Essays in Time Series Econometrics and Machine Learning

Inauguraldissertation zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von

Giovanni Ballarin

im Frühjahrs-/Sommersemester 2024

Abteilungssprecher:	Prof. Klaus Adam, Ph.D.
Referent:	Prof. Dr. Carsten Trenkler
Koreferent:	Prof. Lyudmila Grigoryeva, Ph.D.
Vorsitzender der Disputation:	Prof. Dr. Markus Frölich
Tag der Disputation:	24.05.2024

This page is intentionally left blank.

Contents

Acknowledgments

First and foremost, I would like to thank my main advisor, Carsten Trenkler, who has supported me and taught me so much throughout my research years at the Chair of Empirical Economic Research. Indeed, it was because of his lectures in econometrics and time series that I decided to work within these fields in the first place. Without his advice, I would not have been able to complete any of the work contained in this dissertation. It is often said that support from ones' doctoral advisor is vital; Carsten has been unwavering in his, and for this, and all else, I will be forever grateful.

The endless energy that Lyudmila Grigoryeva, my co-advisor, has channeled and shared with me has been another pillar of my work over the past couple of years. Her passion, knowledge and kindness have helped push me further than I could have ever imagined on my own – like the rising wind spiriting a boat fast across the sea. Our collaboration has been and is a high point of my research, and I only wish we could keep the pace going for many years to come.

The research in this manuscript owes much to collaboration: Petros Dellaportas, Marcel Hirt, Sophie van Huellen, and Juan-Pablo Ortega have been all fantastically generous in sharing their time and expertise when working together. I am further thankful to Petros and Juan-Pablo for their aid and encouragement during the job market.

It was a great luck for me to be able to receive feedback from and have discussion with the faculty in Mannheim, and for this I would like to thank Cathrine Aeckerle-Willems, Mengshan Xu, Matthias Meier and Yoshiyasu Rai. Moreover, I am especially indebted to Christoph Rothe: working as a teaching assistant for his Advanced Econometrics class has been a formative experience for me, given that, as it is often said, one does not truly know a topic until one has tried teaching it first. And over the past few years, Christoph's thoughtful advice has deeply helped me develop the way I approach research and, in earnest, I believe I have grown into a much better researcher because of it.

I am most grateful for the administrative support that I received throughout my time in Mannheim – from both the GESS and the Department of Economics – from Golareh Khalilpour, Marion Lehnert, Caroline Mohr, Ulrich Kehl, and Sylvia Rosenkranz. Special thanks go to Anja Dostert and Regina Mannsperger for their outstanding help and patience in dealing with the many bureaucratic procedures that are naturally involved in academic work and travel: without them I would have quickly gotten lost in the thick of it. I gratefully acknowledge the support of the state of Baden-Württemberg through the bwHPC cluster; the CDSE for providing travel funding through the 2020 CDSE Teaching Award; the generosity of the University of Warwick, KOF ETH Zürich and the University of St. Gallen for hosting me on several occasions; and financial support from UK Research and Innovation (grant number ES/V006347/1).

One can hardly choose one's colleagues, and, as chance goes, I am convinced that in my graduate studies I have had more luck than anyone can ever hope for. Indeed, it was the friendship and

support of many of my peers within the CDSE doctoral program that allowed me to succeed in my graduate studies: Andrés, Boris, Laura, Lukas, Oliver, Jacopo, Jonathan, Suzanne, Tommaso, Jasmina, Valentina – it was a great gift to be in the same cohort as you all. I will always look back fondly at all the fun we had together in the office and outside. I thank Claudia Noack and Tomasz Olma for welcoming me in the Econometrics group, as well as Jonas Krampe for his insights and kindness in sharing his deep expertise in statistics with me. So Jin, I am terribly grateful for your friendship, and the many (work) conversations, discussions, laughs and dinners together over the years. Among my friends outside work, I must thank Anna, Love, Yael and Benjamin for their continued affection, fellowship and camaraderie: these precious things above others make life worth its while.

My family's love has been a true constant throughout my studies – both undergraduate and graduate – and has guided me over the good and bad times alike. For them, and my mother and father especially, a few words as acknowledgement are not quite enough. Thus, I dedicate this work in earnest to them, and all the affection, guidance and counsel that they have given me during my whole life, in all their uncountable forms.

Lastly and most importantly, I thank Leonard, who has stuck with me over these years and who has actually made it possible to reach this achievement. In sharing his time – the most precious of resources – with me, I can only hope he has found as plentiful and wondrous a spring of joy as I have in him: my whole heart is with you.

Preface

Good walkers leave no track.
Good talkers don't stammer.
Good counters don't use their fingers.
The best door's unlocked and
unopened.
The best knot's not in a rope and
can't be untied.

Tao Te Ching, Book I, Verse 27
Lao Tzu
(Adapted by Ursula K. Le Guin)

This dissertation collects three works developed on the broad topic of time series analysis, with a specific focus on machine learning, non- and semi-parametric methods, and regularization. In particular, the discussion will take an econometric perspective with respect to the three key problems of estimation, forecasting and inference.

A quite important thing to note is the fact that until the mid-2010s there was somewhat of a chasm between machine (statistical) learning[1] (ML) and econometrics (Stapleford, 2021). This is somewhat natural: most economic data has historically been (and still often is) costly to obtain,[2] therefore a *practical* desire or need to sieve through Big Data has emerged only relatively recently. Moreover, *theoretical* analysis of many ML techniques simply lags behind empirical implementation due to how comparably cheaper and easier it is for most researchers to gain access to ever-improving computer hardware and software. However, in the past few years there has been a growing push towards marrying empirically successful learning techniques with the theoretical rigor of econometric and statistical analysis. Early reticence and subsequent enthusiasm can thus be read as normal tides within a field (economics) which – as many others in the social and natural sciences alike – may come across as quite conservative in its methods, yet also rather intent in keeping up with broader trends. Indeed, what economists and econometricians most often demand are specific conceptual, statistical, and empirical *guarantees*, which ensure that a given methodological approach is able to effectively and correctly inform policy-making. In the vast majority of interesting settings, such guarantees are not trivial to derive. The contents of this dissertation fall within this context and attempt to bridge the gap between learning ideas and econometric theory and practice, with a specific focus on data that has a prominent time structure.

Chapter 1 develops a new ML approach to forecast economic time series – focusing on US GDP

[1] Intended here as a subject with specific and often rather distinct characteristics within the broader fields of *both* computer science and statistics.

[2] Macroeconomic data is, in fact, *the* prime example of "expensive" economic data to collect, since by definition it is a byproduct of the massive accounting, financial and actuarial work done by private and public institutions.

growth – within an environment consisting on many series with observations sampled at different frequencies. We introduce a method that is based on a reservoir computing approach, which, broadly speaking, leverages the universal approximation properties of nonlinear state-space models with *random* coefficients matrices. Our proposed scheme is computationally efficient, empirically effective – reaching or surpassing state-of-the-art forecasting performance – and straightforward to implement even when there are many different data frequencies.

Chapter 2 deals instead with the important question of regularization in the estimation of linear time series. Vector autoregressive models (VARs) are a fundamental benchmark and foundational analytical tool of modern econometrics. Yet, even in moderate data environments with a few dozen series, estimation of VARs can be severely impacted by efficiency issues – that is, too many parameters need to be recovered compared to the sample size. This is true even in settings that do not fall within the category of high-dimensional processes. Drawing a comparison with Bayesian methods, I propose to apply *anisotropic ridge regression* as an estimation procedure in order to effectively exploit prior information or beliefs on the structure of the VAR model. The theory for inference on impulse responses functions and cross-validation is developed, and in simulations I find that the trade-off of ridge penalization can be positive whenever one is correctly informed about the nature of the underlying data generating process.

Finally, in Chapter 3 I provide a semi-nonparametric approach for the estimation of impulse responses og nonlinear autoregressive models. Impulse response functions (IRFs) are widely studied objects in macroeconometrics, because they quantify the response of a model economy to an unforeseen shock. For example, central banks are often interested in studying the potential effects of credibly exogenous changes in monetary policy over short and long horizons. If one also wants to incorporate *nonlinear* relationships in a model, I prove that estimating the linear and nonlinear (functional) autoregressive coefficients with a semi-nonparametric series approach is a uniformly consistent strategy. In turn, this allows the constructions asymptotically consistent nonlinear IRF estimates – meaning that IRFs can be correctly recovered in large samples. The empirical applications I provide showcase the potential impact of nonlinear IRFs on policy: comparing pointwise linear and nonlinear estimates suggest that linear models can underestimate to varying degrees the negative effects of contractionary monetary policy. This, in turn, provides evidence that proper estimation of nonlinear interactions may lead to better quantitative analysis of macroeconomic dynamics.

Chapter 1

Reservoir Computing for Macroeconomic Forecasting with Mixed Frequency Data

Joint with Petros Dellaportas, Lyudmila Grigoryeva, Marcel Hirt, Sophie van Huellen, and Juan-Pablo Ortega.

1.1 Introduction

The availability of timely and accurate forecasts of key macroeconomic variables is of crucial importance to economic policymakers, businesses, and the banking sector alike. Fundamental macroeconomic figures, such as GDP growth, become available at low frequency with a considerable time lag and are subject to various rounds of revisions after their release. This is particularly problematic in a fast-changing and uncertain economic environment, as experienced during the Great Recession of 2007-2008 (Hindrayanto et al. 2016) and the recent pandemic (Buell et al. 2021, Huber et al. 2021). However, a large number of the potentially predictive financial market (and other macroeconomic) indicators are available at a daily or even higher frequency (Andreou et al. 2013). The desire to utilize such high-frequency data for macroeconomic forecasting has led to the exploration of techniques that can deal with large-scale datasets and series with unequal release periods (see Borio 2011, 2013, Morley 2015; we also refer the reader to Fuleky 2020a for more details regarding high-dimensional data and to Armesto et al. 2010 and Bańbura et al. 2013 for a review on mixed-frequency data).

We contribute to the existing literature by proposing a new macroeconomic forecasting framework that utilizes high-dimensional and mixed-frequency input data, the Multi-Frequency Echo State Network (MFESN). The MFESN originates from a machine learning paradigm called Reservoir Computing (RC). RC is a family of learning models that take advantage of the information processing capabilities of complex dynamical systems (see Maass et al. 2002, Legenstein and Maass 2007, Crutchfield et al. 2010, and Lukoševičius and Jaeger 2009, Tanaka et al. 2019 for reviews). Generally speaking, RC is a versatile class of recurrent neural network (RNN) models (see Salehinejad et al. (2017) for a detailed survey). Although conventional RNNs are well-suited for handling sequence data and dynamic problems, estimating their weights during the training phase is inherently difficult (Pascanu et al. 2013, Doya 1992). Reservoir networks stand out due to the fact that their inner weights can be *randomly generated* and *fixed*, and only the output (readout) layer weights are subject to estimation (supervised training). Echo State Network (ESN) is one of the most popular instances of RC models with provable universality, generalization properties (see Grigoryeva

and Ortega 2018b,a, 2019, Gonon et al. 2020b, 2023a, Gonon and Ortega 2021, and references therein for more details), and excellent performance in forecasting, classification, and learning of dynamical systems (see Hart et al. 2021, Grigoryeva et al. 2021). While conventional RNNs have been adopted for macroeconomic forecasting in a few instances (see, for example, Paranhos 2021), to the best of our knowledge, we are the first to explore easily-trainable reservoir models in this context.

Our main contribution is three-fold. First, inspired by the remarkable empirical success of ESNs in prediction tasks, we propose the so-called Multi-Frequency Echo State Network (MFESN) framework, which allows multistep forecasting of the target variable at lower or the same frequencies as those of the input series. Second, we introduce two different approaches to predicting within the MFESN framework, namely *Single-Reservoir MFESN* (S-MFESN) and *Multi-Reservoir MFESN* (M-MFESN). S-MFESN is determined by modifying the ESN architecture to accommodate input and target variables of mixed frequencies. In M-MFESN, several Echo State Networks are adopted to handle input time series, each ESN corresponding to a group of input variables quoted at one given frequency. Finally, our third contribution consists of an extensive empirical comparative analysis of the forecasting capability of the proposed approaches in a concrete task of predicting the quarterly U.S. output growth. We inspect the forecasting capabilities of the MFESN framework compared to two well-established benchmarks widely used in the macroeconomic literature and among practitioners and show its empirical superiority in several thoroughly conducted forecasting exercises. Moreover, as a bi-product, we propose a new data aggregation scheme that allows bridging these two standard forecasting approaches, which is not available in the literature.

In our empirical study, we evaluate the multistep forecasting performance of the MFESN framework targeting quarterly U.S. output growth – Gross Domestic Product (GDP) growth – and utilizing a small- and medium-sized set of monthly and daily financial and macroeconomic variables. We compare the MFESN approach against two state-of-the-art methods, MIDAS and DFM, known for their ability to incorporate data of heterogeneous frequencies and utilize high-dimensional data inputs. The MIxed DAta Sampling (MIDAS) model developed in Ghysels et al. (2004, 2007) has been adopted widely for macroeconomic forecasting with mixed-frequency data (see for instance Clements and Galvão 2008, 2009, Ghysels and Wright 2009, Francis et al. 2011, Monteforte and Moretti 2012, Galvão and Marcellino 2010, Galvão 2013, Andreou et al. 2013, Ghysels 2016, Jardet and Meunier 2022). However, MIDAS is prone to curse-of-dimensionality problems and performs poorly when the set of predictors is of even moderate size (Clements and Galvão 2009, Kostrov 2021) due to optimization-related issues. Recently, some attempts have been made in the literature to overcome these issues by employing variable selection techniques under some additional assumptions. For instance, Babii et al. (2022) proposes the MIDAS projection approach, which is more amenable to high-dimensional data environments under the assumption of sparsity. Even with these improvements, practical high-dimensional implementations of MIDAS remain challenging. This is in part caused by the ragged edges of the "raw" macroeconomic data, incomplete observations, and uneven sampling frequencies. The relative inflexibility of MIDAS regression lag specifications makes integrating daily and weekly data at true calendar frequencies (that is, without interpolation or aggregation) very complex. State-space models effectively mitigate these issues.

A strong state-of-the-art state-space competitor for our MFESN framework is the Dynamic

Factor Model (DFM), which has been first introduced in Geweke (1977) and Sargent et al. (1977). DFMs have become the standard workhorse for macroeconomic nowcasting and prediction (for more details, we refer the reader to Stock and Watson 1996, 2002, 2016, Giannone et al. 2008, Bańbura and Rünstler 2011, Chauvet et al. 2015, Hindrayanto et al. 2016). Conventional DFMs for data of multiple sampling frequencies are linear state-space models with a latent low-frequency process of interest and high-dimensional input time series. Although their linear structure lends itself to inference with likelihood-based methods and Kalman filtering, using DFMs in the high-dimensional setting is limited by the associated computational effort. For Gaussian state-space models, some of these issues are proposed to be handled with a more compact matrix representation as in Delle Monache and Petrella (2019). Still, in the particular settings of nowcasting and forecasting of GDP growth, the computational complexity is one of the main reasons why DFMs are rarely used with daily input series, see Bańbura et al. (2013) for a detailed review and Aruoba et al. (2009) for a mixed-frequency DFM wherein the latent factor process is updated daily, with the highest input frequency being weekly. We address these numerical difficulties using novel Python libraries for auto-differentiation and using GPUs for parallel computing, which allow the estimation of DFMs even in instances of high-frequency input observations. Further, to adapt the DFM to mixed frequency tasks, we propose a new DFM aggregation scheme with Almon polynomial structure that bridges MIDAS and the DFM for our forecasting comparison. To our knowledge, we are the first to present this aggregation scheme which reduces the number of parameters subject to estimation. In contrast, previous DFM – such as in Mariano and Murasawa (2003), Bańbura and Rünstler (2011), Camacho and Pérez-Quirós (2010), Frale et al. (2011) – commonly assume a fixed aggregation scheme a-priori depending on whether the macroeconomic variable is a flow or stock variable.

To carry out a fair comparison of our MFESN framework with the state-of-the-art MIDAS and DFM models, we designed two model evaluation settings that differ regarding whether the financial crisis of 2007-2008 is included in the estimation period or not. In the first forecasting setting, all the competing models are estimated using the data from January 1st, 1990, until December 31st, 2007. Their performance in the forecasting into and after the financial crisis period is assessed. In the second evaluation setting, fitting is done with data largely encompassing the crisis period, again from January 1st, 1990 but now up to December 31st, 2011. In both cases, the forecasting (testing) period spans time up to the COVID-19 pandemic events, namely the fourth quarter of 2019. Along with the two state-of-the-art DFM and MIDAS models, we use the unconditional mean of the sample as a baseline benchmark against the reservoir models. We find that our ESN-inspired models attain comparable or much better performance than DFMs at a much lower computational cost, even for a relatively long forecasting horizon of four quarters. Additionally, ESNs do not suffer from curse-of-dimensionality problems, which are known to be pervasive for MIDAS models and hence consistently outperform them in a number of forecasting exercises.

The remainder of the paper is structured as follows. Section 1.2 presents reservoir models and discusses their advantages, as well as estimation, hyperparameter tuning, penalization and nonlinear multistep forecasting. In Section 1.3, we introduce the Multi-Frequency Echo State Network (MFESN) framework, propose the single-reservoir and multi-reservoir MFESN models, and spell out their defining features. Section 1.4 contains the empirical study of the comparative

GDP forecasting performance of MFESNs with respect to the set of benchmark models. We assess one-step and multistep forecasting results in several setups, with a small and a medium-sized set of regressors. We fit models with data before and after the 2007-08 financial crisis, and with different estimation windows. Section 1.5 concludes and discusses future research avenues and applications. Finally, the Appendix contains information regarding data sources, forecasting figures and formal details regarding our forecasting setups. Finally, the appendices give detailed information on the implementation of all models, robustness checks and provide additional figures.

1.1.1 Notation

We use the symbol $\mathbb{N}$ (respectively, $\mathbb{N}^+$) to denote the set of natural numbers with the zero element included (respectively, excluded). $\mathbb{Z}$ denotes the set of all integers. We use $\mathbb{R}$ (respectively, $\mathbb{R}_+$) to denote the set of all (respectively, positive excluding zero element) reals. We abbreviate the set $[n] = \{1, \dots, n\}$, with $n \in \mathbb{N}^+$.

Vector notation. A column vector is denoted by a bold lowercase symbol like $\boldsymbol{r}$ and $\boldsymbol{r}^\top$ indicates its transpose. Given a vector $\boldsymbol{v} \in \mathbb{R}^n$, we denote its entries by v_i, with $i \in \{1, \dots, n\}$; we also write $\boldsymbol{v} = (v_i)_{i \in \{1,\dots,n\}}$. The symbols $\boldsymbol{i}_n, \boldsymbol{0}_n \in \mathbb{R}^n$ stand for the vectors of length n consisting of ones and of zeros, respectively. Additionally, given $n \in \mathbb{N}^+$, $\boldsymbol{e}_n^{(i)} \in \mathbb{R}^n$, $i \in \{1, \dots, n\}$ denotes the canonical unit vector of length n determined by $\boldsymbol{e}_n^{(i)} = (\delta_{ij})_{j \in \{1,\dots,n\}}$. For any $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|$ denotes its Euclidean norm.

Matrix notation. We denote by $\mathbb{M}_{n,m}$ the space of real $n \times m$ matrices with $m, n \in \mathbb{N}^+$. When $n = m$, we use the symbols $\mathbb{M}_n$ and $\mathbb{D}_n$ to refer to the space of square and diagonal matrices of order n, respectively. Given a matrix $A \in \mathbb{M}_{n,m}$, we denote its components by A_{ij} and we write $A = (A_{ij})$, with $i \in \{1, \dots, n\}$, $j \in \{1, \dots m\}$. The symbol $\mathbb{I}_n \in \mathbb{D}_n$ denotes the identity matrix, and the symbol $\mathbb{O}_n$ stands for the zero matrix of dimension n. For any $A \in \mathbb{M}_{n,m}$, $\|A\|_2$ denotes its matrix norm induced by the Euclidean norms in $\mathbb{R}^m$ and $\mathbb{R}^n$, and $\|A\|_2 = \sigma_{\max}(A)$, with $\sigma_{\max}(A)$ the largest singular value of A.

Input and target stochastic processes. We fix a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which all random variables are defined. The input and target signals are modeled by discrete-time stochastic processes $\boldsymbol{z} = (\boldsymbol{z}_t)_{t \in \mathbb{Z}}$ and $\boldsymbol{y} = (\boldsymbol{y}_t)_{t \in \mathbb{Z}}$ taking values in $\mathbb{R}^K$ and $\mathbb{R}^J$, respectively. Moreover, we write $\boldsymbol{z}(\omega) = (\boldsymbol{z}_t(\omega))_{t \in \mathbb{Z}}$ and $\boldsymbol{y}(\omega) = (\boldsymbol{y}_t(\omega))_{t \in \mathbb{Z}}$ for each outcome $\omega \in \Omega$ to denote the realizations or sample paths of $\boldsymbol{z}$ and $\boldsymbol{y}$, respectively. Since $\boldsymbol{z}$ can be seen as a random sequence in $\mathbb{R}^K$, we write interchangeably $\boldsymbol{z} : \mathbb{Z} \times \Omega \longrightarrow \mathbb{R}^K$ and $\boldsymbol{z} : \Omega \longrightarrow (\mathbb{R}^K)^{\mathbb{Z}}$. The same applies to the analogous assignments involving $\boldsymbol{y}$.

Temporal notation. Let $(u_t)_{t \in I}$, $u_t \in \mathbb{R}$ be a (scalar) time series with I some index set (in this paper it will always be discrete). Time series $(u_t)_{t \in I}$ will be denoted just as (u_t) when the index set I is specified by the context. We write $u_{s_1:s_2} = (u_t)_{t \in \{s_1,\dots,s_2\}}$ for integers $s_1 < s_2$ and time series (u_t). To define the concept of the sampling frequency, we must introduce an additional series, call it $(v_s)_{s \in J}$. The time index J is not the same as I. We assume that u_t is sampled at the coarsest

rate; equivalently, it has the *lowest* sampling frequency, which we call in what follows the *reference frequency*. In practice, this means that in the same window of time, u_t will be observed at most as frequently as v_s. The case when the sampling frequency of v_s is strictly higher than that of u_t is of primary interest.

We assume that all sampling happens in instants that are evenly spaced in time. Series other than the reference one and with higher sampling frequencies are given an additional time index, the *tempo index*, written $t, *|\kappa$, where κ is the *frequency multiplier*. Our tempo notation assumes that low- and high-frequency series are sampled with temporal *alignment*: this means that the reference time index t and the tempo index $*|\kappa$ have the following properties.

Definition 1.1.1. *A reference time index $t \in \mathbb{N}$ and a tempo index $*|\kappa$ for a given high-frequency $\kappa \in \mathbb{N}^+$ are such that the following relations hold*

(i) $t, 0|\kappa \equiv t$

(ii) $t, \kappa|\kappa \equiv t + 1$

(iii) $t, s|\kappa \equiv t + \lfloor s/\kappa \rfloor, (s \bmod \kappa)|\kappa \quad \text{for } \forall s \in \mathbb{N}$

(iv) $t, -s|\kappa \equiv (t-1) - \lfloor s/\kappa \rfloor, \kappa - (s \bmod \kappa)|\kappa \quad \text{for } \forall s \in \mathbb{N},$

where mod *is the modulo operation and for any $x \in \mathbb{R}$ the floor operator $\lfloor x \rfloor$ outputs the greatest $z \in \mathbb{N}$ such that $z \leq x$.*

Since we can exchange "frequency" and "frequency multiplier" in the tempo notation, we will make no distinction between the two terms in what follows.

Forecasting schemes. The theoretical setup and design of the forecasting exercises conducted in this paper are carefully discussed in Appendix 1.B. There, we formally distinguish between the so-called high-frequency and low-frequecy forecasting in the presence of mixed-frequency data. For more details regarding time series forecasting with economic data, we also refer the reader to Clements and Galvão (2008, 2009), Chen and Ghysels (2010), Jardet and Meunier (2022) and references therein.

1.2 Reservoir Models

In this section, we introduce *reservoir computing* models (Jaeger and Haas, 2004) for forecasting of stochastic time series of a single frequency. We focus on a family of RC systems called *Echo State Networks* (ESNs), which have been successfully applied to forecasting of deterministic dynamical systems (Pathak et al., 2017, 2018, Wikner et al., 2021, Arcomano et al., 2022). In the following, we discuss the linear estimation of ESN model parameters, the hyperparameters tuning, the loss penalty selection, and how to carry out nonlinear forecasting.

1.2.1 Reservoir Models

Reservoir computing (RC) models are nonlinear state-space systems that, in the forecasting setting, are defined by the following equations:

$$\boldsymbol{x}_t = F(\boldsymbol{x}_{t-1}, \boldsymbol{z}_t), \tag{1.1}$$

$$\boldsymbol{y}_{t+1} = h_{\boldsymbol{\theta}}(\boldsymbol{x}_t) + \boldsymbol{\epsilon}_t, \tag{1.2}$$

for all $t \in \mathbb{Z}$, where the *state map* $F : \mathbb{R}^N \times \mathbb{R}^K \to \mathbb{R}^N$, $N, K \in \mathbb{N}^+$ is called also the *reservoir map*, and the *observation map* $h_{\boldsymbol{\theta}} : \mathbb{R}^N \to \mathbb{R}^J$, $J \in \mathbb{N}^+$ is referred to as the *readout* layer, parametrized by $\boldsymbol{\theta} \in \Theta$. Sequences $(\boldsymbol{z}_t)_{t\in\mathbb{Z}}$, $\boldsymbol{z}_t \in \mathbb{R}^K$, and $(\boldsymbol{y}_t)_{t\in\mathbb{Z}}$, $\boldsymbol{y}_t \in \mathbb{R}^J$, stand for the *input* and the *output (target)* of the system, respectively, and $(\boldsymbol{x}_t)_{t\in\mathbb{Z}}$, $\boldsymbol{x}_t \in \mathbb{R}^N$, are the associated *reservoir states*. In (1.2), $(\boldsymbol{\epsilon}_t)_{t\in\mathbb{Z}}$ are J-dimensional independent zero-mean innovations with variance $\sigma_\epsilon^2 \mathbb{I}_J$ that are also independent of $\boldsymbol{x}_t$ across all t. Importantly, many families of RC systems have been proven to have universal approximation properties for L^p-integrable stochastic processes (Gonon and Ortega, 2020), and estimation and generalization error bounds have been established in Gonon et al. (2020b, 2023a).

