline{f}_2 = \sup c_f + \underline{Z}_{22} \sup G_{22} + \underline{Z}_{21} \sup G_{23} \tag{2.39}$$

$$\underline{d}_2(f_{\theta 2}) = \frac{1}{\underline{\alpha}_1} f_{\theta 2} + \sup \left[c_d - \frac{1}{\alpha_{1K}} \left(c_f - \frac{\det G}{g_{h2}(h_1) g_{y2}(h_1)} \underline{Z}_{21} \right) \right] \tag{2.40}$$

$$\begin{aligned} \overline{d}_2(f_{\theta 2}) = {} & \left(\frac{1}{\underline{\alpha}_1} + \frac{1}{\overline{\beta}_1} \right) f_{\theta 2} + \\ & \inf \left[c_d - \frac{1}{\alpha_{1K}\beta_{1K}} \left((\alpha_{1K} + \overline{\beta}_1) c_f - \frac{\det G}{g_{h2}(h_1) g_{y2}(h)} \underline{Z}_{22} \right) \right], \end{aligned} \tag{2.41}$$

and $\sup$ and $\inf$ are taken with respect to $\theta \in \Theta$, $x \in S_{2\epsilon}$, and $h_1 \in [\underline{h}_1, \underline{h}_1 + 1]$.

2.13 Auxiliary Results

Lemma 2.13.1. *Suppose A.2.2.1 and A.2.2.2 hold. Then,*

(I) If $\mathcal{Y} = \mathbb{R}_+$, all losses listed in Table 2.2 satisfy Condition 2.2.1.

(II) If $\mathcal{Y} = \mathbb{R}$, only the losses in Table 2.2 with $\mathcal{S} = \mathbb{R}$ satisfy Condition 2.2.1.

Proof.

Condition 2.2.1(i). It is obvious that Condition 2.2.1(i) holds when $\mathcal{Y} = \mathbb{R}_+$. If $\mathcal{Y} = \mathbb{R}$, the losses with $\mathcal{S} \subseteq \mathbb{R}_+$ do not satisfy the condition because by A.2.2.1(i), (ii), and (iii) we have

$$\begin{aligned}
&\mathbb{P}(Y_1 < 0) \\
&\geqslant \int_{-2}^{-1} \int_{0}^{1} \frac{1}{g_{y2}(h_1)} \phi_Y \left(\frac{y_1 - g_{y1}(h_1)}{g_{y2}(h_1)} \right) \frac{1}{g_{h2}(h)} \phi_H \left(\frac{h_1 - g_{h1}(h)}{g_{h2}(h)} \right) dy_1 dh_1 \\
&\geqslant \inf_{\substack{y_1 \in [-2,-1] \\ h_1 \in [0,1]}} \frac{1}{g_{y2}(h_1)} \phi_Y \left(\frac{y_1 - g_{y1}(h_1)}{g_{y2}(h_1)} \right) \frac{1}{g_{h2}(h)} \phi_H \left(\frac{h_1 - g_{h1}(h)}{g_{h2}(h)} \right) > 0.
\end{aligned}$$

Condition 2.2.1(ii). The case $\psi(u) = u^2$ is obvious. For the case $\psi(u) = -\log(u)$, we can write $L(f_{\theta_1 t}, f_{\theta_2 t}) = \frac{f_{\theta_1 t}}{f_{\theta_2 t}} - \log \frac{f_{\theta_1 t}}{f_{\theta_2 t}} - 1 \leqslant \frac{1}{2\underline{\alpha_0}^2} \left(f_{\theta_1 t} - f_{\theta_2 t} \right)^2$, which follows by A.2.2.2 and Taylor's Remainder formula. By the same arguments, cases $\psi(u) = u \log u - u$, $\psi(u) = u \log \frac{u}{1+u} - \log(1+u)$ and $\psi(u) = u \tan^{-1}(u) - \frac{1}{2} \log(1+u^2)$ hold with $C_\psi = \frac{1}{2\underline{\alpha}_0}$, $C_\psi = \frac{1}{2\underline{\alpha}_0(1+\underline{\alpha}_0)}$ and $C_\psi = \frac{1}{2}$, respectively.

Condition 2.2.1(iii). It is understood that $Y_t \in \mathcal{S}$ a.s. for this part of the proof.

(a) $\psi(u) = u^2$: We have $\|L(Y_t, f_{\theta t})\|_{L_{r_m}} = \|Y_t - f_{\theta t}\|_{L_{2r_m}} \leqslant \|Y_t\|_{L_{2r_m}} + \|f_{\theta t}\|_{L_{2r_m}}$. The claim follows after taking the supremum over Θ on both sides

of the inequality and Proposition 2.7.1.

(b) $\psi(u) = u\tan^{-1}(u) - \frac{1}{2}\log(1+u^2)$: We have $\|L(Y_t, f_{\theta t})\|_{L_{r_m}} \leqslant (1/2 + \pi)\|Y_t\|_{L_{r_m}} + (1/2)\|f_{\theta t}\|_{L_{r_m}}$, since the range of $\tan^{-1}(u)$ is $(-\pi/2, \pi/2)$ and $\log(1+u^2) \leqslant |u|$. The result follows by taking the supremum on both sides and Proposition 2.7.1.

(c) $\psi(u) = -\log(u)$: We have $\|L(Y_t, f_{\theta t})\|_{L_{r_m}} \leqslant \underline{\alpha}_0^{-1}\|Y_t\|_{L_{r_m}} + \|\log Y_t\|_{L_{r_m}} + \|\log f_{\theta t}\|_{L_{r_m}} + 1$ by A.2.2.2. Note that $\|\log Y_t\|_{L_{r_m}} \leqslant \|\log Y_t \cdot \mathbb{1}_{\{Y_t \leqslant 1\}}\|_{L_{r_m}} + \|Y_t\|_{L_{r_m}}$, where $\|Y_t\|_{L_{r_m}} < \infty$ by Proposition 2.7.1. Now,

$$\begin{aligned}
&\left\|\log Y_t \cdot \mathbb{1}_{\{Y_t \leqslant 1\}}\right\|_{L_{r_m}} \\
&\overset{A.2.2.1(ii)}{\leqslant} \left\|\log g_{y2}(H_t) \cdot \mathbb{1}_{\{Y_t \leqslant 1\}}\right\|_{L_{r_m}} + \left\|\log \epsilon_{Y\,t} \cdot \mathbb{1}_{\{Y_t \leqslant 1\}}\right\|_{L_{r_m}} \\
&\leqslant \|\log g_{y2}(H_t)\|_{L_{r_m}} + \|\log \epsilon_{Y\,t}\|_{L_{r_m}} \ ,
\end{aligned}$$

where $\|\log \epsilon_{Y\,t}\|_{L_{r_m}} < \infty$ by A.2.2.1(iii) and

$$\|\log g_{y2}(H_t)\|_{L_{r_m}} \leqslant \log\left(\inf_h g_{y2}(h)\right)^{-1} + C_{y2}\|H_t\|_{L_{r_m}} < \infty$$

by A.2.2.1(ii) and Proposition 2.7.1, where $\inf_h g_{y2}(h) \leqslant 1$ without loss of generality. Similarly, for every $\theta \in \Theta$ we can write $\|\log f_{\theta t}\|_{L_{r_m}} \leqslant \log \underline{\alpha}_0^{-1} + \left\|\log f_{\theta t}\mathbb{1}_{\{f_{\theta t} > 1\}}\right\|_{L_{r_m}} \leqslant \log \underline{\alpha}_0^{-1} + \|f_{\theta t}\|_{L_{r_m}}$, where we have assumed $\underline{\alpha_0} \leqslant 1$ without loss of generality. The claim follows after taking the supremum over Θ on both sides of the inequality and Proposition 2.7.1.

(d) $\psi(u) = u\log(u) - u$: We have $\|L(Y_t, f_{\theta t})\|_{L_{r_m}} \leqslant \|Y_t\|_{L_{2r_m}}(\|\log Y_t\|_{L_{2r_m}} + \|\log f_{\theta t}\|_{L_{2r_m}} + 1) + \|f_{\theta t}\|_{L_{r_m}}$, and the claim follows by the arguments applied in case (c).

(e) $\psi(u) = u \log \frac{u}{1+u} - \log(1+u)$: We have $\|L(Y_t, f_{\theta\, t})\|_{L_{rm}} \leqslant Y_t(\|\log Y_t\|_{L_{rm}} + \|\log f_{\theta\, t}\|_{L_{rm}}) + \|(1+Y_t)(Y_t + f_{\theta\, t})\|_{L_{rm}}$, and the claim follows by arguments similar to those used in case (d).

□

Lemma 2.13.2. *Consider the setup given in the proof of Lemma 2.7.2. Let*

$$\mathbf{q}_X(x, x^G) = \frac{\tilde{q}_X(x) + \tilde{q}_X(x^G)}{2}\,, \quad \textit{and} \quad \overline{R} = (1-\tilde{c}_*)^{-2}\tilde{\gamma}_2(2-\tilde{\gamma}_1)\,.$$

Then,

$$\begin{aligned}&\int_{\mathcal{X}}\int_{\mathcal{X}} \mathbf{q}_X(x_2, x_2^G)\left[P_X^2(x, dx_2) - \tilde{c}_*\varphi(dx_2)\right]\left[P_X^2(x^G, dx_2^G) - \tilde{c}_*\varphi(dx_2^G)\right]\\ &\leqslant \overline{R}(1-\tilde{c}_*)^2\end{aligned}$$

for all $(x, x^G) \in \tilde{S}_{2\,\epsilon} \times \tilde{S}_{2\,\epsilon}$.

Proof. We note that $\tilde{c}_*$ is a strictly positive constant that does not depend on x_2 nor x_2^G. Hence by the symmetry of $\mathbf{q}_X$ we may focus on bounding

$$\int_{\mathcal{X}}\int_{\mathcal{X}} \tilde{q}_X(x_2)\left[P_X^2(x, dx_2) - \tilde{c}_*\varphi(dx_2)\right]\left[P_X^2(x^G, dx_2^G) - \tilde{c}_*\varphi(dx_2^G)\right]\,. \quad (2.42)$$

Using the fact that $\{X_t^G, t \geqslant 0\}$ is an independent copy of $\{X_t, t \geqslant 0\}$, we can re-write (2.42) as

$$\begin{aligned}&\int_{\mathcal{X}} \tilde{q}_X(x_2)\left[P_X^2(x, dx_2) - \tilde{c}_*\varphi(dx_2)\right]\int_{\mathcal{X}}\left[P_X^2(x^G, dx_2^G) - \tilde{c}_*\varphi(dx_2^G)\right]\\ &\quad\leqslant \int_{\mathcal{X}} \tilde{q}_X(x_2)\left[P_X^2(x, dx_2) - \tilde{c}_*\varphi(dx_2)\right] \leqslant \mathbb{E}(\tilde{q}_X(X_2)|X_0 = x)\,,\end{aligned}$$

where we use $\tilde{c}_* > 0, \varphi(\cdot) \geqslant 0, P_X^2(x,\cdot) \geqslant \tilde{c}_* \varphi(\cdot)$, and $\int_{\mathcal{X}} P_X^2(x, dx_2) = 1$. We now write

$$\begin{aligned}\mathbb{E}(\tilde{q}_X(X_2)|X_0 = x) &\leqslant (1-\tilde{\gamma}_1)\mathbb{E}(\tilde{q}_X(X_1)|X_0 = x) + \tilde{\gamma}_2 \mathbb{E}(\mathbb{1}_{\{X_1 \in \tilde{S}_{2\epsilon}\}}|X_0 = x) \\ &\leqslant (1-\tilde{\gamma}_1)\tilde{\gamma}_2 + \tilde{\gamma}_2 = \tilde{\gamma}_2(2-\tilde{\gamma}_1)\,,\end{aligned}$$

which follows by the law of iterated expectations and by repeated application of the drift criterion, where the last inequality uses $x \in \tilde{S}_{2\epsilon}$. Applying the same arguments for the second term, the claim follows. □

Lemma 2.13.3. *Consider the same setup of Theorem 2.3.1. Let* $\Theta_i = \{\theta \in \mathbb{R}^p : \|\theta - \theta_i\|_2 \leqslant \delta\}$ *with* $\Theta_i \in \Theta$ *for* $i = 1, \ldots, N_\delta$ *denote a* δ*-covering of the set* Θ *for any* $\delta \in (0, 1]$. *Define the function* $U_{\theta t} = L(Y_t, f_{\theta t})$ *and let* $U_{it} = U_{\theta_i t}$.

Then (i) *we have that* $\sup_{\theta \in \Theta_i} |U_{\theta t} - U_{it}| \leqslant \delta V_{it} = C_\psi \delta (d_{it}^2 + 2|Y_t - f_{it}|d_{it})$ *a.s. , where* C_ψ *is a positive constant,* $f_{it} = f_{\theta_i t}$ *and* $d_{it} = 1 + |Y_{t-1}| + |f_{it-1}| + \overline{\beta}_1 d_{it-1}$, *with* $d_{i0} = 1$ *and* (ii) *there exists a positive constant* $C_d < \infty$ *(that does not depend on* i *or* δ*) such that* $\mathbb{E}V_{it} \leqslant \delta C_d$.

Proof. (i) Using the generalized triangular equality for Bregman losses we get $|U_{\theta t} - U_{it}| \leqslant L(f_{it}, f_{\theta t}) + |(Y_t - f_{it})(\nabla\psi(f_{\theta t}) - \nabla\psi(f_{it}))|$, which holds almost surely when $\psi(Y_t) = -\log(Y_t)$, since by Condition 2.2.1(iii) we have that $Y_t > 0$ holds almost surely. It follows from the identity $|\nabla\psi(f_{\theta t}) - \nabla\psi(f_{it})||f_{\theta t} - f_{it}| = L(f_{\theta t}, f_{it}) + L(f_{it}, f_{\theta t})$ and Condition 2.2.1(ii) that $|\nabla\psi(f_{\theta t}) - \nabla\psi(f_{it})| \overset{a.s.}{\leqslant} 2C_\psi|f_{\theta t} - f_{it}|$. This inequality and Condition 2.2.1(ii) imply that

$$|U_{\theta t} - U_{it}| \overset{a.s.}{\leqslant} C_\psi \left[(f_{\theta t} - f_{it})^2 + 2|Y_t - f_{it}| \cdot |f_{\theta t} - f_{it}|\right]. \quad (2.43)$$

Set $d_{i\,0} = 1$ and note that $\delta d_{i\,0} > |f_{\theta\,0} - f_{i\,0}| = 0$, and by induction for all $t \geqslant 1$, we have

$$\begin{aligned}|f_{\theta\,t} - f_{i\,t}| &\leqslant \sum_{k=1}^{K} \{|\alpha_{0\,k} - \alpha_{0\,ki}| + |(\alpha_{1\,k} - \alpha_{1\,ki})Y_{t-1}| \\ &\quad + |\beta_{1\,k}(f_{\theta\,t-1} - f_{i\,t-1}) + (\beta_{1\,k} - \beta_{1\,ki})f_{i\,t-1}|\}\, \mathbb{1}_{t-1\,k} \\ &\leqslant \delta + \delta|Y_{t-1}| + \delta|f_{i\,t-1}| + \overline{\beta}_1|f_{\theta\,t-1} - f_{i\,t-1}| \leqslant \delta d_{i\,t}\ . \end{aligned} \tag{2.44}$$

The result in (i) follows by combining (2.43) and (2.44) and noting that $\delta \in (0, 1]$.

(ii) It suffices to show that $\sup_{t\geqslant 1} \|Y_t\|_{L_2}$, $\sup_{t\geqslant 1} \sup_{\theta\in\Theta} \|f_{\theta\,t}\|_{L_2}$, and $\sup_{t\geqslant 1} \sup_{\theta\in\Theta} \|d_{\theta\,t}\|_{L_2}$ exist, which holds by Proposition 2.7.1. □

Definition 2.13.1 (Covering and packing.)**.** *Consider the metric space* $(\mathbb{R}^p, \|\cdot\|_2)$.

(i) *The* δ*-covering number of* Θ *is*

$$N_\delta = \min\{\mathsf{card}(\tilde{\Theta}) : \forall\theta \in \Theta\ \exists\tilde{\theta} \in \tilde{\Theta}\ \textit{s.t.}\ \|\theta - \tilde{\theta}\|_2 \leqslant \delta\}\ .$$

(ii) *The* δ*-packing number of* Θ *is*

$$M_\delta = \max\{\mathsf{card}(\check{\Theta}) : \|\check{\theta}_1 - \check{\theta}_2\|_2 > \delta\ \ \forall\check{\theta}_1, \check{\theta}_2 \in \check{\Theta} \subset \Theta\}\ .$$

Lemma 2.13.4. *Consider the metric space* $(\mathbb{R}^p, \|\cdot\|_2)$*, and suppose A.2.2.2 holds. Then, for any* $\delta > 0$*, we have* $N_\delta \leqslant \left(1 + \frac{2C_\Theta}{\delta}\right)^p$*, where* $C_\Theta = \sup_{\theta\in\Theta} \|\theta\|_2$.

Proof. Consider a maximal δ-packing $\{\theta_i\}_{i=1}^{M_\delta}$ of size M_δ. Since it is a packing, the balls $\{\theta \in \mathbb{R}^p : \|\theta - \theta_i\|_2 \leqslant \delta/2\}$, $i = 1, \ldots, M_\delta$ are disjoint. Each of these balls is contained in $\{\theta \in \mathbb{R}^p : \|\theta\|_2 \leqslant C_\Theta + \delta/2\}$. Thus,

$$\bigcup_{i=1}^{M_\delta} \{\theta \in \mathbb{R}^p : \|\theta - \theta_i\| \leqslant \delta/2\} \subseteq \{\theta \in \mathbb{R}^p : \|\theta\|_2 \leqslant C_\Theta + \delta/2\}\ .$$

It follows that

$$M_\delta \text{vol}\left(\{\theta \in \mathbb{R}^p : \|\theta\| \leqslant \delta/2\}\right) \leqslant \text{vol}\left(\{\theta \in \mathbb{R}^p : \|\theta\|_2 \leqslant C_\Theta + \delta/2\}\right)$$

$$M_\delta (\delta/2)^p \text{vol}(B) \leqslant (C_\Theta + (\delta/2))^p \text{vol}(B) \,,$$

where $\text{vol}(\cdot)$ denotes volume and B is the unit ball. The result follows by noting that $N_\delta \leqslant M_\delta$. □

2.14 Proofs of Applications

Forecasting an AR(1) plus noise with an ARMA(1,1). Note that we can write model as (2.1) and (2.2), where $g_{y1}(h) = h,\ g_{y2}(h) = 1,\ g_{h1}(h) = \mu_H(1-\varrho) + \varrho h,\ g_{h2}(h) = 1$. Assumptions for g_{y1} and g_{y2} are satisfied with $C_{y1} = 1$ and $C_{y2} = 1$. Set $a = |\varrho|$. Then,

$$|g_{h1}(h)| - a|h| \leqslant |\mu_H| \cdot |1-\varrho| + |\varrho| \cdot |h| - a|h| = |\mu_H| \cdot |1-\varrho| = o(|h|).$$

Set $b \in (0,1)$ such that $\mathbb{E}(a + b|\epsilon_{H\,t}|)^{2r_m} < 1$. For that choice, we can always write that $|g_{h2}(h)| - b|h| \leqslant 1 = o(|h|)$. Hence the assumptions for g_{h1} and g_{h2} are satisfied as well.

First we show that the optimal forecaster is the conditional mean $\mu_t = \mathbb{E}_{t-1}(Y_t)$, where $\mathbb{E}_t(\cdot) = \mathbb{E}(\cdot|Y_t, \ldots, Y_1)$. Since $\mathbb{E}_{t-1}(f_{\theta\,t}) = f_{\theta\,t}$, we have that for all $t = T+1, \ldots, T+M$,

$$\begin{aligned}\mathbb{E}_T[L(Y_t, f_{\theta\,t}) - L(Y_t, \mu_t)] &= \mathbb{E}_T[\mathbb{E}_{t-1}[L(Y_t, f_{\theta\,t}) - L(Y_t, \mu_t)]] \\ &= \mathbb{E}_T[L(\mu_t, f_{\theta\,t})] \geqslant 0 \,, \end{aligned} \tag{2.45}$$

where the last inequality holds with equality if and only if $f_{\theta\, t} = \mu_t$ by properties of Bregman divergences. Second, note that for this DGP we can compute its conditional mean μ_t and (unconditional) error variance P_t via the Kalman filter. Since $H_0 = \mu_H$, this amounts to setting $\mu_1 = \mu_H$, $P_1 = \sigma_H^2$, and for $t \geqslant 2$ we have

$$\mu_t = \mu_H(1-\varrho) + K_{t-1}Y_{t-1} + (\varrho - K_{t-1})\mu_{t-1}\,, \quad P_t = \frac{\varrho^2}{\frac{1}{\sigma_Y^2} + \frac{1}{P_{t-1}}} + \sigma_H^2,$$

where $K_t = \varrho\frac{P_t}{P_t+\sigma_Y^2}$ is the Kalman gain, and the recursion for the error variance is known as the Ricatti equation. It is well known that the Kalman filter has a steady-state solution if there exists a time-invariant error variance that satisfies the Ricatti equation. If such a solution exists, we can set $P_t = P_{t-1}$, thereby obtaining the algebraic Riccati equation $\overline{P} - \varrho^2\overline{P} + \frac{(\varrho\overline{P})^2}{\overline{P}+\sigma_Y^2} - \sigma_H^2 = 0\,,\ \overline{P} \geqslant 0\,$, which is uniquely solved by

$$\overline{P} = \frac{1}{2}\left(\sigma_H^2 - (1-\varrho^2)\sigma_Y^2 + \sqrt{\sigma_H^4 + (1-\varrho^2)^2\sigma_Y^4 + 2\sigma_H^2\sigma_Y^2(1+\varrho^2)}\right)\,. \quad (2.46)$$

Note that in fact, $\overline{P} \geqslant \sigma_H^2$, which accomodates the case where H_0 has an initial distribution with zero variance. Therefore, the steady-state Kalman gain is $\overline{K} = \varrho\frac{\overline{P}}{\overline{P}+\sigma_Y^2}$. At the steady state, the filter is time-invariant, which naturally has the form of (2.3) with $K = 1$, $f_{\theta\,0} = \mathbb{E}(Y_t)$ and $\theta = (\mu_H(1-\varrho), \overline{K}, \varrho - \overline{K})'$. Finally, we check that the restrictions given by A.2.2.2 are satisfied for suitable choices of the parameter bounds. The restriction for the intercept is trivially satisfied for any choice of $\underline{\alpha}_0$ and $\overline{\alpha}_0$ that contains the intercept. The restriction for $\alpha_{1\,1}$ implies that $\underline{\alpha}_1 \leqslant \overline{K} \leqslant \overline{\alpha}_1$. The lower bound is satisfied for any $\underline{\alpha}_1 \in (0, \varrho\sigma_H^2/(\sigma_H^2+\sigma_Y^2)]$, and the upper bound holds for any $\overline{\alpha}_1 \geqslant \varrho$ (e.g. $\overline{\alpha}_1 = 1$). The restriction for $\beta_{1\,1}$

implies that $0 \leqslant \varrho\sigma_Y^2/(\overline{P} + \sigma_Y^2) \leqslant \overline{\beta}_1$, which holds for any $\varrho \in [0, 1)$ and any $\overline{\beta}_1 \in [\varrho\sigma_Y^2/(\overline{P} + \sigma_Y^2), 1)$. □

Forecasting a stochastic volatility model with a GARCH(1,1). Set $g_{y1} \equiv 0$, $g_{y2}(h) \equiv h$ and $g_{h1} \equiv 0$, and $g_{h2}(h) \equiv \exp\{\mu_H + \varrho(\log h - \mu_H)\}$. The functions g_{y1}, g_{y2} and g_{h1} trivially satisfy A.2.2.1(i) and A.2.2.1(ii) with e.g. $C_{y1} = 1$, $C_{y2} = 1$ and $a = 0$. To see that $g_{h2}(h) \equiv \exp\{\mu_H + \varrho(\log h - \mu_H)\}$ satisfies A.2.2.1(ii), note that for any $b > 0$, we have that $g_{h2}(h) - bh \leqslant \exp\{\mu_H(1 - \varrho)\}h^\varrho = o(h)$ as $h \to \infty$ since $\varrho \in (0, 1)$.

Define $\epsilon_{Y\,t} = z_t^2$ and $\epsilon_{H\,t} = e^{\eta_t}$. Then, $\epsilon_{Y\,t}$ follows a chi-square distribution with one degree of freedom and $\epsilon_{H\,t}$ follows a log-normal distribution. Clearly, $\epsilon_{H\,t}$ and $\epsilon_{Y\,t}$ are independent and are both supported on $(0, \infty)$ as required by A.2.2.1(iii) with $\underline{\epsilon} = 0$. It is well known that $\mathbb{E}\epsilon_{H\,t}^{2r_m} = \exp(2 \cdot r_m^2) < \infty$ and $\mathbb{E}\epsilon_{Y\,t}^{2r_m} = 2^{2r_m}\Gamma(2r_m + 1/2)/\sqrt{\pi} < \infty$. Furthermore, the cumulant generating function of $\log \epsilon_{Y\,t}$ is given by

$$K_{\log \epsilon_Y}(\lambda) = \log \mathbb{E}e^{\lambda \log \epsilon_{Y\,t}} = \log \mathbb{E}\epsilon_{Y\,t}^{\lambda} = \lambda \log 2 + \log \Gamma\left(\lambda + \frac{1}{2}\right) - \frac{1}{2}\log \pi,$$

and the cumulants are given by $\kappa_1 = \log 2 + \Psi(1/2)$ and $\kappa_n = \Psi^{(n-1)}(1/2)$ for $n \geqslant 2$, where $\Psi^{(n)}$ denotes the n-th derivative of the digamma function Ψ. Since all cumulants of $\log \epsilon_{Y\,t}$ exist, it follows that all moments of $\log \epsilon_{Y\,t}$ exist as well. Therefore, A.2.2.1(iii) is satisfied. If $0 < b < e^{-r_m}$, then we have that $\mathbb{E}(a_h + b|\epsilon_{H\,t}|)^{2r_m}) = b^{2r_m}\mathbb{E}(\epsilon_{H\,t}^{2r_m}) < 1$, which proves that A.2.2.1(iv) holds. □

Recursive prediction as a solution of a sequential optimization problem. The first

order condition of the optimization problem given in (2.31) can be written as

$$0 = -\nabla_f^2 \psi(f) \left[w_1(\bar{f} - f) + w_2(Y_{t-1} - f) + w_3(f_{t-1} - f) \right] ,$$

which is uniquely solved by (2.33). Clearly, $f_t \in \mathcal{F}$. The second derivative of the objective function evaluated at f_t

$$\nabla_f^2 Q_t(f)\big|_{f=f_t} = -\nabla_f^3 \psi(f)(f_t - f)\big|_{f=f_t} + \nabla_f^2 \psi(f)\big|_{f=f_t} = \nabla_f^2 \psi(f_t)$$

is strictly positive by the strict convexity of ψ (we are implicitly assuming that the third derivative exists), which verifies that this is a local minimum. Moreover, note that $\nabla_f Q_t(f) = -\nabla_f^2 \psi(f)(f_t - f)$ is positive whenever $f > f_t$ and negative whenever $f < f_t$ for all $f \in \text{int}(\mathcal{S})$. This verifies that the solution is a global minimum. Finally, note that by recursive substitution in (2.33) we have that

$$\begin{aligned} f_t &= w_1 \bar{f} \sum_{i=1}^{t} w_3^{i-1} + w_2 \sum_{i=1}^{t} w_3^{i-1} Y_{t-i} + w_3^t f_0 \\ &= \left(1 - w_3^t - w_2 \sum_{i=1}^{t} w_3^{i-1} \right) \bar{f} + w_2 \sum_{i=1}^{t} w_3^{i-1} Y_{t-i} + w_3^t \bar{f} , \\ &= \left(1 - w_2 \sum_{i=1}^{t} w_3^{i-1} \right) \bar{f} + w_2 \sum_{i=1}^{t} w_3^{i-1} Y_{t-i} , \end{aligned}$$

since $f_0 = \bar{f}$. Note that this corresponds to the first order condition for a minimum of (2.32), since by simple algebra we have that (2.32) is equal to $w_2^{-1} \tilde{Q}_t(f)$, where

$$\tilde{Q}_t(f) = w_2 \sum_{i=1}^{t} w_3^{i-1} L(Y_{t-i}, f) + \left(1 - w_2 \sum_{i=1}^{t} w_3^{i-1} \right) L(\bar{f}, f) .$$

□

Chapter 3

AN ORACLE INEQUALITY FOR MULTIVARIATE DYNAMIC QUANTILE FORECASTING

3.1 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 esti-

mation. 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 chapter 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 chapter 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.

An important reason to adopt a nonparametric perspective in the analysis of

the performance of this class of dynamic quantile forecasts is that it is in general quite challenging to find a realistic data generating process that justifies this methodology. The theoretical framework in Engle and Manganelli (2004) operates under an additive error structure assumption where the quantile of interest of the error is zero.[1] For example, if the error is asymmetric Laplace distributed, the CAViaR estimated via regression quantiles is a maximum likelihood estimator. Nevertheless, this type of additive error DGP seems unrealistic for financial and macroeconomic time series. Yet, CAViaR forecasting yields satisfactory results in those applications, suggesting that it is robust to misspecification.

The theoretical framework of this chapter 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 stable. In contrast, (White *et al.*, 2015) assume the existence of some set over which the VFV is stable.

[1] As pointed out by the authors, the symmetric absolute value and asymmetric absolute value CAViaR also arise naturally from a GARCH process with i.i.d. errors where the standard deviation (rather than the variance) is modeled symmetrically or asymmetrically. However, in that paper the theoretical framework does not use a multiplicative error DGP.

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 chapter 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 chapter. 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 chapter, identification and correct specification assumptions are not required.

Quasi-maximum likelihood. The oracle inequality derived in this chapter 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 chapter 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.

This paper is not the first to use the framework of statistical learning theory in time series 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 chapter is structured as follows. Section 3.2 lays out the notation and presents the class of forecasts and the estimation procedure. Section 3.3 introduces the theoretical framework under which the main result is derived, and section 3.4 highlights the main steps followed to prove the claim. Section 3.5 contains the empirical application to Growth-at-Risk, and section 3.6 concludes. All proofs and additional tables are gathered in sections 3.8-3.14.

3.2 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, $|||A|||_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^m |a_{ij}|$, 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 $\mathrm{vec}(A)$ 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 chapter, $I(\cdot)$ 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 $\mathbb{E}_t(\cdot) = \mathbb{E}(\cdot|X_{t-1}, \ldots, X_1)$. 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.

3.2.1 Definition of the class of forecasts

The main goal of this chapter is out-of-sample conditional quantile forecasting of a 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 multi-step ahead forecasting as well (see sections 3.11 and 3.12). More specifically, for some $\tau_i \in [0, 1]$ and $i = 1, \ldots, N$, let $q_{i\,t}^{\tau_i}$ denote the conditional τ_i-quantile of $Y_{i\,t}$ given information up to time $t-1$. That is, $q_{i\,t}^{\tau_i}$ is implicitly defined as $\Pr(Y_{i\,t} \leqslant q_{i\,t}^{\tau_i} | Y_{t-1}, \ldots, Y_1) = \tau_i$.

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

$$f_{\theta\, t} = \omega + A s_\lambda(Y_{t-1}) + B f_{\theta\, t-1} \,, \tag{3.1}$$

where $f_{\theta\, t} \in \mathbb{R}^N$, $\theta = (\omega', \mathrm{vec}(A)', \mathrm{vec}(B)', \lambda')'$, $\omega \in \Theta_\omega \subset \mathbb{R}^{p_\omega}$, $\mathrm{vec}(A) \in \Theta_A \subset \mathbb{R}^{p_A}$, $\mathrm{vec}(B) \in \Theta_B \subset \mathbb{R}^{p_B}$, $\lambda \in \Theta_\lambda \subset \mathbb{R}^{p_\lambda}$, $p = p_\omega + p_A + p_B + p_\lambda$ and $s_\lambda(\cdot)$ is shorthand for $s(\cdot, \lambda)$, where $s : \mathbb{R}^N \times \mathbb{R}^{p_\lambda} \to \mathbb{R}^N$.[2] The precise assumptions on the parameters and the function s_λ are spelled out in what follows. In practice, the forecaster chooses a fixed value $f_{\theta\, 1}$ to start the recursion.

For example, a simple bivariate version of the above relates the conditional quantile forecasts of both random variables according to a vector autoregressive structure (VAR)[3]

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

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

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

[2] 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.

[3] 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).

multivariate versions of the symmetric and asymmetric absolute value specifications of Engle and Manganelli (2004).[4] 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, $q_t^\tau := (q_{1t}^{\tau_1}, \ldots, q_{Nt}^{\tau_N})'$ 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.[5]

3.2.2 Loss function

The focus of this chapter 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.[6] It is well known that this loss function elicits the τ-quantile of a random variable. Technically, the forecasting problem in this chapter (and in the CAViaR literature) is formulated as forecasting Y_t on the basis of 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

[4] Section 3.7 provides a list of examples of data transformations allowed by A.3.3.2.

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

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

out of the scope of this chapter. The interested reader can 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.

3.2.3 Estimation

As usual in the CAViaR literature, the parameter θ in (3.1) 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[7]

$$\hat{\theta}_{T,\tau} \in \arg\min_{\Theta} R_T(\theta, \tau) \,, \quad R_T(\theta, \tau) = \frac{1}{T} \sum_{t=1}^{T} l_t(\theta, \tau) \,, \tag{3.2}$$

[7] In practice, the forecaster needs to choose a suitable initial value $f_{\theta\,1}$ to initiate the recursion, which is computed using $Y_0 = y$ and $f_{\theta\,0} = f$ that are fixed, known and do not depend on θ. A typical choice is the unconditional quantiles of Y_t.

and

$$l_t(\theta, \tau) = \frac{1}{N} \sum_{i=1}^{N} \rho_{\tau_i}(Y_{i\,t} - f_{\theta\,i\,t}). \tag{3.3}$$

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 (3.2) 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 chapter 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.

3.3 Theory

As it is clear from section 3.2, $f_{\theta\,t}$ 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.

3.3.1 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

$$R(\hat{\theta}_{T,\tau}, \tau) := \mathbb{E}\left[\frac{1}{M} \sum_{t=T+1}^{T+M} l_t(\hat{\theta}_{T,\tau}, \tau) \middle| Y_T, \ldots, Y_1 \right] . \tag{3.4}$$

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.

Note that if the data is independent and identically distributed, 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 $\theta, \dot{\theta} \in \Theta$ and every $t \geqslant 1$. The dominating process in question is given by the following recursion

$$d_{\theta\, t} = 1 + C_s \left(1 + \overline{A}\right) \|Y_{t-1}\|_1 + \|f_{\theta\, t-1}\|_1 + \overline{B} d_{\theta\, t-1} + \epsilon_{d\, t}\,, \tag{3.5}$$

where $d_{\theta\, 0} \geqslant 1$, C_s and $\overline{A}$ are positive finite constants, and $\{\epsilon_{d\, t}\}$ is an i.i.d. sequence of non-negative random variables. It follows that

$$\left| l_t(\theta, \tau) - l_t(\dot{\theta}, \tau) \right| \leqslant \frac{1}{N} \delta d_{\dot{\theta}\, t} \tag{3.6}$$

holds for all $t \geqslant 1$. 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.[8]

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 forecasts considered in this chapter.

Condition 3.3.1 (Moments and mixing)**.** *The following conditions are satisfied by* $\{l_t(\theta, \tau)\}$ *and* $\{d_{\theta\, t}\}$*, which are given by* (3.2) *and* (3.5)*:*

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

[8] For instance, 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 (3.6) together with a suitable uniform moment requirement on $d_{\theta\, t}$ are enough to verify the smoothness condition A3.