In the case of an ESN model, the state and observation equations (1.1)-(1.2) are given by

$$\boldsymbol{x}_t = \alpha \boldsymbol{x}_{t-1} + (1-\alpha)\sigma(A\boldsymbol{x}_{t-1} + C\boldsymbol{z}_t + \boldsymbol{\zeta}) \tag{1.3}$$

$$\boldsymbol{y}_{t+1} = \boldsymbol{a} + W^\top \boldsymbol{x}_t + \boldsymbol{\epsilon}_t, \tag{1.4}$$

where $A \in \mathbb{M}_N$ is the *reservoir matrix*, $C \in \mathbb{M}_{N,K}$ is the *input matrix*, $\boldsymbol{\zeta} \in \mathbb{R}^N$ is the *input shift*, $\alpha \in [0, 1)$ is the *leak rate* and $W \in \mathbb{M}_{N,J}$ are the *readout coefficients*. The map $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function applied elementwise, which in what follows we take to be the hyperbolic tangent. We refer to A, C, $\boldsymbol{\zeta}$ as *state parameters* that are randomly generated. Notice that if $A = 0$ and $\alpha = 0$ the state equation reduces to a nonlinear regression model with random coefficients (or a feedforward neural network with random weights) which is usually referred to as an *Extreme Learning Machine* (Cao et al., 2018, Gonon et al., 2023a).

Properties of ESN models. We focus on ESNs with the so-called *echo state property (ESP)*, that is, when for any $\boldsymbol{z} \in (\mathbb{R}^K)^{\mathbb{Z}}$ there exists a unique $\boldsymbol{y} \in (\mathbb{R}^J)^{\mathbb{Z}}$ such that (1.3)-(1.4) hold (see Grigoryeva and Ortega (2018b,a, 2019) and references therein). One can require that the ESP holds only on the level of the state equation, that is for any input sequence $\boldsymbol{z} \in (\mathbb{R}^K)^{\mathbb{Z}}$ there exists a unique state sequence $\boldsymbol{x} \in (\mathbb{R}^N)^{\mathbb{Z}}$ such that (1.3) holds. The result in Corollary 3.2 in Grigoryeva and Ortega (2018a), which is also valid for the case of ESNs with the leak rate, shows that the sufficient condition of the ESP associated with (1.3) to hold is $\|A\|_2 L_\sigma < 1$ where L_σ is the Lipschitz constant of the activation function σ (in our setting, $L_{tanh} = 1$). This sufficient ESP condition has been extensively studied in the ESN literature; see Jaeger (2010), Jaeger and Haas (2004), Buehner and Young (2006), Bai Zhang et al. (2012), Yildiz et al. (2012), Wainrib and Galtier (2016), Manjunath and Jaeger (2013) for more details. The result in Corollary 3.2 in Grigoryeva and Ortega (2018a) also shows that this condition implies the so-called *fading memory property* (Boyd and Chua, 1985), which from the practical point of view means that the impact of

initial $\boldsymbol{x}_0$ is negligible for sufficiently long samples.

In the stochastic setting, part (i) of Proposition 4.2 in Grigoryeva and Ortega (2021) proves that the condition $\|A\|_2 < 1$ guarantees variance stationarity of the states associated with variance stationary inputs. Moreover, Manjunath and Ortega (2023) show that this condition implies the so-called stochastic state contractivity ensuring a stochastic analog of the ESP. Notably, violations of $\|A\|_2 < 1$ do not have detrimental implications for the performance of ESNs in various learning tasks, as reported in multiple empirical studies.

Computational advantages of ESNs. We emphasize that the core computational advantage of ESNs is that state parameters A, C, and $\boldsymbol{\zeta}$ are randomly sampled and need not be estimated. Additionally, since observation equation (1.4) is linear in $\boldsymbol{x}_t$, coefficients W can be estimated via (penalized) least squares regression, as we explain in the following subsection. The choice of properties of state parameters determines memory properties and forecasting performance of linear (Ballarin et al., 2023) and nonlinear ESNs (Gonon et al., 2020a) as we discuss in Section 1.2.2.

1.2.2 Estimation

We now discuss in detail the estimation of coefficients W in (1.4). Let a sample $(\boldsymbol{z}_t, \boldsymbol{y}_t)_{t=1}^T$ of input and target pairs be available. Given an initial state $\boldsymbol{x}_0$, the reservoir states can be computed iteratively according to state equation (1.3) as:

$$\boldsymbol{x}_1 = \alpha\boldsymbol{x}_0 + (1-\alpha)\sigma(A\boldsymbol{x}_0 + C\boldsymbol{z}_1 + \boldsymbol{\zeta}), \quad \ldots, \quad \boldsymbol{x}_T = \alpha\boldsymbol{x}_{T-1} + (1-\alpha)\sigma(A\boldsymbol{x}_{T-1} + C\boldsymbol{z}_T + \boldsymbol{\zeta}).$$

Collect the states and the targets into the state and the observation matrices, respectively, as

$$X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{T-1})^\top \in \mathbb{M}_{T-1,N}, \qquad Y = (\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_T)^\top \in \mathbb{M}_{T-1,J}.$$

Consider the ridge regression estimator for W given by

$$\widehat{W}_\lambda := \underset{W\in\mathbb{R}^N}{\arg\min} \sum_{t=1}^{T-1} \left\| \boldsymbol{y}_{t+1} - W^\top \boldsymbol{x}_t \right\|_2^2 + \lambda\|W\|_2^2 = \left(X^\top X + \lambda((T-1)\,\mathbb{I}_N)\right)^{-1} X^\top Y, \tag{1.5}$$

where $\lambda \in \mathbb{R}_+$ is the ridge penalty strength. When $\lambda \to 0$, the estimator $\widehat{W}_\lambda$ converges to the minimum-norm least squares solution (Ishwaran and Rao, 2014). In applications, ridge regression is the most commonly used estimation method applied to ESNs, as it provides a straightforward regularization scheme both when $N < T$ and $N \geq T$. This is especially important since in practice the ESN state dimension is often chosen to be 10^3–10^4 (see for example Pathak et al. (2017)). Additionally, a virtue of the ridge regression problem is the fact that the associated objective function is convex and, hence, it can be efficiently solved using stochastic gradient descent even when $\min\{N, T\}$ is large and one decides against the closed-form solution (1.5). Finally, as mentioned in the properties of reservoir systems in Subsection 1.2.1, we notice that in the presence of the fading memory property, the estimation does not depend significantly on the choice of $\boldsymbol{x}_0$ as sample size T increases.

We refer to (1.5) as the *fixed-parameter* estimator. In our empirical analyses, we also implement *expanding* and *rolling window* estimation strategies which update $\widehat{W}_\lambda$ as new observations become

available (we refer the reader to Appendix 1.C.1 for details). In the rest of the paper, for brevity, we use $\widehat{W}$ to denote the ridge estimator of coefficients W assuming that the appropriate choice of the penalty strength λ is made for each concrete situation.

Hyperparameter Tuning

As discussed in Subsection 1.2.1, the performance of ESNs depends on the choice of randomly drawn state parameters A, C, $\boldsymbol{\zeta}$. Much work has been put into determining optimal specifications (see for example Rodan and Tino 2011, Goudarzi et al. 2016, Farkas et al. 2016, Grigoryeva et al. 2015, 2016, Gonon et al. 2020a). We construct these parameters by first sampling $\widetilde{A}$, $\widetilde{C}$ and $\widetilde{\boldsymbol{\zeta}}$ from appropriately chosen laws. Then, we normalize each element of the tuple such that

$$\overline{A} = \widetilde{A}/\rho(\widetilde{A}), \quad \overline{C} = \widetilde{C}/\left\|\widetilde{C}\right\|, \quad \overline{\boldsymbol{\zeta}} = \widetilde{\boldsymbol{\zeta}}/\left\|\widetilde{\boldsymbol{\zeta}}\right\|, \tag{1.6}$$

where $\rho(\widetilde{A})$ denotes the spectral radius of $\widetilde{A}$. As discussed in the properties of reservoir systems in Subsection 1.2.1, the sufficient condition of the ESP is $\|A\|_2 < 1$. By this normalizing choice, we allow for some more flexibility in terms of marginal violations of the non-sharp ESP constraint. Finally, defining $A = \rho\overline{A}$, $C = \gamma\overline{C}$, and $\boldsymbol{\zeta} = \omega\overline{\boldsymbol{\zeta}}$, we can rewrite state equation (1.3) as

$$\boldsymbol{x}_t = \alpha\boldsymbol{x}_{t-1} + (1-\alpha)\sigma(\rho\overline{A}\boldsymbol{x}_{t-1} + \gamma\overline{C}\boldsymbol{z}_t + \omega\overline{\boldsymbol{\zeta}}). \tag{1.7}$$

We refer to tuple $\boldsymbol{\varphi} := (\alpha, \rho, \gamma, \omega)$ as the *hyperparameters* of the ESN. Specifically, $\alpha \in [0, 1)$ is the leak rate and $\rho \in \mathbb{R}_+$ is called the *spectral radius* of the reservoir matrix, $\gamma \in \mathbb{R}_+$ is the *input scaling*, and $\omega \in \mathbb{R}_+$ is the *shift scaling*. The choice of the hyperparameters determines the properties of the state map. For simplicity, in Section 1.4, we choose the hyperparameters based on the empirical ESN literature. In Appendix 1.C.2, we also propose a general though more computationally intensive procedure to select hyperparameters in a data-driven way that could be interesting to practitioners.

Penalty Selection

To apply ridge estimator (1.5), it is necessary to first select a penalty λ. Cross-validation (CV) is a common selection procedure for regularization strength in penalized methods such as ridge, LASSO, and Elastic Net. CV techniques have also been applied in the time series context (Kock et al., 2020, Ballarin, 2023) with their validity established in Bergmeir et al. (2018a).

In our empirical study, to account for temporal dependence, we use a sequential CV strategy with ten validation folds. More precisely, we reserve the last 50 observations for validation and all other previous data points for training. The first fold consists of the first five observations out of the validation set, and the model is fitted using all training data. The following validation fold comprises the next five subsequent validation observations while the training set is expanded by five data points (from the previous fold). This procedure is repeated ten times and the CV loss is the average of the one-step-ahead forecast MSE on each fold. In expanding or rolling window setups, we rerun the CV penalty selection to ensure that estimated ESN coefficients do not induce oversmoothing. We refer the reader to Appendix 1.C.3 for additional details.

1.2.3 Relation to Nonparametric Regression

Together with hyperparameters and penalty strength selection, the choice of the state dimension N is a key ingredient of an ESN model. A large state space generally implies better approximation bounds (Gonon et al., 2023a,b). Although it is customary in the empirical literature to take N as large as possible (Lukoševičius, 2012), some recent literature discusses both the statistical risk bounds and the approximation-risk trade-off bounds for various RC families (see Gonon et al. 2020b and Gonon et al. 2023b for details). Under simplified assumptions that $\alpha = 0$ and $\rho = 0$ in (1.7), ESNs have a natural connection to random-weights neural networks (Cao et al., 2018) and random projection regression (Maillard and Munos, 2012), and are thus comparable to nonparametric sieve methods. If the data were independently sampled, known results on sieve estimation would require that at most $N/T = o(1)$ up to logarithmic factors for consistency (Belloni et al., 2015a). Chen and Christensen (2015) have extended this result to β-mixing data with B-spline and wavelet sieves. Sieve rates appear to suggest that choosing $N = O(T)$ in echo state networks could lead to nontrivial forecasting bias owing to poor approximation properties. Unfortunately, this comparison relies on neglecting the dynamic component of the ESN model, and as such it is only qualitative. It is, therefore, an important topic for future research.

A different but related problem is the potential degradation of forecasting performance when a model is at the interpolation threshold in the overparametrized regime, $N \geq T$. Ridge regression is also commonly applied to address generalization concerns in statistical learning (see Hastie et al. (2009)). Recent work has studied more in-depth the link between regularization and generalization: Hastie et al. (2022a) show that "ridgeless", that is interpolation, solutions can be optimal in some scenarios. However, in our empirical evaluations in Section 1.4, cross-validation consistently selects non-zero ridge penalties, confirming that ridge penalization plays an important role in ESN forecasting performance.

1.2.4 ESN Forecasting

We are primarily interested in using ESN models to construct conditional forecasts of target variables. Given that the conditional mean is the best mean square error estimator for h-step-ahead target $\boldsymbol{y}_{t+h}$, $h \geq 1$, our main focus is approximating

$$\widehat{\boldsymbol{y}}_{t+h|t} := \mathbb{E}\left[\boldsymbol{y}_{t+h}|\boldsymbol{x}_{0:t}, \boldsymbol{z}_{0:t}\right].$$

The case $h = 1$ is trivial, since the ESN model is estimated by regressing $\boldsymbol{y}_{t+1}$ on state $\boldsymbol{x}_t$, and thus we can set $\widetilde{\boldsymbol{y}}_{t+1|t} = \widehat{W}^\top \boldsymbol{x}_t$. However, when $h > 1$ the nonlinear state dynamics precludes a direct computation of the conditional mean. This is in contrast to linear models like VARMAs or DFMs, where the assumption of linearity implies that conditional expectations reduce to simple matrix-vector operations. In particular, linear models are such that the variance (and any other higher-order moments) of the noise term do not impact the conditional mean forecast.

Let $p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{z}_t)$ and $g_\theta(\boldsymbol{y}_{t+1}|\boldsymbol{x}_t)$ be the state transition and observation densities, respec-

tively. Then, for $h > 1$,

$$\widehat{\boldsymbol{y}}_{t+h|t} = \int \boldsymbol{y}_{t+h}\, g_\theta(\boldsymbol{y}_{t+h}|\boldsymbol{x}_{t+h-1}) \prod_{j=1}^{h-1} p_\theta(\boldsymbol{x}_{t+j}|\boldsymbol{x}_{t+j-1}, \boldsymbol{z}_{t+j})\nu(\boldsymbol{z}_{t+j}|\boldsymbol{x}_{t+j-1})\mathrm{d}\boldsymbol{z}_{t+j}\mathrm{d}\boldsymbol{x}_{t+j}\mathrm{d}\boldsymbol{y}_{t+h}, \quad (1.8)$$

where $\nu(\boldsymbol{z}_{t+j}|\boldsymbol{x}_{t+j-1})$ is the conditional density of inputs. Here, we introduce the additional assumption that $\boldsymbol{x}_{t+j-1}$ is sufficient to condition on past states and inputs, that is

$$\nu(\boldsymbol{z}_{t+j}|\boldsymbol{x}_{t+j-1}) \equiv \nu(\boldsymbol{z}_{t+j}|\boldsymbol{x}_{0:t+j-1}, \boldsymbol{z}_{0:t+j-1}). \quad (1.9)$$

Some elements in the expectation integral are not directly available. Specifically, while an ESN explicitly models both $p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{z}_t)$ and $g_\theta(\boldsymbol{y}_{t+1}|\boldsymbol{x}_t)$, the density $\nu(\boldsymbol{z}_{t+j}|\boldsymbol{x}_{t+j-1})$ is unavailable.

In the remaining part of this subsection, we present a novel ESN-based approach to forecasting the target variable. Our idea is to enrich the ESN model with an auxiliary observation equation for the input covariates. As we demonstrate in Section 1.4, our proposed method shows superior performance with respect to the standard state-of-the-art benchmarks.

Multi-step Forecasting of Targets via Iterative Forecasting of Inputs

In general, we are interested in constructing forecasts of target variables that are not the same as the model inputs. To do so, we resolve the issue of the intractability of (1.8) while simultaneously capitalizing on the available results using ESNs in the forecasting of dynamical systems. More explicitly, we add to the ESN specification (1.3)-(1.4) an equation that allows sidestepping modeling the density ν directly, thus making the computation of $\widehat{\boldsymbol{y}}_{t+h|t}$ feasible even when $h > 1$.

Consider the ESN where the reservoir states $(\boldsymbol{x}_t)_{t\in\mathbb{Z}}$ follow (1.3), while the target sequence is the same as the input sequence $(\boldsymbol{z}_t)_{t\in\mathbb{Z}}$,

$$\boldsymbol{x}_t = \alpha\boldsymbol{x}_{t-1} + (1-\alpha)\sigma(A\boldsymbol{x}_{t-1} + C\boldsymbol{z}_t + \boldsymbol{\zeta}) \quad (1.10)$$

$$\boldsymbol{z}_{t+1} = \mathcal{W}^\top \boldsymbol{x}_t + \boldsymbol{u}_{t+1}. \quad (1.11)$$

Here, we use symbol $\mathcal{W}$ for the output coefficients to separate this case from the general ESN equations (1.3)-(1.4). In (1.11), $(\boldsymbol{u}_t)_{t\in\mathbb{Z}}$ are K-dimensional independent zero-mean innovations with variance $\sigma_u^2\mathbb{I}_K$ that are also independent of $\boldsymbol{x}_t$ across all t.

In this case, the reservoir map $F(\boldsymbol{x}_{t-1}, \boldsymbol{z}_t)$ in (1.1) is determined by (1.10), and it is possible to re-feed the forecasted variables back into the state equation as inputs. This yields the following state recursion:

$$\boldsymbol{x}_t = F(\boldsymbol{x}_{t-1}, \mathcal{W}^\top\boldsymbol{x}_{t-1} + \boldsymbol{u}_t) =: G_\theta(\boldsymbol{x}_{t-1}, \boldsymbol{u}_t),$$

where the subscript θ denotes the dependence on the model coefficients. In the reservoir computing literature, regimes, where the ESN state equation is iteratively fed with the model outputs, are called "autonomous" (Gonon et al., 2020a). They are widely and successfully utilized for the prediction of deterministic dynamical systems. Indeed, in those instances, provided that the ridge estimate $\widehat{\mathcal{W}}$ is available from data according to Subsection 1.2.2, the $h > 1$ steps autonomous state iteration is given by

$$F_\theta^*(\boldsymbol{x}_t) := \alpha\boldsymbol{x}_t + (1-\alpha)\sigma((A + C\widehat{\mathcal{W}}^\top)\boldsymbol{x}_t + \boldsymbol{\zeta})$$

and

$$\boldsymbol{x}_{t+h} = \underbrace{F_\theta^* \circ F_\theta^* \circ \cdots \circ F_\theta^*}_{h \text{ times}} (\boldsymbol{x}_t).$$

Hence one can directly obtain the h-steps ahead predictions of the input time series as $\boldsymbol{z}_{t+h} = \widehat{\mathcal{W}}^\top \boldsymbol{x}_{t+h-1}$.

In the case of stochastic target variables, assuming (1.9), we notice that for the conditional forecast of the states, it holds that

$$\widehat{\boldsymbol{x}}_{t+1|t} = \mathbb{E}\left[\boldsymbol{x}_{t+1}|\boldsymbol{x}_{0:t}, \boldsymbol{z}_{0:t}\right] = \int \boldsymbol{x}_t \, p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{z}_t)\nu(\boldsymbol{z}_t|\boldsymbol{x}_{t-1})\mathrm{d}\boldsymbol{z}_t = \int G_\theta(\boldsymbol{x}_{t-1}, \boldsymbol{u}_t)\phi(\boldsymbol{u}_t)\mathrm{d}\boldsymbol{u}_t, \quad (1.12)$$

where density ϕ of $\boldsymbol{u}_t$ is, again, unavailable. Note that, even under the assumption $\boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \Sigma_{\boldsymbol{u}})$, which is standard in the filtering literature, the presence of nonlinear map G_θ makes the computation of the forecasts of $\boldsymbol{z}_{t+h}$ a non-straightforward exercise. Nevertheless, this forecast construction can be readily used when one is interested exclusively in predicting the time series $\boldsymbol{z}_t$.

Whenever the final goal of the exercise is forecasting some other explained variable $\boldsymbol{y}_{t+h}$ h-steps ahead, additional issues arise. In this case, one needs to compute the conditional expectation in (1.8) which is intractable even under Gaussian assumptions on the innovations. One option is to apply particle filtering techniques such as bootstrap sampling or sequential importance sampling (SIS) to evaluate the expectation (Doucet et al., 2001). We emphasize that the state dimension is usually chosen to be large, and hence implementing filtering techniques requires some care.

Our approach is to avoid dealing with the nonlinear densities involved in (1.8) with the help of (1.12) and, instead, to reduce the computation of the conditional expectation $\widehat{\boldsymbol{y}}_{t+h|t}$ to a composition of functions. By the linearity of observation equation (1.4) and the assumption of independence in the zero-mean noise $\boldsymbol{\epsilon}_{t+h}$, we write

$$\widehat{\boldsymbol{y}}_{t+h|t} = W^\top \widehat{\boldsymbol{x}}_{t+h-1|t} = \int W^\top \boldsymbol{x}_{t+h-1} \prod_{j=1}^{h-1} p_\theta(\boldsymbol{x}_{t+j}|\boldsymbol{x}_{t+j-1}, \boldsymbol{z}_{t+j})\nu(\boldsymbol{z}_{t+j}|\boldsymbol{x}_{t+j-1})\mathrm{d}\boldsymbol{x}_{t+j}\mathrm{d}\boldsymbol{z}_{t+j}$$

and use the approximation

$$\widehat{\boldsymbol{y}}_{t+h|t} \approx \widetilde{\boldsymbol{y}}_{t+h} = W^\top \underbrace{F_\theta^* \circ F_\theta^* \circ \cdots \circ F_\theta^*}_{h-1 \text{ times}} (\boldsymbol{x}_t), \quad (1.13)$$

which originates from

$$\widehat{\boldsymbol{x}}_{t|t-1} = \int G_\theta(\boldsymbol{x}_{t-1}, \boldsymbol{u}_t)\phi(\boldsymbol{u}_t)\mathrm{d}\boldsymbol{u}_t \approx G_\theta(\boldsymbol{x}_{t-1}, \mathbb{E}[\boldsymbol{u}_t]) = F(\boldsymbol{x}_{t-1}, \mathcal{W}^\top \boldsymbol{x}_{t-1}) \equiv F_\theta^*(\boldsymbol{x}_{t-1}), \quad (1.14)$$

where $\boldsymbol{u}_t$ is assumed to be zero-mean. The validity of (1.14) itself requires implicit assumptions on the nature of the distribution of $\boldsymbol{u}_t$, but here we want to keep the analysis of $\widehat{\boldsymbol{y}}_{t+h|t}$ to a minimum, and just use the insights from the dynamical systems ESN literature. We are hence not delving deeper into alternative approaches to estimate forecasts or, more generally, to compute conditional expectations of ESN models with stochastic inputs.

1.3 Multi-Frequency Echo State Models

In this subsection, we construct a broad class of ESN models that can accommodate input and target time series sampled at distinct sampling frequencies. We call this family of reservoir models the *Multi-Frequency Echo State Networks* (MFESNs). The state-space structure of MFESNs is naturally amenable to the setting of time series with mixed frequencies. Additionally, the prediction strategy discussed in Section 1.2.4 is straightforward to extend to MFESNs.

We present two groups of MFESN architectures. The first family is based on a single echo state network architecture and we call these models *Single-Reservoir Multi-Frequency Echo State Networks* (S-MFESNs). The second group, referred to as *Multi-Reservoir Multi-Frequency Echo State Networks* (M-MFESNs), allows for as many state equations as the number of distinct sampling frequencies present in the input data.

1.3.1 Single-Reservoir MFESN

Recall that, in the temporal notation of Definition 1.1.1, we reserve t to be the reference time index, which is also used for the target variable, and all other frequencies will be measured with respect to the reference frequency.

Consider L collections of different time series. We assume that the lth collection, $l \in [L]$, consists of n_l time series that are sampled at a common frequency κ_l and contain observations $(\boldsymbol{z}^{(l)}_{t,s|\kappa_l})_{t,s}$ with $\boldsymbol{z}^{(l)}_{t,s|\kappa_l} \in \mathbb{R}^{n_l}$ for all $t \in \mathbb{Z}$ and $s \in \{0, \dots, \kappa_l - 1\}$. Let $\kappa_{\max} = \max_l \kappa_l$ be the highest sampling frequency among the L time series groups and let $q_l := \kappa_{\max}/\kappa_l$ indicate how low each κ_l sampling frequency is with respect to $\kappa_{\max}$. We can now stack together and repeat the observations in a way that is consistent with the high-frequency index by defining

$$\boldsymbol{z}_{t,s|\kappa_{\max}} := \left(\boldsymbol{z}^{(1)\top}_{t,\lfloor s/q_1 \rfloor|\kappa_1},\ \boldsymbol{z}^{(2)\top}_{t,\lfloor s/q_2 \rfloor|\kappa_2},\ \dots,\ \boldsymbol{z}^{(L)\top}_{t,\lfloor s/q_L \rfloor|\kappa_L}\right)^\top \in \mathbb{R}^{\sum_{l=1}^L n_l},\ \ s \in \{0, \dots, \kappa_{\max} - 1\},$$

where for all $l \in [L]$, $\boldsymbol{z}^{(l)}_{0,0|\kappa_l} = \mathbf{0}_{n_l}$. Thus, it is possible to write a single high-frequency ESN as

$$\boldsymbol{x}_{t,s|\kappa_{\max}} = \alpha \boldsymbol{x}_{t,s-1|\kappa_{\max}} + (1-\alpha)\sigma(A\boldsymbol{x}_{t,s-1|\kappa_{\max}} + C\boldsymbol{z}_{t,s|\kappa_{\max}} + \boldsymbol{\zeta}), \tag{1.15}$$

$$\boldsymbol{z}_{t,s+1|\kappa_{\max}} = \mathcal{W}^\top \boldsymbol{x}_{t,s|\kappa_{\max}} + \boldsymbol{u}_{t,s+1|\kappa_{\max}}, \tag{1.16}$$

where $\mathcal{W} \in \mathbb{M}_{N,\sum_{l=1}^L n_l}$ and $s > 0$. We term this class of MFESN models the *Single-Reservoir Multi-Frequency ESNs* (S-MFESNs).

Notice that equations (1.15)-(1.16) of the S-MFESN model prescribe the dynamics at the highest frequency, $\kappa_{\max}$. In order to forecast a lower frequency target, we map high-frequency states $\boldsymbol{x}_{t,s|\kappa_{\max}}$ to low-frequency targets $\boldsymbol{y}_{t+1} \in \mathbb{R}^J$ by introducing a *state alignment* scheme. An *aligned* S-MFESN uses the most recent state with respect to the reference time index t to construct the forecast. More precisely, the state equation of an S-MFESN is iterated $\kappa_{\max}$ times until the state $\boldsymbol{x}_{t-1,\kappa_{\max}|\kappa_{\max}} = \boldsymbol{x}_{t,0|\kappa_{\max}}$ is obtained and then target $\boldsymbol{y}_{t+1}$ is forecast with observation equation

$$\boldsymbol{y}_{t+1} = W^\top \boldsymbol{x}_{t,0|\kappa_{\max}} + \boldsymbol{\epsilon}_{t+1}, \quad W \in \mathbb{M}_{N,J}. \tag{1.17}$$

ESTIMATION OF ALIGNED S-MFESN. Both coefficient matrices W and $\mathcal{W}$ can be estimated as explained in Subsection 1.2.2 under appropriate choices of corresponding penalty strengths. In particular, in order to obtain $\widehat{\mathcal{W}}$, the state and the observation matrices in (1.5) are given by

$$X_{\kappa_{\max}} = (\boldsymbol{x}_{1,0|\kappa_{\max}}, \ldots, \boldsymbol{x}_{1,\kappa_{\max}-1|\kappa_{\max}}, \ldots, \boldsymbol{x}_{T-1,0|\kappa_{\max}}, \ldots, \boldsymbol{x}_{T-1,\kappa_{\max}-1|\kappa_{\max}})^\top \in \mathbb{M}_{(T-1)\kappa_{\max}-1,N},$$
$$Y_{\kappa_{\max}} = (\boldsymbol{z}_{1,1|\kappa_{\max}}, \ldots, \boldsymbol{z}_{1,\kappa_{\max}|\kappa_{\max}}, \ldots, \boldsymbol{z}_{T-1,1|\kappa_{\max}}, \ldots, \boldsymbol{z}_{T-1,\kappa_{\max}|\kappa_{\max}})^\top \in \mathbb{M}_{(T-1)\kappa_{\max}-1,\sum_{l=1}^{L} n_l},$$

while

$$X = \left(\boldsymbol{x}_{1,0|\kappa_{\max}}, \boldsymbol{x}_{2,0|\kappa_{\max}}, \ldots, \boldsymbol{x}_{T-1,0|\kappa_{\max}}\right)^\top \in \mathbb{M}_{T-1,N},$$
$$Y = (\boldsymbol{y}_2, \ldots, \boldsymbol{y}_T)^\top \in \mathbb{M}_{T-1,J},$$

are used for the estimation of $\widehat{W}$. We note that the state equation (1.15) of S-MFESN can be initialized by $\boldsymbol{x}_{0,0|\kappa_{\max}}$, which under the fading memory property is inconsequential for long enough samples (see the discussion in Subsection 1.2).

FORECASTING WITH ALIGNED S-MFESN. Let $\widehat{W}$ and $\widehat{\mathcal{W}}$ be the sample estimates of the readout matrices as explained above. The fitted high-frequency autonomous state transition map associated with (1.15) is given by

$$F_{\kappa_{\max}}(\boldsymbol{x}_{t,s-1|\kappa_{\max}}) := \alpha \boldsymbol{x}_{t,s-1|\kappa_{\max}} + (1-\alpha)\sigma\left((A + C\widehat{\mathcal{W}}^\top)\boldsymbol{x}_{t,s-1|\kappa_{\max}} + \boldsymbol{\zeta}\right), \tag{1.18}$$

which, composed with itself exactly $\kappa_{\max}$ times, yields the target-frequency-aligned autonomous state transition map

$$F(\boldsymbol{x}_{t,0|\kappa_{\max}}) := \underbrace{F_{\kappa_{\max}} \circ F_{\kappa_{\max}} \circ \cdots \circ F_{\kappa_{\max}}}_{\kappa_{\max} \text{ times}}(\boldsymbol{x}_{t,0|\kappa_{\max}}). \tag{1.19}$$

Finally, from (1.13) the h-steps ahead low-frequency forecasts, $h \in \mathbb{N}$, can be computed as

$$\widetilde{y}_{T+h|T} = \widehat{W}^\top\big(\underbrace{F \circ F \circ \cdots \circ F}_{h-1 \text{ times}}(\boldsymbol{x}_{T,0|\kappa_{\max}})\big). \tag{1.20}$$

Figure 1.1 gives a graphical diagram of the 1-step forecasting procedure for an S-MFESN. Additionally, Figure 1.12 in Appendix 1.J provides a similar diagram for the case of multistep forecasts.

The following example illustrates this proposed forecasting strategy for the case of quarterly GDP forecasting using monthly and daily series inputs.

Example 1.3.1. Suppose that we wish to use an aligned S-MFESN model to forecast a quarterly one-dimensional target (y_t) using $n_{(\mathtt{m})}$ monthly and $n_{(\mathtt{d})}$ daily series, $(\boldsymbol{z}^{(\mathtt{m})}_{t,s|\kappa_1})$ and $(\boldsymbol{z}^{(\mathtt{d})}_{t,s|\kappa_2})$, respectively. We adopt the assumption that daily data is released 24 days over each calendar month and hence $\kappa_1 = 3$, $\kappa_2 = 72$ and $\kappa_{\max} = 72$, while $q_1 = 24$ and $q_2 = 1$. Let $t, *|72$ be the temporal index with a quarterly reference frequency. The input vector for the S-MFESN state equation consistent

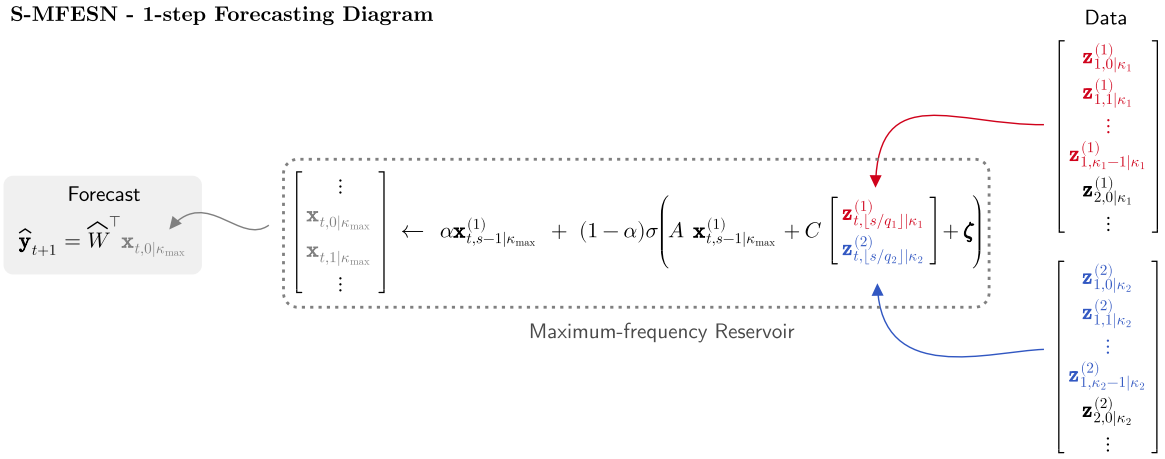


Figure 1.1: Scheme of a Single-Reservoir MFESN (S-MFESN) model combining input data sampled at two frequencies with state alignment and estimation for one-step ahead forecasting of the target series.

with the daily frequency is given by

$$\boldsymbol{z}^{(\mathtt{m},\mathtt{d})}_{t,s|72} := (\boldsymbol{z}^{(\mathtt{m})}_{t,\lfloor s/24 \rfloor|3}{}^{\top}, \boldsymbol{z}^{(\mathtt{d})}_{t,s|72}{}^{\top})^{\top} \in \mathbb{R}^{n_{(\mathtt{m})}+n_{(\mathtt{d})}} \text{ with } \boldsymbol{z}^{(\mathtt{d})}_{0,0|3} = \mathbf{0}_{n_{(\mathtt{d})}} \text{ and } \boldsymbol{z}^{(\mathtt{m})}_{0,0|24} = \mathbf{0}_{n_{(\mathtt{m})}}.$$

The complete S-MFESN model with the state space dimension N can be written as:

$$\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s|72} = \alpha \boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s-1|72} + (1-\alpha)\sigma(A\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s-1|72} + C\boldsymbol{z}^{(\mathtt{m},\mathtt{d})}_{t,s|72} + \boldsymbol{\zeta}), \tag{1.21}$$

$$\boldsymbol{z}^{(\mathtt{m},\mathtt{d})}_{t,s+1|72} = \mathcal{W}^{\top}\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s|72} + \boldsymbol{u}_{t,s+1|72}, \tag{1.22}$$

$$y_{t+1} = W^{\top}\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,0|\kappa_{\max}} + \boldsymbol{\epsilon}_{t+1}, \tag{1.23}$$

where the state equations (1.21)-(1.22) are run in their own maximum frequency temporal index $s > 0$, and only the states $\boldsymbol{x}_{t-1,\kappa_{\max}|\kappa_{\max}} = \boldsymbol{x}_{t,0|\kappa_{\max}}$ are used in the observation equation (1.23). Provided the input-target pairs sample of length T, the coefficient matrices $\mathcal{W} \in \mathbb{M}_{N,n_{(\mathtt{m})}+n_{(\mathtt{d})}}$ in (1.22) and $W \in \mathbb{R}^N$ in (1.23) can be estimated via ridge regression as explained above.

From (1.18) the high-frequency autonomous state transition map is given by

$$F^{(\mathtt{m},\mathtt{d})}_{72}(\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s-1|72}) := \alpha \boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s-1|72} + (1-\alpha)\sigma\left((A + C\widehat{\mathcal{W}}^{\top})\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s-1|72} + \boldsymbol{\zeta}\right),$$

which, composed with itself exactly 72 times, by (1.19) yields the target-frequency-aligned autonomous state transition map

$$F^{(\mathtt{m},\mathtt{d})}(\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,0|72}) := \underbrace{F^{(\mathtt{m},\mathtt{d})}_{72} \circ F^{(\mathtt{m},\mathtt{d})}_{72} \cdots \circ F^{(\mathtt{m},\mathtt{d})}_{72}}_{72 \text{ times}}(\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,0|72}).$$

By applying $F^{(\mathtt{m},\mathtt{d})}$ to state $\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,0|72}$ we iterate the S-MFESN forward in time to provide an estimate for $\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t+1,0|72}$, which can then be linearly projected using $\widehat{W}$ to yield a forecast for y_{t+2}. For the target variable, as well as forecasts, we do not use our temporal notation for the sake of compactness and clarity of exposition. Finally, the quarterly forecasts for $h \in \mathbb{N}$ can be computed using (1.20)

as

$$\widetilde{y}_{T+h|T} = \widehat{W}^\top (\underbrace{F^{(\mathtt{m},\mathtt{d})} \circ F^{(\mathtt{m},\mathtt{d})} \circ \cdots \circ F^{(\mathtt{m},\mathtt{d})}}_{h-1 \text{ times}} (\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{T,0|72})).$$

1.3.2 Multi-Reservoir MFESN

Constructing an MFESN with a single reservoir is not necessarily the most effective modeling strategy. Having more than one reservoir allows more flexible modeling of state dynamics for different subsets of input variables sampled at common frequencies. For example, suppose quarterly and monthly data are used as regressors. Our presentation is general enough to accommodate other types of partitioning of series into the corresponding reservoir models. We leave it to future research to test other approaches based, for instance, on markets or data types as done in van Huellen et al. (2020).

Assume again L groups of series with input observations $(\boldsymbol{z}^{(l)}_{t,s|\kappa_l})_{t,s}$ with $\boldsymbol{z}^{(l)}_{t,s|\kappa_l} \in \mathbb{R}^{n_l}$, $l \in [L]$, for all $t \in \mathbb{Z}$ and $s \in \{0, \ldots, \kappa_l - 1\}$ sampled at common frequencies $\{\kappa_1, \ldots, \kappa_L\}$, respectively. For each of the L groups of input series we define the corresponding ESN model as

$$\boldsymbol{x}^{(l)}_{t,s|\kappa_l} = \alpha_l \boldsymbol{x}^{(l)}_{t,s-1|\kappa_l} + (1-\alpha_l)\sigma(A_l \boldsymbol{x}^{(l)}_{t,s-1|\kappa_l} + C_l \boldsymbol{z}^{(l)}_{t,s|\kappa_l} + \boldsymbol{\zeta}_l), \tag{1.24}$$

$$\boldsymbol{z}^{(l)}_{t,s+1|\kappa_l} = \mathcal{W}_l^\top \boldsymbol{x}^{(l)}_{t,s|\kappa_l} + \boldsymbol{u}^{(l)}_{t,s+1|\kappa_l}, \quad l \in [L], \tag{1.25}$$

with $s > 0$, $\mathcal{W}_l \in \mathbb{M}_{N_l,n_l}$ with N_l the dimension of the state space. Notice that the time index s is different for each l according to our temporal notation introduced in Definition 1.1.1 and each state equation runs at its own frequency κ_l. The dimensions $\{N_1, N_2, \ldots, N_L\}$ of the state spaces can be chosen for the L reservoir models individually. Additionally, multiple reservoirs have the associated hyperparameter tuples $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_L\}$ to be tuned. This requires some care whenever one wants to optimize all hyperparameters jointly. Since there are L reservoir state equations, we call this class of MFESN models *Multi-Reservoir Multi-Frequency ESN* (M-MFESN).

Similar to S-MFESN, all L state equations are iterated each κ_l times respectively until the states $\boldsymbol{x}^{(l)}_{t-1,\kappa_l|\kappa_l} = \boldsymbol{x}^{(l)}_{t,0|\kappa_l}$ are obtained. The *aligned* M-MFESN observation equation is given by

$$\boldsymbol{y}_{t+1} = W^\top \boldsymbol{x}_{t,L} + \boldsymbol{\epsilon}_{t+1}, \text{ with } \boldsymbol{x}_{t,L} = \begin{pmatrix} \boldsymbol{x}^{(1)}_{t,0|\kappa_1} \\ \vdots \\ \boldsymbol{x}^{(L)}_{t,0|\kappa_L} \end{pmatrix} \in \mathbb{R}^{\sum_{l=1}^L N_l}, \; W \in \mathbb{M}_{\sum_{l=1}^L N_l, J}. \tag{1.26}$$

Estimation of aligned M-MFESN. The coefficient matrices W_l, $l \in [L]$, and $\mathcal{W}$ can be estimated similarly to the case of S-MFESN. The state and observation matrices for the estimation of $\widehat{\mathcal{W}}_l$, $l \in [L]$, in (1.5) are constructed as

$$X^{(l)} = (\boldsymbol{x}^{(l)}_{1,0|\kappa_l}, \ldots, \boldsymbol{x}^{(l)}_{1,\kappa_l-1|\kappa_l}, \ldots, \boldsymbol{x}^{(l)}_{T-1,0|\kappa_l}, \ldots, \boldsymbol{x}^{(l)}_{T-1,\kappa_l-1|\kappa_l})^\top \in \mathbb{M}_{(T-1)\kappa_l-1,N_l},$$
$$Y^{(l)} = (\boldsymbol{z}^{(l)}_{1,1|\kappa_l}, \ldots, \boldsymbol{z}^{(l)}_{1,\kappa_l|\kappa_l}, \ldots, \boldsymbol{z}^{(l)}_{T-1,1|\kappa_l}, \ldots, \boldsymbol{z}^{(l)}_{T-1,\kappa_l|\kappa_l})^\top \in \mathbb{M}_{(T-1)\kappa_l-1,n_l},$$

while with the notation as in (1.26)

$$X = (\boldsymbol{x}_{1,L}, \boldsymbol{x}_{2,L}, \dots, \boldsymbol{x}_{T-1,L})^\top \in \mathbb{M}_{T-1,\sum_{l=1}^{L} N_l},$$
$$Y = (\boldsymbol{y}_2, \dots, \boldsymbol{y}_T)^\top \in \mathbb{M}_{T-1,J},$$

are used for the estimation of $\widehat{W}$. Again, the state equations (1.24) of M-MFESN can be started with $\boldsymbol{x}^{(l)}_{0,0|\kappa_l} = \mathbf{0}_{N_l}$ (see Subsection 1.2 for more details).

Forecasting with aligned M-MFESN. Let $\widehat{W}$ and $\widehat{\mathcal{W}}_l$, $l \in [L]$, be the sample estimates of the readout matrices. For any $l \in [L]$ the κ_l-frequency autonomous state transition map is given by

$$F^{(l)}_{\kappa_l}(\boldsymbol{x}^{(l)}_{t,s-1|\kappa_l}) := \alpha_l \boldsymbol{x}^{(l)}_{t,s-1|\kappa_l} + (1-\alpha_l)\sigma\left((A_l + C_l\widehat{\mathcal{W}}_l^\top)\boldsymbol{x}^{(l)}_{t,s-1|\kappa_l} + \boldsymbol{\zeta}_l\right). \tag{1.27}$$

The target-frequency-aligned autonomous state transition map associated with each frequency l is hence defined as

$$F^{(l)}(\boldsymbol{x}_{t,0|\kappa_l}) := \underbrace{F^{(l)}_{\kappa_l} \circ F^{(l)}_{\kappa_l} \circ \cdots \circ F^{(l)}_{\kappa_l}}_{\kappa_l \text{ times}}(\boldsymbol{x}^{(l)}_{t,0|\kappa_l}). \tag{1.28}$$

Finally, from (1.13) the h-steps ahead forecasts can be computed as

$$\widetilde{y}_{T+h|T} = \widehat{W}^\top \begin{pmatrix} \underbrace{F^{(1)} \circ F^{(1)} \circ \cdots \circ F^{(1)}}_{h-1 \text{ times}}(\boldsymbol{x}^{(1)}_{T,0|\kappa_1}) \\ \vdots \\ \underbrace{F^{(L)} \circ F^{(L)} \circ \cdots \circ F^{(L)}}_{h-1 \text{ times}}(\boldsymbol{x}^{(L)}_{T,0|\kappa_L}) \end{pmatrix}. \tag{1.29}$$

In Figure 1.2 we provide a diagram for the case of 1-step ahead forecasting with an aligned M-MFESN involving regressors of only two frequencies. Figure 1.13 in Appendix 1.J provides a similar diagram for the case of multistep forecasting.

Example 1.3.2. Similar to Example 1.3.1, we aim to forecast a quarterly target with monthly and daily series, but this time we use an M-MFESN model. We have to define two independent state equations, one for monthly and one for daily series; in the observation equations, two states must be aligned temporally and stacked to form the full set of regressors. The data consists again of quarterly (y_t), $n_{(\mathtt{m})}$ monthly series ($\boldsymbol{z}^{(\mathtt{m})}_{t,s|3}$) and $n_{(\mathtt{d})}$ daily series ($\boldsymbol{z}^{(\mathtt{d})}_{t,s|72}$).

The aligned M-MFESN model with two reservoirs of dimensions $N_{(\mathtt{m})}$ and $N_{(\mathtt{d})}$, respectively, is given by

$$\boldsymbol{x}^{(\mathtt{m})}_{t,s|3} = \alpha_1 \boldsymbol{x}^{(\mathtt{m})}_{t,s-1|3} + (1-\alpha_1)\sigma(A_1 \boldsymbol{x}^{(\mathtt{m})}_{t,s-1|3} + C_1 \boldsymbol{z}^{(\mathtt{m})}_{t,s|3} + \boldsymbol{\zeta}_1), \tag{1.30}$$
$$\boldsymbol{z}^{(\mathtt{m})}_{t,s+1|3} = \mathcal{W}^\top_{(\mathtt{m})} \boldsymbol{x}^{(\mathtt{m})}_{t,s|3} + \boldsymbol{u}^{(\mathtt{m})}_{t,s+1|3}, \tag{1.31}$$
$$\boldsymbol{x}^{(\mathtt{d})}_{t,s|72} = \alpha_2 \boldsymbol{x}^{(\mathtt{d})}_{t,s-1|72} + (1-\alpha_2)\sigma(A_2 \boldsymbol{x}^{(\mathtt{d})}_{t,s-1|72} + C_2 \boldsymbol{z}^{(\mathtt{d})}_{t,s|72} + \boldsymbol{\zeta}_2), \tag{1.32}$$
$$\boldsymbol{z}^{(\mathtt{d})}_{t,s+1|72} = \mathcal{W}^\top_{(\mathtt{d})} \boldsymbol{x}^{(\mathtt{d})}_{t,s|72} + \boldsymbol{u}^{(\mathtt{d})}_{t,s+1|72}, \tag{1.33}$$
$$y_{t+1} = W^\top \begin{pmatrix} \boldsymbol{x}^{(\mathtt{m})}_{t,0|3} \\ \boldsymbol{x}^{(\mathtt{d})}_{t,0|72} \end{pmatrix} + \boldsymbol{\epsilon}_{t+1}, \tag{1.34}$$

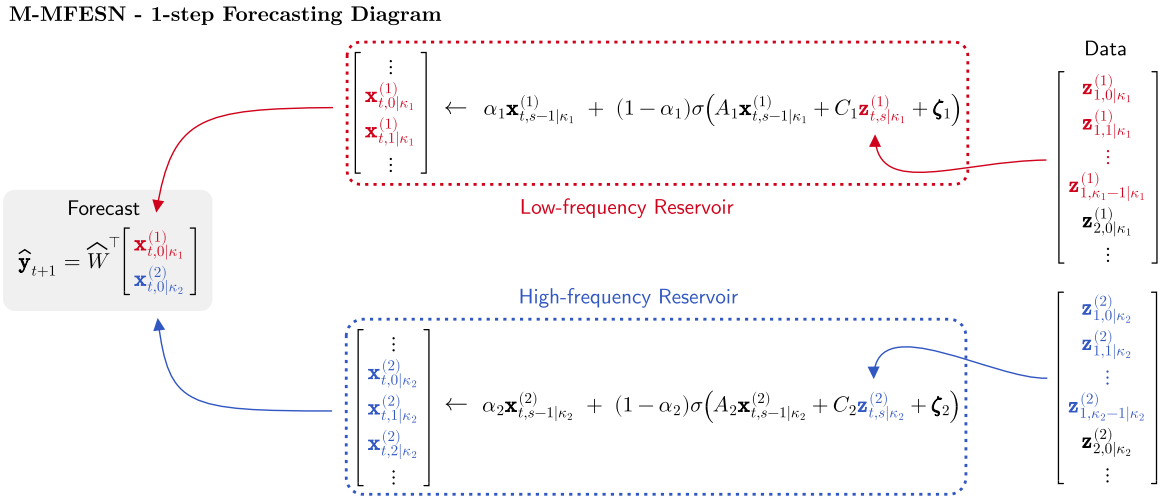


Figure 1.2: Scheme of a Multi-Reservoir MFESN (M-MFESN) model combining input data sampled at two frequencies with state alignment and estimation for one-step ahead forecasting of the target series.

where $s > 0$, $\mathcal{W}_{(\mathtt{m})} \in \mathbb{M}_{N_{(\mathtt{m})},n_{(\mathtt{m})}}$, $\mathcal{W}_{(\mathtt{d})} \in \mathbb{M}_{N_{(\mathtt{d})},n_{(\mathtt{d})}}$ and $W \in \mathbb{R}^{N_{(\mathtt{m})}+N_{(\mathtt{d})}}$. Here, the monthly reservoir $(\boldsymbol{x}^{(\mathtt{m})}_{t,s|3})$ has the temporal index of frequency 3, while the daily reservoir $(\boldsymbol{x}^{(\mathtt{d})}_{t,s|72})$ of 72; the high-frequency index s is different for the two models. Notice that in an M-MFESN model it is necessary to introduce 2 additional observation equations for the states, that is (1.31) and (1.33). Notice that the state equations are iterated each κ_l times to collect the states to be aligned in the observation equation (1.34). Again, the sample-based estimates of coefficient matrices $\widehat{\mathcal{W}}_{(\mathtt{m})}$, $\widehat{\mathcal{W}}_{(\mathtt{d})}$ and $\widehat{W}$ in (1.31), (1.32), and in (1.34), respectively, can be obtained via the ridge regression as discussed above.

Exactly as in Example 1.3.1, using (1.27) we can introduce high-frequency autonomous state maps $F^{(\mathtt{m})}_3$ and $F^{(\mathtt{d})}_{72}$ as

$$F^{(\mathtt{m})}_3(\boldsymbol{x}^{(\mathtt{m})}_{t,s-1|3}) := \alpha_1 \boldsymbol{x}^{(\mathtt{m})}_{t,s-1|3} + (1-\alpha_1)\sigma\left((A_1 + C_1\widehat{\mathcal{W}}^{\top}_{(\mathtt{m})})\boldsymbol{x}^{(\mathtt{m})}_{t,s-1|3} + \boldsymbol{\zeta}_1\right),$$
$$F^{(\mathtt{d})}_{72}(\boldsymbol{x}^{(\mathtt{d})}_{t,s-1|72}) := \alpha_2 \boldsymbol{x}^{(\mathtt{d})}_{t,s-1|72} + (1-\alpha_2)\sigma\left((A_2 + C_2\widehat{\mathcal{W}}^{\top}_{(\mathtt{d})})\boldsymbol{x}^{(\mathtt{d})}_{t,s-1|72} + \boldsymbol{\zeta}_2\right),$$

as well as their target-frequency aligned counterparts $F^{(\mathtt{m})}$ and $F^{(\mathtt{d})}$, by (1.28), as

$$F^{(\mathtt{m})}(\boldsymbol{x}^{(\mathtt{m})}_{t,0|3}) := \underbrace{F^{(\mathtt{m})}_3 \circ F^{(\mathtt{m})}_3 \circ F^{(\mathtt{m})}_3}_{3 \text{ times}}(\boldsymbol{x}^{(\mathtt{m})}_{t,0|3}),$$
$$F^{(\mathtt{d})}(\boldsymbol{x}^{(\mathtt{d})}_{t,0|72}) := \underbrace{F^{(\mathtt{d})}_{72} \circ F^{(\mathtt{d})}_{72} \circ \cdots \circ F^{(\mathtt{d})}_{72}}_{72 \text{ times}}(\boldsymbol{x}^{(\mathtt{d})}_{t,0|72}).$$

The h-step ahead forecasts can be computed using the approximation in (1.29) as

$$\widetilde{y}_{T+h|T} = \widehat{W}^{\top}\begin{pmatrix} \underbrace{F^{(\mathtt{m})} \circ F^{(\mathtt{m})} \circ \cdots \circ F^{(\mathtt{m})}}_{h-1 \text{ times}}(\boldsymbol{x}^{(\mathtt{m})}_{T,0|3}) \\ \underbrace{F^{(\mathtt{d})} \circ F^{(\mathtt{d})} \circ \cdots \circ F^{(\mathtt{d})}}_{h-1 \text{ times}}(\boldsymbol{x}^{(\mathtt{d})}_{T,0|72}) \end{pmatrix}.$$

In this case, it is important to note that while both $F^{(\mathtt{m})}$ and $F^{(\mathtt{d})}$ are composed $h-1$ times at step h, the underlying number of autonomous reservoir iterations is different for the monthly and daily reservoirs, namely 3 and 72, and depends on their own frequencies. This also suggests that one should take into account the different time dynamics when, for example, tuning M-MFESN hyperparameters $\boldsymbol{\varphi}^{(\mathtt{m})}$ and $\boldsymbol{\varphi}^{(\mathtt{d})}$, as proposed in Appendix 1.C.2.

1.4 Empirical Study

In this section, we compare the forecasting performance of our proposed MFESN to state-of-the-art benchmarks. We use a combination of macroeconomic and financial data sampled at low and high-frequency intervals, respectively. Our empirical exercises encompass several setups, with a small and a medium-sized set of regressors, fitting models with data before and after the 2007-08 crisis, and with fixed, rolling, and expanding estimation windows.