(ii) For every $\theta \in \Theta$ *and* $\tau \in [0,1]^N$*, the processes* $\{l_t(\theta,\tau)\}$ *and* $\{d_{\theta\, t}\}$ *are* 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*.*[9]

(iii) There exists $C_L < \infty$ *such that* $\sup_{\tau\in[0,1]^N} \sup_{t\geqslant 1} \sup_\Theta \|l_t(\theta,\tau)\|_{L_k} \leqslant C_L$ *and* $\sup_{t\geqslant 1} \sup_\Theta \|d_{\theta\, t}\|_{L_k} \leqslant C_L$*, for some* $k > p+2$*.*

(iv) There exists a stationary process $\{l_t^G(\theta,\tau)\}$ *with* $\sup_\Theta \|l_1^G(\theta,\tau)\|_{L_k} < \infty$ *such that* $\sup_\Theta \sum_{t=1}^\infty |\,\mathbb{E}\, l_t(\theta,\tau) - \mathbb{E}\, l_1^G(\theta,\tau)| \leqslant C_0$ *for all* $t \geqslant 1$*, where* $C_0 < \infty$*.*

(v) 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 3.3.1 deserves some discussion.

The first thing to note is that Condition 3.3.1 can be verified for a large class of parameter-driven DGP's (Cox, 1981). For instance, A.3.3.1, A.3.3.2 and A.3.3.3 imply Condition 3.3.1. This is established in this chapter by application of Markov chain theory. 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 3.3.1. Section 3.9 contains a full derivation of these results.

Condition 3.3.1(i) is a standard compactness requirement on the parameter space. Condition 3.3.1(ii) is a strong mixing assumption (Doukhan, 1994). Although strong mixing assumptions are not the most general type of condition, they

[9] See Definition 3.9.1 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.

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 (3.7) and (3.8). Condition 3.3.1(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 (3.9). Condition 3.3.1(iv) requires that the first moment of the loss process, whose forecast is initialized at a fixed value $f_{\theta 1}$, converges to its stationary counterpart, which is assumed to exist. For example, A.3.3.1, and A.3.3.2 below are enough to satisfy this requirement. Condition 3.3.1(v) 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 3.3.1. 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 chapter this is not required. The CAViaR literature assumes (inter alia) that the loss process satisfies a uniform law of large numbers (ULLN). Instead, Condition 3.3.1 can be seen as a sufficient condition to obtain the assumed ULLN from the CAViaR literature. Furthermore, Condition 3.3.1 is sufficient to establish a rate of convergence. In summary, Condition 3.3.1 is easier to verify and tailored to the goal of this chapter – which is out-of-sample forecasting.

3.3.2 Assumptions

This sub-section gives a list of sufficient conditions under which Condition 3.3.1 holds.

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

$$Y_t = g_{y1}(H_t) + g_{y2}(H_t)\epsilon_{Y\,t} \tag{3.7}$$

$$H_t = g_{h1}(H_{t-1}) + g_{h2}(H_{t-1})\epsilon_{H\,t}\,, \tag{3.8}$$

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 jointly i.i.d. sequences of random vectors supported in $\mathcal{Y}$ and $\mathcal{H}$, respectively. The process is initialized at $H_0 = h_0 \in \mathcal{H}$, i.e. h_0 is a fixed initial value. To simplify notation, take $\mathcal{Y} = \mathbb{R}^N$ and $\mathcal{H} = \mathbb{R}^{p_h}$.

A.3.3.1. *The process given by equations* (3.7) *and* (3.8) *satisfies the following:*
(i) The functions g_{h1} and g_{h2} are bounded on bounded subsets of $\mathbb{R}^{p_h}$. Moreover, $\|g_{h1}(h)\|_1 \leqslant a\|h\|_1 + o(\|h\|_1)$ and $\|\!|g_{h2}(h)|\!\|_1 \leqslant b\|h\|_1 + o(\|h\|_1)$ as $\|h\|_1 \to \infty$. The 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 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) The random process $\{(\epsilon'_{Y\,t-1}, \epsilon'_{H\,t})'\}$ is i.i.d. and $(\epsilon'_{Y\,t-1}, \epsilon'_{H\,t})'$ has a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{N+p_h}$,

and is supported on $\mathbb{R}^{N+p_h}$*. The joint density* ϕ *of the random vector* $(\epsilon'_{Y\,t-1}, \epsilon'_{H\,t})'$ *satisfies* $\phi(\epsilon_{Y\,t-1}, \epsilon_{H\,t}) = \phi_Y(\epsilon_{Y\,t-1})\phi_H(\epsilon_{H\,t})$*, where* ϕ_Y *and* ϕ_H *are densities that are bounded away from zero on compact subsets of* $\mathbb{R}^N$ *and* $\mathbb{R}^{p_h}$*, respectively. The random variables* $\epsilon_{Y\,t}$ *and* $\epsilon_{H\,t}$ *satisfy* $\|\|\epsilon_{Y\,t}\|_1\|_{L_k} < \infty$ *and* $\|\|\epsilon_{H\,t}\|_1\|_{L_k} < \infty$ *(resp.) for some* $k > p + 2$*.*

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

Class of forecasts

A.3.3.2. *The class of forecasts given by* (3.1) *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) $\theta = (\omega', \mathrm{vec}(A)', \mathrm{vec}(B)', \lambda')' \in \Theta \subseteq \mathbb{R}^p$*, 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.3.3.3. *The dominating process given by* (3.5) *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 (ii)* $\epsilon_{d\,t}$ *is supported in* $[0, 1]$ *for all* $t \geqslant 1$*, with density* ϕ_d *that is bounded away from zero on compact subsets of* $[0, 1]$*.*

Remarks. A.3.3.1 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) and it allows for a fairly broad class of parameter-driven processes. A.3.3.1(*i*) is similar to Assumption 3.2 in Masry and Tjøstheim (1995) and it implies that (3.8) 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. A.3.3.1(*ii*) allows for a fair amount of flexibility in equation (3.7). In particular, it requires $\|Y_t\|_1$ to be bounded from above by a linear function of $\|H_t\|_1$. A.3.3.1(*iii*) imposes conditions on the random variables ϵ_{Ht} and ϵ_{Yt} that are analogous to standard conditions used in the literature. A.3.3.1(*iv*) is a stability condition analogous to the one assumed in Masry and Tjøstheim (1995) or Lanne and Saikkonen (2005).

A.3.3.2(*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.3.3.2(*i*) implies that the spectral radius of B is strictly less than unity. A.3.3.2(*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. A.3.3.2(*iii*), A.3.3.2(*iv*) and A.3.3.2(*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 section 3.7 for examples of s_λ that satisfy A.3.3.2).

A.3.3.3 is an auxiliary assumption that is useful to simplify the proof of irreducibility and aperiodicity of the "companion Markov chain" defined in (2.11). 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 3.3.1 leads to an oracle inequality for the class of forecasts introduced in (3.1), with out-of-sample performance defined as in equation (3.4).

Theorem 3.3.1. *Suppose Condition 3.3.1 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 3.3.1 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 3.4.3 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 $|R(\hat{\theta}_{T,\tau}, \tau) - \inf_{\Theta} R(\theta, \tau)| \xrightarrow{\text{p}} 0$ 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 chapter do not allow p nor N to diverge to infinity. This limits the extent to which Theorem 3.3.1 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 3.3.1 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 3.4, 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 chapter 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.

3.3.3 Additional Discussion

This paper provides a list of sufficient conditions which involve a data generating process given by a hidden Markov model. 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 chapter. However, the Markov chain analysis would differ and the analysis of this case is left for future research.

The theoretical framework of this chapter 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

$$R^*(\tau) = \mathbb{E}\left[\frac{1}{M}\frac{1}{N}\sum_{t=T+1}^{T+M}\sum_{i=1}^{N}\rho_{\tau_i}(Y_{it} - q_{it}^{\tau_i})\middle| Y_T, \ldots, Y_1\right].$$

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 chapter.

The focus of this chapter 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 (3.6) 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 (3.5) can be derived so that $\|\theta - \dot{\theta}\|_2 \leqslant \delta$ implies that $\|f_{\theta\, t} - f_{\dot{\theta}\, t}\|_2 \leqslant \delta d_{\dot{\theta}\, t}$ for every pair $\theta, \dot{\theta} \in \Theta$. However, notation and proofs do require modifications which are not pursued here.

The check loss elicits *marginal* quantile forecasts, but one may also be interested in extending the setup to *multivariate* quantiles. For example, the framework can also accommodate the notion of geometric quantiles introduced by Chaudhuri (1996). By letting $\tau \in \{x \in \mathbb{R}^N : \|x\|_2 \leqslant 1\}$ instead of $[0,1]^N$, a geometric τ-quantile is obtained by minimizing the criterion in (3.2) with[10]

$$l_t(\theta, \tau) = \|Y_t - f_{\theta\, t}\|_2 + \tau'(Y_t - f_{\theta\, t}), \quad \|\tau\|_2 \leqslant 1.$$

3.4 Sketch of proof of Theorem 3.3.1

This section explains the main steps to derive the proof of Theorem 3.3.1, which are broken down in four propositions. Proofs can be found in section 3.8.

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 3.4.1. *Let $\overline{R}(\theta, \tau) = \mathbb{E}\, l_t^G(\theta, \tau)$, where $\{l_t^G(\theta, \tau)\}$ is the process de-*

[10] Note that for every pair $\theta, \dot{\theta} \in \Theta$, if $\|\theta - \dot{\theta}\|_2 \leqslant \delta$, then $|l_t(\theta, \tau) - l_t(\dot{\theta}, \tau)| \leqslant 2\|f_{\theta\, t} - f_{\dot{\theta}\, t}\|_2 \leqslant 2\delta d_{\dot{\theta}\, t}$, where $d_{\theta\, t}$ is a dominating process but defined with ℓ_2-norms instead of ℓ_1.

fined in Condition 3.3.1. Then,

$$R(\hat{\theta}_{T,\tau}, \tau) - \inf_{\Theta} R(\theta, \tau)$$
$$\leqslant 2 \sup_{\Theta} |R_T(\theta, \tau) - \overline{R}(\theta, \tau)| + 2 \sup_{\Theta} |R(\theta, \tau) - \overline{R}(\theta, \tau)| . \qquad (3.9)$$

It is important to emphasize that Proposition 3.4.1 is a general result that only requires the "ghost" stationary loss process $\{l_t^G(\theta, \tau)\}$ to exist. Note that when the data is i.i.d., $R(\theta, \tau) = \overline{R}(\theta, \tau)$ and the inequality in Proposition 3.4.1 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.

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

Proposition 3.4.2. *Suppose Condition 3.3.1 is satisfied. Then, for any $\varepsilon > 0$, any $T \geqslant 4C_0\varepsilon^{-1}$, and any $M \geqslant 4C_0\varepsilon^{-1}$, it holds that*

$$\Pr\left(\sup_{\Theta} |R_T(\theta, \tau) - \overline{R}(\theta, \tau)| > \frac{\varepsilon}{2} \right)$$
$$\leqslant \left(1 + \frac{48 C_\Theta C_d}{N\varepsilon} \right)^p \sup_{\Theta} \left[P_1^T \left(l_t(\theta, \tau), \frac{\varepsilon}{8} \right) + P_1^T \left(d_{\theta\, t}, C_d \right) \right] ,$$

and

$$\Pr\left(\sup_{\Theta} |R(\theta, \tau) - \overline{R}(\theta, \tau)| > \frac{\varepsilon}{2} \right)$$
$$\leqslant \left(1 + \frac{48 C_\Theta C_d}{N\varepsilon} \right)^p \sup_{\Theta} \left[P_{T+1}^{T+M} \left(\mathbb{E}_T\, l_t(\theta, \tau), \frac{\varepsilon}{8} \right) + P_{T+1}^{T+M} \left(\mathbb{E}_T\, d_{\theta\, t}, C_d \right) \right] ,$$

where $P_a^b(U_t, \varepsilon) = \Pr\left(\left| \frac{1}{b-a+1} \sum_{t=a}^{b} [U_t - \mathbb{E}\, U_t] \right| > \varepsilon \right)$, $C_\Theta = \sup_\Theta \|\theta\|_1$ *and* $C_d = \sup_{t \geqslant 1} \sup_\Theta \|d_{\theta\, t}\|_{L_1}$.

Proposition 3.4.2 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 3.4.3 formalizes the result.

Proposition 3.4.3. *Suppose Condition 3.3.1 is satisfied. Then, for all T sufficiently large and for $\varepsilon_T = \sigma\sqrt{\frac{p \log T}{NT}}$, it holds that*

$$\left(1 + \frac{48 C_\Theta C_d}{N \varepsilon_T}\right)^p \sup_\Theta \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}[l_t(\theta,\tau) - \mathbb{E}\, l_t(\theta,\tau)]\right| > \frac{\varepsilon_T}{8}\right) \leqslant \frac{1}{\log T} \text{ and}$$

$$\left(1 + \frac{48 C_\Theta C_d}{N \varepsilon_T}\right)^p \sup_\Theta \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}[d_{\theta\, t} - \mathbb{E}\, d_{\theta\, t}]\right| > C_d\right) \leqslant o\left(\frac{1}{\log T}\right)$$

as $T \to \infty$, where $\sigma^2 = 16\frac{k}{k-2}C_L^2\left(1 + 2\sum_{m=1}^{\infty}\exp\left(-C_\alpha m^{r_\alpha}\right)^{1-\frac{2}{k}}\right)$.

Step 4: Concentration inequality (part II). The fourth step – summarized in Proposition 3.4.4 – 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 3.4.4. *Suppose Condition 3.3.1 is satisfied. Then, for all T sufficiently large and for $\varepsilon_T = \sigma\sqrt{\frac{p \log T}{NT}}$, it holds that*

$$\left(1 + \frac{48 C_\Theta C_d}{N \varepsilon_T}\right)^p \sup_\Theta \Pr\left(\left|\frac{1}{M}\sum_{t=T+1}^{T+M}[\mathbb{E}_T\, l_t(\theta,\tau) - \mathbb{E}\, l_t(\theta,\tau)]\right| > \frac{\varepsilon_T}{8}\right) \leqslant \frac{1}{\log T}$$

and

$$\left(1+\frac{48C_\Theta C_d}{N\varepsilon_T}\right)^p \sup_\Theta \Pr\left(\left|\frac{1}{M}\sum_{t=T+1}^{T+M}[\mathbb{E}_T\, d_{\theta\, t}-\mathbb{E}\, d_{\theta\, t}]\right|>C_d\right)\leqslant o\left(\frac{1}{\log T}\right)$$

as $T\to\infty$, where σ^2 is defined in Proposition 3.4.3.

It follows from Propositions 3.4.2, 3.4.3 and 3.4.4 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 3.4.1 imply Theorem 3.3.1.

Proof of Theorem 3.3.1. Follows from Condition 3.3.1 and Propositions 3.4.1, 3.4.2, 3.4.3 and 3.4.4. □

3.5 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 chapter. 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 3.1 – 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 (3.1) are implemented, labeled as pooled VFV in Table 3.1. 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 3.7.3 for a verification of A.3.3.2 for this choice of the

function s_λ. The restrictions on the parameters α, β are naturally deduced from A.3.3.2.[11]

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 (3.2). 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 3.1; (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-

[11] Estimating the parameters in practice is a challenging optimization problem and convergence issues may arise even with small p. This is addressed in two ways which lead to a very fast and scalable implementation: (i) replace $\rho_\tau(u)$ in the objective function with its differentiable approximation $b(\sqrt{1+(u/b)^2}-1)\ |\tau - I(u<0)|$ with $\tau = \tau$ and $b = 10^{-4}$, and (ii) set $\omega_i = f_i(1-\beta)$, where f_i is the sample τ-quantile of the i-th country's in-sample observations. It turns out that the in-sample objective is very close compared to $R_T(\hat{\theta}_{T,\tau}, \tau)$, which can be obtained via more computationally intensive optimizers.

sample period, that is,

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

For completeness, Table 3.1 also reports Coverage and Length, which are defined as

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

and

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

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.

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 3.14.

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

Table 3.1: 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	93.06	5.08	11.51
Pooled VFV	Sym	93.53	5.28	−0.53
Pooled VFV	Asym	94.82	5.37	8.50
Pooled VFV	AR-VFV Sym	93.94	5.20	11.52
Pooled VFV	AR-VFV Asym	95.36	5.40	2.91
Pooled GARCH-VFV	GARCH-VFV Sym	93.09	5.18	4.54
Pooled GARCH-VFV	GARCH-VFV Asym	94.07	5.24	12.59
Pooled GARCH-VFV	AR-GARCH-VFV Sym	92.99	5.09	12.77
Pooled GARCH-VFV	AR-GARCH-VFV Asym	93.43	5.13	12.49
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.

specification 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 3.1 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.

3.6 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 not specified. 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 chapter 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 chapter hopefully paves the way for the development of new forecasting strategies for time series applications with minimal assumptions on how the data is generated.

3.7 Data transformations allowed by A.3.3.2

Example 3.7.1. *Suppose that* $N = 1$ *and let* $s_\lambda(u) = s(u) = \sqrt{1+u^2} - 1$.