1.4.1 Data

Two sets of predictors of different sizes are compiled: Small-MD with 9 predictors and Medium-MD with 33 predictors in monthly and daily frequency. The reference frequency is quarterly: this is the frequency at which the target variable, US GDP growth, is available. Seasonally adjusted quarterly and monthly data is obtained from the Federal Reserve Bank of St. Louis Monthly (FRED-MD) and Quarterly (FRED-QD) Databases for Macroeconomic Research (see McCracken and Ng 2016, 2020 for detail). Daily data is obtained from Refinitiv Datastream, a subscription-based data service. All data is the last revised vintage data. The macroeconomic target and predictors, their transformations, and availability are provided in full detail in Table 1.9 in Appendix 1.A.

The selection of predictors follows the seminal work by Stock and Watson (1996, 2006) in which the FRED-MD and FRED-QD data are proposed. Variations of their dataset have been used profusely in the literature (for example, see Boivin and Ng 2005, Marcellino et al. 2006, Hatzius et al. 2010). Indicators from ten macroeconomic and financial categories are considered: (1) output and income, (2) labor market, (3) housing, (4) orders and inventories, (5) price indices, (6) money and credit, (7) interest rates, (8) exchange rates, (9) equity, and (10) derivatives. The latter five categories represent financial market conditions and are sourced at daily frequency. The exception is interest rates, which move relatively slowly and enter as monthly aggregates, available in the FRED-MD data. We refer to this dataset as Medium-MD. A subset of predictors is selected for the Small-MD dataset by choosing variables that have been identified as leading indicators in the empirical literature (Ingenito and Trehan, 1996, Clements and Galvão, 2008, Andreou et al., 2013, Marsilli, 2014, Ferrara et al., 2014, Carriero et al., 2019, Jardet and Meunier, 2022). Data availability is an additional criterion, and predictors unavailable before 1990 are not considered. This excludes the VIX volatility index, which has been identified as a leading indicator in some studies, for example in Andreou et al. (2013), Jardet and Meunier (2022).

We follow instructions by McCracken and Ng (2016, 2020) on pre-processing macroeconomic predictors before they are used as input for forecasting. These are mainly differenced for detrending. We further transform financial predictors to capture market disequilibrium and volatility.

Disequilibrium indicators, such as interest rate spreads, have been found to be more relevant for macroeconomic prediction than routine changes captured by differencing (see Borio and Lowe 2002, Gramlich et al. 2010, Qin et al. 2022). In addition to disequilibrium indicators, realized stock market volatility has been found to improve macroeconomic predictions (Chauvet et al., 2015). In the absence of intraday trading data from the 1990s onward, which prevents us from utilizing conventional daily realized volatility indicators, we extract volatility indicators from daily price series by fitting a GARCH(1,1) by Bollerslev (1986).[1] In addition to volatility of stock and commodity prices, term structure indicators are used. The term structure is forward-looking, capturing information about future demand and supply, and has been found to be a leading predictor of GDP growth (see for example Hong and Yogo 2012, Kang and Kwon 2020).

The data spans the period January 1st, 1990 to December 31st, 2019.[2] We are interested in evaluating model performance under two stylized settings. First, a researcher fits all models up until the Great Recession, including data from Q1 1990 to Q4 2007. Second, fitting is done with data largely encompassing the crisis period, again from Q1 1990 but now up to Q4 2011. In both cases, the testing sample ranges from the next GDP growth observation after fitting up to Q4 2019. All exercises exclude the global COVID-19 economic depression, as we consider it as an extreme, unpredictable event that induces significant structural changes in the underlying macroeconomic dynamics.[3]

To avoid having to handle the many edge cases that daily data in its "raw" calendar releases involves, we use an interpolation approach. We set *ex ante* the number of working days in *any* month to be exactly 24: given that in forecasting the most recent information sets are more relevant, when interpolating daily data over months with less than calendar 24 observations, we linearly interpolate the "missing" data starting from a months' beginning (using the previous months' last observation). The choice of 24 as a daily frequency is transparent by noting that this is the closest number to actual commonly observed data releases, whilst also being a multiple of both 4 (approximate number of weeks per month) and 6 (upper bound on the number of working days per week).

1.4.2 Models

In this section, we present the set of models that we use throughout our empirical exercises. For a general overview, Table 1.1 summarizes all models, including hyperparameters. In our analysis, we compare the competing models based on several performance measures, which we introduce in Appendix 1.D.

[1] We include a control `scale = 1` to ensure convergence of the optimization algorithm and only include a constant mean term in the return process for simplicity.

[2] In the Small-MD dataset experiments we make a small variation and instead include data starting from 1st January 1975, but *only* for the initial CV selection of ridge penalties for MFESN models. Our aim is to make sure that at least for the fixed window estimation strategy – where λ is cross-validated once and only one $\widehat{W}$ is estimated – the ridge estimator is robust. In practice, when we compare to expanding and rolling window estimators, where λ is re-selected at each window, we find that extending the initial CV data window has little impact on out-of-sample performance.

[3] In the macroeconomic literature this falls under the category of "natural disaster" events, and should not be naïvely modeled together with previous observations. In this section, we therefore avoid dealing with post-COVID-19 macroeconomic data altogether.

Model Name	Description	Specification
Mean	Unconditional mean of target series over estimation sample.	None
AR(1)	Autoregressive model of target series estimated using OLS.	None
MIDAS	Almon-weighted MIDAS regression, linear (unconstrained) autoregressive component.	Autoregressive lags: 3 Monthly freq. lags: 9 Daily freq. lags: 30
DFM [A]	Stock aggregation, VAR(1) factor process.	Factors: 5 for Small-MD 10 for Medium-MD
DFM [B]	Almon aggregation, VAR(1) factor process	Factors: 5 for Small-MD 10 for Medium-MD
singleESN [A]	S-MFESN model: Sparse-normal $\widetilde{A}$, sparse-uniform $\widetilde{C}$, $\widetilde{\boldsymbol{\zeta}} = \mathbf{0}$. Isotropic ridge regression fit.	Reservoir dim: 30 Sparsity: 33.3% $\rho = 0.5$, $\gamma = 1$, $\alpha = 0.1$
singleESN [B]	S-MFESN model: Sparse-normal $\widetilde{A}$, sparse-uniform $\widetilde{C}$, $\widetilde{\boldsymbol{\zeta}} = \mathbf{0}$. Isotropic ridge regression fit.	Reservoir dim: 120 Sparsity: 8.3% $\rho = 0.5$, $\gamma = 1$, $\alpha = 0.1$
multiESN [A]	M-MFESN model: Monthly and daily frequency reservoirs. Sparse-normal $\widetilde{A}_1$, $\widetilde{A}_2$, sparse-uniform $\widetilde{C}_1$, $\widetilde{C}_2$, $\widetilde{\boldsymbol{\zeta}}_1 = \mathbf{0}$, $\widetilde{\boldsymbol{\zeta}}_2 = \mathbf{0}$. Isotropic ridge regression fit.	Reservoir dims: M=100, D=20 Sparsity: M=10%, D=50% M: $\rho = 0.5$, $\gamma = 1.5$, $\alpha = 0$ D: $\rho = 0.5$, $\gamma = 0.5$, $\alpha = 0.1$
multiESN [B]	M-MFESN model: Monthly and daily frequency reservoirs. Sparse-normal $\widetilde{A}_1$, $\widetilde{A}_2$, sparse-uniform $\widetilde{C}_1$, $\widetilde{C}_2$, $\widetilde{\boldsymbol{\zeta}}_1 = \mathbf{0}$, $\widetilde{\boldsymbol{\zeta}}_2 = \mathbf{0}$. Isotropic ridge regression fit.	Reservoir dims: M=100, D=20 Sparsity: M=10%, D=50% M: $\rho = 0.08$, $\gamma = 0.25$, $\alpha = 0.3$ D: $\rho = 0.01$, $\gamma = 0.01$, $\alpha = 0.99$

Table 1.1: Table of models used in applied forecasting exercises. MFESN hyperparameters are defined with respect to normalized state parameters c.f. (1.6).

Benchmarks

Unconditional mean. We use the unconditional mean of the sample used for fitting as a baseline benchmark. For GDP growth forecasting, there is evidence that the unconditional mean produces forecasts that are competitive with linear models such as VARs in terms of mean square forecasting errors (MSFE), even at relatively short horizons (Arora et al., 2013). It is therefore an important reference for the performance of all other models and we report relative MSFE with respect to the unconditional mean in the tables below.

AR(1) MODEL. A simple autoregressive process of order one on the target variable is included as a benchmark model.[4] This is also a common benchmark in the literature, as AR(1) models are often able to capture key dynamics and produce meaningful forecasts for macroeconomic variables (Stock and Watson, 2002, Bai and Ng, 2008). We emphasize that since AR(1) model is fit to the series of quarterly GDP targets and does not use any additional information, its forecasts are identical for both the Small-MD and Medium-MD samples.

MIXED DATA SAMPLING (MIDAS). The first mixed-frequency model benchmark is given by a MIDAS model (Ghysels et al., 2004, 2007). Our dynamic MIDAS specification includes autoregressive lags of the target series and uses an Almon weighting scheme. As shown in Bai et al. (2013), exponential Almon MIDAS regressions are related to dynamic factor models, which we also consider as benchmarks. The MIDAS model includes three lags of quarterly GDP target variable, and 30 daily and 9 monthly lags for all daily and monthly series, respectively. This model prescription allows for some parsimony as the Almon polynomial weighing reduces the number of daily and monthly lag coefficients.

A thorough description of our MIDAS implementation can be found in Appendix 1.F. To make optimization more efficient, we use explicit expressions for MIDAS loss gradients as in Kostrov (2021). The MIDAS estimation can be hard to perform in practice due to the complexity of nonlinear optimization. First, exponential weighting schemes might require computing floating-point numbers that exceed numerical precision. Therefore, it is a better choice to start the gradient descent close to the origin of the parameter space. Second, even with this choice of starting points, one may encounter issues with optimization results since the Almon-scheme MIDAS loss can have a large number of distinct local minima. In Appendix 1.I.1 we document, using a simple replication experiment, that even small changes in the initial conditions can result in different local minima picked by the numerical optimization algorithm.[5] These important robustness issues are present even when using closed-form gradients and multi-start optimization routines for the MIDAS models. The computational issues become more pronounced as the number of MIDAS parameters increases unless a careful model/variable selection step is performed. We, therefore, do not include any MIDAS model specifications in the Medium-MD setup.

DYNAMIC FACTOR MODEL (DFM). The dynamic factor model framework has been extensively applied in macroeconometrics, starting with Geweke (1977) and Sargent et al. (1977). A DFM specification assumes that predictable dynamics of a large set of time series can be explained by a small number of factors with an autoregressive dependence (see for example Forni et al. (2005), Doz et al. (2011), Stock and Watson (2016)). We generalize the standard two-frequency DFM modeling setup (Mariano and Murasawa, 2003, Bańbura and Modugno, 2014) to a flexible mixed-frequency DFM that encompasses any number of data frequencies. Moreover, we derive a novel weighting scheme that effectively links the MIDAS and DFM approaches. For a detailed discussion of our factor model setup, we refer the reader to Appendix 1.G. Two distinct DFM specifications are used. The first one termed DFM [A] uses the standard linear aggregation scheme, as provided

[4] Suggested by an anonymous referee.

[5] We set the initial coefficient values to zero in all empirical exercises.

in Example 1.G.1, while the second is a variation that implements an Almon weighting scheme as presented in Example 1.G.2 (we name it DFM [B]). The latter is similar to a MIDAS-type aggregation scheme (Marcellino and Schumacher, 2010): the factor structure effectively mitigates the parameter proliferation.

A key choice for a DFM model is the dimension of the factor process. While a number of methods have been developed over the years to systematically derive the number of factors (see, for example, the review of Stock and Watson 2016), commonly used macroeconomic panels feature a number of challenges, such as weak factors (Onatski, 2012). Moreover, as mentioned in Appendix 1.G.1, factor number selection in the mixed-frequency setting has not been sufficiently addressed in the literature. To sidestep these issues, we construct both DFM models with 5 unobserved factors for Small-MD and 10 for Medium-MD, respectively, and assume that they follow a VAR(1) process.

One extant issue with integrating daily data is its very high release frequency compared to monthly and especially quarterly releases: computationally this can be extremely taxing, which might be one of the reasons why to our knowledge we are *the first to provide DFM forecasts that include daily data.* Our solution is to reduce aggregate daily data every 6 days by averaging, thus leaving 4 observations per month. This eases the computational burden to estimate coefficients and latent states considerably (12 versus 72 daily observations per quarter).

Multi-Frequency ESNs

The first set of ESNs we propose is given by two S-MFESN models, based on Example 1.3.1. One model uses a reservoir of 30 neurons (we call it singleESN [A]); the other has a larger reservoir of dimension 120 (named singleESN [B]). The sparsity degree of state parameters for both models is set to be $10/N$, where N is the reservoir size. Both MFESNs share the same hyperparameters, $\rho = 0.5$, $\gamma = 1$, $\alpha = 0.1$ (see (1.7)). These values have not been tuned but are presumed credible given other ESN implementations in the literature. To make a fair comparison with DFMs, we fit the S-MFESN models using 6-day-averaged daily data. Note here that for MFESN models the computational gains of averaging are negligible, and are most apparent when tuning the ridge penalty via cross-validation.

Our second set of proposed models consists of two M-MFESNs according to Example 1.3.2. Both models have two reservoirs, one for each data frequency – monthly and daily – with 100 and 20 neurons, respectively; sparsity degrees are again adjusted to be $10/N$, where N is the reservoir state dimension. The first M-MFESN has hyperparameters that are hand-selected among reasonable values: we note that the monthly-frequency reservoir has no state leak and a larger input scaling, while the daily frequency reservoir features smaller scaling than usual (to avoid compressing high volatility events with the activation function) and the same leak rate as in S-MFESN models (we call this specification multiESN [A]). For the second M-MFESN, we change hyperparameters more radically: we aim to set up a model that has a very high input memory (Ballarin et al., 2023), and that also features long-term smoothing of states. Note that here input scaling values are small, spectral radii are an order of magnitude smaller than in previous models, and leak rates are large (we term this model multiESN [B]).

Execution Time (Seconds) for Model Estimation

				DFM		singleESN		multiESN	
Dataset	Mean	AR(1)	MIDAS	[A]	[B]	[A]	[B]	[A]	[B]
Small-MD	0.1	0.7	1.3	40.5	85.5	2.6	4.5	15.3	14.6
Medium-MD	0.1	0.8	–	48.0	226.5	2.5	5.7	17.7	14.7

Table 1.2: Execution time in seconds for model estimation measured over a single run on a quad-core computer. MFESN models timing includes ridge penalty cross-validation. MIDAS estimation time refers to optimization from a single initial value. DFM models were estimated on a single-core server and times are adjusted by a factor of 1/4 for comparison.

1.4.3 Results

We start by commenting on the computational efficiency of competing models and report execution times in seconds in Table 1.2. Firstly, DFM models appear to be the most computationally effortful models among all specifications. For the Small-MD dataset, the simplest MFESN models, that is, singleESN [A] and [B], have execution times which are at most 3.5 times higher than the MIDAS model, while still being at least 15.6 times computationally cheaper than any of the DFM models. The more resource-demanding models MFESN, multiESN [A] and [B], are nevertheless at least 2.6 times faster to run than the best DFM model (DFM [A]). When moving to the Medium-MD dataset, where the MIDAS model is not, as explained earlier, a feasible choice, the most inefficient MFESN model (singleESN [B]) still outperforms the best DFM model, DFM [A], by 8.4 times, while the same holds for multiESN [A] model versus DFM [A] model by 2.7 times. We can conclude that our proposed MFESN architectures provide an attractive and computationally efficient framework for GDP forecasting in the multifrequency framework which is feasible for computations on low-cost machine configurations available to practitioners.

Competing forecasts are compared using the Model Confidence Set (MCS) test derived in Hansen et al. (2011). One should note that due to the intrinsic nature of data availability of macroeconomic time series and panels, our sample sizes are modest. This implies that the small sample sensitivities of the MCS test need to be taken into account when evaluating our comparisons. Recent analyses of the finite sample properties of the MCS methodology have shown that it requires signal-to-noise ratios which are unattainable in most empirical settings, an issue that undermines its applicability (Aparicio and de Prado, 2018). Given this fact, we also conduct pairwise model comparison tests with the Modified Diebold-Mariano (MDM) test for predictive accuracy (Diebold and Mariano, 2002, Harvey et al., 1997).

As we also provide multiple-steps-ahead forecasts, we test for the best subset of models uniformly across all horizons using the Uniform Multi-Horizon MCS (uMCS) test proposed by Quaedvlieg (2021). Since there is relatively little systematic knowledge regarding the power properties of the uMCS test in small samples, our inclusion of this procedure is meant as a statistical counterpoint to simple relative forecasting error comparisons, which provide limited information about the significance of performance differences. We provide more details on our implementation of the test in Appendix 1.E. Finally, we do not report uMCS test outcomes for the expanding window setup, as

1-Step-ahead GDP Forecasting - Small-MD Dataset

Model	Fixed Parameters				Expanding Window				Rolling Window			
	2007		2011		2007		2011		2007		2011	
	MSFE	MCS	MSFE	MCS	MSFE	MCS	MSFE	MCS	MSFE	MCS	MSFE	MCS
Mean	1.000	*	1.000	**	1.000	**	1.000	**	1.000	**	1.000	**
AR(1)	0.758	*	1.230	**	0.789	**	1.226	**	0.824	**	1.196	**
MIDAS	**0.533**	**	1.300		0.596	**	1.129	*	0.709	**	1.170	*
DFM [A]	0.799	*	1.337		0.980	*	1.320		0.919	*	1.226	
DFM [B]	0.885		1.221	**	0.982	*	1.022	**	0.948		1.028	**
singleESN [A]	0.721	**	1.015	**	0.597	**	0.867	**	**0.529**	**	0.863	**
singleESN [B]	0.758	*	**0.921**	**	0.602	**	**0.844**	**	0.561	**	0.930	**
multiESN [A]	0.802	*	1.250		0.635	**	0.874	**	0.621	**	**0.859**	**
multiESN [B]	0.590	**	0.969	**	**0.552**	**	0.895	**	0.530	**	0.921	**

Table 1.3: Relative MSFE and Model Confidence Set (MCS) comparison between models in 1-step-ahead forecasting exercises. Unconditional mean MSFE is used as a reference. MCS columns show inclusion among best models: $*$ indicates inclusion at 90%, $**$ indicates inclusion at 75% confidence.

Quaedvlieg (2021) argues that in such context the test is invalid.

Small Dataset

We begin our discussion of the Small-MD forecasting results by reviewing Table 1.3. For both sample setups (2007 and 2011) and all three estimation strategies (fixed, expanding, and rolling windows) we provide relative MSFE metrics, with the unconditional mean being used as a reference. Plots of each of the model's forecasts are given in Figures 1.3 and 1.4; additional plots for cumulative SFE, cumulative RMSFEs and other metrics can be found in Appendix 1.J.

The overall finding is that MFESN models perform excellent, and, when we exclude the 2007 fixed parameters setup, they perform the best. It is easy to see from Figure 1.3 (a) why the 2007 fixed window estimation case is different from other cases: the 2008 Financial Crisis induced a deep drop in quarter-to-quarter GDP growth that was in stark contrast with previous business cycle fluctuations. By keeping model parameters fixed, and using only information from 1990 to 2007 – periods where systematic fluctuations are small – DFM and MFESN models are fit to produce smooth, low-volatility forecasts. MIDAS, on the other hand, yields an exponential smoothing which can be more responsive to changes in monthly and daily series. From Figure 1.3 (b) and (c) it is possible to see that expanding and rolling window estimation resolves this weakness of state-space models. At the same time, the AR(1) model outperforms the unconditional mean only in the 2007 sample with fixed parameters, losing to the MIDAS model in all but one scenario.

Table 1.3 shows that MFESN models always perform better than the mean in terms of MSFE, something which no other model class achieves across all setups. In both expanding and rolling window setups they also always outperform the AR(1) model. Furthermore, at least one MFESN model for each subclass (single or multi-reservoir) is always included in the model confidence set at the highest confidence level. We remind again that the MCS test of Hansen et al. (2011) might be distorted due to the modest sample sizes considered, even more so in the 2011 test sample. To

complement the MCS, we provide graphical tables for pairwise Modified Diebold-Mariano (MDM) tests, with 10% level rejections highlighted in Figure 1.14, Appendix 1.J. The MDM tests broadly agree with the results of Table 1.3, although they do not account for multiple testing, and therefore cannot be interpreted as yielding subsets of the most accurate forecasting models in a statistical sense.

For multiple-steps-ahead forecasts, relative RMSFE and uMCS are reported in Tables 1.4 and 1.5: we constrain our exercise to $h \in \{1, \dots, 8\}$ steps, since we are interested in GDP growth forecasts within 2 years. Note that for $h = 1$ our results are similar, but do not reduce to the one-step-ahead results. To make correct multistep RMSFE evaluations and execute the uMCS procedure one must select h different vectors of residuals of the same length: this implies that residuals at the end of the forecasting sample must be trimmed off to compute short-term multistep RMSFEs that are comparable to the long-term ones. Generally, we notice that MIDAS, as well as S-MFESNs, provide the worst-performing multistep forecasts, with RMSFEs considerably exceeding the unconditional mean baseline after horizon 1. Figures 1.5 and 1.6 reproduce the RMSFE numbers of the aforementioned tables graphically.

For MIDAS, we have already discussed how the existence of multiple loss minima can generate numerical instabilities. Model re-fitting at each horizon can amplify this problem, as the loss landscape itself changes as new observations are added to the fitting sample. We provide more discussions in Appendix 1.I.1. In the case of S-MFESN models, the reason is structural: we have discussed how in our framework multistep MFESN forecasting entails iterating the state map, which can have multiple attraction (stable) points. If the hyperparameters and estimated full model $\widehat{W}$s jointly do not define a contraction, the limit of the multistep forecast does not have to be the estimated MFSEN model intercept. However, Figures 1.5 and 1.6 show that our M-MFESN models, multiESN [A] and multiESN [B], both perform on par or better than DFM models even after horizon $h = 4$. For example, in the 2007 expanding and rolling window experiments, multiESN [B] is able to outperform both DFMs and an unconditional mean forecast by meaningful margins for forecasts up to a year into the future.

Medium Dataset

We now present the results for the Medium-MD dataset, which includes more than 30 regressors and many high-frequency daily series. The same metrics as in the previous subsection are used for this dataset to evaluate the relative performance of different methods.

The main difference in our empirical exercises is that now we a priori exclude MIDAS from the set of forecasting methods as explained in detail in Subsection 1.4.2. Table 1.6 showcases the relative performance of DFM and MFESN models in the Medium-MD forecast setup. We find that the MFESN model multiESN [B] performs best in all setups, particularly under fixed parameters, where MCS testing reveals that it is the only model included at a 75% confidence level. Of course, for the MCS results we must again take into account the relatively small sample size, which could distort the selection of best model subsets. MDM tests of Figure 1.16 in Appendix 1.J largely agree with the MCS results: in the fixed parameter setup any pairwise comparison of an alternative model against MFESN multiESN [B] is rejected in favor of the latter. A visual inspection of one-step-

Multistep-ahead GDP Forecasting - Small-MD Dataset - 2007 Sample

Setup	Model	Horizon								uMCS
		1	2	3	4	5	6	7	8	
FIX	Mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
FIX	AR(1)	0.870	**0.950**	**0.982**	0.991	0.992	0.991	0.992	0.992	**
FIX	MIDAS	0.823	1.672	2.737	1.816	2.213	2.791	1.888	1.921	
FIX	DFM [A]	0.890	0.969	1.014	1.077	1.341	1.701	2.001	2.180	*
FIX	DFM [B]	0.937	1.069	1.202	1.344	1.799	2.310	2.638	2.801	
FIX	singleESN [A]	0.852	0.994	0.995	0.995	0.993	0.991	0.991	0.991	*
FIX	singleESN [B]	0.871	0.986	0.989	**0.989**	**0.985**	**0.981**	**0.981**	**0.981**	**
FIX	multiESN [A]	0.898	0.980	0.990	0.991	0.988	0.985	0.985	0.985	**
FIX	multiESN [B]	**0.767**	0.954	0.983	0.991	0.991	0.990	0.991	0.991	**
EW	Mean	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**	**1.000**	-
EW	AR(1)	0.887	0.922	0.951	0.962	0.957	**0.981**	1.001	1.008	-
EW	MIDAS	0.814	1.283	1.518	1.596	1.697	1.391	1.951	1.800	-
EW	DFM [A]	0.985	1.109	1.123	1.114	1.217	1.226	1.241	1.539	-
EW	DFM [B]	0.989	1.082	1.149	1.199	1.315	1.412	1.373	1.425	-
EW	singleESN [A]	0.771	1.260	1.485	1.564	2.070	2.728	2.550	2.834	-
EW	singleESN [B]	0.772	1.031	1.135	1.319	1.831	2.279	2.449	2.556	-
EW	multiESN [A]	0.792	0.897	0.941	0.976	1.015	1.240	1.377	1.227	-
EW	multiESN [B]	**0.740**	**0.853**	**0.894**	**0.911**	**0.873**	0.993	1.020	1.020	-
RW	Mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	*
RW	AR(1)	0.898	0.943	0.968	0.974	0.963	**0.968**	0.970	**0.962**	**
RW	MIDAS	0.933	1.438	1.642	1.993	1.794	1.661	1.816	1.973	*
RW	DFM [A]	0.931	1.017	1.033	1.020	1.024	1.003	**0.918**	1.062	*
RW	DFM [B]	0.942	0.973	0.970	1.045	1.059	1.203	1.225	1.263	*
RW	singleESN [A]	**0.714**	1.320	1.693	1.972	2.733	3.669	3.391	3.719	*
RW	singleESN [B]	0.737	1.100	1.248	1.667	2.327	2.765	2.842	2.792	*
RW	multiESN [A]	0.773	0.972	1.053	1.111	1.187	1.293	1.505	1.131	*
RW	multiESN [B]	0.716	**0.895**	**0.916**	**0.926**	**0.890**	1.041	1.102	1.105	**

Table 1.4: Relative RMSFE and Uniform Multi-Horizon Model Confidence Set (uMCS) comparison between models in multiple-steps-ahead forecasting exercises. Unconditional mean RMSFE used as reference. FIX: Fixed parameters, EW: Expanding window, and RW: Rolling window. uMCS columns show inclusion among best models: $*$ indicates inclusion at 90% confidence, $**$ indicates inclusion at 75% confidence.

Multistep-ahead GDP Forecasting - Small-MD Dataset - 2011 Sample

Setup	Model	Horizon								uMCS
		1	2	3	4	5	6	7	8	
FIX	Mean	1.000	1.000	1.000	1.000	**1.000**	1.000	1.000	1.000	**
FIX	AR(1)	1.119	1.031	1.008	1.001	1.001	0.999	**0.999**	**0.998**	*
FIX	MIDAS	1.090	1.721	1.793	2.203	2.363	1.997	2.846	2.328	
FIX	DFM [A]	1.112	1.051	0.999	1.079	1.084	1.025	1.020	1.061	*
FIX	DFM [B]	1.058	**0.945**	**0.916**	1.003	1.012	**0.970**	1.038	1.033	**
FIX	singleESN [A]	0.978	1.705	2.561	2.704	3.314	3.151	2.999	3.316	
FIX	singleESN [B]	**0.930**	1.095	1.885	2.356	2.650	2.704	2.880	2.844	**
FIX	multiESN [A]	1.059	1.148	1.262	1.312	1.339	1.409	1.424	1.162	
FIX	multiESN [B]	0.981	1.007	0.985	**0.994**	1.008	0.999	**0.999**	**0.998**	**
EW	Mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-
EW	AR(1)	1.117	1.033	1.011	1.002	1.007	1.003	1.004	1.003	-
EW	MIDAS	1.005	1.382	1.339	1.354	1.609	1.444	1.803	1.263	-
EW	DFM [A]	1.144	1.132	1.057	1.093	1.076	1.067	1.038	1.016	-
EW	DFM [B]	0.985	**0.940**	**0.918**	0.995	1.010	**0.980**	1.050	**0.971**	-
EW	singleESN [A]	0.935	1.645	2.184	1.929	2.388	1.959	1.810	2.266	-
EW	singleESN [B]	**0.911**	1.092	1.101	1.529	2.195	1.843	1.847	2.060	-
EW	multiESN [A]	0.922	0.965	1.089	0.978	**0.977**	1.043	1.278	0.995	-
EW	multiESN [B]	0.944	0.992	0.978	**0.977**	0.991	0.985	**0.990**	0.996	-
RW	Mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
RW	AR(1)	1.080	1.000	0.984	**0.989**	0.982	0.976	**0.963**	0.968	
RW	MIDAS	1.051	1.303	1.310	1.674	1.762	1.467	1.643	1.463	
RW	DFM [A]	1.061	1.033	1.012	1.088	1.077	1.015	1.040	1.069	
RW	DFM [B]	0.947	**0.893**	**0.901**	1.009	1.040	**0.966**	1.030	**0.949**	**
RW	singleESN [A]	0.919	1.788	2.359	2.483	2.981	2.401	2.234	2.690	
RW	singleESN [B]	0.944	1.132	1.214	1.762	2.608	2.552	2.517	2.541	
RW	multiESN [A]	**0.896**	1.047	1.222	1.124	1.122	1.410	1.666	1.316	
RW	multiESN [B]	0.940	1.003	0.969	**0.989**	**0.979**	0.972	0.967	0.961	**

Table 1.5: Relative RMSFE and Uniform Multi-Horizon Model Confidence Set (uMCS) comparison between models in multiple-steps-ahead forecasting exercises. Unconditional mean RMSFE used as reference. FIX: Fixed parameters, EW: Expanding window, and RW: Rolling window. uMCS columns show inclusion among best models: $*$ indicates inclusion at 90%, $**$ indicates inclusion at 75% confidence.

1-Step-ahead GDP Forecasting - Medium-MD Dataset

Model	Fixed Parameters				Expanding Window				Rolling Window			
	2007		2011		2007		2011		2007		2011	
	MSFE	MCS	MSFE	MCS	MSFE	MCS	MSFE	MCS	MSFE	MCS	MSFE	MCS
Mean	1.000		1.000	**	1.000	**	1.000	**	1.000	**	1.000	**
AR(1)	0.758	*	1.230	**	0.789	**	1.226	**	0.824	*	1.196	**
DFM [A]	0.841	*	1.325	*	0.682	**	1.272	**	0.747	*	1.517	**
DFM [B]	1.118	*	1.408	**	0.821	*	1.117	**	0.926		1.186	**
singleESN [A]	0.967	*	1.717	*	0.775	**	1.072	**	0.791	*	1.493	*
singleESN [B]	0.826	*	1.278	**	0.655	**	1.028	**	0.561	**	0.944	**
multiESN [A]	0.901	*	1.080	**	0.618	**	0.913	**	0.556	**	0.884	**
multiESN [B]	**0.682**	**	**0.748**	**	**0.587**	**	**0.774**	**	**0.547**	**	**0.728**	**

Table 1.6: Relative MSFE and Model Confidence Set (MCS) comparison between models in 1-step-ahead forecasting exercises. Unconditional mean MSFE is used as a reference. MCS columns show inclusion among best models: $*$ indicates inclusion at 90% confidence, $**$ indicates inclusion at 75% confidence.

ahead forecasts in Figures 1.7 and 1.8 also shows that DFM models estimated over the Medium-MD datasets produce forecasts with larger variability than MFESN methods, which is likely the key driver of the difference in performance.

The multistep-ahead experiments are run as for the Small-MD dataset, with a maximum horizon of 8 quarters. Tables 1.7 and 1.8 present the relative RMSFE performance of multistep forecasts for all models, and we use Figures 1.9 and 1.10 of RMSFEs as references for our discussion. What can be seen visually – and is also reproduced in the Tables – is that multi-reservoir MFESN models and DFM model [A] have the best performance up to 4 quarters ahead; overall, taking into account also the longer term, expanding or rolling window estimation of model multiESN [B] yields the best forecasting results in the 2007 sample setup. The post-crisis 2011 sample setup makes comparison harder, as DFM and M-MFESN models largely produce results in line with the unconditional sample mean. This evaluation is confirmed by uMCS tests, consistently with the multistep results obtained with the Small-MD dataset.

1.5 Conclusions

Macroeconomic forecasting – especially long-term forecasting of macroeconomic aggregates – is a topic of crucial importance for institutional policymakers, private companies, and economic researchers. Given the modern-day availability of "big data" resources, methods capable of integrating heterogeneous data sources are increasingly sought to provide more precise and robust forecasts.

This paper presents a new methodological framework inspired by the Reservoir Computing literature to deal with data sampled at multiple frequencies and with multiple-step-ahead forecasts. We have then taken Echo State Networks – a type of RC models – and formally extended them to allow the modeling of data with multiple release frequencies. Our discussion encompasses model fitting, hyperparameter tuning, and forecast computation. As a result, we provide two classes of models, single- and multiple reservoir multi-frequency ESNs, that can be effectively applied to our empirical setup: forecasting US GDP growth using monthly and daily data series. Along

Multistep-ahead GDP Forecasting - Medium-MD Dataset - 2007 Sample

Setup	Model	Horizon								uMCS
		1	2	3	4	5	6	7	8	
FIX	Mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	*
FIX	AR(1)	0.870	0.950	0.982	0.991	0.992	0.991	0.992	0.992	
FIX	DFM [A]	0.914	**0.947**	**0.955**	0.988	1.015	1.027	1.034	0.995	**
FIX	DFM [B]	1.046	1.204	1.293	1.341	1.649	1.984	2.101	2.070	*
FIX	singleESN [A]	0.985	0.995	0.995	0.995	0.994	0.992	0.992	0.992	*
FIX	singleESN [B]	0.912	0.985	0.985	**0.985**	**0.980**	**0.976**	**0.976**	**0.976**	*
FIX	multiESN [A]	0.950	0.993	0.994	0.994	0.992	0.990	0.990	0.990	*
FIX	multiESN [B]	**0.826**	0.972	0.988	0.990	0.989	0.986	0.985	0.985	*
EW	Mean	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**	**1.000**	-
EW	AR(1)	0.887	0.922	0.951	0.962	0.957	0.981	1.001	1.008	-
EW	DFM [A]	0.805	0.916	0.978	1.038	1.077	1.126	1.077	1.073	-
EW	DFM [B]	0.893	1.134	1.418	1.567	2.238	2.964	3.375	3.629	-
EW	singleESN [A]	0.879	1.125	1.305	1.442	1.860	2.166	2.361	2.443	-
EW	singleESN [B]	0.802	1.174	1.439	1.744	2.305	2.869	2.935	3.167	-
EW	multiESN [A]	0.780	0.935	1.012	1.005	1.093	1.337	1.328	1.313	-
EW	multiESN [B]	**0.760**	**0.874**	**0.911**	**0.891**	**0.863**	**0.971**	1.030	1.051	-
RW	Mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
RW	AR(1)	0.898	0.943	0.968	0.974	0.963	**0.968**	**0.970**	**0.962**	
RW	DFM [A]	0.837	0.913	0.924	0.954	1.012	0.997	1.018	1.005	
RW	DFM [B]	0.932	1.116	1.232	1.414	1.952	2.704	3.183	3.294	
RW	singleESN [A]	0.873	1.274	1.530	1.652	2.095	2.575	2.786	3.014	
RW	singleESN [B]	0.732	1.190	1.490	1.712	2.218	2.861	2.967	3.094	
RW	multiESN [A]	0.732	0.914	0.960	1.011	1.202	1.618	1.683	1.572	
RW	multiESN [B]	**0.731**	**0.871**	**0.875**	**0.844**	**0.771**	0.971	1.014	1.014	**

Table 1.7: Relative RMSFE and Uniform Multi-Horizon Model Confidence Set (uMCS) comparison between models in multiple-steps-ahead forecasting exercises. Unconditional mean RMSFE used as reference. FIX: Fixed parameters, EW: Expanding window, and RW: Rolling window. uMCS columns show inclusion among best models: $*$ indicates inclusion at 90% confidence, $**$ indicates inclusion at 75% confidence.

Multistep-ahead GDP Forecasting - Medium-MD Dataset - 2011 Sample

Setup	Model	Horizon								uMCS
		1	2	3	4	5	6	7	8	
FIX	Mean	1.000	1.000	1.000	1.000	**1.000**	1.000	1.000	1.000	*
FIX	AR(1)	1.119	1.031	1.008	1.001	1.001	0.999	**0.999**	**0.998**	**
FIX	DFM [A]	1.126	0.987	0.962	1.054	1.031	**0.988**	1.001	1.002	**
FIX	DFM [B]	1.149	0.987	**0.885**	1.064	1.142	1.134	1.273	1.296	
FIX	singleESN [A]	1.283	1.921	2.527	3.038	3.285	3.154	3.193	3.655	
FIX	singleESN [B]	1.059	1.523	1.918	2.417	2.812	2.683	2.703	2.970	
FIX	multiESN [A]	1.011	1.061	1.434	1.477	1.748	2.030	2.023	1.994	
FIX	multiESN [B]	**0.841**	**0.945**	0.997	**0.978**	1.004	1.015	1.013	1.014	**
EW	Mean	1.000	1.000	1.000	1.000	**1.000**	1.000	1.000	1.000	-
EW	AR(1)	1.117	1.033	1.011	1.002	1.007	1.003	1.004	1.003	-
EW	DFM [A]	1.092	0.942	**0.944**	1.049	1.026	**0.994**	**0.996**	**0.999**	-
EW	DFM [B]	0.971	1.046	1.031	1.114	1.238	1.116	1.223	1.310	-
EW	singleESN [A]	1.039	1.451	1.980	2.385	2.699	2.353	2.506	2.608	-
EW	singleESN [B]	0.992	1.828	2.465	3.072	3.547	3.357	3.368	3.610	-
EW	multiESN [A]	0.934	1.014	1.391	1.252	1.371	1.369	1.228	1.279	-
EW	multiESN [B]	**0.857**	**0.931**	1.003	**0.973**	1.002	1.009	1.025	1.029	-
RW	Mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**
RW	AR(1)	1.080	1.000	0.984	**0.989**	**0.982**	**0.976**	**0.963**	**0.968**	**
RW	DFM [A]	1.113	0.982	**0.927**	1.038	1.030	0.997	1.016	1.028	*
RW	DFM [B]	0.881	0.996	1.021	1.098	1.150	1.114	1.114	1.212	**
RW	singleESN [A]	1.193	2.267	3.265	3.580	4.090	3.790	4.015	4.562	
RW	singleESN [B]	0.927	1.933	2.612	3.265	3.753	3.567	3.556	3.792	
RW	multiESN [A]	0.900	1.049	1.500	1.465	1.789	1.707	1.505	1.462	
RW	multiESN [B]	**0.816**	**0.916**	0.977	1.009	**0.982**	0.988	0.974	0.981	**

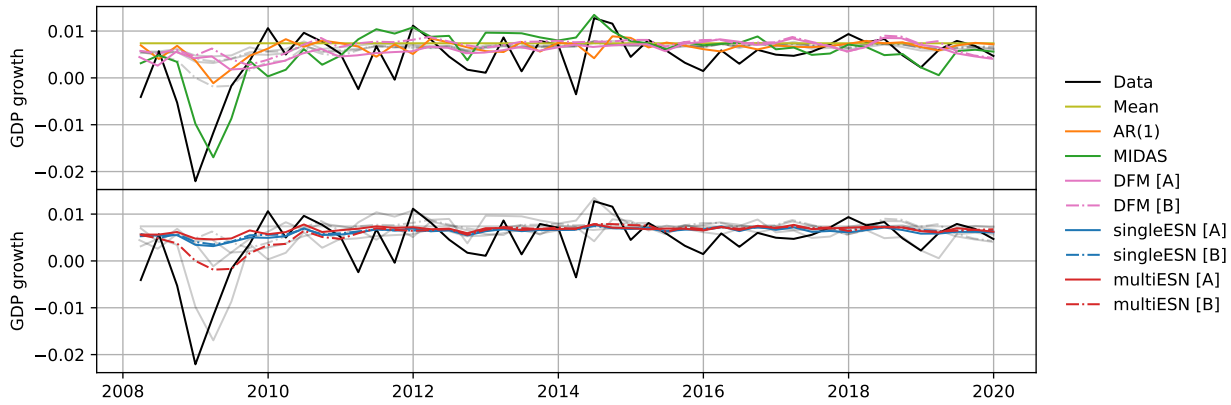
Table 1.8: Relative RMSFE and Uniform Multi-Horizon Model Confidence Set (uMCS) comparison between models in multiple-steps-ahead forecasting exercises. Unconditional mean RMSFE used as reference. FIX: Fixed parameters, EW: Expanding window, and RW: Rolling window. uMCS columns show inclusion among best models: $*$ indicates inclusion at 90% confidence, $**$ indicates inclusion at 75% confidence.

with the unconditional mean and AR(1) model, we considered two well-known methods, MIDAS and DFMs, as the current benchmarks available in the literature. In our applications, we find that MFESN models are computationally more efficient and easier to implement than DFMs and MIDAS, respectively, and perform better than or as well as benchmarks in terms of MSFE. These improvements are statistically significant in a number of setups, as shown by our MCS and MDM tests. Thus, we argue that our machine learning-based methodology can be a useful addition to the toolbox of contemporary macroeconomic forecasters.

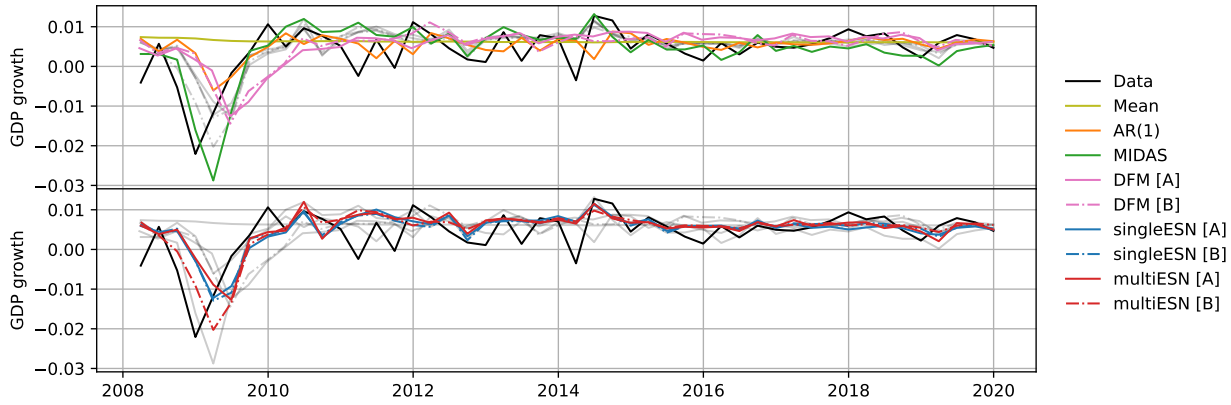
Lastly, we wish to highlight the many potential areas of research that we believe would be interesting to explore in the future. We have not discussed the role of the distribution from which we sample the entries of the reservoir matrices. While it is known that these can have significant effects on the forecasting capacity of an ESN model, the literature lacks definitive theoretical results (even for dynamical systems applications) or systematic studies with stochastic inputs and targets. The hyperparameter tuning routine we have developed neither allows separating individual hyperparameters nor does it tackle the identification problem. Moreover, we assume that the ridge regression penalty strength, λ, is tuned *ex ante*: it would be interesting and desirable to understand if it is possible to jointly tune λ and $\boldsymbol{\varphi}$, or rather if one can fully separate their selection. In our preliminary experiments, we have noticed that the roles of the ridge penalty and the input scaling, for example, cannot be trivially disentangled – thus prompting the ψ-form normalization. Model selection for the dimension of MFESN models is another question that would be key to exploring and designing more efficient and effective ESN models, especially when dealing with multiple frequencies and reservoirs. Finally, practitioners may be interested in identifying the combination of frequencies in the regressor series that would lead to the most accurate GDP forecasts produced by MFESN models.

Figure 1.3: 1-Step-ahead GDP Forecasting – 2007 Sample – Small-MD Dataset

(a) Fixed



(b) Expanding



(c) Rolling

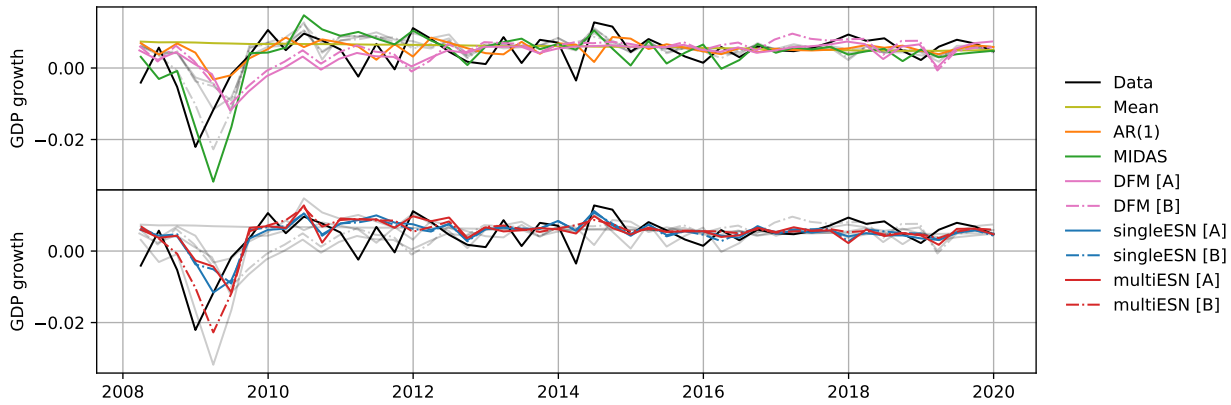
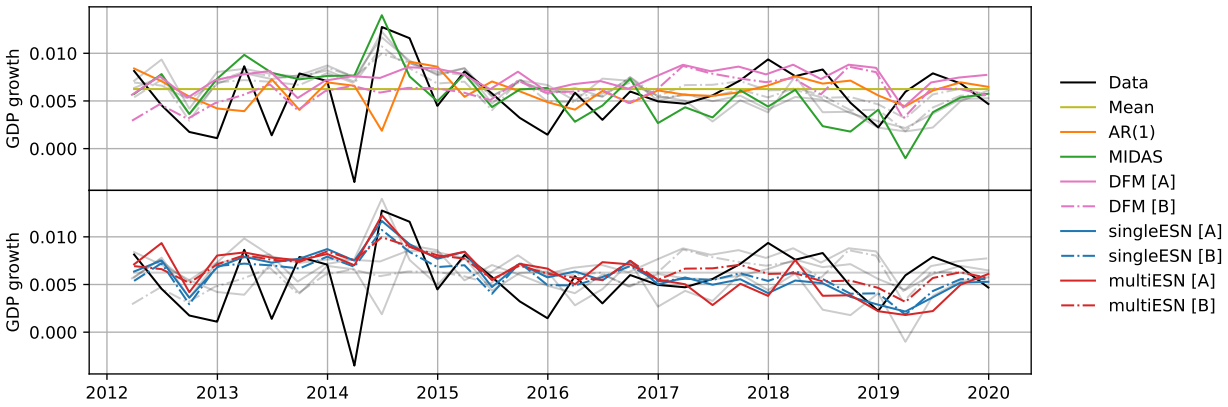
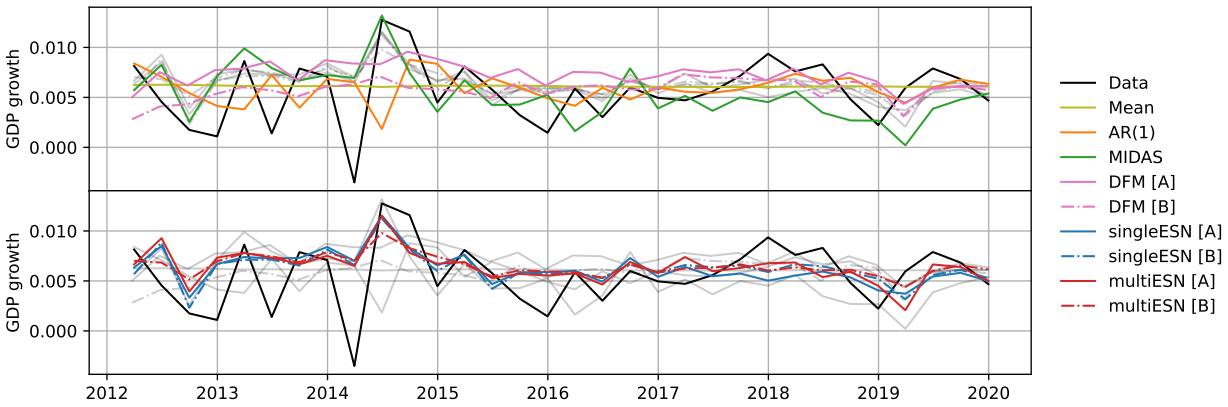


Figure 1.4: 1-Step-ahead GDP Forecasting – 2011 Sample – Small-MD Dataset

(a) Fixed



(b) Expanding



(c) Rolling

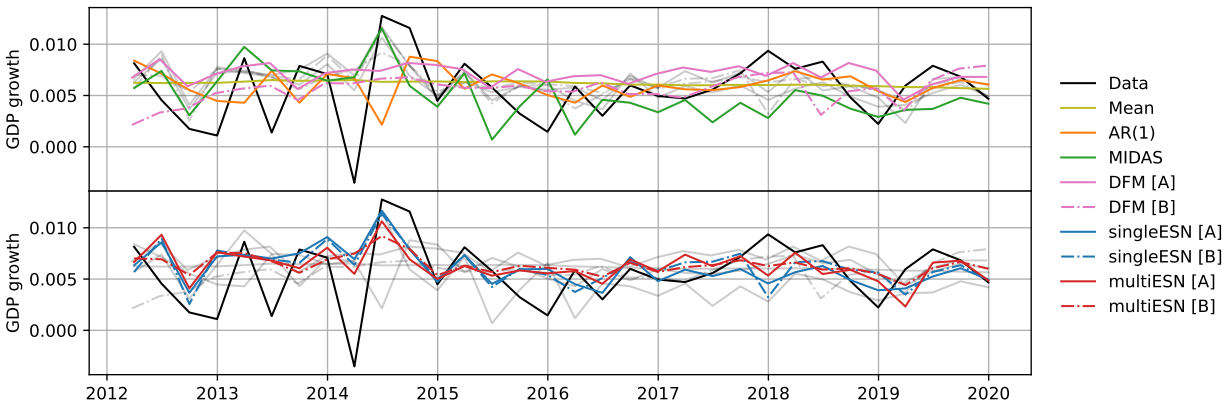
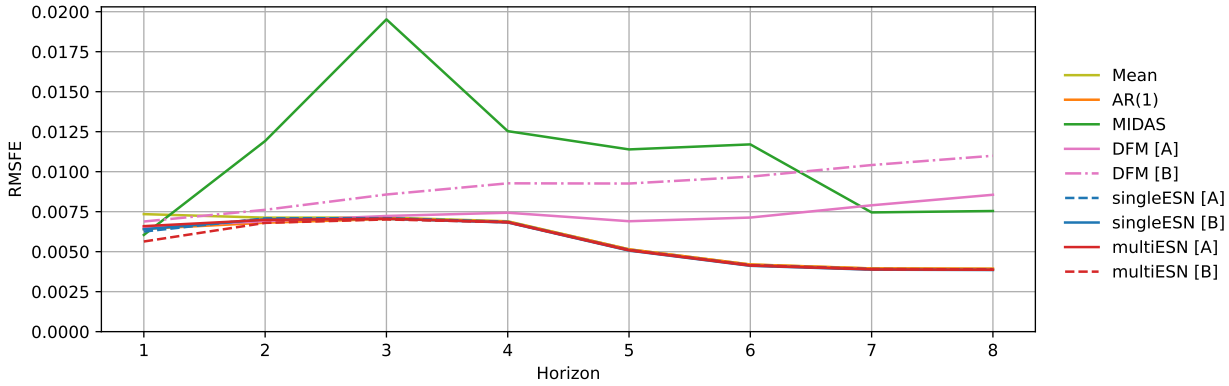
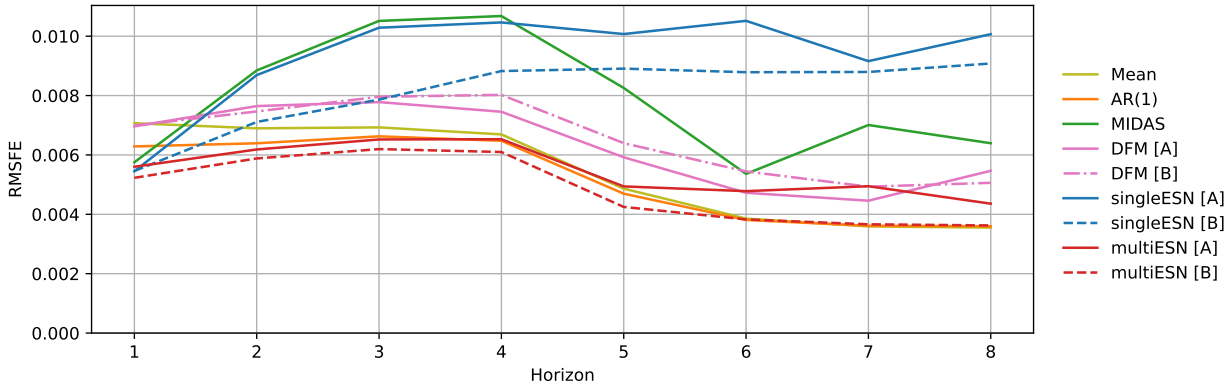