Proof of validity of A.3.3.2 for Example 3.7.1. By the chain rule, for every $h \in \mathbb{R}^{p_h}$,

$$\frac{\partial \tilde{s}_\lambda(h,z)}{\partial z} = \frac{[g_{y1}(h) + g_{y2}(h)z]g_{y2}(h)}{\sqrt{1+[g_{y1}(h)+g_{y2}(h)z]^2}},$$

which can only be zero when the numerator is zero. If $g_{y2}(h) \neq 0$ (which holds by A.3.3.1(*ii*)), A.3.3.2(*iii*) holds for any $z \neq -[g_{y2}(h)]^{-1}g_{y1}(h)$. A.3.3.2(*iv*) holds

with $C_s = 1$, since $\sqrt{1+u^2}-1 \leqslant |u|$ for all $u \in \mathbb{R}$. Furthermore, A.3.3.2(*vi*) holds (for example) by letting $D_f = [0, \infty)$, since in D_f the inverse of $s(u)$ is $s^{-1}(v) = \sqrt{(1+v)^2 - 1}$ for all $v \geqslant 0$. The same arguments generalize to the multivariate case with every component of the vector function s_λ defined as above. □

Example 3.7.2. *Suppose that $N = 1$ and consider the example $s_\lambda(u) = s_b(u)|\tau - I(u < 0)|$, where $s_b(u) = b^2(\sqrt{1+(u/b)^2} - 1)$ and $\lambda = (\tau, b)' \in [0,1] \times [\underline{b}, \overline{b}]$, $\underline{b} > 0$, and $1 < \overline{b} < \infty$. The function s_b is also known as the Pseudo-Huber loss function.*

Proof of validity of A.3.3.2 for Example 3.7.2. By the chain rule, for every $h \in \mathbb{R}^{p_h}$ and $b \in \mathbb{R}_+$,

$$\frac{\partial \tilde{s}_\lambda(h,z)}{\partial z} = [\tau I(u \geqslant 0) + (1-\tau) I(u < 0)] \frac{u}{\sqrt{1+(u/b)^2}} g_{y2}(h),$$

where $u = g_{y1}(h) + g_{y2}(h)z$. Note that $\partial \tilde{s}_\lambda(h,z)/\partial z$ can only be zero when the numerator is zero. If $g_{y2}(h) \neq 0$ (which holds by A.3.3.1(*ii*)), A.3.3.2(*iii*) holds for any $z \neq -[g_{y2}(h)]^{-1} g_{y1}(h)$. A.3.3.2(*iv*) holds with $C_s = \overline{b}$. To see this, write

$$|s_\lambda(u)| \leqslant |s_b(u)| = b|\sqrt{b^2+u^2} - b| = b(\sqrt{b^2+u^2} - b) \leqslant b(b + |u| - b) \leqslant \overline{b}|u|,$$

which verifies the first part of A.3.3.2(*iv*). The second part of A.3.3.2(*iv*) can be verified as follows. First, note that for all $b > 0$,

$$\left| \frac{\partial s_b(u)}{\partial b} \right| = \left| 2b \left(\sqrt{1+(u/b)^2} - 1 \right) - \frac{u^2}{b\sqrt{1+(u/b)^2}} \right| \leqslant |u|,$$

which implies that $s_b(u)$ is $|u|$-Lipschitz w.r.t. b by the mean value theorem. Now, note that $||\tau - I(u<0)| - |\dot{\tau} - I(u<0)|| \leqslant |\tau - \dot{\tau}|$ by the inequality $||x|-|y|| \leqslant$

$|x-y|$. Then,

$$
\begin{aligned}
&|s_\lambda(u) - s_{\dot\lambda}(u)| \\
&\quad = \left|\left[s_b(u) - s_{\dot b}(u)\right]|\tau - I(u<0)| - s_{\dot b}(u)\left[|\dot\tau - I(u<0)| - |\tau - I(u<0)|\right]\right| \\
&\quad \leqslant |u||b-\dot b| + \bar b|u||\tau - \dot\tau| \\
&\quad \leqslant \bar b|u|\left(|b-\dot b| + |\tau - \dot\tau|\right) = \bar b|u|\|\lambda - \dot\lambda\|_1 \,,
\end{aligned}
$$

which holds since $1 < \bar b < \infty$. This verifies the second part of A.3.3.2(*iv*). To verify A.3.3.2(*vi*), let $D_f = [0,\infty)$ and note that in D_f the inverse of $s_\lambda(u)$ (w.r.t. u) is

$$
s_\lambda^{-1}(v) = b\sqrt{\left(1 + \frac{v}{\tau b^2}\right)^2 - 1}\,, \quad v \geqslant 0,
$$

which is differentiable w.r.t. v. The same arguments generalize to the multivariate case with every component of the vector function s_λ defined as above. □

Example 3.7.3. *Consider again Example 3.7.2 but with* $s_b(u) = b(\sqrt{1+(u/b)^2} - 1)$*, and let* $0 < \underline b < 1$*. It is easy to see that* $\lim_{b\downarrow 0} s_\lambda(u) = |u||\tau - I(u<0)|$.

Proof of validity of A.3.3.2 for Example 3.7.3. By the chain rule, for every $h \in \mathbb{R}^{p_h}$ and $b \in \mathbb{R}_+$,

$$
\frac{\partial \tilde s_\lambda(h,z)}{\partial z} = \left[\tau I(u \geqslant 0) + (1-\tau)I(u<0)\right]\frac{u}{\sqrt{u^2+b^2}} g_{y2}(h),
$$

where $u = g_{y1}(h) + g_{y2}(h)z$. Again, $\partial \tilde s_\lambda(h,z)/\partial z$ can only be zero when the numerator is zero. If $g_{y2}(h) \neq 0$ (which holds by A.3.3.1(*ii*)), A.3.3.2(*iii*) holds for any $z \neq -[g_{y2}(h)]^{-1} g_{y1}(h)$. A.3.3.2(*iv*) holds with $C_s = \underline b^{-1}$. To see this,

write

$$|s_\lambda(u)| \leqslant |s_b(u)| = |\sqrt{b^2+u^2} - b| = \sqrt{b^2+u^2} - b \leqslant b + |u| - b = |u| \leqslant \underline{b}^{-1}|u|,$$

which verifies the first part of A.3.3.2(*iv*). The second part of A.3.3.2(*iv*) can be verified as follows. First, note that for all $b \geqslant \underline{b} > 0$,

$$\left|\frac{\partial s_b(u)}{\partial b}\right| = \left|\frac{b}{\sqrt{b^2+u^2}} - 1\right| \leqslant \underline{b}^{-1}|u|,$$

which implies that $s_b(u)$ is $(\underline{b}^{-1}|u|)$-Lipschitz w.r.t. b by the mean value theorem. Again, recall that $||\tau - I(u < 0)| - |\dot{\tau} - I(u < 0)|| \leqslant |\tau - \dot{\tau}|$ by the reverse triangular inequality. Then,

$$\begin{aligned}
&|s_\lambda(u) - s_{\dot{\lambda}}(u)| \\
&\quad = \left|\left[s_b(u) - s_{\dot{b}}(u)\right]|\tau - I(u<0)| - s_{\dot{b}}(u)\left[|\dot{\tau} - I(u<0)| - |\tau - I(u<0)|\right]\right| \\
&\quad \leqslant \underline{b}^{-1}|u||b - \dot{b}| + \underline{b}^{-1}|u||\tau - \dot{\tau}| \\
&\quad \leqslant \underline{b}^{-1}|u|\left(|b - \dot{b}| + |\tau - \dot{\tau}|\right) = \underline{b}^{-1}|u|\|\lambda - \dot{\lambda}\|_1 .
\end{aligned}$$

This verifies the second part of A.3.3.2(*iv*). To verify A.3.3.2(*vi*), let $D_f = [0, \infty)$ and note that in D_f the inverse of $s_\lambda(u)$ (w.r.t. u) is

$$s_\lambda^{-1}(v) = b\sqrt{\left(1 + \frac{v}{\tau b}\right)^2 - 1}\,, \quad v \geqslant 0,$$

which is differentiable w.r.t. v. The same arguments generalize to the multivariate case with every component of the vector function s_λ defined as above. □

Example 3.7.4. *Suppose* $N = 1$ *and consider smooth transitions of the form*

$$s_\lambda(u) = [1 + aG(u)]s_b(u), \quad \lambda = (a, b)',$$

where $G : \mathbb{R} \to (0,1)$ *is differentiable with* $G'(u) > 0$, $a \in [0, \overline{a}]$, $\overline{a} < \infty$, *and* $s_b(u)$ *is defined in Example 3.7.3.*

Proof of validity of A.3.3.2 for Example 3.7.4. By the chain and product rules, for every $h \in \mathbb{R}^{p_h}$ and $b \in \mathbb{R}_+$,

$$\frac{\partial \tilde{s}_\lambda(h,z)}{\partial z} = \left[aG'(u)s_b(u) + [1 + aG(u)]\frac{u}{\sqrt{u^2 + b^2}} \right] g_{y2}(h),$$

where $u = g_{y1}(h) + g_{y2}(h)z$. Note that if $g_{y2}(h) \neq 0$ (which holds by A.3.3.1(*ii*)), then $\partial \tilde{s}_\lambda(h,z)/\partial z$ can only be zero if the first term is zero. Note that for every $h \in \mathbb{R}^{p_h}$, there is always some $z \in \mathbb{R}$ such that the first term is not zero, which verifies A.3.3.2(*iii*). A.3.3.2(*iv*) holds with $C_s = \underline{b}^{-1}(1 + \overline{a})$. To see this, write

$$|s_\lambda(u)| = |1 + aG(u)||s_b(u)| \leqslant (1 + a)|s_b(u)| \leqslant \underline{b}^{-1}(1 + \overline{a})|u|,$$

which verifies the first part of A.3.3.2(*iv*). The second part of A.3.3.2(*iv*) can be verified using the fact that $|\partial s_b(u)/\partial b| \leqslant \underline{b}^{-1}|u|$ and the mean value theorem, and noting that $|1 + aG(u) - 1 - \dot{a}G(u)| = G(u)|a - \dot{a}| \leqslant |a - \dot{a}|$. Combining both results leads to

$$\begin{aligned} |s_\lambda(u) - s_{\dot{\lambda}}(u)| &= |[1 + aG(u)][s_b(u) - s_{\dot{b}}(u)] + s_{\dot{b}}(u)G(u)(a - \dot{a})| \\ &\leqslant \underline{b}^{-1}(1 + \overline{a})|u||b - \dot{b}| + \underline{b}^{-1}|u||a - \dot{a}| \leqslant C_s|u|\|\lambda - \dot{\lambda}\|_1 \,, \end{aligned}$$

as expected. This verifies the second part of A.3.3.2(*iv*). A.3.3.2(*vi*) is verified by noting that for all $u \in D_f = (0, \infty)$, we have that

$$\frac{\partial s_\lambda(u)}{\partial u} = aG'(u)s_b(u) + [1 + aG(u)]\frac{u}{\sqrt{u^2 + b^2}} > 0,$$

hence s_λ^{-1} exists and is differentiable. □

Remark. Note that for lower quantiles, it may be more interesting to consider

$$s_\lambda(u) = [1 - aG(u)]s_b(u)$$

instead. The proof of the validity of A.3.3.2 above holds with $\overline{a} < 1$ and letting $D_f = (-\infty, 0)$ instead of $(0, \infty)$.

Example 3.7.5. *Let* $s_\lambda(u) = g_\tau(u)s_b(u)$*, where* $\lambda = (\tau, b)'$*,*

$$g_\tau(u) = \sqrt{\tau^2 + (1-\tau)^2}\left(\frac{1}{1-\tau}I(u > 0) + \frac{1}{\tau}I(u \leqslant 0)\right),$$

and $\tau \in [\underline{\tau}, \overline{\tau}]$*,* $\underline{\tau} > 0$*,* $\overline{\tau} < 1$*, which corresponds to the "improved CAViaR" of Huang* et al. *(2009) but with the absolute value replaced by its differentiable approximation* $s_b(u)$ *defined in Example 3.7.3.*

Proof of validity of A.3.3.2 for Example 3.7.5. By the chain rule, for every $h \in \mathbb{R}^{p_h}$ and $b \in \mathbb{R}_+$,

$$\frac{\partial \tilde{s}_\lambda(h, z)}{\partial z} = g_\tau(u)\frac{u}{\sqrt{u^2 + b^2}}g_{y2}(h),$$

where $u = g_{y1}(h) + g_{y2}(h)z$. Again, $\partial \tilde{s}_\lambda(h, z)/\partial z$ can only be zero when the numerator is zero. If $g_{y2}(h) \neq 0$ (which holds by A.3.3.1(*ii*)), A.3.3.2(*iii*) holds for any $z \neq -[g_{y2}(h)]^{-1}g_{y1}(h)$. A.3.3.2(*iv*) holds with $C_s = \underline{b}^{-1}C_{0\tau}$, where

$$C_{0\tau} = \max\left\{\underline{\tau}^{-1}\sqrt{\underline{\tau}^2 + (1-\underline{\tau})^2}, (1-\overline{\tau})^{-1}\sqrt{\overline{\tau}^2 + (1-\overline{\tau})^2}\right\}.$$

To see this, note that $|g_\tau(u)| \leqslant C_{0\tau}$. Then, $|s_\lambda(u)| \leqslant C_{0\tau}|s_b(u)| \leqslant \underline{b}^{-1}C_{0\tau}|u|$, where the last inequality follows from the same arguments used in Example 3.7.3. This verifies the first part of A.3.3.2(*iv*). The second part is verified as follows.

First, let $\Delta_\tau g(\tau, u) := g(\tau, u) - g(\dot\tau, u)$ for every pair $\tau, \dot\tau \in [\underline\tau, \overline\tau]$, $u \in \mathbb{R}$ and any function g, and define

$$g_1(\tau) = \sqrt{\tau^2 + (1-\tau)^2}\left(\frac{1}{1-\tau} - \frac{1}{\tau}\right) \quad \text{and} \quad g_2(\tau) = \frac{\sqrt{\tau^2 + (1-\tau)^2}}{\tau}.$$

Observe that $g_\tau(u) = I(u > 0)g_1(\tau) + g_2(\tau)$. Then,

$$|\Delta_\tau g_\tau(u)| = |I(u > 0)\Delta_\tau g_1(\tau) + \Delta_\tau g_2(\tau)| \leqslant |\Delta_\tau g_1(\tau)| + |\Delta_\tau g_2(\tau)|.$$

Notice that for all $\tau \in [\underline\tau, \overline\tau]$,

$$\begin{aligned}
\left|\frac{\partial g_1(\tau)}{\partial \tau}\right| &\leqslant \max\left\{\frac{3\underline\tau^2 - 3\underline\tau + 1}{\underline\tau^2\sqrt{2\underline\tau^2 - 2\underline\tau + 1}(1-\underline\tau)^2}, \frac{3\overline\tau^2 - 3\overline\tau + 1}{\overline\tau^2\sqrt{2\overline\tau^2 - 2\overline\tau + 1}(1-\overline\tau)^2}\right\} \\
&:= C_{1\tau} < \infty \\
\left|\frac{\partial g_2(\tau)}{\partial \tau}\right| &\leqslant \frac{1-\underline\tau}{\underline\tau^2\sqrt{2\underline\tau^2 - 2\underline\tau + 1}} := C_{2\tau} < \infty,
\end{aligned}$$

and let $C_\tau = C_{0\tau} \vee C_{1\tau} \vee C_{2\tau}$. Thus, by the mean value theorem it holds that

$$|\Delta_\tau g_\tau(u)| \leqslant C_{1\tau}|\tau - \dot\tau| + C_{2\tau}|\tau - \dot\tau| \leqslant C_\tau|\tau - \dot\tau|.$$

Recall from Example 3.7.3 that $s_b(u)$ is $(\underline b^{-1}|u|)$-Lipschitz w.r.t. b by the mean value theorem. Then,

$$\begin{aligned}
|s_\lambda(u) - s_{\dot\lambda}(u)| &= |[s_b(u) - s_{\dot b}(u)]\, g_\tau(u) - s_{\dot b}(u)\Delta_\tau g_\tau(u)| \\
&\leqslant \underline b^{-1}C_\tau|u||b - \dot b| + \underline b^{-1}C_\tau|u||\tau - \dot\tau| \\
&= \underline b^{-1}C_\tau|u|\|\lambda - \dot\lambda\|_1\,,
\end{aligned}$$

as expected. This verifies the second part of A.3.3.2(*iv*). To verify A.3.3.2(*iv*), let $D_f = (-\infty, 0)$ and note that in D_f the inverse of $s_\lambda(u)$ (w.r.t. u) is

$$s_\lambda^{-1}(v) = b\sqrt{\left(1 + \frac{v}{bg_2(\tau)}\right)^2 - 1}, \quad v \geqslant 0,$$

which is differentiable w.r.t. v. The same arguments generalize to the multivariate case with every component of the vector function s_λ defined as above. □

3.8 Proofs of Propositions 1-4

Proof of Proposition 3.4.1. Define $\overline{R}(\hat{\theta}_{T,\tau}, \tau) = \mathbb{E} l_1^G(\hat{\theta}_{T,\tau}, \tau)$. 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

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

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

Proof of Proposition 3.4.2. Adding and subtracting $\mathbb{E}\, l_t(\theta, \tau)$, we have

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

But by Condition 3.3.1(*iv*), we have that

$$\Pr\left(\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}|\,\mathbb{E}\,l_t(\theta,\tau)-\mathbb{E}\,l_t^G(\theta,\tau)|>\frac{\varepsilon}{4}\right)\leqslant\Pr\left(\frac{C_0}{T}>\frac{\varepsilon}{4}\right),$$

so $\Pr\left(\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}|\,\mathbb{E}\,l_t(\theta,\tau)-\mathbb{E}\,l_t^G(\theta,\tau)|>\frac{\varepsilon}{4}\right)=0$ for all $T\geqslant 4C_0\varepsilon^{-1}$. The next step 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

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

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

$$\begin{aligned}&\Pr\left(\sup_{\theta\in\Theta_j}\left|\frac{1}{T}\sum_{t=1}^{T}l_t(\theta,\tau)-\mathbb{E}\,l_t(\theta,\tau)\right|>\frac{\varepsilon}{4}\right)\\&\leqslant\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}l_t(\theta_j,\tau)-\mathbb{E}\,l_t(\theta_j,\tau)\right|>\frac{\varepsilon}{8}\right)\\&\quad+\Pr\left(\sup_{\theta\in\Theta_j}\left|\frac{1}{T}\sum_{t=1}^{T}[l_t(\theta,\tau)-l_t(\theta_j,\tau)]-\mathbb{E}[l_t(\theta,\tau)-l_t(\theta_j,\tau)]\right|>\frac{\varepsilon}{8}\right).\end{aligned}$$

Now, by (3.6) we have that $|l_t(\theta,\tau)-l_t(\theta_j,\tau)|\leqslant\frac{\delta}{N}d_{\theta_j t}$, which is proven in Lemma 3.10.2. By the triangular inequality, the second term is bounded above by

$$\begin{aligned}&\Pr\left(\sup_{\theta\in\Theta_j}\frac{1}{T}\sum_{t=1}^{T}|l_t(\theta,\tau)-l_t(\theta_j,\tau)|+|\,\mathbb{E}[l_t(\theta,\tau)-l_t(\theta_j,\tau)]|>\frac{\varepsilon}{8}\right)\\&\quad\leqslant\Pr\left(\frac{1}{T}\sum_{t=1}^{T}d_{\theta_j t}+\mathbb{E}\,d_{\theta_j t}>\frac{N\varepsilon}{8\delta}\right).\end{aligned}$$

Furthermore, by Condition 3.3.1(*iii*) we have $\sup_{t\geqslant 1}\sup_{\Theta}\mathbb{E}(d_{\theta_j\,t})\leqslant C_d$ for some $C_d<\infty$ that does not depend on j nor t, by choosing $\delta=N\varepsilon/(24C_d)$ 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

$$N_\delta\leqslant\left(1+\frac{2C_\Theta}{\delta}\right)^p=\left(1+\frac{48C_\Theta C_d}{N\varepsilon}\right)^p.$$