Figure 1.5: Multistep-ahead GDP Forecasting, RMSFE – 2007 Sample – Small-MD Dataset

(a) Fixed



(b) Expanding



(c) Rolling

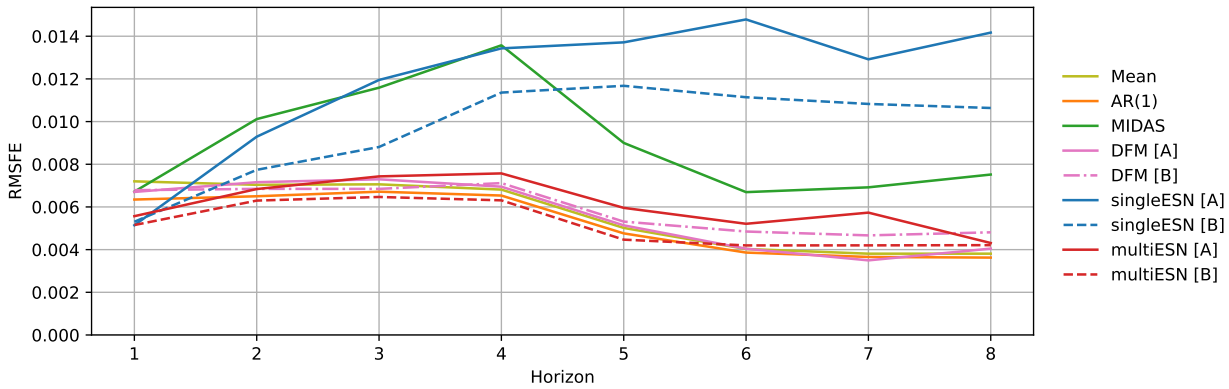
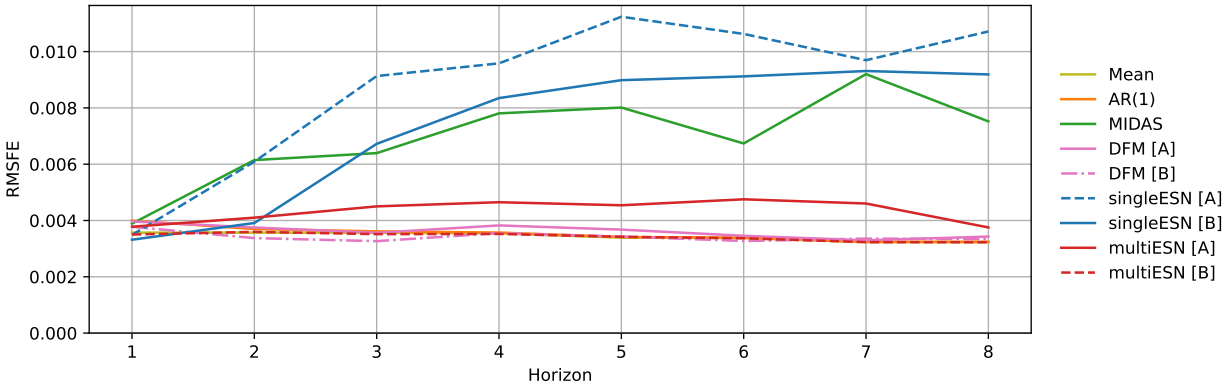
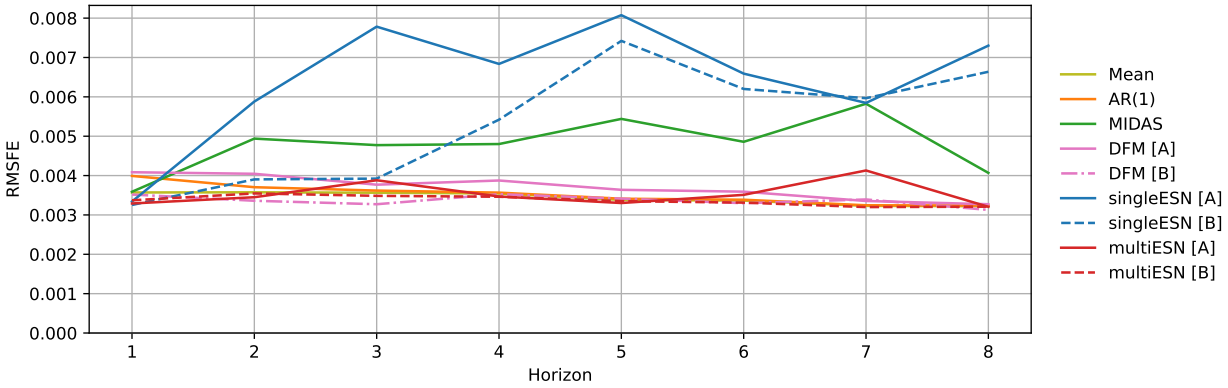


Figure 1.6: Multistep-ahead GDP Forecasting, RMSFE – 2011 Sample – Small-MD Dataset

(a) Fixed



(b) Expanding



(c) Rolling

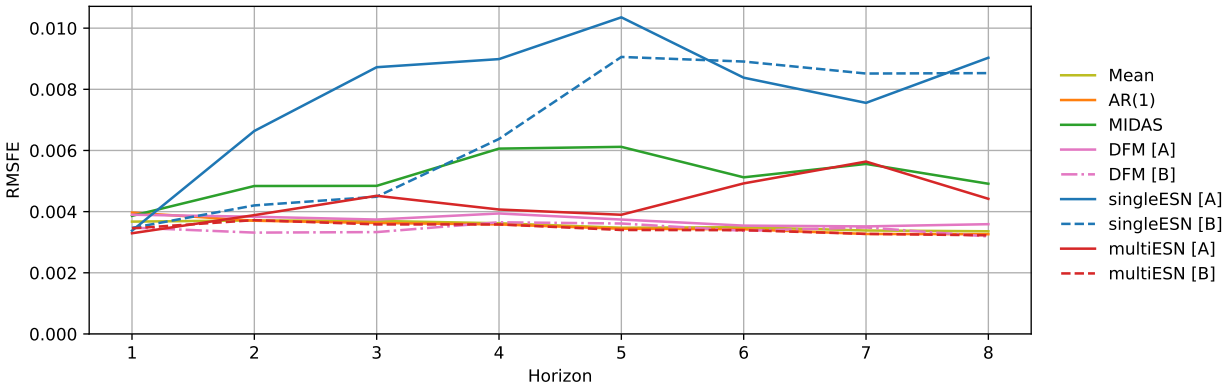
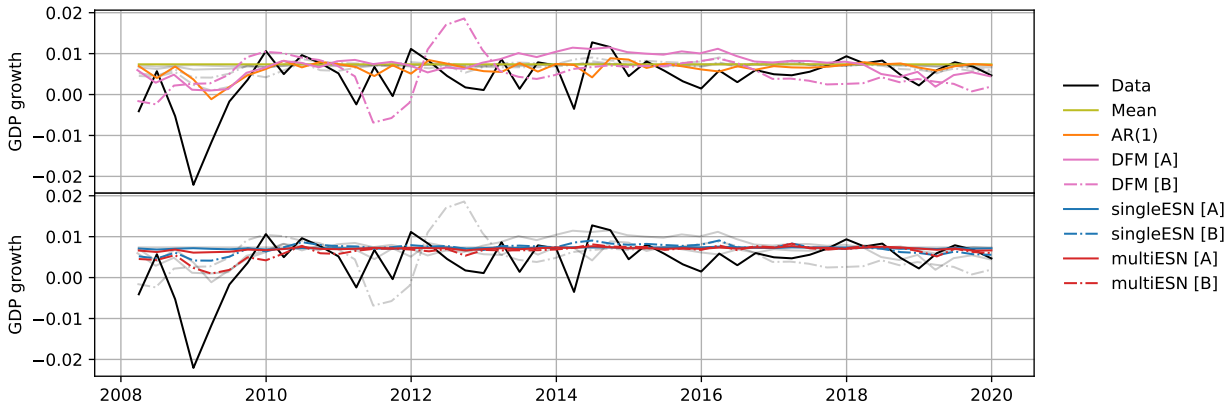
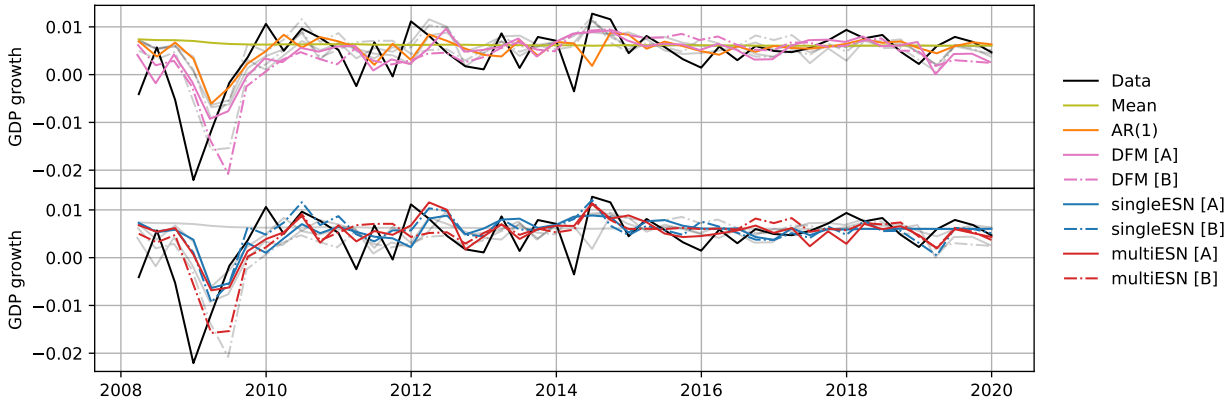


Figure 1.7: 1-Step-ahead GDP Forecasting – 2007 Sample – Medium-MD Dataset

(a) Fixed



(b) Expanding



(c) Rolling

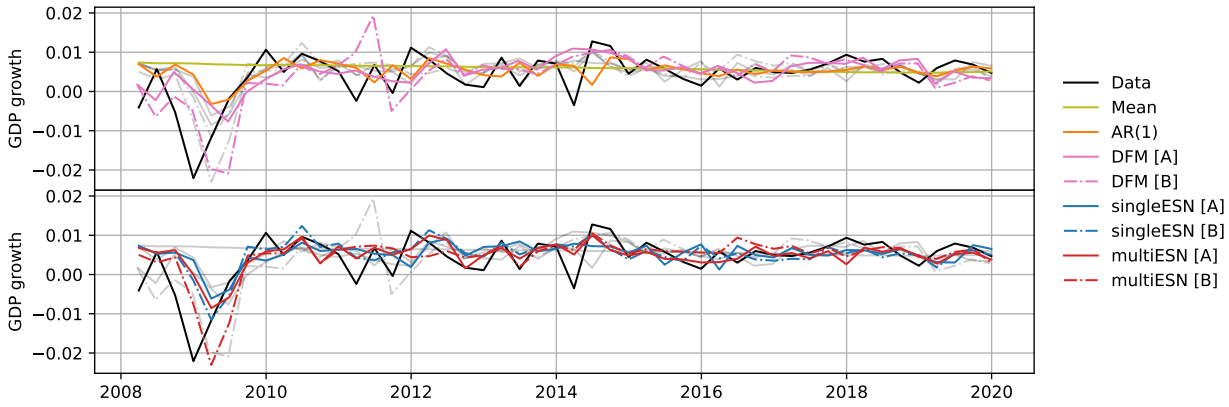
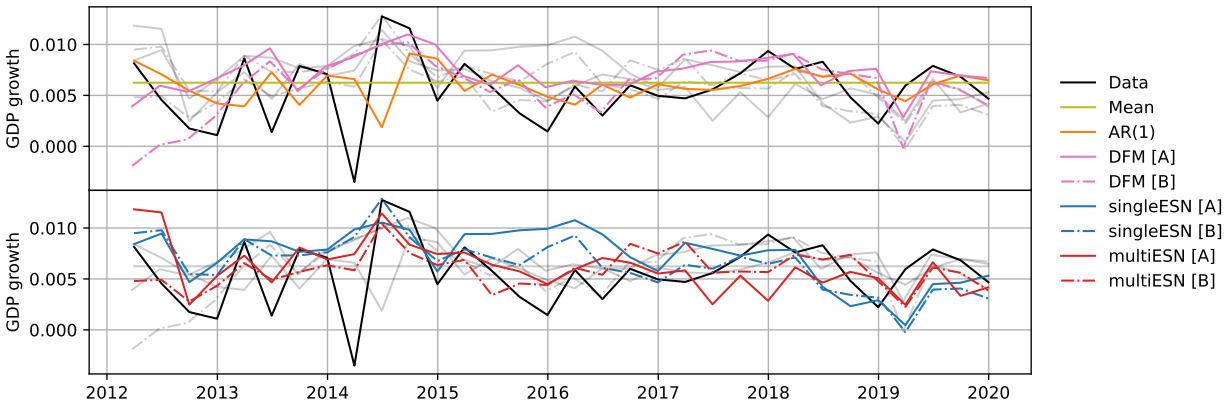
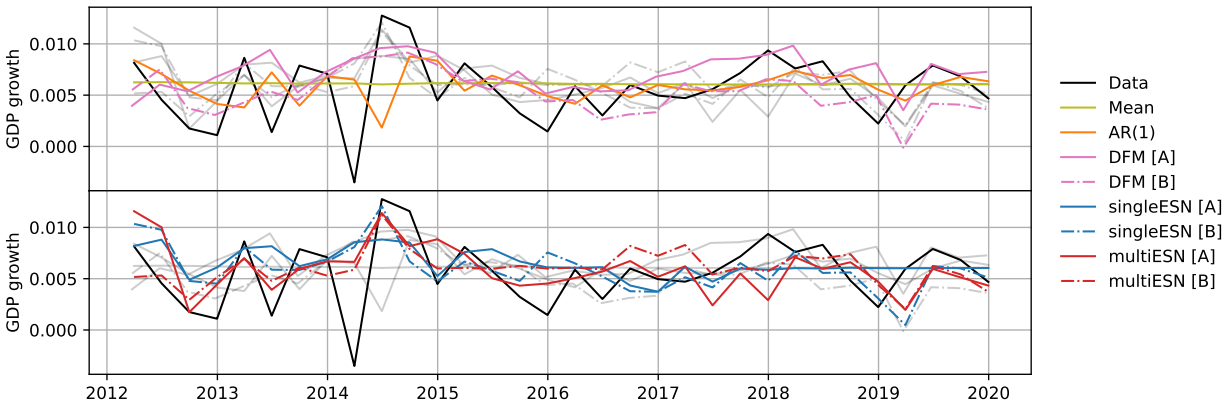


Figure 1.8: 1-Step-ahead GDP Forecasting – 2011 Sample – Medium-MD Dataset

(a) Fixed



(b) Expanding



(c) Rolling

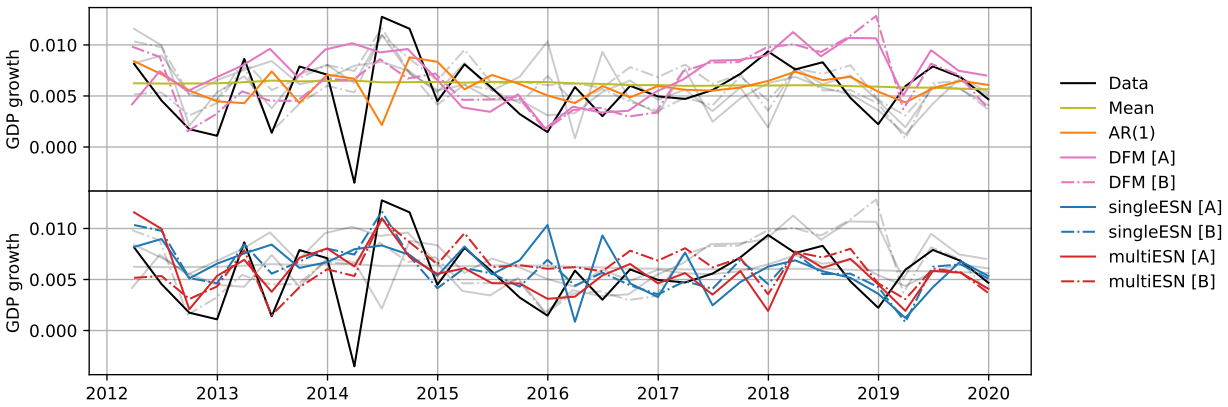
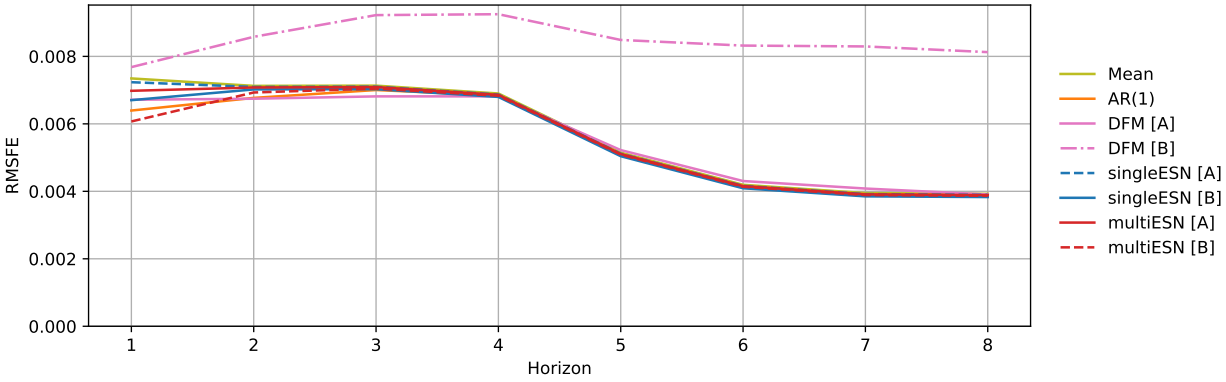
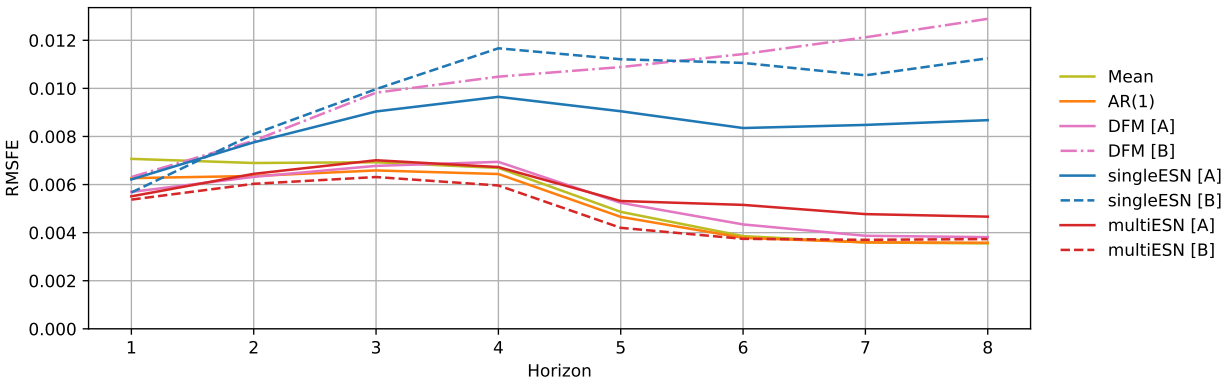


Figure 1.9: Multistep-ahead GDP Forecasting, RMSFE – 2007 Sample – Medium-MD Dataset

(a) Fixed



(b) Expanding



(c) Rolling

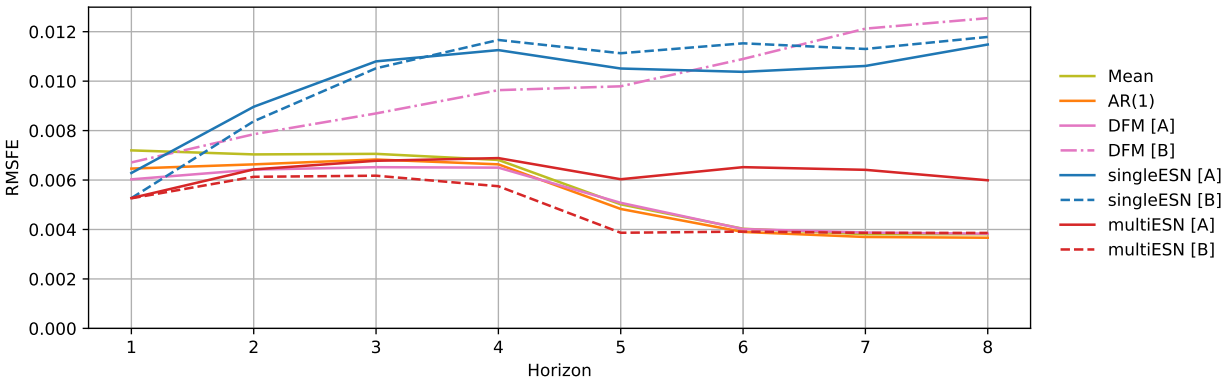
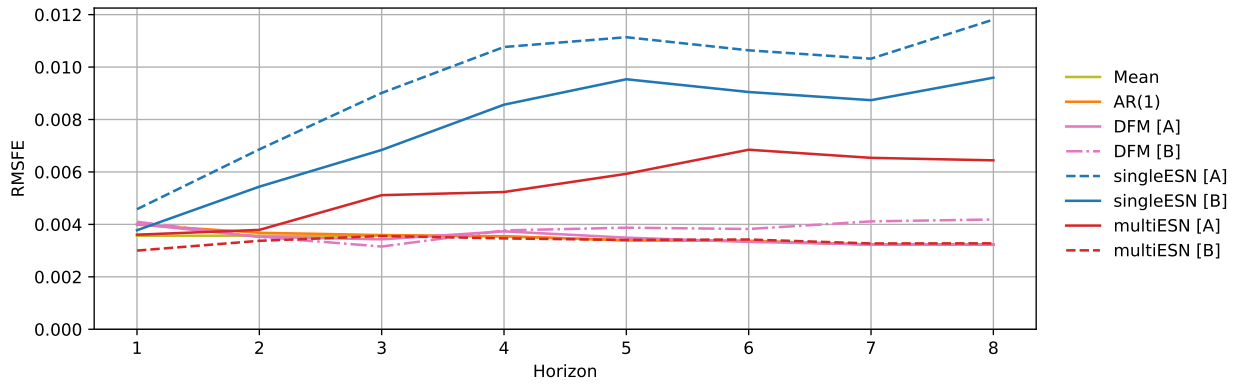
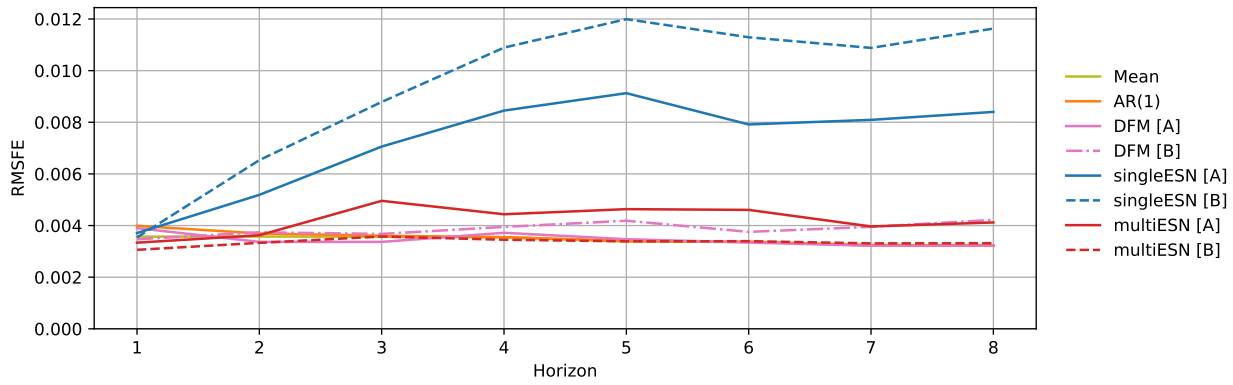


Figure 1.10: Multistep-ahead GDP Forecasting, RMSFE – 2011 Sample – Medium-MD Dataset

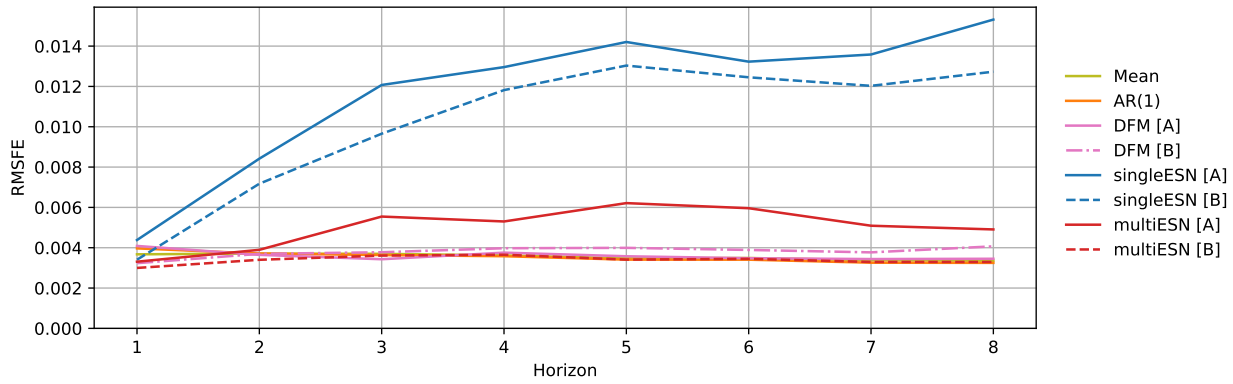
(a) Fixed



(b) Expanding



(c) Rolling



Appendix

1.A Data Table

Table 1.9: Variables, Frequencies and Transformations for Small and Medium

S M	Start Date	T	Code	Name	Description
Quarterly					
XX	31/03/1959	5	GDPC1	Y	Real Gross Domestic Product
Monthly					
XX	30/01/1959	5	INDPRO	XM1	Industrial Production Index
XX	30/01/1959	5	PAYEMS	XM4	Payroll All Employees: Total nonfarm
XX	30/01/1959	4	HOUST	XM5	Housing Starts: Total New Privately Owned
XX	30/01/1959	5	RETAILx	XM7	Retail and Food Services Sales
XX	31/01/1973	5	TWEXMMTH	XM11	Nominal effective exchange rate US
XX	30/01/1959	2	FEDFUNDS	XM12	Effective Federal Funds Rate
XX	30/01/1959	1	BAAFFM	XM14	Moody's Baa Corporate Bond Minus FEDFUNDS
XX	30/01/1959	1	COMPAPFFx	XM15	3-Month Commercial Paper Minus FEDFUNDS
X	30/01/1959	2	CUMFNS	XM2	Capacity Utilization: Manufacturing
X	30/01/1959	2	UNRATE	XM3	Civilian Unemployment Rate
X	30/01/1959	5	DPCERA3M086SBEA	XM6	Real personal consumption expenditures
X	30/01/1959	5	AMDMNOx	XM8	New Orders for Durable Goods
X	31/01/1978	2	UMCSENTx	XM9	Consumer Sentiment Index
X	30/01/1959	6	WPSFD49207	XM10	PPI: Finished Goods
X	30/01/1959	1	AAAFFM	XM13	Moody's Aaa Corporate Bond Minus FEDFUNDS
X	30/01/1959	1	TB3SMFFM	XM16	3-Month Treasury C Minus FEDFUNDS
X	30/01/1959	1	T10YFFM	XM17	10-Year Treasury C Minus FEDFUNDS
X	30/01/1959	2	GS1	XM18	1-Year Treasury Rate
X	30/01/1959	2	GS10	XM19	10-Year Treasury Rate
X	30/01/1959	1	GS10-TB3MS	XM20	10-Year Treasury Rate - 3-Month Treasury Bill
Daily					
XX	30/01/1959	8	DJINDUS	XD3	DJ Industrial price index
X	31/12/1963	8	S&PCOMP	XD1	S&P500 price index
X	01/05/1982	1	ISPCS00-S&PCOMP†	XD2	S&P500 basis spread
X	11/09/1989	8	SP5EIND	XD4	S&P Industrial price index
X	31/12/1969	8	GSCITOT	XD5	Spot commodity price index
X	10/01/1983	8	CRUDOIL	XD6	Spot price oil
X	02/01/1979	8	GOLDHAR	XD7	Spot price gold
X	30/03/1982	8	WHEATSF	XD8	Spot price wheat
X	01/11/1983	8	COCOAIC,COCINUS‡	XD9	Spot price cocoa
X	30/03/1983	1	NCLC.03-NCLC.01	XD10	Futures price oil term structure
X	30/10/1978	1	NGCC.03-NGCC.01	XD11	Futures price gold term structure
X	02/01/1975	1	CWFC.03-CWFC.01	XD12	Futures price wheat term structure
X	02/01/1973	1	NCCC.03-NCCC.01	XD13	Futures price cocoa term structure

Notes: S and M stand for small and medium datasets, respectively. An 'X' indicates selection into the dataset. 'Start Date' is the date for which the series is first available (before data transformations). Following McCracken and Ng (2016, 2020), the transformation codes in column 'T' indicate with D for difference and log for natural logarithm 1: none, 2: D, 3: DD, 4: Log, 5: Dlog, 6: DDlog, 7: percentage change, 8: GARCH volatility. 'Codes' are the codes in the FRED-QD and FRED-MD datasets for quarterly and monthly data and Datastream mnemonic for the remaining frequencies. Missing values due to public holidays are interpolated by averaging over the previous five observations. ‡Available until 20/09/2021. ‡Average before 29/12/2017, COCINUS mean adjusted thereafter.