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

$$|\,\mathbb{E}_T\,l_t(\theta,\tau)-\mathbb{E}_T\,l_t(\theta_j,\tau)|\leqslant\mathbb{E}_T\,|l_t(\theta,\tau)-l_t(\theta_j,\tau)|\leqslant\frac{\delta}{N}\,\mathbb{E}_T\,d_{\theta_j\,t}$$

by Jensen's inequality and the order-preserving property of the conditional expectation. □

Proof of Proposition 3.4.3. 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

$$M_T=\lfloor T^{\frac{1}{2}-\frac{p+1}{2(k-1)}}\log^{-\frac{1}{2}}T\rfloor,\quad\text{and}\quad b_T=C_bT^{\frac{p+1}{2(k-1)}}(\log T)^{-\frac{p-1}{2(k-1)}},$$

where C_b is a positive constant to be chosen in what follows. Let $\widetilde{U}_t=U_t'+U_t''$ where $U_t'=l_t(\theta,\tau)I(l_t(\theta,\tau)\leqslant b_T)-\mathbb{E}\,(l_t(\theta,\tau)I(l_t(\theta,\tau)\leqslant b_T))$ and $U_t''=l_t(\theta,\tau)I(l_t(\theta,\tau)>b_T)-\mathbb{E}\,(l_t(\theta,\tau)I(l_t(\theta,\tau)>b_T))$. Then,

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\widetilde{U}_t\right|>\frac{\varepsilon_T}{8}\right)\leqslant\Pr\left(\left|\sum_{t=1}^{T}U_t'\right|>\frac{T\varepsilon_T}{16}\right)+\Pr\left(\left|\sum_{t=1}^{T}U_t''\right|>\frac{T\varepsilon_T}{16}\right).$$

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 $T\varepsilon_T/16 > 4M_T b_T$. By application of that theorem and noting that $\{l_t(\theta,\tau)\}$ is non-negative,

$$\begin{aligned}
&\Pr\left(\left|\sum_{t=1}^{T} U_t'\right| > \frac{T\varepsilon_T}{16}\right) \\
&\leqslant 4\exp\left(-\frac{T\varepsilon_T^2}{\frac{16384}{M_T}\sup_{0\leqslant t\leqslant T-1}\mathbb{E}(\sum_{s=t+1}^{(t+M_T)\wedge T} U_s')^2 + \frac{128}{3}M_T b_T \varepsilon_T}\right) \\
&\quad + 4\frac{T}{M_T}\exp\left(-C_\alpha M_T^{r_\alpha}\right) .
\end{aligned}$$

Let $\gamma_t(m) = |\mathrm{Cov}(U_t', U_{t+m}')|$ for $m = 0,\ldots,T-1$. Then,

$$\sup_{0\leqslant t\leqslant T-1}\mathbb{E}(\sum_{s=t+1}^{(t+M_T)\wedge T} U_s')^2 \leqslant M_T \sup_{t\geqslant 1}(\gamma_t(0) + 2\sum_{m=1}^{\infty}\gamma_t(m)).$$

Noting that $l_t(\theta,\tau) \geqslant 0$ and $k \geqslant 2$, Davydov's inequality (Bosq, 1998, Corollary 1.1) implies

$$\gamma_t(m) \leqslant 2\frac{k}{k-2}2^{1-2/k}\alpha(m)^{1-2/k}\|U_t'\|_{L_k}\|U_{t+m}'\|_{L_k}$$

for $m = 0,\ldots,T-1$. Also note that for any $k > 1$ we have $\|U_t'\|_{L_k} \leqslant 2C_L$ by Jensen's inequality, and the last inequality holds by Condition 3.3.1(*iii*). Thus,

$$\begin{aligned}
\sup_{0\leqslant t\leqslant T-1}\mathbb{E}\left(\sum_{s=t+1}^{(t+M_T)\wedge T} U_s'\right)^2 &\leqslant M_T 16\frac{k}{k-2}C_L^2\left(1 + 2\sum_{m=1}^{\infty}\exp\left(-C_\alpha m^{r_\alpha}\right)^{1-\frac{2}{k}}\right) \\
&:= M_T\sigma^2,
\end{aligned}$$

where the sum converges by Condition 3.3.1(*ii*). Then, for all T sufficiently large,

since $p, k > 1$ are such that $p < k - 2$, it holds that

$$\left(1 + \frac{48C_\Theta C_d}{N\varepsilon_T}\right)^p \Pr\left(\left|\sum_{t=1}^{T} U'_t\right| > \frac{T\varepsilon_T}{16}\right) = o(\log^{-1} T) \, .$$

Furthermore,

$$\begin{aligned}
&\left(1 + \frac{48C_\Theta C_d}{N\varepsilon_T}\right)^p \Pr\left(\left|\sum_{t=1}^{T} U''_t\right| > \frac{T\varepsilon_T}{16}\right) \overset{(a)}{\leqslant} \left(1 + \frac{48C_\Theta C_d}{N\varepsilon_T}\right)^p \frac{16}{T\varepsilon_T} \mathbb{E}\left|\sum_{t=1}^{T} U''_t\right| \\
&\quad \leqslant \left(1 + \frac{48C_\Theta C_d}{N\varepsilon_T}\right)^p \frac{32}{\varepsilon_T} \sup_{t \geqslant 1} \mathbb{E}\left[l_t(\theta, \tau) I(l_t(\theta, \tau) > b_T)\right] \\
&\quad \overset{(b)}{\leqslant} \left(1 + \frac{48C_\Theta C_d}{N\varepsilon_T}\right)^p \frac{32}{\varepsilon_T} \frac{C_L^k}{b_T^{k-1}} \\
&\quad \overset{(c)}{\leqslant} \log^{-1} T \, ,
\end{aligned}$$

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 3.3.1(*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 3.4.4. By Proposition 3.4.2, we have that

$$\begin{aligned}
&\Pr\left(\sup_\Theta \left|R(\theta, \tau) - \overline{R}(\theta, \tau)\right| > \frac{\varepsilon}{2}\right) \\
&\quad \leqslant \left(1 + \frac{48C_\Theta C_d}{\varepsilon}\right)^p \sup_\Theta \left\{P_{T+1}^{T+M}\left(\mathbb{E}_T \, l_t(\theta, \tau), \frac{\varepsilon}{8}\right) + P_{T+1}^{T+M}\left(\mathbb{E}_T \, d_{\theta t}, C_d\right)\right\} \, .
\end{aligned}$$

By Markov's inequality,

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

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

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

where $\sup_{t\geqslant 1} \sup_{\Theta} \left\| l_t(\theta,\tau) \right\|_{L_k} \leqslant C_L < \infty$ by Condition 3.3.1(*iii*). Consequently, and by Condition 3.3.1(*ii*),

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

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

$$\left(1 + \frac{48 C_\Theta C_d}{N \varepsilon_T} \right)^p \sup_{\Theta} P_{T+1}^{T+M} \left(\mathbb{E}_T \, l_t(\theta,\tau), \frac{\varepsilon_T}{8} \right) \leqslant \frac{C}{\gamma^p N^p T^p \varepsilon_T^{2p}} = O\left(\log^{-p} T \right) \, .$$

for some $C < \infty$. By Condition 3.3.1(*ii*) and 3.3.1(*iii*), $\{d_{\theta\, t}\}$ is also α-mixing with exponentially decaying coefficients and $\sup_{t\geqslant 1} \sup_{\Theta} \left\| d_{\theta\, t} \right\|_{L_k} < \infty$. The same arguments as above lead to the bound

$$\left(1 + \frac{48 C_\Theta C_d}{N \varepsilon_T} \right)^p \sup_{\Theta} P_{T+1}^{T+M} \left(\mathbb{E}_T \, d_{\theta\, t}, C_d \right) \leqslant \frac{C}{\gamma^p N^p T^p \varepsilon_T^p} = o(\log^{-p} T)$$

for all T sufficiently large. □

3.9 Verification of Condition 3.3.1

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 $\pi_X(\mathcal{A}) = \int_{\mathcal{X}} \pi_X(dx) P_X(x, \mathcal{A})$.

3.9.1 Companion Markov chain

Let $X_t = (X_{1\,t}', X_{2\,t}', X_{3\,t})'$ be defined as $X_0 = x \in \mathcal{X}$, and

$$\begin{bmatrix} X_{1\,t} \\ X_{2\,t} \\ X_{3\,t} \end{bmatrix} = \begin{bmatrix} g_{h1}(X_{1\,t-1}) + g_{h2}(X_{1\,t-1}) Z_{1\,t} \\ \omega + A\tilde{s}_\lambda(X_{1\,t-1}, Z_{2\,t}) + B X_{2\,t-1} \\ 1 + C_s\left(1 + \overline{A}\right) \|\tilde{Y}(X_{1\,t-1}, Z_{2\,t})\|_1 + \|X_{2\,t-1}\|_1 + \overline{B} X_{3\,t-1} + Z_{3\,t} \end{bmatrix}, \tag{3.10}$$

where

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

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 $p_x = p_h + N + 1$.

3.9.2 V-geometric ergodicity

The concept of V-geometric ergodicity used in this chapter is the same as in Meitz and Saikkonen (2008a) and Definition 2.7.1. Note that this is stronger than $\mathcal{Q}$-

geometric ergodicity (Liebscher, 2005). Verification of Condition 3.3.1 begins by establishing the V-geometric ergodicity of the companion Markov chain $\{X_t\}$. The proof follows by Lemmas 3.9.1 and 3.9.2 (Meyn and Tweedie, 1993).

Lemma 3.9.1 (Irreducibility and Aperiodicity of X_t.)**.** *Let X_t be the Markov chain defined in* (3.10)*. Then, X_t is irreducible and aperiodic.*

Proof. Start by noting that X_t in (3.10) can be cast as a nonlinear state space model NSS(F) (Meyn and Tweedie, 1993), i.e. $X_t = F(X_{t-1}, (Z'_{1\,t}, Z'_{2\,t}, Z_{3t})')$ with F defined in an obvious way.[12] 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 $Z^* = (Z_1^*, \ldots, Z_n^*) \in \bigtimes_{i=1}^{n} (\mathbb{R}^{p_h} \times \mathbb{R}^N \times \mathbb{R}_+)$ such that 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 A.3.3.1(*i*), A.3.3.2(*ii*) and A.3.3.2(*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

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

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

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 A.3.3.1(*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 A.3.3.2(*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 3.9.2 (Drift Criterion for X_t)**.** *Let X_t be the Markov chain defined in* (3.10)*. 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 A.3.3.1(*i*), for every $\epsilon > 0$ there exists $M'_\epsilon < \infty$ such that

$$\|g_{h1}(x_1) + g_{h2}(x_1)Z_{1\,t}\|_1 \leqslant (a + b^\epsilon \|Z_{1\,t}\|_1 + \epsilon)\, \|x_1\|_1 \tag{3.11}$$

holds for all $\|x_1\|_1 > M'_\epsilon$, where $b^\epsilon = b + \epsilon$. In particular, $\epsilon > 0$ is chosen small enough such that $\mathbb{E}(a + b^\epsilon \|Z_{1\,t}\|_1 + \epsilon)^k < 1$ and $\overline{B} + \epsilon < 1$. Such a choice is possible by A.3.3.1(*iv*) and A.3.3.2(*i*), respectively.

Now, let $S_{2\,\epsilon} = \{x \in \mathcal{X} : \kappa'\dot{x} \leqslant M_\epsilon\}$, which is compact, and $S_{1\,\epsilon} = \mathcal{X}\backslash S_{2\,\epsilon}$.[13] The proof proceeds by analyzing the cases $\|x_1\|_1 > M'_\epsilon$ and $\|x_1\|_1 \leqslant M'_\epsilon$ separately.[14]

Case $\|x_1\|_1 > M'_\epsilon$. By A.3.3.2(*iv*) and A.3.3.1(*ii*),

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

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

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

[13] 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.

[14] Note that the conclusions in both cases hold for any choice of $\kappa \in \times_{i=1}^{3}(0,1)$.

where $\overline{A} < \infty$ is a uniform upper bound for $|||A|||_1$ over Θ by A.3.3.2(*v*). Also note that $|||B|||_1 \leqslant \overline{B} < 1$ by A.3.3.2(*i*). Similarly,

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

where the first inequality uses A.3.3.3(*ii*). Let $\rho_{Z\,\epsilon} = a + b^{\epsilon}\|Z_{1\,t}\|_1 + \epsilon$, and $C^{\epsilon}_{y\,Z} = C_y(1+\epsilon+\|Z_{2\,t}\|_1)$. It follows that[15]

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

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

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

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

$$\mathbb{E}(q_X(X_t)|X_{t-1}=x) \leqslant (\kappa^{\otimes k})'\,\mathbb{E}\left(\mathbf{C}_{\epsilon}(Z_t)^{\otimes k}\right)\dot{x}^{\otimes k}\,. \tag{3.12}$$

Case $\|x_1\|_1 \leqslant M'_{\epsilon}$**.** Note that by A.3.3.1(*i*),

$$\|g_{h1}(x_1) + g_{h2}(x_1)Z_{1\,t}\|_1 \leqslant \underbrace{\overline{g}^{\epsilon}_h\left(1+\|Z_{1\,t}\|_1\right)}_{C_1}\,, \tag{3.13}$$

[15]Note that M'_{ϵ} can be enlarged if necessary to absorb the constant 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

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

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 $\kappa \in \bigtimes_{i=1}^3 (0, 1)$, 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 (3.12) 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.[16] 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 A.3.3.1, A.3.3.2 and A.3.3.3 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. □

Lemma 3.9.3. *Suppose A.3.3.1, A.3.3.2 and A.3.3.3 are satisfied. Then, there exist positive constants $\rho \in (0,1)$ and $R < \infty$ that do not depend on θ such that*

[16] 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.

$\{X_t\}$ *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* $\tilde{V}_X(x) = 1 + \|x\|_1$.

Proof. The claim follows by invoking Theorem 12 in Roberts and Rosenthal (2004), which allows us to analyze the geometric ergodicity constants R and ρ explicitly. The theorem holds under the following conditions: (i) that there exists $C \subseteq \mathcal{X}$, such that C is "small" (Meyn and Tweedie, 1993, Ch. 5) and (ii) that there exists $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$ such that the following bivariate drift condition

$$\begin{aligned}
&\overline{\mathbb{E}}\left[h(X_n, X_n^G) | X_{n-1} = x, X_{n-1}^G = x^G\right] \\
&:= \int_{\mathcal{X}} \int_{\mathcal{X}} h(x_n, x_n^G) P(x_{n-1}, dx_n) P(x_{n-1}^G, dx_n^G) \\
&\leqslant \alpha^{-1} h(x, x^G)
\end{aligned}$$

holds for all $(x, x^G) \notin C \times C$ and for some $\alpha > 1$, where $\{X_t^G\}$ is an independent copy of $\{X_t\}$ initialized at the stationary distribution and $\overline{\mathbb{E}}$ is the expectation under the product measure. Choose $h(x, x^G) = \frac{1}{2}(\tilde{q}_X(x) + \tilde{q}_X(x^G))$, where $\tilde{q}_X(x) = 1 + \kappa' \dot{x}$.

Condition (i) holds by choosing $C = S_{2\epsilon}$ in Lemma 3.9.2. This is a consequence of the fact that for T-chains such as $\{X_t\}$, every compact set is small (Meyn and Tweedie, 1993). Verification of Condition (ii) proceeds by cases.

Case $\alpha^{-1} < 1$**.** Proposition 11 in Roberts and Rosenthal (2004) establishes that the univariate drift condition in Lemma 3.9.2 (with $k = 1$) implies the bivariate

drift with the same C and $\alpha^{-1} = 1 - \tilde{\gamma}_1 + \tilde{\gamma}_2/(1+\tilde{M})$, where $\tilde{M} = \inf_{x \in C^c} \tilde{q}_X(x)$ and $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are analogous to γ_1 and γ_2 in Lemma 3.9.2 but with $k = 1$. Inspection of the proof of Lemma 3.9.2 reveals that γ_1 and γ_2 (as well as $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$) do not depend on θ.

Case $\alpha^{-1} > 1$**.** One can find an enlargement of $S_{2\,\epsilon}$ for which the result in case $\alpha^{-1} < 1$ still holds. More specifically, enlarge M_ϵ such that $2 + M_\epsilon > \frac{\tilde{\gamma}_2}{\tilde{\gamma}_1}$. Note that for such an enlargement all arguments used to obtain the univariate drift are still valid.

Thus, in both cases conditions (i) and (ii) of Theorem 12 cited above hold and this implies that

$$\begin{aligned}
&\frac{1}{2} \sup_{v:|v| \leqslant 1} \left| \int_{\mathcal{X}} [P_X^n(x, dx_n) - \pi_X(dx_n)] v(x_n) \right| \\
&\leqslant (1 - \epsilon_*)^j + \alpha^{-n} \beta^{j-1} \overline{\mathbb{E}} \left[h(X_0, X_0^G) \right]
\end{aligned}$$

holds for all $x \in \mathcal{X}$, all $n \geqslant 1$ and for any integer $1 \leqslant j \leqslant n$, where ν and ϵ_* are defined in Lemma 3.9.4.Furthermore, define

$$\beta = \max \left[1, \alpha(1 - \epsilon_*) \sup_{C \times C} \overline{R} h(x, x^G) \right],$$

where

$$\begin{aligned}
&\overline{R} h(x, x^G) = \\
&\int_{\mathcal{X}} \int_{\mathcal{X}} \frac{h(x_1, x_1^G)}{(1 - \epsilon_*)^2} (P_X(x, dx_1) - \epsilon_* \nu(dx_1))(P_X(x^G, dx_1^G) - \epsilon_* \nu(dx_1^G)).
\end{aligned}$$

It follows that there exists $\check{R} < \infty$ that does not depend on θ such that $\sup_{C \times C} \overline{R} h \leqslant \check{R}$. This is verified in the same way as in Lemma 2.13.2. Furthermore, that

ϵ_* can be chosen so that it does not depend on θ is true by Lemma 3.9.4. It follows that $\overline{\beta} = \max[1, \alpha(1-\epsilon_*)\check{R}]$ exists and does not depend on θ. Finally, $\overline{\mathbb{E}}\left[h(X_0, X_0^G)\right]$ can be upper bounded by a finite constant that does not depend on θ. Noting that $\{X_t\}$ is initialized at a fixed value $x \in \mathcal{X}$, we have $\overline{\mathbb{E}}\left[h(X_0, X_0^G)\right] = \frac{1}{2}\left(\tilde{q}_X(x) + \mathbb{E}\,\tilde{q}_X(X_0^G)\right)$, and $\mathbb{E}\,\tilde{q}_X(X_0^G) \leqslant \mathbb{E}\,\tilde{V}_X(X_0^G) \leqslant \overline{V}$ where

$$\overline{V} = 1 + \|\|X_{10}^G\|_1\|_{L_1} + \sup_{\Theta} \|\|X_{20}^G\|_1\|_{L_1} + \sup_{\Theta} \|X_{30}^G\|_{L_1}.$$

The finiteness of $\|\|X_{10}^G\|_1\|_{L_1}$ follows from the V_X-geometric ergodicity of $\{X_t\}$. Moreover, by A.3.3.1 and A.3.3.2,

$$\|\|X_{20}^G\|_1\|_{L_1} \leqslant \frac{\overline{\omega} + C_s C_y \overline{A}(1 + \|\|Z_{20}\|_1\|_{L_1})\|\|X_{10}^G\|_1\|_{L_1}}{1 - \overline{B}}$$

which holds by stationarity. Since the upper bound is finite and independent of θ, we have that $\sup_\Theta \|\|X_{20}^G\|\|_{L_1} < \infty$. By analogous steps and by A.3.3.1, A.3.3.2 and A.3.3.3 one obtains that

$$\sup_{\Theta} \|X_{30}^G\|_{L_1} \leqslant \frac{2 + C_s C_y(1 + \overline{A})(1 + \|\|Z_{20}\|_1\|_{L_1})\|\|X_{10}^G\|_1\|_{L_1}}{1 - \overline{B}} < \infty.$$

Therefore, $\overline{V} < \infty$. Set $j = \lfloor rn \rfloor$ for sufficiently small $r > 0$ such that the bound converges to zero at a geometric rate. Thus, the claim holds with $\rho = (1-\epsilon_*)^{r/2} \vee (\alpha^{-1}\overline{\beta}^r) < 1$ and $R = 4\overline{V}$ for all $n \geqslant r^{-1}$, and $R = 4\overline{V}\rho^{-1/r}$ and any $\rho \in (0,1)$ for all $1 \leqslant n < r^{-1}$. The proof is complete since ρ and R do not depend on θ. □

The proof of Lemma 3.9.3 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 3.9.3 is that the dependence properties of the companion Markov chain $\{X_t\}$ can be characterized independently of θ.

Lemma 3.9.4 ("Irreducibility constant" independent of θ.)**.** *Consider the setup of the proof of Lemma 3.9.3. Let ν denote the Lebesgue measure in $\mathcal{X}$ restricted to the set $D \in \mathcal{X}$, where D is any compact subset of $\mathcal{H} \times D_f \times [1, \infty)$. Then, there exists $\epsilon_* > 0$ (independent of θ) such that*

$$P_X(x, \mathcal{A}) \geqslant \epsilon_* \nu(\mathcal{A}) \quad \textit{for all} \quad x \in C, \quad \mathcal{A} \in \mathcal{B}(\mathcal{X}).$$

Proof. For all $\mathcal{A} \in \mathcal{B}(\mathcal{X})$ and for all t,

$$P_X(x, \mathcal{A}) = \int_{\mathcal{H}} \int_{\mathcal{Y}} \int_0^1 I(X_t \in \mathcal{A}) \phi_H(Z_1) \phi_Y(Z_2) \phi_d(Z_3) dZ_1 dZ_2 dZ_3.$$

Note that for all $x \in C$ and all $X_t \in D$, the inverse map of (3.10) (with respect to Z) can be written explicitly as

$$\begin{aligned}
Z_1(X_t, x) &= g_{h2}(x_1)^{-1}(X_{1\,t} - g_{h1}(x_1)) \\
Z_2(X_t, x) &= g_{y2}(x_1)^{-1} \left(s_\lambda^{-1} \left(A^{-1}(X_{2\,t} - \omega - B x_2) \right) - g_{y1}(x_1) \right) \\
Z_3(X_t, x) &= X_{3\,t} - 1 - C_s(1 + \overline{A}) \| g_{y1}(x_1) + g_{y2}(x_1) Z_2(X_t, x) \|_1 - \| x_2 \|_1 - \overline{B} x_3,
\end{aligned}$$

which exists by A.3.3.1, A.3.3.2 and A.3.3.3. Hence the map is a diffeomorphism in the restriction $\mathcal{A} \cap D$. Changing variables in the restriction, one obtains

$$P_X(x, \mathcal{A}) \geqslant \int_{\mathcal{A} \cap D} |J(X_t, x)| \phi_H(Z_1(X_t, x)) \phi_Y(Z_2(X_t, x)) \phi_d(Z_3(X_t, x)) dX_t,$$

where $J(X_t, x) = \det[\partial Z(X_t, x)/\partial X_t']$. By A.3.3.1, A.3.3.2 and A.3.3.3, the densities ϕ_H, ϕ_Y and ϕ_d are bounded away from zero on bounded subsets of their domains, and the same holds for $J(X_t, x)$. Therefore, the following exists

$$\epsilon_* = \inf_{\substack{x \in C \\ \theta \in \Theta \\ X_t \in \mathcal{A} \cap D}} |J(X_t, x)| \phi_H(Z_1(X_t, x)) \phi_Y(Z_2(X_t, x)) \phi_d(Z_3(X_t, x)) > 0,$$

and it follows that $P_X(x, \mathcal{A}) \geqslant \epsilon_* \int_{\mathcal{A} \cap D} dX_t = \epsilon_* \nu(\mathcal{A})$, which completes the proof. □

The next step of the analysis consists of using the properties of the companion Markov chain $\{X_t\}$ to establish the properties of $W_t = \{(Y_t', S_t')'\} = \{(Y_t', H_t', f_{\theta t}', d_{\theta t})'\}$.[17] The following lemma establishes the connection between the transition kernels of $\{X_t\}$ and $\{W_t\}$.

Lemma 3.9.5. *Consider the Markov chain $\{W_t\}$. Let $\pi_{Y|S}(dy|s)$ denote the (invariant) conditional distribution of Y_t given $S_t = s$. 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

$$\tilde{x} = \tilde{x}(w, h_1, \epsilon_{d1}) = (h_1, \omega + As_\lambda(y) + Bf, 1 + C_s \left(1 + \overline{A}\right) \|y\|_1 + \|f\|_1 + \overline{B}d + \epsilon_{d1})'.$$

Proof. For all $n \geqslant 2$ it holds that

$$\begin{aligned} P_W^n(w, dw_n) &= \pi_{Y|S}(dy_n|s_n) \Pr(ds_n|w) \\ &= \pi_{Y|S}(dy_n|s_n) \int_0^1 \int_{\mathcal{H}} \Pr(ds_n|w, h_1, \epsilon_{d1}) P_H(h, dh_1) \Pr(d\epsilon_{d1}) \,, \end{aligned}$$

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

where the last equality is true because the H_t component of W_t is a Markov chain of its own. Note that by A.3.3.1(*iii*) and A.3.3.3(*i*) the innovations $\epsilon_{H\,t}$, $\epsilon_{Y\,t}$ and $\epsilon_{d\,t}$ are i.i.d., which leads to the observation that S_t has an n-step transition mechanism which is identical to the $(n-1)$-step transition mechanism of the companion Markov chain defined in (3.10) with initial value given by $\tilde{x} = \tilde{x}(w, h_1, \epsilon_{d\,1})$ (Meitz and Saikkonen, 2008a, cf. Assumption $1(b)$). Thus, $\Pr(ds_n|w, h_1, \epsilon_{d\,1}) = P_X^{n-1}(\tilde{x}, ds_n)$, which completes the proof. □

The proof of the lemma builds upon the analysis of GARCH models of Meitz and Saikkonen (2008a). The structure given by equations (3.1), (3.5), (3.7) and (3.8) 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 3.9.6. *Suppose A.3.3.1, A.3.3.2 and A.3.3.3 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*

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

Proof. (i) This proof is an adaptation of Meitz and Saikkonen (2008a, Proposition 1). First, $\{X_t\}$ is V_X-geometrically ergodic by Lemmas 3.9.1 and 3.9.2. The next

step is to show that for all $s \in \mathcal{X}$, $\mathbb{E}(V_W(W_t)|S_t = s) \leqslant C\, V_X(s)$. To see this, note that by A.3.3.1 one can write

$$\begin{aligned}
\mathbb{E}(V_W(W_t)|S_t = s) &= 1 + \|s\|_1^k + \mathbb{E}\,\|Y_t\|_1^k \\
&= V_X(s) + \mathbb{E}\,\|g_{y1}(h) + g_{y2}(h)\epsilon_{Y\,t}\|_1^k \\
&\leqslant V_X(s) + \mathbb{E}\,(\|g_{y1}(h)\|_1 + \|\!|g_{y2}(h)|\!\|_1\|\epsilon_{Y\,t}\|_1)^k \\
&\leqslant V_X(s) + 2^{k-1}\,\mathbb{E}\left(\|g_{y1}(h)\|_1^k + \|\!|g_{y2}(h)|\!\|_1^k\|\epsilon_{Y\,t}\|_1^k\right) \\
&\leqslant V_X(s) + 2^{k-1}C_y^k\|h\|_1^k\left(1 + \mathbb{E}\,\|\epsilon_{Y\,t}\|_1^k\right) \\
&= V_X(s) + C\|h\|_1^k \leqslant V_X(s) + C\|s\|_1^k \leqslant C\, V_X(s)\,,
\end{aligned}$$

where C is a generic positive constant that may change from line to line. To satisfy V_W-geometric ergodicity we must have that $\int_{\mathcal{Y}\times\mathcal{X}} V_W(W_t)\pi_W(dw_t) < \infty$. This holds by the previous inequality, since

$$\begin{aligned}
\int_{\mathcal{Y}\times\mathcal{X}} V_W(W_t)\pi_W(dw_t) &= \int_{\mathcal{X}} \pi_S(ds_t)\int_{\mathcal{Y}} V_W(W_t)\pi_{Y|S}(dy_t|s_t) \\
&\leqslant C\int_{\mathcal{X}} \pi_X(ds_t)V_X(s_t) < \infty,
\end{aligned}$$

and note that $\pi_S = \pi_X$. Now, for any $w \in \mathcal{Y} \times \mathcal{X}$ and all $n \geqslant 2$,

$$
\begin{aligned}
&\sup_{v:|v|\leqslant V_W} \left| \int_{\mathcal{Y}\times\mathcal{X}} [P_W^n(w, dw_n) - \pi_W(dw_n)] v(w_n) \right| \\
&\quad = \sup_{v:|v|\leqslant V_W} \left| \int_{\mathcal{X}} \int_{\mathcal{Y}} \left[\pi_{Y|S}(dy_n|s_n) \int_0^1 \int_{\mathcal{H}} P_X^{n-1}(\tilde{x}, ds_n) P_H(h, dh_1) \Pr(d\epsilon_{d\,1}) - \pi_W(dw_n) \right] v(w_n) \right| \\
&\quad = \sup_{v:|v|\leqslant V_W} \left| \int_{\mathcal{X}} \left(\int_0^1 \int_{\mathcal{H}} P_X^{n-1}(\tilde{x}, ds_n) P_H(h, dh_1) \Pr(d\epsilon_{d\,1}) - \pi_S(ds_n) \right) \int_{\mathcal{Y}} \pi_{Y|S}(dy_n|s_n) v(w_n) \right| \\
&\quad \leqslant C \sup_{v':|v'|\leqslant V_X} \left| \int_{\mathcal{X}} \left(\int_0^1 \int_{\mathcal{H}} P_X^{n-1}(\tilde{x}, ds_n) P_H(h, dh_1) \Pr(d\epsilon_{d\,1}) - \pi_S(ds_n) \right) v'(s_n) \right| \\
&\quad = C \sup_{v':|v'|\leqslant V_X} \left| \int_0^1 \int_{\mathcal{H}} P_H(h, dh_1) \Pr(d\epsilon_{d\,1}) \int_{\mathcal{X}} \left[P_X^{n-1}(\tilde{x}, ds_n) - \pi_S(ds_n) \right] v'(s_n) \right| \\
&\quad \leqslant C \int_0^1 \int_{\mathcal{H}} P_H(h, dh_1) \Pr(d\epsilon_{d\,1}) \sup_{v':|v'|\leqslant V_X} \left| \int_{\mathcal{X}} \left[P_X^{n-1}(\tilde{x}, ds_n) - \pi_S(ds_n) \right] v'(s_n) \right| \\
&\quad \leqslant C \int_0^1 \int_{\mathcal{H}} P_H(h, dh_1) \Pr(d\epsilon_{d\,1}) R_\theta V_X(\tilde{x}) \rho_\theta^{n-1},
\end{aligned}
$$

where the first equality is true by Lemma 3.9.5, the second uses the fact that $\pi_W(dw_n) = \pi_{Y|S}(dy_n|s_n)\pi_S(ds_n)$, the first inequality uses $\mathbb{E}(v(W_t)|S_t = s) \leqslant C\ V_X(s)$ for all v such that $|v| \leqslant V_W$, the third equality follows by simple rearrengement and changing the order of integration, the second inequality uses the convexity of the supremum, and the last inequality is implied by the drift criterion used in the proof of Lemma 3.9.2 (note that $\pi_S = \pi_X$), where the θ subscript in the constants $R_\theta < \infty$ and $\rho_\theta < 1$ is used to emphasize that they may depend on θ. Now,

$$
\begin{aligned}
&\int_0^1 \int_{\mathcal{H}} P_H(h, dh_1) \Pr(d\epsilon_{d\,1}) V_X(\tilde{x}) \\
&\quad = \mathbb{E}(V_X(S_t)|W_{t-1} = w) \\
&\quad = 1 + \mathbb{E}[(\|H_t\|_1 + \|f_{\theta\,t}\|_1 + d_{\theta\,t})^k | W_{t-1} = w] \\
&\quad \leqslant 1 + 3^{k-1} \left(\mathbb{E}[\|H_t\|_1^k | W_{t-1} = w] + \mathbb{E}[\|f_{\theta\,t}\|_1^k | W_{t-1} = w] + \mathbb{E}[d_{\theta\,t}^k | W_{t-1} = w] \right).
\end{aligned}
$$

Notice that by the same arguments and notation as in the proof of Lemma 3.9.2 (which hold by A.3.3.1),

$$\mathbb{E}[\|H_t\|_1^k|W_{t-1}=w] \leqslant \begin{cases} \mathbb{E}[(a_h + b_h^\epsilon \|Z_{1\,t}\|_1 + \epsilon)^k]\|h\|_1^k & \|h\|_1 > M'_\epsilon \\ (\overline{g}_h^\epsilon)^k\, \mathbb{E}[(1+\|Z_{1\,t}\|_1)^k] & \|h\|_1 \leqslant M'_\epsilon \end{cases}$$
$$\leqslant C\|h\|_1^k$$

for some $C < \infty$, and recall that $\epsilon > 0$ is such that $\mathbb{E}[(a_h + b_h^\epsilon \|Z_{1\,t}\|_1 + \epsilon)^k] < 1$. Also note that

$$\mathbb{E}[\|f_{\theta\,t}\|_1^k|W_{t-1}=w] = \|\omega + As_\lambda(y) + Bf\|_1^k \leqslant (\overline{\omega} + \overline{A}C_s\|y\|_1 + \overline{B}\|f\|_1)^k$$

where $\overline{\omega} < \infty$ by the compactness of Θ. Similarly,

$$\mathbb{E}[d_{\theta\,t}^k|W_{t-1}=w] = \mathbb{E}[(1 + C_s(1+\overline{A})\|y\|_1 + \|f\|_1 + \overline{B}d + \epsilon_{d\,t})^k|W_{t-1}=w]$$
$$\leqslant (2 + C_s(1+\overline{A})\|y\|_1 + \|f\|_1 + \overline{B}d)^k.$$

Putting it all together, one may redefine $R_\theta < \infty$ to absorb the constants and ρ_θ^{-1} so that

$$\sup_{v:|v|\leqslant V_W} \left| \int_{\mathcal{Y}\times\mathcal{X}} [P_W^n(w, dw_n) - \pi_W(dw_n)]v(w_n) \right| \leqslant R_\theta V_X(\check{s})\rho_\theta^n,$$

which proves part (i).