1.B Forecasting Schemes

To clarify the design of the forecasting experiments conducted in this paper, we present two different types of prediction illustrated in Figure 1.11.

Let t denote time in the reference frequency of the target series (y_t) and suppose a regressor (z_r) of frequency κ is included in the forecasting model. The notation can be readily extended to include multiple regressors. Let $h \geq 0$ be a *low-frequency* prediction horizon counted from the last available observation of (y_t). Let $l \geq 0$ be a *high-frequency* horizon with respect to frequency κ.

Low-frequency forecasting. We call an h-steps ahead forecast *low-frequency* when predictions for the target variable are constructed only at the end of the low-frequency periods. The information set which is used at the time of h-steps ahead low-frequency forecasting at t is the σ-algebra defined as

$$\mathcal{F}_t = \sigma\left(\left\{y_t, y_{t-1}, y_{t-2}, \ldots, z_{t,0|\kappa}, z_{t,-1|\kappa}, z_{t,-2|\kappa}, \ldots\right\}\right) \tag{1.35}$$

and, when using the mean square error as a loss, the optimal forecast is given by

$$\widehat{y}_{t+h} = \mathbb{E}\left[y_{t+h}|\mathcal{F}_t\right]. \tag{1.36}$$

High-frequency forecasting. In this forecasting scheme, one may also use high-frequency regressors to produce additional high-frequency forecasts of the low-frequency target variable. For example, in the case of a target released at the end of each year and having monthly quoted covariates, the low-frequency forecasting scheme would correspond to constructing forecasts always at the end of the last month of the year (December). At the same time, with all the information collected up to the end of December, there are other possibilities to construct forecasts. In particular, the forecaster could consider placing herself at the end of any other month of the year instead and construct predictions for the monthly proxy of the yearly variable for the next hth year.

In this scheme, one often artificially *reduces* the information set. Although not all the available information is exploited, this procedure has its benefits: first, it renders high-frequency forecast instances; second, it allows taking into account misspecification due to a seasonal response of (y_t) to (z_r). This is especially important whenever multiple time series with different sampling frequencies are combined in one model and seasonality effects are either difficult to detect or impossible to avoid. In the context of macroeconomic forecasting, we again refer the reader to Clements and Galvão (2008, 2009), Chen and Ghysels (2010) and Jardet and Meunier (2022), where these questions are carefully discussed.

Let the forecaster place herself at time t: she wishes to construct a high-frequency forecast for some $t, l|\kappa$ with $l \in \mathbb{N}$. The maximal information set available at t is $\mathcal{F}_t$ as in (1.35). However, if she uses $\mathcal{F}_t$ then the forecast for $t, l|\kappa$ coincides with the low-frequency forecast and is given by (1.36) for any l. Notice that the forecasts can be constructed using the reduced information sets instead. Let $h = \lceil l/\kappa \rceil$, $\ell = l \bmod \kappa$, and $m = h - \lfloor l/\kappa \rfloor$, and define

$$\begin{aligned}\mathcal{F}_{t-m,\ell} &= \sigma\left(\left\{y_{t-m}, y_{t-1-m}, \ldots, z_{t-m,\ell|\kappa},\ z_{t-m,(\ell-1)|\kappa},\ z_{t-m,(\ell-2)|\kappa}, \ldots\right\}\right) \\ &= \sigma\left(\left\{y_{t-m}, y_{t-1-m}, \ldots, z_{t+1-m,-(\kappa-\ell)|\kappa}, z_{t+1-m,-(\kappa-\ell)+1|\kappa}, z_{t+1-m,-(\kappa-\ell+2)|\kappa}, \cdots\right\}\right).\end{aligned}$$

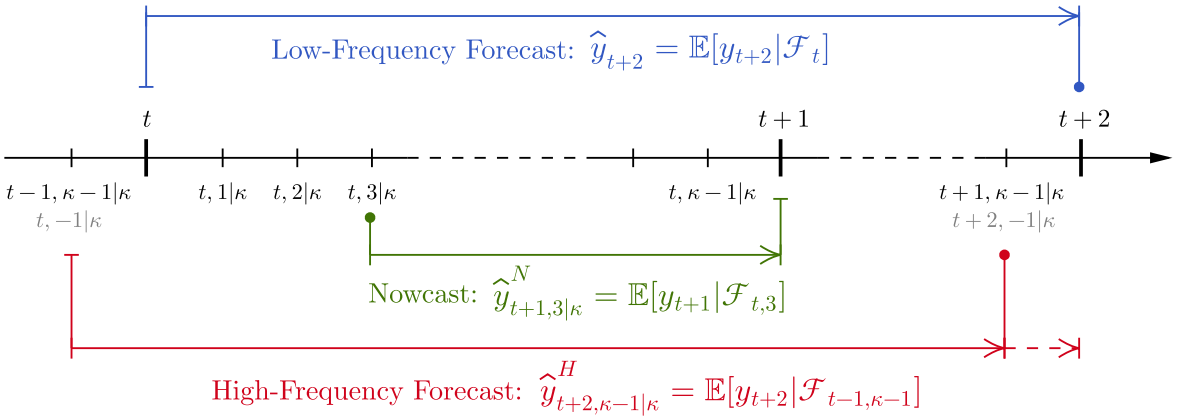


Figure 1.11: Diagram of the low-/high-frequency forecasting and nowcasting schemes in tempo notation. Arrows point to time indices of the forecast target, solid dots indicate the high-frequency time placeholder for the constructed high-frequency forecasts.

The high-frequency forecast information sets nest the low-frequency forecasting setup since $\mathcal{F}_{t-m,\ell} \equiv \mathcal{F}_t$ if $l = \kappa h$ for $h \in \mathbb{N}$ and the forecast for the high-frequency proxy constructed for the moments $t, l|\kappa$ for the low-frequency variable is provided by the conditional expectation

$$\widehat{y}^{H}_{t+h,\ell|\kappa} = \mathbb{E}\left[y_{t+h}|\mathcal{F}_{t-m,\ell}\right].$$

It is easy to see that if the forecaster is interested in nowcasting, it can be readily obtained by taking $m = 0$ and writing for all $0 < \ell \leq \kappa - 1$:

$$\widehat{y}^{N}_{t+1,\ell|\kappa} = \mathbb{E}\left[y_{t+1}|\mathcal{F}_{t,\ell}\right].$$

Nowcasting. We call *nowcasting* the setup in which one constructs a high-frequency proxy for a yet-unobserved target which will be available at the end of the *current* low-frequency period. As such, we construct a nowcast only for horizons $0 < l \leq \kappa - 1$; notice that $l = \kappa$ yields a contemporaneous regression at $t+1$, while $l = 0$ falls into the category of low-frequency forecasting considered in Section 1.B, hence both these two cases are excluded. The σ-algebras that are used in order to construct nowcasts $\widehat{y}_{t+1,\ell|\kappa}$ are given by

$$\begin{aligned}\mathcal{F}_{t,\ell|\kappa} &= \sigma\left(\left\{y_t, y_{t-1}, \ldots, z_{t,\ell|\kappa},\ z_{t,(\ell-1)|\kappa},\ z_{t,(\ell-2)|\kappa}, \ldots\right\}\right) \\ &= \sigma\left(\left\{y_t, y_{t-1}, \ldots, z_{t+1,-(\kappa-\ell)|\kappa}, z_{t+1,-(\kappa-\ell)+1|\kappa}, z_{t+1,-(\kappa-\ell+2)|\kappa}, \ldots\right\}\right).\end{aligned}$$

The l-steps nowcast for the high-frequency proxy constructed at moments $t, \ell|\kappa$ of the current period for the low-frequency variable which becomes available at $t+1, 0|\kappa \equiv t+1$ is provided by the conditional expectation

$$\widehat{y}^{N}_{t+1,\ell|\kappa} = \mathbb{E}\left[y_{t+1}|\mathcal{F}_{t,\ell}\right].$$

Multicasting. One always aims to construct one-step and multistep forecasts by using all the available information at a given point in time. It is, therefore, natural to compare models by constructing high-frequency nowcasts for the target variable to be released at the end of the current period and its high-frequency proxy forecasts for the next periods. To avoid confusion, we refer to this situation as *multicasting*. More explicitly, provided that the forecaster finds herself at time index $t, s|\kappa$ and is interested in all the forecasts up to some maximal low-frequency horizon $H \geq 1$, for each $1 \leq l \leq H\kappa$ the multicasting scheme yields the following combination:

(a) *Nowcasting* when $0 < l \leq \kappa - 1$ and $\ell = l$: $\widehat{y}^{N}_{t+1,\ell|\kappa} = \mathbb{E}\left[y_{t+1}|\mathcal{F}_{t,\ell}\right]$.

(b) Forecasting when $l > \kappa - 1$:

- *Low-frequency forecasting* if l satisfies $l \bmod \kappa = 0$: $\widehat{y}_{t+h} = \mathbb{E}\left[y_{t+h}|\mathcal{F}_{t}\right]$.
- *High-frequency forecasting* if $l \bmod \kappa \neq 0$: $\mathcal{F}_{t,\ell}$: $\widehat{y}^{H}_{t+h,\ell|\kappa} = \mathbb{E}\left[y_{t+h}|\mathcal{F}_{t,\ell}\right]$.

1.C ESN Implementation

1.C.1 Fixed, Expanding and Rolling Window Estimation

Model parameter stability is an important and well-studied question in linear time series analysis. Indeed, identifying and explaining structural breaks play a key role in macroeconomic modeling. To account for this possibility, we compare multiple estimation setups which may reflect possible changes in model parameters.

Suppose again that a sample $Y = (\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_T)^\top \in \mathbb{M}_{T-1,J}$ of targets is available, an initial state $\boldsymbol{x}_0$ is given and regressors $Z = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{T-1})^\top \in \mathbb{M}_{T-1,K}$ are observed. Additionally, the researcher has available an out-of-sample dataset, $Y^\dagger = (\boldsymbol{y}_{T+1}, \boldsymbol{y}_{T+2}, \ldots, \boldsymbol{y}_{T+S})^\top \in \mathbb{M}_{S,J}$, $Z^\dagger = (\boldsymbol{z}_T, \boldsymbol{z}_{T+1}, \ldots, \boldsymbol{z}_{T+S-1})^\top \in \mathbb{M}_{S,K}$ for $S \geq 1$. We now define the estimation setups which can be used for subsequent forecasting for $h \in \mathbb{N}^+$ steps ahead and can be adjusted for the multi-frequency setup. We consider the following estimation strategies:

(i) **Fixed parameters**: An estimator $\widehat{W}$ is computed strictly over sample observations Y and Z with some penalty λ chosen with data available up to time T. Model parameters are kept fixed when the estimated model is applied to construct out-of-sample forecasts $\widehat{\boldsymbol{y}}_{T+1}, \widehat{\boldsymbol{y}}_{T+2}, \ldots, \widehat{\boldsymbol{y}}_{T+S}$ as out-of-sample regressors $\boldsymbol{z}_T, \boldsymbol{z}_{T+1}, \ldots, \boldsymbol{z}_{T+S-1}$ are added to the information set.

(ii) **Expanding window:** For each out-of-sample time step $s = 0, \ldots, S$, define $\widehat{W}^{\mathrm{EW}}_s$ as the estimate computed by "expanding" the sample window up to time $T + s$, given by $Y^{\mathrm{EW}}_s := (\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_T, \boldsymbol{y}_{T+1}, \ldots, \boldsymbol{y}_{T+s})^\top$ and $Z^{\mathrm{EW}}_s := (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{T-1}, \boldsymbol{z}_T, \ldots, \boldsymbol{z}_{T+s-1})^\top$. Coefficients $\widehat{W}^{\mathrm{EW}}_s$ are re-estimated and penalty strength λ is re-validated over windows Y^{EW}_s, Z^{EW}_s.

(iii) **Rolling window:** In this setup the within-window sample size is kept fixed across windows – that is, the sample window "rolls" over the data – by defining $\widehat{W}^{\mathrm{RW}}_s$ as the estimate over $Y^{\mathrm{RW}}_s := (\boldsymbol{y}_{2+s}, \boldsymbol{y}_{3+s}, \ldots, \boldsymbol{y}_{T+s-1}, \boldsymbol{y}_{T+s})^\top$ and $Z^{\mathrm{RW}}_s := (\boldsymbol{z}_{1+s}, \boldsymbol{z}_{2+s}, \ldots, \boldsymbol{z}_{T+s-2}, \boldsymbol{z}_{T+s-1})^\top$ for

$s = 0, \ldots, S$. Coefficients $\widehat{W}_s^{\mathrm{RW}}$ are re-estimated and penalty strength λ is re-validated over windows Y_s^{RW}, Z_s^{RW}.

In all three strategies, hyperparameters $\boldsymbol{\varphi} := (\alpha, \rho, \gamma, \omega)$ could also be re-tuned on corresponding windows as in Appendix 1.C.2. The fixed-parameter setup is the most rigid one. It builds upon the idea that the initial sample contains sufficient information for correct model estimation and forecasting and that the model parameters are constant. Its theoretical analysis is relatively easy as there is no need to discuss the stability of the penalty and the hyperparameters across sample windows. An expanding window setup is based on the belief that newly available data contains key information to produce forecasts and, therefore, must be continuously incorporated. In essence, forecasters do this when they re-estimate a model at each data release cycle. In the case of a rolling window estimation strategy, one can theoretically handle model changes. Although proper structural break modeling would require a consistent identification of breakpoints, rolling window estimation can potentially accommodate slow drifts in model parameters over time by directly discarding old data, unlike with an expanding window. We do not explore the selection of an optimal window size, which in rolling window estimation has been shown to improve forecasting performance (Inoue et al. (2017)).

1.C.2 Hyperparameter Tuning

We now propose a general scheme for selection of hyperparameters $\boldsymbol{\varphi} := (\alpha, \rho, \gamma, \omega)$ in (1.7) for a model of the form (1.3)-(1.4). Our approach builds on the idea of leave-one-out cross-validation for time series models. Using a fixed, expanding, or rolling window over the training data, one can always compute the one-step forecasting errors committed by the ESN, given fixed normalized model matrices $(\overline{A}, \overline{C}, \overline{\boldsymbol{\zeta}})$ and a hyperparameter vector $\boldsymbol{\varphi}$. By choosing an appropriate loss function $\boldsymbol{\ell} : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}_+$, $J \in \mathbb{N}^+$, we can thus compute the empirical ESN forecasting error

$$\mathcal{L}_T(\boldsymbol{\varphi}) := \sum_{t=T_0}^{T-1} \boldsymbol{\ell}(\boldsymbol{y}_{t+1}, \widehat{W}_t(\boldsymbol{\varphi})^\top \boldsymbol{x}_t),$$

where $\widehat{W}_t(\boldsymbol{\varphi})$ is the readout coefficients estimator involving data available up to time t and $1 < T_0 < T - 1$ is the minimum number of observations used for fitting. Notice that if $\boldsymbol{\ell}(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|_2^2$, $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^J$, then $\mathcal{L}_T(\boldsymbol{\varphi})$ is the cumulative squared error that is minimized in training (modulo a ridge penalty term). Here, however, the interest is not in estimating W, which minimizes $\mathcal{L}_T$, but rather finding the optimal hyperparameter vector

$$\boldsymbol{\varphi}^* \in \underset{\boldsymbol{\varphi} \in [0,1) \times [0,\overline{\rho}] \times [0,\overline{\gamma}] \times [0,\overline{\omega}]}{\arg\min} \mathcal{L}_T(\boldsymbol{\varphi}),$$

where upper bounds $\overline{\rho}$, $\overline{\gamma}$, and $\overline{\omega}$ can be appropriately chosen (in our empirical exercises we use 10 and verify that solutions are never on the boundary). We highlight that to tune $\boldsymbol{\varphi}$ one may choose $\boldsymbol{\ell}$ that is different from the one used in the estimation of the readout coefficients W.

We present the entire hyperparameter optimization routine in Algorithm 1. Note that step (i) might entail re-normalizing inputs and targets at each window t. This setup is general and allows applying any global optimization routine to minimize $\mathcal{L}_T(\boldsymbol{\varphi})$. We construct the loss $\mathcal{L}_T(\boldsymbol{\varphi}_j)$

Algorithm 1: Hyperparameter tuning

Data: Sample $\boldsymbol{y}_{2:T} = \{\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_T\}$, $\boldsymbol{z}_{1:T-1} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{T-1}\}$, initial state $\boldsymbol{x}_0$, initial guess $\boldsymbol{\varphi}_0$, convergence criterion Crit, maximal algorithm iterations MaxIter. If ridge regression is used to estimate W, fixed regularization strength $\lambda > 0$.

Result: $\boldsymbol{\varphi}^*$

Fix T and determine the model fit windows for $t = T_0, \ldots, T-1$. Choose whether the ESN model is estimated with a fixed or rolling window;

$j = 0$;

while (***not*** Crit) ***and*** ($j <$ MaxIter) **do**

(i) Given $\boldsymbol{\varphi}_j$, estimate coefficient matrices

$$\left(\widehat{W}_t(\boldsymbol{\varphi}_j)\right)_{T_0:T-1},$$

where possibly $\widehat{W}_t(\boldsymbol{\varphi}_j)$ does not depend on t, e.g. in the fixed estimation setup;

(ii) Compute

$$\mathcal{L}_T(\boldsymbol{\varphi}_j) := \sum_{t=T_0}^{T-1} \boldsymbol{\ell}(\boldsymbol{y}_{t+1}, \widehat{W}_t(\boldsymbol{\varphi}_j)^\top \boldsymbol{x}_t),$$

the cumulative one-step-ahead forecasting loss;

(iii) Update $\boldsymbol{\varphi}_{j+1} \leftarrow \boldsymbol{\varphi}_j$ with an appropriate rule (for example, the gradient descent of $\mathcal{L}_T$ in the direction of $\boldsymbol{\varphi}_j$; in our applications, we use variants L-BFGS-B and pattern search);

(iv) $j \leftarrow j + 1$, update Crit;

sequentially, that is by summing squared residuals of the model estimated in step (i) of Algorithm 1 when $\boldsymbol{\ell}$ is a quadratic loss. One can program $\mathcal{L}_T(\boldsymbol{\varphi}_j)$ via TensorFlow so that the gradient can be evaluated by backpropagation in Algorithm 1 (iii). Since there is no guarantee that the objective function is convex or even everywhere smooth, we suggest applying optimizers known to explore the parameter space efficiently. We emphasize that the lack of convexity guarantees is much more consequential for the other benchmarks, in particular for the MIDAS model (see Appendix 1.I.1 for more details).

One issue with the state formulation in (1.3) and thus with the hyperparameter optimization routine in Algorithm 1, is that $\boldsymbol{\varphi}$ can not always be point identified. For example, if one considers identity activation σ and lets $\alpha = \omega = 0$, it is obvious that the ESN model is system isomorphic Grigoryeva and Ortega (2021) to $\boldsymbol{x}_t^* = d\rho\overline{A}\boldsymbol{x}_{t-1}^* + d\gamma\overline{C}\boldsymbol{z}_t$, $\boldsymbol{y}_t = d^{-1}W\boldsymbol{x}_t^* + \boldsymbol{\epsilon}_t$ for all $d \neq 0$. This issue also arises in nonlinear models, for example when σ is taken as a hyperbolic tangent and γ is sufficiently small. Parameter identification in nonlinear models has been extensively studied in semi- and nonparametric cross-sectional regressions. For instance, it is known that in certain setups, point identification requires a proper normalization to be imposed. The interested reader can refer to Section 6.3 of Horowitz (2009) for a discussion in a similar vein regarding nonparametric transformation models. Since often $\omega = 0$ is used, hyperparameter identification can be a significant issue when attempting model tuning. Whenever $\omega = 0$ we propose a helpful reparametrization given by

$$\boldsymbol{x}_t = \alpha\boldsymbol{x}_{t-1} + (1-\alpha)\sigma\left(\psi\overline{A}\boldsymbol{x}_{t-1} + \overline{C}\boldsymbol{z}_t\right),$$

where $\psi = \rho/\gamma$. This prescription allows decoupling ρ and γ at the cost of the constant input scaling, which may be undesirable whenever one wants to attenuate the nonlinearity induced by the sigmoid map without also reducing the spectral radius.[6] It is immediate to modify the optimization scheme to deal with the case $\widetilde{\boldsymbol{\varphi}} = (\alpha, \psi)$. In the sequel, we assume that the ESN models are estimated using the approaches proposed in this subsection and use the conventional ESN specification as in (1.3)-(1.4) to discuss the forecasting strategy.

1.C.3 Cross-validation

Because the initial cross-validation of λ uses an extended sample to try and improve generalization – specifically, our concern is for the fixed estimation setups – we use two slightly different approaches:

- In all setups – fixed, expanding, rolling – the *initial* ridge penalty cross-validation is done on the extended sample (starting January 1st, 1975 instead of January 1st, 1990). We construct 10 folds with 5 out-of-sample observations starting from the end of the sample. Each fold and out-of-sample observation set is re-normalized.

- Only in the expanding and rolling setups, for each subsequent window (the ones that now include at least one testing observation), we use the true sample (starting January 1st, 1990) and construct 5 folds, again with 5 out-of-sample observations. This is done to keep cross-validation computational complexity low and avoid making some folds too small, which could hurt larger MFESN models.

In practice, simple experiments show that there is not much difference between using 5 or 10 folds in the initial cross-validation.

1.D Performance measures

In this section we define the performance measures used throughout the paper to quantify the quality of forecasts produced by competing models. Suppose that a given model is used to produce a collection of forecasts $\{\widehat{\boldsymbol{y}}_s\}_{s\in S}$, $\widehat{\boldsymbol{y}}_s \in \mathbb{R}^J$. The ordered index set $S = \{s_1, \ldots, s_{|S|}\}$, where $|S|$ is the number of indices in S, can change depending on the setup. For example, in the case of 1-step ahead forecasting, $S = \{T+1, T+2, \ldots, T+\overline{T}\}$ where the 1-step ahead forecasts are constructed using the data up to $T, T+1, \ldots, T+\overline{T}-1$, respectively. For h-step ahead forecasts, we set $S = \{T+h, T+h+1, \ldots, T+\overline{T}-H+h\}$ where H is the maximal forecasting horizon. This ensures that the same number of forecasts are produced at each horizon and can be compared, for example, using the uniform Model Confidence Set (MCS) test described in Appendix 1.E.