(ii) By the same arguments as in part (i) with $k=1$ and with $\sup_{v:|v|\leqslant 1}$ instead of $\sup_{v:|v|\leqslant V_W}$, one can use Lemma 3.9.3 in the previous derivation instead of the standard drift criterion to obtain constants $\rho \in (0,1)$ and $R < \infty$ that do not depend on θ. □

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

Definition 3.9.1. *For a process $\{X_t\}$, its α-mixing coefficients are defined by*

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

where $\mathcal{F}_0^s$ 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 3.9.1. *Suppose A.3.3.1, A.3.3.2 and A.3.3.3 are satisfied. Then, the process $\{W_t\}$ (i) satisfies $\sup_{t \geqslant 1} \| \|Y_t\|_1 \|_{L_k} < \infty$, $\sup_{t \geqslant 1} \| \|H_t\|_1 \|_{L_k} < \infty$, $\sup_{t \geqslant 1} \sup_{\Theta} \| \|f_{\theta\, t}\|_1 \|_{L_k} < \infty$ and $\sup_{t \geqslant 1} \sup_{\Theta} \|d_{\theta\, t}\|_{L_k} < \infty$; and (ii) has 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; and (iii) its distribution converges to the invariant measure π_W, which admits k moments.*

Proof. (i) By the same arguments used to arrive at (3.11) and (3.13), which rely on A.3.3.1(*i*) and A.3.3.1(*iii*), we have that there exists $\epsilon > 0$ such that for all $t \geqslant 1$,

$$\|H_t\|_1 \leqslant (a + b^{\epsilon} \|\epsilon_{H\, t}\|_1 + \epsilon) \|H_{t-1}\|_1 I(\|H_{t-1}\|_1 > M'_{\epsilon}) + C_1,$$

where $C_1 < \infty$. Taking the L_k-norm on both sides, we get

$$\| \|H_t\|_1 \|_{L_k} \leqslant \underbrace{\|a + b^{\epsilon} \|\epsilon_{H\, t}\|_1 + \epsilon\|_{L_k}}_{\varrho_\epsilon < 1} \| \|H_{t-1}\|_1 \|_{L_k} + C_1,$$

where we have used A.3.3.1(*iv*) and $\|H_{t-1}\|_1 I(\|H_{t-1}\|_1 > M'_\epsilon) \leqslant \|H_{t-1}\|_1$. Thus, $\sup_{t\geqslant 1} \| \|H_t\|_1 \|_{L_k} \leqslant \|H_0\|_1 + \frac{C_1}{1-\varrho_\epsilon} < \infty$, where $H_0 = h \in \mathbb{R}^{p_h}$. By A.3.3.1(*ii*) and A.3.3.1(*iii*), we have that there exists $C_{\epsilon_Y} < \infty$ such that

$$\sup_{t\geqslant 1} \| \|Y_t\|_1 \|_{L_k} \leqslant C_y \sup_{t\geqslant 1} [\| \|H_t\|_1 \|_{L_k} + \| \|H_t\|_1 \|_{L_k} C_{\epsilon_Y}] < \infty.$$

Moreover, by A.3.3.2 it holds that

$$\| \|f_{\theta\, t}\|_1 \|_{L_k} \leqslant \overline{\omega} + C_s \overline{A} \sup_{t\geqslant 1} \| \|Y_{t-1}\|_1 \|_{L_k} + \overline{B} \, \| \|f_{\theta\, t-1}\|_1 \|_{L_k}$$

where $\overline{\omega} < \infty$ by A.3.3.2. Therefore,

$$\sup_{t\geqslant 1} \sup_{\Theta} \| \|f_{\theta\, t}\|_1 \|_{L_k} \leqslant \|f\|_1 + \frac{\overline{\omega} + C_s \overline{A} \sup_{t\geqslant 1} \| \|Y_{t-1}\|_1 \|_{L_k}}{1 - \overline{B}} < \infty$$

where it is recalled that $f_{\theta\, 0} = f \in \mathbb{R}^N$, $\overline{A} < \infty$ and $\overline{B} < 1$ by A.3.3.2. Similarly, by A.3.3.2 and A.3.3.3, we have

$$\|d_{\theta\, t}\|_{L_k} \leqslant 2 + C_s \left(1 + \overline{A}\right) \sup_{t\geqslant 1} \| \|Y_{t-1}\|_1 \|_{L_k} + \sup_{t\geqslant 1} \sup_{\Theta} \| \|f_{\theta\, t-1}\|_1 \|_{L_k} + \overline{B} \|d_{\theta\, t-1}\|_{L_k}.$$

Thus, for any $d_{\theta\, 0} = d \in [1, \infty)$,

$$\begin{aligned} &\sup_{\Theta} \|d_{\theta\, t}\|_{L_k} \\ &\leqslant d + \frac{2 + C_s \left(1 + \overline{A}\right) \sup_{t\geqslant 1} \| \|Y_{t-1}\|_1 \|_{L_k} + \sup_{t\geqslant 1} \sup_{\Theta} \| \|f_{\theta\, t-1}\|_1 \|_{L_k}}{1 - \overline{B}} < \infty \, . \end{aligned}$$

(ii) It is enough to show that $\{W_t\} = \{(Y'_t, S'_t)'\}$ is geometrically β-mixing, since $\alpha(l) \leqslant \beta(l)$, where

$$\beta(l) = \sup_{t\geqslant 0} \frac{1}{2} \sup \sum_{i=1}^{I} \sum_{j=1}^{J} \left| \Pr(\mathcal{A}_i \cap \mathcal{B}_j) - \Pr(\mathcal{A}_i) \Pr(\mathcal{B}_j) \right| ,$$

and the supremum is taken over all pairs of finite partitions $\{\mathcal{A}_1, \ldots, \mathcal{A}_I\}$ and $\{\mathcal{B}_1, \ldots, \mathcal{B}_J\}$ of Ω such that $\mathcal{A}_i \in \sigma\{W_{t'} : 0 \leqslant t' \leqslant t\}$, $i = 1, \ldots, I$, and $\mathcal{B}_j \in \sigma\{W_{t'} : t' \geqslant t + l\}$, $j = 1, \ldots, J$. Let $\delta_w(\mathcal{A}) = \mathbb{1}\{w \in \mathcal{A}\}$ for any $\mathcal{A} \in \mathcal{B}(\mathcal{Y} \times \mathcal{X})$. By Proposition 4 in Liebscher (2005), $\{W_t\}$ is β-mixing with geometrically decaying mixing numbers if (a) $\int_{\mathcal{Y}\times\mathcal{X}} V_X(s_0)\delta_w(dw_0) = V_X(s) < \infty$, and (b) $\{W_t\}$ is Q-geometrically ergodic in the sense of Liebscher (2005) with $Q(w) = V_X(s)$. Condition (a) holds for all $w = (y', s')' \in \mathcal{Y} \times \mathcal{X}$. For condition (b), we first need to show that $\int_{\mathcal{Y}\times\mathcal{X}} V_X(s_n)\pi_W(dw_n) < \infty$. This follows from

$$\int_{\mathcal{Y}\times\mathcal{X}} V_X(s_n)\pi_W(dw_n) = \int_{\mathcal{X}} V_X(s_n)\pi_X(ds_n) \int_{\mathcal{Y}} \pi_{Y|S}(dy_n|s_n) < \infty,$$

where the last inequality follows from the V_X-geometric ergodicity of $\{X_t\}$. As for the remaining part of condition (b), notice that from Lemma 3.9.6(ii) we have that

$\left\|P_W^l(w,\cdot) - \pi_W\right\|_{TV} \leqslant R\tilde{V}_X(\check{s})\rho^l \wedge 1$, where

$$\left\|P_W^l(w,\cdot) - \pi_W\right\|_{TV} = \frac{1}{2} \sup_{v:|v|\leqslant 1} \left| \int_{\mathcal{Y}\times\mathcal{X}} \left[P_W^l\left(w, dw_l\right) - \pi_W(dw_l)\right] v(w_l) \right| ,$$

which completes the proof of condition (b). It remains to be shown that the rate of decay does not depend on θ. For any probability measure τ on $\mathcal{Y} \times \mathcal{X}$, define $\xi_l(\tau) = \int_{\mathcal{Y}\times\mathcal{X}} \left\|P_W^l(w,\cdot) - \pi_W\right\|_{TV} \cdot \tau(dw)$. By virtue of part (ii) of Lemma 3.9.6 we compute that $\xi_l(\pi_W) \leqslant R\check{V}\rho^l$, where

$$\begin{aligned}\check{V} = 2 &+ \left\|\left\|H_1^G\right\|_1\right\|_{L_1} + (1 + 2\overline{A})C_s\left\|\left\|Y_1^G\right\|_1\right\|_{L_1} + (1 + \overline{B}) \sup_{\theta\in\Theta} \left\|\left\|f_{\theta 1}^G\right\|_1\right\|_{L_1} \\ &+ \overline{\beta}_1 \sup_{\theta\in\Theta} \left\|d_{\theta 1}^G\right\|_{L_1} ,\end{aligned}$$

and $\xi_l(\delta_w) = \left\|P_W^l(w,\cdot) - \pi_W\right\|_{TV} \leqslant R\tilde{V}_X(\check{s})\rho^l \wedge 1$. Now, by Proposition 3 in Liebscher (2005) we have that for any $w \in \mathcal{Y} \times \mathcal{X}$, and $m = \lfloor l/2 \rfloor$, $\beta(l) \leqslant 3\xi_m(\delta_w) + \xi_m(\pi_W) \leqslant R\left(\check{V} + 3\tilde{V}_X(\check{s})\right)\rho^m \wedge 1$. It is not difficult to verify that $\alpha(l) \leqslant \beta(l) \leqslant \exp\left(-C_\alpha l^{r_\alpha}\right) \wedge 1$ for all $l \geqslant 1$. The choice of C_α and r_α depends on $R, \check{V}, \rho$ and $\tilde{V}_X(\check{s})$. Note that the rate of decay of the uniform bound for the α-mixing coefficients does not depend on θ (Lemma 3.9.3). The claim follows by redefining R and noting that $\tilde{V}_X \geqslant 1$.

(iii) The existence of the stationary distribution with k moments of $\{(Y_t, S_t')'\}$ follows from its V_W-geometric ergodicity, which is established in Lemma 3.9.6.

□

The verification of Condition 3.3.1 concludes with the following result.

Lemma 3.9.7. *Suppose Proposition 3.9.1 holds. Then, Condition 3.3.1 holds.*

Proof. Condition 3.3.1(*i*) is verified by finding a suitable compact set $\Theta \subset \mathbb{R}^p$ compatible with A.3.3.2(*i*) and A.3.3.2(*ii*). For example, let $\Theta = \Theta_\omega \times \Theta_A \times \Theta_B \times \Theta_\lambda$, where

$$\Theta_\omega = \{\omega \in \mathbb{R}^{p_\omega} : \|\omega\|_1 \leqslant \overline{\omega} < \infty\},$$

$$\Theta_A = \{\mathrm{vec}(A) \in \mathbb{R}^{p_A} : 0 < \underline{A} \leqslant |\det(A)|, \|\!|A|\!\|_1 \leqslant \overline{A}\},$$

$$\Theta_B = \{\mathrm{vec}(B) \in \mathbb{R}^{p_B} : \|\!|B|\!\|_1 \leqslant \overline{B}\},$$

$$\Theta_\lambda = \{\lambda \in \mathbb{R}^{p_\lambda} : \|\lambda\|_1 \leqslant \overline{\lambda} < \infty\},$$

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.[18] Condition 3.3.1(*ii*) holds because $l_t(\theta, \tau) = \frac{1}{N}\sum_{i=1}^N \rho_{\tau_i}(Y_{it} - f_{\theta\, it})$

[18]The fact that Θ_A is compact and nonempty is verified in section 3.13.

and $d_{\theta t}$ are both measurable functions of W_t, which is 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 3.9.1. To verify Condition 3.3.1(*iii*), note that by (3.3), one can write

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

Thus, for all $t \geqslant 1$,

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

but by Proposition 3.9.1, $\sup_{t\geqslant 1}\left\|\|Y_t\|_1\right\|_{L_k} < \infty$, $\sup_{t\geqslant 1}\sup_\Theta \left\|\|f_{\theta t}\|_1\right\|_{L_k} < \infty$ and $\sup_{t\geqslant 1}\sup_\Theta \|d_{\theta t}\|_{L_k} < \infty$, which completes the proof. □

3.10 Dominating process

Lemma 3.10.1. *Suppose A.3.3.1, A.3.3.2 and A.3.3.3 hold. Then, for all $t \geqslant 1$, we have $\|f_{\theta t} - f_{\dot\theta t}\|_1 \leqslant \delta d_{\dot\theta t}$, where $\delta > 0$.*

Proof. Let $a_{\theta t} := 1 + C_s\left(1 + \overline{A}\right)\|Y_t\|_1 + \|f_{\theta t}\|_1$, where the constants $C_s < \infty$ and $\overline{A} < \infty$ are given in A.3.3.2. By (3.1), A.3.3.2(*i*), A.3.3.2(*ii*) and A.3.3.2(*iv*) and adding and subtracting $(As_{\dot\lambda}(Y_{t-1}) + Bf_{\dot\theta\, t-1})$, it holds that[19] $\|\theta - \dot\theta\|_1 \leqslant \delta$

[19]Note that

$$\begin{aligned}\|As_\lambda(Y_{t-1}) - \dot{A}s_{\dot\lambda}(Y_{t-1})\|_1 &\leqslant \|A\|_1\|s_\lambda(Y_{t-1}) - s_{\dot\lambda}(Y_{t-1})\|_1 + \|A - \dot{A}\|_1\|s_{\dot\lambda}(Y_{t-1})\|_1 \\ &\leqslant \delta C_s\left(1 + \overline{A}\right)\|Y_{t-1}\|_1.\end{aligned}$$

implies that

$$\begin{aligned}
&\|f_{\theta t} - f_{\dot{\theta} t}\|_1 \\
&= \|\omega - \dot{\omega} + As_{\lambda}(Y_{t-1}) - \dot{A}s_{\dot{\lambda}}(Y_{t-1}) + B(f_{\theta t-1} - f_{\dot{\theta} t-1}) - (\dot{B} - B)f_{\dot{\theta} t-1}\|_1 \\
&\leqslant \delta a_{\dot{\theta} t-1} + \overline{B}\|f_{\theta t-1} - f_{\dot{\theta} t-1}\|_1
\end{aligned}$$

by the triangular inequality and properties of the ℓ_1-norm.[20]

Note that for $t = 1$, $\|f_{\theta 1} - f_{\dot{\theta} 1}\|_1 = 0 \leqslant \delta \cdot 1 \leqslant \delta d_{\dot{\theta} 0}$, and by induction for all $t > 1$,

$$\|f_{\theta t} - f_{\dot{\theta} t}\|_1 \leqslant \delta \left(a_{\dot{\theta} t-1} + \overline{B} d_{\dot{\theta} t-1} + \epsilon_{dt}\right) = \delta d_{\dot{\theta} t},$$

where the last inequality holds because ϵ_{dt} is non-negative. □

Lemma 3.10.2. *Suppose Lemma 3.10.1 holds. Then, the inequality in* (3.6) *is satisfied.*

Proof of Lemma 3.10.2. First, by (3.3) one can write

$$\begin{aligned}
\left|l_t(\theta, \tau) - l_t(\dot{\theta}, \tau)\right| &\leqslant \frac{1}{N}\sum_{i=1}^{N} |\rho_{\tau_i}(Y_{it} - f_{\theta it}) - \rho_{\tau_i}(Y_{it} - f_{\dot{\theta} it})| \\
&\leqslant \frac{1}{N}\sum_{i=1}^{N} |f_{\dot{\theta} it} - f_{\theta it}| = \frac{1}{N}\|f_{\theta t} - f_{\dot{\theta} t}\|_1 \leqslant \frac{1}{N}\delta d_{\dot{\theta} t}\,,
\end{aligned}$$

where the second line follows by noting that for all $i = 1, \ldots, N$, ρ_{τ_i} is $(\tau_i \vee (1 - \tau_i))$-Lipschitz with $\tau_i \in [0, 1]$ (so $\max_{i=1,\ldots,N}\{\tau_i \vee (1 - \tau_i)\} \leqslant 1$) and by Lemma 3.10.1. □

[20] In particular, that for any $m \times n$ matrix A and n-vector x, it holds that $\|Ax\|_1 \leqslant |||A|||_1 \cdot \|x\|_1$.

3.11 Multi-step ahead direct forecasts

For finite $\mathsf{h} \geqslant 1$, the direct forecast strategy is based on equation (3.1) but modifying the loss function to be

$$l_{t,\mathsf{h}}(\theta, \tau) = \frac{1}{N} \sum_{i=1}^{N} \rho_{\tau_i}(Y_{i\,t+\mathsf{h}-1} - f_{\theta\,i\,t}), \quad \text{for } 1 \leqslant t \leqslant T - \mathsf{h} + 1.$$

In other words, we compute

$$\hat{\theta}^{\mathsf{h}}_{T,\tau} \in \arg\min_{\theta \in \Theta} \frac{1}{T - \mathsf{h} + 1} \sum_{t=1}^{T-\mathsf{h}+1} l_{t,\mathsf{h}}(\theta, \tau).$$

It is clear that $l_{t,\mathsf{h}}(\theta, \tau)$ is a measurable transformation of $W_t = (Y_t', H_t', f_{\theta\,t}', d_{\theta\,t})'$ and $W_{t+\mathsf{h}}$, and hence it also satisfies Condition 3.3.1.

3.12 Multi-step ahead iterated forecasts

For finite $\mathsf{h} \geqslant 1$, consider the following h-step ahead forecast strategy:

$$\begin{aligned}
f_{\theta\,t+1|t} &= f_{\theta\,t+1} = \omega + As_\lambda(Y_t) + Bf_{\theta\,t} \\
f_{\theta\,t+2|t} &= \omega + As_\lambda(f_{\theta\,t+1|t}) + Bf_{\theta\,t+1|t} = \omega + As_\lambda(f_{\theta\,t+1}) + Bf_{\theta\,t+1} \\
f_{\theta\,t+3|t} &= \omega + As_\lambda(f_{\theta\,t+2|t}) + Bf_{\theta\,t+2|t} \\
&\vdots \\
f_{\theta\,t+\mathsf{h}|t} &= \omega + As_\lambda(f_{\theta\,t+\mathsf{h}-1|t}) + Bf_{\theta\,t+\mathsf{h}-1|t}.
\end{aligned}$$

Let $W_t = (Y_t', H_t', f_{\theta\,t}', d_{\theta\,t})'$. Suppose that $\{W_t\}$ is α-mixing with mixing coefficients $\alpha_W(m)$. From the above, it follows that there exists a measurable transfor-

mation g such that

$$V_{t+\mathsf{h}} = \begin{bmatrix} Y_{t+\mathsf{h}} \\ \\ f_{\theta\, t+\mathsf{h}|t} \end{bmatrix} = g(W_{t+\mathsf{h}}, W_t).$$

Thus, $\{V_t\}$ is also α-mixing by Davidson (1994, Theorem 14.1), with $\alpha_V(m) \leqslant \alpha_W(m-\mathsf{h})$ for $m \geqslant \mathsf{h}$. Thus, the loss function defined as

$$l_{t,\mathsf{h}}(\theta,\tau) := \frac{1}{N}\sum_{i=1}^{N} \rho_{\tau_i}(Y_{i\,t} - f_{\theta\, i\, t|t-\mathsf{h}}), \quad \text{for } t \geqslant \mathsf{h}$$

is also α-mixing with coefficients $\alpha_V(m) \leqslant \alpha_W(m-\mathsf{h})$ for $m \geqslant \mathsf{h}$. It still needs to be shown that Condition 3.3.1(*iii*) holds. To see this, write

$$|l_{t,\mathsf{h}}(\theta,\tau)| \leqslant \frac{1}{N}\|Y_t\|_1 + \frac{1}{N}\|f_{\theta\, t|t-\mathsf{h}}\|_1.$$

By Proposition 3.9.1, $\sup_{t\geqslant 1} \|\|Y_t\|_1\|_{L_k} < \infty$, and

$$\sup_{t\geqslant 1}\sup_{\Theta} \|\|f_{\theta\, t}\|_1\|_{L_k} \leqslant \frac{\overline{\omega} + C_s\overline{A}\sup_{t\geqslant 1}\|\|Y_t\|_1\|_{L_k}}{1-\overline{B}} := C_0 < \infty.$$

Thus, by A.3.3.2,

$$\begin{aligned}
&\sup_{t\geqslant 1}\sup_{\Theta} \|\|f_{\theta\, t+1|t}\|_1\|_{L_k} \\
&\leqslant \overline{\omega} + \overline{A}C_s \sup_{t\geqslant 1}\|\|Y_t\|_1\|_{L_k} + \overline{B}\sup_{t\geqslant 1}\sup_{\Theta}\|\|f_{\theta\, t}\|_1\|_{L_k} := C_1 < \infty, \\
&\sup_{t\geqslant 1}\sup_{\Theta} \|\|f_{\theta\, t+2|t}\|_1\|_{L_k} \leqslant \overline{\omega} + \left(\overline{A}C_s + \overline{B}\right) C_1 := C_2 < \infty, \\
&\vdots \\
&\sup_{t\geqslant 1}\sup_{\Theta} \|\|f_{\theta\, t+\mathsf{h}|t}\|_1\|_{L_k} \leqslant \overline{\omega} + \left(\overline{A}C_s + \overline{B}\right) C_{\mathsf{h}-1} := C_{\mathsf{h}} < \infty.
\end{aligned}$$

By combining the results above, it follows that $\sup_{t\geqslant 1}\sup_{\Theta}\|l_{t,\mathsf{h}}(\theta,\tau)\|_{L_k} < \infty$, as expected. This shows that Condition 3.3.1 is also satisfied for h-step ahead iterated forecasting.

3.13 Auxiliary Results

Lemma 3.13.1 (Θ_A is compact and nonempty.)**.** *Consider the setup of the proof of Lemma 3.9.7. Then, there exist positive and finite constants $\overline{A}$ and $\underline{A}$ such that Θ_A is compact and nonempty.*

Proof. Recall that

$$\Theta_A = \{\mathrm{vec}(A) \in \mathbb{R}^{p_A^2} : |||A|||_1 \leqslant \overline{A}\} \cap \{\mathrm{vec}(A) \in \mathbb{R}^{p_A^2} : |\det(A)| \geqslant \underline{A}\}.$$

First, it is shown that Θ_A is closed. For any sequence $\{A_k, k = 1, 2, \ldots\}$ with limit A and $A_k \in \Theta_A$, the following inequalities

$$|\det(A)| = \lim_{k\to\infty} |\det(A_k)| \geqslant \underline{A},$$
$$|||A|||_1 = \lim_{k\to\infty} |||A_k|||_1 \leqslant \overline{A}$$

hold by the continuous mapping theorem. Therefore, Θ_A is closed. Furthermore, Θ_A is bounded for any finite $\overline{A}$. To see that the bounds can be chosen so that Θ_A is nonempty, let $\overline{A} = 2$ and $\underline{A} = 1/2$. Then, the identity matrix $\mathbf{I} \in \Theta_A$. □

3.14 Additional Tables

Table 3.2: 95% GaR Marginal Forecast Evaluation: OECD Countries

	Benchmark			AR-GARCH			VFV-A			AR-GARCH-VFV-A		
Country	Cov	Len	CL	Cov	Len	CL	Cov	Len	CL	Cov	Len	CL
AUS	99.24	5.06	0.10	96.21	4.36	21.03	98.48	4.79	7.28	94.70	4.38	21.36
AUT	96.97	4.17	0.09	93.94	3.83	−1.65	95.45	4.11	−0.40	94.70	3.87	4.54
BEL	96.21	2.80	0.09	93.94	2.47	8.63	96.21	2.77	−1.53	93.18	2.66	−0.14
CAN	95.45	3.72	0.09	89.39	3.21	17.78	97.73	3.65	16.82	90.91	3.23	20.86
CHE	98.48	5.10	0.09	95.45	4.42	29.23	99.24	4.94	10.06	95.45	4.43	29.64
DEU	96.21	4.78	0.13	95.45	4.71	3.29	96.21	4.78	−5.16	94.70	4.60	10.74
DNK	93.94	4.41	0.12	93.18	4.32	−0.26	96.21	4.55	4.22	93.18	4.38	−31.73
ESP	90.15	5.18	0.10	92.42	4.91	26.51	93.94	5.19	10.17	94.70	5.08	24.92
FIN	94.70	6.40	0.16	93.18	6.00	5.28	94.70	6.43	−4.79	93.94	6.14	3.42
FRA	93.18	3.80	0.06	93.94	3.84	13.34	92.42	3.74	7.67	95.45	3.89	24.37
GBR	97.73	5.35	0.10	93.94	4.79	9.45	96.97	5.13	21.13	95.45	4.76	26.04
GRC	96.21	12.39	0.26	91.67	10.71	18.42	96.21	12.13	2.82	89.39	11.13	3.96
IRL	87.12	7.09	0.22	91.67	8.01	14.88	91.67	7.55	11.32	91.67	8.04	10.17
ISL	88.64	7.68	0.28	92.42	8.12	6.63	93.94	8.52	7.67	92.42	8.20	4.96
ITA	90.91	3.73	0.09	93.18	3.59	30.98	93.94	3.74	28.67	89.39	3.46	9.23
JPN	87.12	4.61	0.15	88.64	4.83	9.30	85.61	4.70	12.92	90.15	5.03	11.46
KOR	97.73	9.09	0.22	94.70	7.88	9.35	97.73	8.62	3.19	93.94	8.23	5.36
LUX	91.67	4.85	0.18	87.12	4.50	−16.99	90.91	5.06	−25.66	86.36	4.77	−33.96
MEX	93.18	4.17	0.21	93.18	4.03	16.15	96.21	4.78	6.44	91.67	4.50	8.37
NLD	99.24	6.09	0.13	95.45	5.30	15.80	97.73	5.79	4.15	95.45	5.37	16.26
NOR	91.67	4.87	0.15	88.64	4.79	−9.81	95.45	5.54	−11.63	90.15	5.15	−13.58
PRT	94.70	5.49	0.11	93.94	4.91	7.28	96.97	5.46	5.83	95.45	5.00	6.97
SWE	97.73	5.79	0.13	96.21	5.46	6.95	96.97	5.65	3.57	96.21	5.44	9.95
USA	97.73	3.37	0.09	94.70	2.85	20.12	97.73	3.23	8.99	94.70	2.89	24.74

Cov: Average empirical coverage; Len: average empirical length; CL: average check loss (for the Benchmark block) and percentage improvement in average check loss relative to historical benchmark (for the remaining blocks).

Table 3.3: Panel Diebold-Mariano tests statistics of superior predictive ability for the out-of-sample period based on the check loss. The one-sided alternative is that the specification in the row has better predictive ability than the specification in the column. The numbers in parenthesis represent the specifications given in Table 3.1 in order of appearance. For example, Pool. VFV (3) is AR-VFV Sym. The symbols *, **, and *** indicate that results are significant at the 10, 5 and 1% levels.

		Benchmark	Pool. GARCH		Pool. VFV				Pool. GARCH-VFV				QR			
		(1)	(1)	(2)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Benchmark	(1)	–	3.67	5.12	-0.73	5.19	5.34	1.15	3.77	6.20	5.38	5.10	1.40	-0.04	-0.36	-2.42***
Pool. GARCH	(1)	-3.67***	–	3.40	-4.76***	2.16	3.18	-0.58	-0.07	4.43	3.77	3.37	-0.23	-1.31*	-1.63*	-3.02***
	(2)	-5.12***	-3.40***	–	-5.31***	-1.99**	0.01	-4.37***	-3.46***	0.71	2.04	1.59	-2.85***	-3.67***	-4.03***	-4.03***
Pool. VFV	(1)	0.73	4.76	5.31	–	5.13	5.29	1.25	4.95	6.25	5.52	5.12	1.47	0.12	-0.20	-2.35***
	(2)	-5.19***	-2.16**	1.99	-5.13***	–	2.14	-2.77***	-2.20**	4.02	2.74	2.63	-1.74**	-2.69***	-3.04***	-3.59***
	(3)	-5.34***	-3.18***	-0.01	-5.29***	-2.14**	–	-6.16***	-3.28***	0.64	0.99	0.90	-3.30***	-4.08***	-4.47***	-4.06***
	(4)	-1.15	0.58	4.37	-1.25	2.77	6.16	–	0.60	4.00	4.61	4.96	0.45	-1.20	-1.57*	-2.94***
Pool. GARCH-VFV	(1)	-3.77***	0.07	3.46	-4.95***	2.20	3.28	-0.60	–	4.45	3.84	3.43	-0.23	-1.34*	-1.65**	-3.03***
	(2)	-6.20***	-4.43***	-0.71	-6.25***	-4.02***	-0.64	-4.00***	-4.45***	–	0.13	-0.07	-3.00***	-3.76***	-4.10***	-4.12***
	(3)	-5.38***	-3.77***	-2.04**	-5.52***	-2.74***	-0.99	-4.61***	-3.84***	-0.13	–	-0.44	-3.26***	-4.01***	-4.29***	-4.18***
	(4)	-5.10***	-3.37***	-1.59*	-5.12***	-2.63***	-0.90	-4.96***	-3.43***	0.07	0.44	–	-3.30***	-4.02***	-4.36***	-4.17***
QR	(1)	-1.40*	0.23	2.85	-1.47*	1.74	3.30	-0.45	0.23	3.00	3.26	3.30	–	-2.20**	-2.31**	-3.05***
	(2)	0.04	1.31	3.67	-0.12	2.69	4.08	1.20	1.34	3.76	4.01	4.02	2.20	–	-0.88	-2.55***
	(3)	0.36	1.63	4.03	0.20	3.04	4.47	1.57	1.65	4.10	4.29	4.36	2.31	0.88	–	-2.42***
	(4)	2.42	3.02	4.03	2.35	3.59	4.06	2.94	3.03	4.12	4.18	4.17	3.05	2.55	2.42	–

Bibliography

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

Agosto, A., Cavaliere, G., Kristensen, D., and Rahbek, A. (2016). Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX). *Journal of Empirical Finance*, **38**, 640–663.

Aielli, G. P. (2013). Dynamic Conditional Correlation: On Properties and Estimation. *Journal of Business & Economic Statistics*, **31**(3), 282–299.

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.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and Forecasting Realized Volatility. *Econometrica*, **71**, 579–625.

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.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2021). Using Wasserstein Generative Adversarial Networks for the design of Monte Carlo simulations. *Journal of Econometrics*.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005a). Clustering with Bregman Divergences. *Journal of Machine Learning Research*, **6**, 1705–1749.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005b). Clustering with Bregman Divergences. *Journal of Machine Learning Research*, **6**, 1705–1749.

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.

Bauschke, H. G. and Borwein, J. M. (1997). Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, **4**(1), 27–67.

Bauwens, L., Laurent, S., and Rombouts, J. (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, **21**, 79–109.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model. *The Review of Economics and Statistics*, **72**(3), 498–505.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes*. Springer, New York, Second edition.

Bougerol, P. and Picard, N. (1992). Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics*, **52**, 115–127.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, **7**(3), 200 – 217.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, **16**, 199 – 231.

Brownlees, C. and Engle, R. (2017). SRISK: A Conditional Capital Shortfall Measure of Systemic Risk. *Review of Financial Studies*, **30**(1), 48–79.

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.

Carrasco, M. and Chen, X. (2002). Mixing and Moment Properties of various GARCH and Stochastic Volatility Models. *Econometric Theory*, **18**, 17–39.

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*.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA.

Chan, K. (1990). Deterministic stability, stochastic stability, and ergodicity. In H. Tong, editor, *Non-linear Time Series: A Dynamical System Approach*, pages 448–466. Clarendon Press.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, **91**(434), 862–872.

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.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21**(1), C1–C68.

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

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized Autoregressive Score Models with Applications. *Journal of Applied Econometrics*, **28**, 777–795.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press, New York.

De Nard, G., Ledoit, O., and Wolf, M. (2021). Factor models for portfolio selection in large dimensions: The good, the better and the ugly. *Journal of Financial Econometrics*. Forthcoming.

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

Dey, D. K. and Srinivasan, C. (1985). Estimation of a Covariance Matrix under Stein's Loss. *The Annals of Statistics*, **13**(4), 1581–1591.

Dhillon, S. I. and Tropp, A. J. (2007). Matrix Nearness Problems with Bregman Divergences. *SIAM Journal on Matrix Analysis and Applications*, **29**.

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. (2002). Dynamic Conditional Correlation. *Journal of Business and Economic Statistics*, **20**(3), 339–350.

Engle, R. (2009). *Anticipating correlations: a new paradigm for risk management*. Princeton University Press.

Engle, R. and Kelly, B. (2012). Dynamic Equicorrelation. *Journal of Business & Economic Statistics*, **30**(2), 212–228.

Engle, R. and Russell, J. R. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica*, **66**, 1127–1162.

Engle, R. and Sheppard, K. (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate garch. Technical report, NBER. Working Paper No. W8554.

Engle, R. F. and Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, **131**, 3–27.

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.

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business and Economic Statistics*, **37**(2), 363–375.

Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, **107**(498), 592–606.

Francq, C. and Zakoïan, J. (2010). *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley.

Francq, C. and Zakoïan, J.-M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli*, **10**, 605 – 637.

Francq, C. and Zakoïan, J.-M. (2006). Mixing Properties of a General Class of GARCH(1,1) Models without Moment Assumptions on the Observed Process. *Econometric Theory*, **22**, 815–834.

Francq, C., Wintenberger, O., and Zakoian, J.-M. (2013). GARCH models without positivity constraints: Exponential or Log GARCH? *Journal of Econometrics*, **177**(1), 34–46.

Gallant, A. R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Basil Blackwell.

Ghysels, E., Gourieroux, C., and Jasiak, J. (2004). Stochastic volatility duration models. *Journal of Econometrics*, **119**(2), 413–433.

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

Gijbels, I., Pope, A., and Wand, M. P. (1999). Understanding Exponential Smoothing via Kernel Regression. *Journal of the Royal Statistical Society: Series B*, **61**, 39–50.

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

Hansen, P. R. and Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, **131**, 97–121.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails With Applications to Financial and Economic Time Series*. Cambridge University Press.

Harvey, A. C. and Chakravarty, T. (2008). Beta-t-(E)GARCH. Cambridge working papers in economics, Faculty of Economics, University of Cambridge.

Hautsch, N. and Voigt, S. (2019). Large-scale portfolio allocation under trans-

action costs and model uncertainty. *Journal of Econometrics*, **212**(1), 221 – 240.

Hautsch, N., Kyj, L. M., and Malec, P. (2015). Do High-Frequency Data Improve High-Dimensional Portfolio Allocations? *Journal of Applied Econometrics*, **30**(2), 263–290.

Higham, N. J. (2002). Computing the nearest correlation matrix–a problem from finance. *IMA Journal of Numerical Analysis*, **22**(3), 329–343.

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.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, **58**(4), 1651–1683.

Jiang, W. and Tanner, M. (2010). Risk Minimization for Time Series Binary Choice with Variable Selection. *Econometric Theory*, **26**, 1437–1452.

Kendall, S. and Stuart, A. (1979). *The Advanced Theory of Statistics (4ed.), Volume 2*. MacMillan Publishing Co., Inc.

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.

Kristensen, D. (2009). On stationarity and ergodicity of the bilinear model with applications to GARCH models. *Journal of Time Series Analysis*, **30**, 125–144.

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.

Kulis, B., Sustik, M. A., and Dhillon, I. S. (2009). Low-Rank Kernel Learning with Bregman Matrix Divergences. *J. Mach. Learn. Res.*, **10**, 341–376.

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.

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

Laurent, S., Rombouts, J., and Violante, F. (2013). On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics*, **173**(1), 1–10.

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

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, **15**(5), 850–859.

Ledoit, O. and Wolf, M. (2011). Robust Performances Hypothesis Testing With the Variance. *Wilmott*, **2011**(55), 86–89.

Ledoit, O. and Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli*, **24**(4B), 3791 – 3832.

Ledoit, O. and Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, **48**(5), 3043 – 3065.

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

Liebscher, E. (2005). Towards a Unified Approach for Proving Geometric Ergod-

icity and Mixing Properties of Nonlinear Autoregressive Processes. *Journal of Time Series Analysis*, **25**, 669–689.

Ling, S. and McAleer, M. (2003). Asymptotic Theory for a Vector ARMA-GARCH Model. *Econometric Theory*, **19**, 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.

Meddahi, N. (2003). ARMA representation of integrated and realized variances. *The Econometrics Journal*, **6**, 335–356.

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**, 1291–1320.

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

Mendelson, S. (2015). Learning without concentration. *Journal of the ACM*, **62**, 1–25.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*. Springer, London.

Morris, C. N. (1982). Natural Exponential Families with Quadratic Variance Functions. *The Annals of Statistics*, **10**, 65 – 80.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, **31**(2), 87–106.

Nelson, D. B. (1992). Filtering and forecasting with misspecified ARCH models I: Getting the right variance with the wrong model. *Journal of Econometrics*, **52**, 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.

Pakel, C., Shephard, N., Sheppard, K. K., and Engle, R. F. (2018). Fitting Vast Dimensional Time-Varying Covariance Models. *NYU Working Paper No. FIN-08-009*.

Patton, A. (2020). Comparing Possibly Misspecified Forecasts. *Journal of Business and Economic Statistics*, **38**(4), 796–809.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, **160**, 246–256.

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.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. Springer.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Inc.

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.

Silvennoinen, A. and Teräsvirta, T. (2008). Multivariate GARCH models. SSE/EFI Working Paper Series in Economics and Finance 669, Stockholm School of Economics.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, **34**(1), 1373.

Straumann, D. and Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics*, **34**, 2449 – 2495.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Clarendon Press.

Tse, L. D. Y. K. and Tsui, A. K. C. (2002). Evaluating the hedging performance of the constant-correlation GARCH model. *Applied Financial Economics*, **12**, 791–798.

Van Os, B. and Van Dijk, D. (2021). Pooling Dynamic Conditional Correlation Models. Technical report, Erasmus University.

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*. Nauka.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, **28**(2), 3–28.

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*, **52**, 143–162.

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.

Zumbach, G. (2007). The Riskmetrics 2006 Methodology. *SSRN Electronic Journal*.