MSFE and RMSFE. The *root mean squared forecasting error* is given by

$$\text{MSFE}(S) := \frac{1}{|S|}\sum_{s\in S} \|\boldsymbol{y}_s - \widehat{\boldsymbol{y}}_s\|_2^2,$$

[6] One can fix $\overline{C}$ to have a different scaling before optimizing the hyperparameter ψ. However, this amounts to one more *ex ante* model tuning step.

while the *root mean squared forecasting error* is

$$\text{RMSFE}(S) := \sqrt{\text{MSFE}(S)}\,.$$

Cumulative SFE and Cumulative RMSFE. The *cumulative squared forecasting error* is given by the cumulative sum of squared errors. We define for any forecasting index $\tau \in S$,

$$\text{CSFE}(\tau) := \sum_{\substack{s \in S \\ s \leq \tau}} \|\boldsymbol{y}_s - \widehat{\boldsymbol{y}}_s\|_2^2.$$

To define the *cumulative RMSFE* for any $\tau \in S$ we first define $\mathcal{T}_l(\tau) := \{s \in S : s \leq \tau\}$ and then write

$$\text{CRMSFE}(\tau) := \sqrt{\frac{1}{|\mathcal{T}_l(\tau)|} \sum_{s \in \mathcal{T}_l(\tau)} \|\boldsymbol{y}_s - \widehat{\boldsymbol{y}}_s\|_2^2}.$$

Ahead RMSFE and 1-Year-Ahead RMSFE. If one wants to the evaluate performance *ahead* of a certain point of time, it is also possible to define the *ahead RMSFE*,

$$\text{AheadRMSFE}(\tau) := \sqrt{\frac{1}{|\mathcal{T}_u(\tau)|} \sum_{s \in \mathcal{T}_u(\tau)} \|\boldsymbol{y}_s - \widehat{\boldsymbol{y}}_s\|_2^2},$$

where we introduce $\mathcal{T}_u(\tau) := \{s \in S : s \geq \tau\}$.

In the special case where the indices of S are associated to dates, one may also compare performance after a given amount of time has passed from the current time index. For example, one may evaluate how performance degrades after model estimation when parameters are fixed and not updated. In our empirical exercises where $\{y_t\}_{t \in \mathbb{Z}}$, $y_t \in \mathbb{R}$, is a quarterly GDP series, we can define the *1-year-ahead RMSFE* as

$$\text{1YAheadRMSFE}(\tau) := \sqrt{\frac{1}{|\mathcal{T}_u(\tau+4)|} \sum_{s \in \mathcal{T}_u(\tau+4)} (y_s - \widehat{y}_s)^2}.$$

1.E Uniform Multi-Horizon MCS

We now give details on the implementation of the Uniform Multi-Horizon MCS test for the multi-horizon forecast comparisons in our empirical analysis. Our procedure follows closely the one originally provided by Quaedvlieg (2021): we provide R code for our functions, while the author's code was originally developed in the Ox programming language.

A main difference is that we prefer to use a Bartlett kernel to compute the sample uSPA statistic, whereas Quaedvlieg (2021) uses the quadratic spectral (QS) kernel of Andrews (1991). Our main reason for this choice is that the QS kernel features non-zero weights for all lags, while the Bartlett kernel has finite support. This is especially important since we only have a few forecasts in our case; thus, higher lag autocovariances between model losses can only be poorly estimated. It means our uMCS procedure implements the standard Newey-West HAC estimator. We use $B = 100$ replications for the outer and inner bootstraps. Finally, the inner bootstrap critical value

is set at $\alpha = 0.1$.

1.F MIDAS

A state-of-the-art methodology for incorporating data of heterogeneous frequencies into one model is the MIDAS framework developed in Ghysels et al. (2004, 2007). Here we present MIDAS in its dynamic form, which allows the inclusion of target series autoregressive lags. We use our temporal notation given in Definition 1.1.1 throughout.

If the MIDAS model contains only one explanatory variable (z_r) with frequency multiplier κ, then it can be written as

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \beta \sum_{k=0}^{K} \varphi(\boldsymbol{\theta}, k) z_{t,-k|\kappa} + \epsilon_t, \tag{1.37}$$

where α_0 is a constant term, $\{\alpha_i\}_{i=1}^{p}$ are the autoregressive parameters, β is a scaling parameter, $\{\varphi(\boldsymbol{\theta}, k)\}_{k=0}^{K}$ are the MIDAS weights given as a parametric function of lag k and underlying parameter vector $\boldsymbol{\theta} \in \mathbb{R}^q$, and (ϵ_t) is a martingale difference process relative to the filtration $\{\mathcal{F}_t\}$ generated by $\{y_{t-1-j}, x_{t-j}, \ldots, x_{t-j,-K|\kappa}, \epsilon_{t-1-j} \mid j \geq 0\}$ and such that $\mathbb{E}[\epsilon_t^2] = \sigma_\epsilon^2 < \infty$.

The MIDAS weighting scheme is the core innovation of the model. It borrows parsimony from distributed lag models in the sense that, even if K is large, the vector $\boldsymbol{\theta} \in \mathbb{R}^q$ is usually restricted to contain only a handful of parameters. This greatly reduces the number of coefficients that need to be estimated, and a nonlinear least-squares estimator $\widehat{\boldsymbol{\theta}}$ can be readily implemented. There are alternative formulations of the MIDAS framework where $\varphi(\boldsymbol{\theta}, k) = \theta_k$ so that the above reduces to a full linear model, the so-called unrestricted MIDAS or U-MIDAS (Foroni and Marcellino, 2011).

We follow the literature and use the most commonly applied weighting scheme that is based on the exponential Almon weighting polynomial map $\varphi : \mathbb{R}^q \times \mathbb{N}^+ \longrightarrow \mathbb{R}^+$ (see Almon (1965) for more details). In particular, for the case of $q = 2$, the two-parameter Almon weighting polynomial is given by

$$\varphi(\boldsymbol{\theta}, k) = \varphi((\theta_1, \theta_2), k) = \exp(\theta_1 k + \theta_2 k^2), \;\; k \in \mathbb{N}^+.$$

Since Almon weights need not sum up to a given constant for different values of θ_1 and θ_2, it is often common to consider the normalized Almon scheme

$$\overline{\varphi}(\boldsymbol{\theta}, k) = \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{k=0}^{K} \exp(\theta_1 k + \theta_2 k^2)}, \tag{1.38}$$

which together with (1.37) allows to treat β as a rescaling constant.

Let us now consider a more general model suitable for situations where time series of different frequencies are available and must be integrated into the MIDAS equation. Consider the case of L regressor time series. We assume that the lth time series is sampled at a frequency κ_l and contains observations $(z^{(l)}_{t,s|\kappa_l})_{t,s}$ with $z^{(l)}_{t,s|\kappa_l} \in \mathbb{R}$ for all $t \in \mathbb{Z}$ and $s \in \{0, \ldots, \kappa_l - 1\}$. It happens frequently in practice that κ_l, $l \in [L]$ takes values from a small set of integers. For example, in the case of yearly, quarterly, and monthly data $\kappa_l \in \{1, 4, 12\}$ even though L could be very large (often, hundreds or thousands of series might be of interest). The MIDAS model explaining low-frequency target

variable y_t with L regressors $(z^{(l)}_{t,s|\kappa_l})_{t,s}$, $l \in [L]$ can be written as follows

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{l=1}^{L} \beta_l \sum_{k=0}^{K_l} \varphi(\boldsymbol{\theta}_l, k) z^{(l)}_{t,-k|\kappa_l} + \epsilon_t, \tag{1.39}$$

where the martingale difference process (ϵ_t) is relative to the filtration generated by sets as in (1.37), modified to include all the considered L regressors.

The MIDAS framework produces forecasts of the chosen target variable at the low frequency of the target. Yet, due to the MIDAS multi-frequency structure, *nowcasting* is also a straightforward exercise: if, for example, the high-frequency regressor is a single series (z_r) with frequency multiplier κ, one can construct exactly κ regression equations – one for each high-frequency release within a low-frequency period – and use these to produce high-frequency nowcasts of the target. In fact, due to the convenience of the MIDAS model, it is easy to define high-frequency regression specifications to study high-frequency forecasts and multicasts (see Section 1.B and Appendix 1.B).

In practice, implementing (1.39) demands some care. From a computational point of view, as long as the relevant regression matrices can be constructed, estimation amounts to a nonlinear least-squares problem, which can be readily solved. In Appendix 1.F.1 and Appendix 1.I.1 we discuss the technical aspects of our MIDAS implementation in more detail. One of the important issues of the MIDAS estimation is the non-convexity of the nonlinear least squares loss as a function of parameters. Often, a practitioner may obtain different estimation results depending on initialization and, more importantly, those that lead to a different quality of forecasts. Other weighting schemes that allow for convex estimation problems can be used. For example, one may adopt the Almon lag polynomial parametrization (Ghysels, 2016, Pettenuzzo et al., 2016) using a discrete polynomial basis for the transformation of high-frequency regressors. This specification allows for standard OLS estimation but requires careful choice of the polynomial order hyperparameter.

Another crucial disadvantage of the MIDAS specification is that practical implementations can be very challenging. This is caused mainly by the ragged edges of the "raw" macroeconomic data, incomplete observations, and uneven sampling frequencies. The relative inflexibility of MIDAS regression lag specifications makes integrating daily and weekly data at true calendar frequencies (i.e. without interpolation or aggregation) very complex.[7] State-space models effectively mitigate these issues.

Finally, as shown in Bai et al. (2013), exponential Almon MIDAS regressions have inherent connections to dynamic factor models, which we discuss in the next section. When the factor structure is not trivial, MIDAS can, however, only yield a finite-order approximation to a DFM data-generating process. Furthermore, Bai et al. (2013) prove that in well-identified setups the mapping between exponential Almon and factor model coefficients is highly nonlinear. Given the robustness evaluations in Appendix 1.I.1, in practice, it appears hard to formally relate MIDAS and DFM forecasting performance.

[7] One could set up a MIDAS regression with the full yearly calendar of weeks and working days as lags. However, ragged edges arising from holidays, leap years, etc. would still be non-trivial to handle coherently without resorting to downsampling, data re-alignment, or interpolation.

1.F.1 MIDAS Implementation

While the MIDAS regression framework is straightforward to discuss in terms of equations, some care must be taken when implementing it computationally. A key assumption that can be imposed is that the integer frequencies $\boldsymbol{\kappa} := \{\kappa_1, \ldots, \kappa_L\}$ of L regressors are such that $\kappa_{\max} := \max(\boldsymbol{\kappa})$ is a multiple of each of the κ_l, $l \in [L]$. In this case, MIDAS parameter estimation can be written in matrix form, which allows for efficient numerical implementation, which we spell out in the following paragraphs.

Let $q_l = \kappa_{\max}/\kappa_l$, $l \in [L]$ denote the frequency ratios and define $\boldsymbol{y} := (y_1, y_2, \ldots, y_T)^\top$ the vector of target observations, where T is the sample length in reference time scale. Additionally, let $\boldsymbol{z}^{(l)} := (z_1^{(l)}, z_2^{(l)}, \ldots, z_{T_l}^{(l)})^\top$ be $T_l = T \cdot \kappa_l$ long vector which consists of observations of the lth covariate $z^{(l)}$ released with frequency κ_l. For the parameters of the MIDAS model in (1.39) to be identifiable, we assume that

$$T > 1 + p + \sum_{l=1}^{L} \left\lceil \frac{K_l}{\kappa_l} \right\rceil .$$

Since $\kappa_{\max}$ is a multiple of each of the L frequencies, for each series we introduce

$$\mathbf{Y} = \boldsymbol{y} \otimes \boldsymbol{i}_{\kappa_{\max}}, \quad \mathbf{Z}^{(l)} = \boldsymbol{z}^{(l)} \otimes \boldsymbol{i}_{q_l},$$

where $\boldsymbol{i}_{q_l}$ and $\boldsymbol{i}_{\kappa_{\max}}$ are vectors of ones of lengths q_l and $\kappa_{\max}$, respectively. In the absence of missing observations, we have that $\mathbf{Y}, \mathbf{Z}^{(l)} \in \mathbb{R}^{T_{\max}}$ with $T_{\max} = T \cdot \kappa_{\max}$ observations. We now construct preliminary regression matrices such that their maximal rows number is $T_{\max}$ without accounting for the lags structure of both the target (autoregressive lags) and regressors (MIDAS lags) and we introduce zeros where no observations are available.[8] Define for $p \geq 1$ and for $K_l \geq 0$

$$Y_p = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ y_1 & 0 & \cdots & 0 \\ y_2 & y_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ y_{T-2} & y_{T-3} & \cdots & y_{T-p-1} \\ y_{T-1} & y_{T-2} & \cdots & y_{T-p} \end{pmatrix} \otimes \boldsymbol{i}_{\kappa_{\max}} \quad \text{and} \quad Z_{K_l} = \begin{pmatrix} z_1^{(l)} & 0 & \cdots & 0 \\ z_2^{(l)} & z_1^{(l)} & \cdots & 0 \\ z_3^{(l)} & z_2^{(l)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ z_{T_l-1}^{(l)} & z_{T_l-2}^{(l)} & \cdots & z_{T_l-K_l-1}^{(l)} \\ z_{T_l}^{(l)} & z_{T_l-1}^{(l)} & \cdots & z_{T_l-K_l}^{(l)} \end{pmatrix} \otimes \boldsymbol{i}_{q_l}.$$

In the special case $p = 0$ (the MIDAS model, in this case, is called *static*, since it does not contain an autoregressive term) we take Y_p as empty. We now follow by noticing that one should not use $Y_p \in \mathbb{M}_{T_{\max},p}$ and $Z_{K_l} \in \mathbb{M}_{T_{\max},K_l+1}$ as autoregressive and mixed-frequency regression matrices, respectively, since some observations are missing. To overcome this we introduce

$$n := \max\left\{p, \left\lceil \frac{K_1}{q_1} \right\rceil, \ldots, \left\lceil \frac{K_L}{q_L} \right\rceil\right\} \cdot \kappa_{\max} \tag{1.40}$$

and the so-called upper truncation (selection) matrix

$$U = \begin{pmatrix} \mathbb{O}_{T_{\max}-n+1,n-1} & \mathbb{I}_{T_{\max}-n+1,T_{\max}-n+1} \end{pmatrix}$$

[8] At the time of implementation of this procedure in any convenient coding environment it is more natural to introduce placeholders instead and to perform the subsequently discussed truncation via matrix manipulation rather than by using matrix multiplication.

with which we obtain the following required response vector and regression matrices

$$\begin{aligned}
\mathbf{Y}^{\text{resp}} &= U\mathbf{Y} \in \mathbb{R}^{T_{\max}-n+1}, \\
Y_p^{\text{reg}} &= UY_p \in \mathbb{M}_{T_{\max}-n+1,p}, \\
Z_{K_l}^{\text{reg}} &= UZ_{K_l} \in \mathbb{M}_{T_{\max}-n+1,K_l+1}, \\
\mathbf{Z}^{\text{reg}} &= \begin{pmatrix} Y_p^{\text{reg}} & Z_{K_1}^{\text{reg}} & \cdots & Z_{K_L}^{\text{reg}} \end{pmatrix} \in \mathbb{M}_{T_{\max}-n+1,\sum_{l=1}^{L} K_l+L},
\end{aligned}$$

where Y_p^{reg} is empty whenever $p = 0$.

We can now observe that $\mathbf{Y}^{\text{resp}}$ and $\mathbf{Z}^{\text{reg}}$ are sufficient to construct all MIDAS forecasting and nowcasting regressions. In practice, some care needs to be taken to make sure that data is correctly aligned: for example, in the case of nowcasting exercise regressors in $\mathbf{Z}^{\text{reg}}$ and targets in $\mathbf{Y}^{\text{resp}}$ have to be aligned differently than in the case of forecasting exercises. Provided the aligned data is executed correctly, the estimation of MIDAS parameters can be carried out efficiently. An important thing to mention is that the truncation with the help of s in (1.40) may be too restrictive, as it may lead to excluding up to $K_{\max} - 1$ rows from $\mathbf{Z}^{\text{reg}}$ that could be used for estimation. This can be avoided at the time of implementation. In our repository available at [the address removed for anonymous submission] we consider this detail and exclude from the final regression matrices only those rows which cannot be used due to the lag requirements in the model. We warn the reader that this comes at a cost, namely the codes are lengthier and less elegant.

1.G Mixed-frequency DFM

Macroeconomic modeling based on dynamic factor models has been popular since their introduction in Geweke (1977) and Sargent et al. (1977). The proposition of DFMs is that a low-dimensional latent factor $(\boldsymbol{f}_t)_{t\in\mathbb{Z}}$, $\boldsymbol{f}_t \in \mathbb{R}^d$, drives a high-dimensional observable stochastic process $(\boldsymbol{y}_t)_{t\in\mathbb{Z}}$, $\boldsymbol{y}_t \in \mathbb{R}^n$. We consider a time-inhomogeneous state-space model with dynamics

$$\boldsymbol{f}_{t+1}|\boldsymbol{f}_{1:t}, \boldsymbol{y}_{1:t} \sim h_{t+1,\boldsymbol{\theta}}(\cdot|\boldsymbol{f}_t) \tag{1.41}$$

$$\boldsymbol{y}_{t+1}|\boldsymbol{f}_{1:t+1}, \boldsymbol{y}_{0:t} \sim g_{t+1,\boldsymbol{\theta}}(\cdot|\boldsymbol{f}_{t+1}) \tag{1.42}$$

for some time-dependent state transition kernels $h_{t,\boldsymbol{\theta}}$ and observation densities $g_{t,\boldsymbol{\theta}}$ and some parameter vector $\boldsymbol{\theta}$ in a parameter space Θ. A common example in the literature (see Watson and Engle (1983) for more details) is linear Gaussian factor models with time-inhomogeneous state transitions that can be represented as

$$\boldsymbol{f}_{t+1} = A_{\boldsymbol{\theta}}\boldsymbol{f}_t + R_{\boldsymbol{\theta}}\boldsymbol{u}_t \tag{1.43}$$

$$\boldsymbol{y}_{t+1} = \Lambda_{t+1,\boldsymbol{\theta}}\boldsymbol{f}_{t+1} + S_{t+1,\boldsymbol{\theta}}\boldsymbol{w}_{t+1} \tag{1.44}$$

with state transition matrix $A_{\boldsymbol{\theta}} \in \mathbb{M}_d$, time-dependent factor loading matrices $\Lambda_t \in \mathbb{M}_{n,d}$, and where $\boldsymbol{u}_t$ and $\boldsymbol{w}_t$ are independent Gaussian vectors with zero mean and identity covariance matrix of dimension p and n, respectively, and $R_{\boldsymbol{\theta}}$ and $S_{t,\boldsymbol{\theta}}$ re matrices of appropriate dimensions. It is often assumed that the dimension p of the state noise vector $\boldsymbol{u}_t$ is smaller than the latent state

space dimension d, which implies that $R_{\boldsymbol{\theta}}R_{\boldsymbol{\theta}}^\top$ is rank deficient, such as for AR(p) factor dynamics (Stock and Watson, 2016, Forni et al., 2005, Doz et al., 2011). In this case, $d = kp$ for some $k \in \mathbb{N}^+$,

$$A_{\boldsymbol{\theta}} = \begin{pmatrix} A_{\boldsymbol{\theta}}^{(1)} & A_{\boldsymbol{\theta}}^{(2)} & \cdots & A_{\boldsymbol{\theta}}^{(p-1)} & A_{\boldsymbol{\theta}}^{(p)} \\ \mathbb{I}_k & \mathbb{O}_k & \cdots & \mathbb{O}_k & \mathbb{O}_k \\ \mathbb{O}_k & \mathbb{I}_k & \cdots & \mathbb{O}_k & \mathbb{O}_k \\ \vdots & & \ddots & & \vdots \\ \mathbb{O}_k & \mathbb{O}_k & \cdots & \mathbb{I}_k & \mathbb{O}_k \end{pmatrix}, \quad \Lambda_{t,\boldsymbol{\theta}} = \begin{pmatrix} \Lambda_{t,\boldsymbol{\theta}}^{(1)} & \Lambda_{t,\boldsymbol{\theta}}^{(2)} & \cdots & \Lambda_{t,\boldsymbol{\theta}}^{(p)} \end{pmatrix} \tag{1.45}$$

with $A_{\boldsymbol{\theta}}^{(j)} \in \mathbb{M}_k$ and $\Lambda_{t,\boldsymbol{\theta}}^{(j)} \in \mathbb{M}_{n,k}$. Setting $\boldsymbol{f}_t = (\boldsymbol{v}_t^\top, \boldsymbol{v}_{t-1}^\top, \ldots, \boldsymbol{v}_{t-p+1}^\top)^\top$ implies that $(\boldsymbol{v}_t)_{t\in\mathbb{Z}}$ is a k-dimensional AR(p) process and it is commonly assumed that $\Lambda_{t,\boldsymbol{\theta}}^{(j)} = \mathbb{O}_{n,k}$ for $j > 1$. Let the initial state $\boldsymbol{f}_0$ be distributed according to ν. The joint density of the latent path $\boldsymbol{f}_{0:T}$ and observations $\boldsymbol{y}_{0:T}$ is then

$$p_{\boldsymbol{\theta},\nu}(\boldsymbol{f}_{0:T}, \boldsymbol{y}_{0:T}) = \nu(\boldsymbol{f}_0) g_{0,\boldsymbol{\theta}}(\boldsymbol{y}_0|\boldsymbol{f}_0) \prod_{t=1}^{T} h_{t,\boldsymbol{\theta}}(\boldsymbol{f}_t|\boldsymbol{f}_{t-1}) g_{t,\boldsymbol{\theta}}(\boldsymbol{y}_t|\boldsymbol{f}_t),$$

while the marginal likelihood of $\boldsymbol{y}_{0:T}$ is $p_{\boldsymbol{\theta},\nu}(\boldsymbol{y}_{0:T}) = \int p_{\boldsymbol{\theta},\nu}(\boldsymbol{f}_{0:T}, \boldsymbol{y}_{0:T}) \mathrm{d}\boldsymbol{f}_{0:T}$. Popular procedures for learning the static parameters $\boldsymbol{\theta} \in \Theta$ are based on gradient descent of the negative log-likelihood function $\ell_T \colon \Theta \to \mathbb{R}, \boldsymbol{\theta} \mapsto -\log p_{\boldsymbol{\theta},\nu}(\boldsymbol{y}_{0:T})$ or on the Expectation Maximization (EM) algorithm introduced in Dempster et al. (1977). We consider here gradient descent algorithmsbased on a sequence of step sizes $\gamma_k > 0$, that update the model parameters based on iterations of the form

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \gamma_{k+1} \nabla_{\boldsymbol{\theta}} \ell_T(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k},$$

for $k \in \mathbb{N}^+$.[9]

Assuming a linear Gaussian setting where the transition density of the latent factor process is given by (1.45) to yield an AR(p) process $(\boldsymbol{v}_t)_{t\in\mathbb{Z}}$, $\boldsymbol{v}_t = (v_{1,t}, \ldots, v_{k,t})^\top$, there remains some flexibility as to how the linear mappings

$$\mathsf{Agg}_{\boldsymbol{\theta},L} \colon \mathbb{M}_{k,p} \to \mathbb{R}, \quad (\boldsymbol{v}_{t-p+1} \ldots, \boldsymbol{v}_t) \mapsto (\Lambda_{t,\boldsymbol{\theta}} \boldsymbol{f}_t)_i = (\Lambda_{t,\boldsymbol{\theta}})_{i,\cdot} \begin{bmatrix} \boldsymbol{v}_t \\ \boldsymbol{v}_{t-1} \\ \vdots \\ \boldsymbol{v}_{t-p+1} \end{bmatrix}$$

for some lag parameter $L \leq p$ are chosen for each dimension $i \in [n]$.[10] We call this linear mapping $\mathsf{Agg}_{\boldsymbol{\theta},L}$ an *aggregation function* and consider specific examples below that yield different models for the factor loadings matrices $\Lambda_{t,\boldsymbol{\theta}}$. Notice that our aggregation functions are linear with respect

[9] Consistency of the maximum log-likelihood estimate for the dynamics (1.43)-(1.44) in the time-homogeneous case has been established for instance in Douc et al. (2011) under regularity assumptions, including, for instance, the full-rank of the noise covariance matrix $S_{\boldsymbol{\theta}}$, of the controllability matrix $C_{\boldsymbol{\theta}} = \big(R_{\boldsymbol{\theta}}|A_{\boldsymbol{\theta}}R_{\boldsymbol{\theta}}|\cdots|A_{\boldsymbol{\theta}}^{d-1}R_{\boldsymbol{\theta}}\big)$, and of the observability matrix $O_{\boldsymbol{\theta}} = \big(\Lambda_{\boldsymbol{\theta}}^\top|(\Lambda_{\boldsymbol{\theta}}A_{\boldsymbol{\theta}})^\top|\cdots|(\Lambda_{\boldsymbol{\theta}}A_{\boldsymbol{\theta}}^{d-1})^\top\big)^\top$. It is also possible to consider an online learning setting using a recursive decomposition of the score function as in LeGland and Mevel (1997). For general latent state dynamics (1.41) and observation densities (1.42) that can be non-linear with non-Gaussian noise, particle filtering algorithms are often utilized that make use of particle approximations in gradient-descent or EM learning approaches, see for instance Kantas et al. (2015).

[10] The Markovian representation (1.41)-(1.42), that is, the companion form, is based on the autoregressive order p, however, one can set $A_{\boldsymbol{\theta}}^{(\ell)} = \mathbb{O}_k$ for $\ell > p$.

to the latent factors in contrast to the non-linear approaches introduced in Proietti and Moauro (2006) that require approximations, such as resorting to extended Kalman filtering techniques.

Example 1.G.1 (Stock aggregation)**.** For $i \in [n]$, let $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ik}) \in \mathbb{R}^k$ and consider

$$\mathsf{Agg}^{\mathrm{S}}_{\boldsymbol{\theta},1}(\boldsymbol{v}_{t-p+1} \dots, \boldsymbol{v}_t)_i = \sum_{m=1}^{k} \beta_{im} v_{m,t},$$

with $\boldsymbol{\theta} = \boldsymbol{\beta}_i$.

Example 1.G.2 (Almon-Lag aggregation)**.** For $i \in [n]$, let $\boldsymbol{\beta}_i \in \mathbb{R}^k$, $\boldsymbol{\psi}_i \in \mathbb{R}^{2k}$ and consider

$$\mathsf{Agg}^{\mathrm{AL}}_{\boldsymbol{\theta},L}(\boldsymbol{v}_{t-p+1} \dots, \boldsymbol{v}_t)_i = \sum_{m=1}^{k} \beta_{im} \sum_{\ell=0}^{L-1} \overline{\varphi}(\psi_{im}, \ell) v_{m,t-\ell},$$

with $\boldsymbol{\theta} = (\boldsymbol{\beta}_i, \boldsymbol{\psi}_i, \boldsymbol{\beta}_i, \boldsymbol{\psi}_i)$ and Almon-Lag weights $\overline{\varphi}$ given in (1.38).

Example 1.G.3 (Trigonometric aggregation)**.** For $i \in [n]$, let $\boldsymbol{\beta}_i \in \mathbb{R}^k$, and for $K \in \mathbb{N}$, let $\boldsymbol{\lambda} \in \mathbb{R}^K_+$, $\boldsymbol{\omega} \in [0,1]^K$, $\boldsymbol{\gamma} \in [-\pi, \pi]^K$ and $\tau \in \mathbb{R}_+$. Define

$$\mathsf{Agg}^{\sin}_{\boldsymbol{\theta},L}(\boldsymbol{v}_{t-p+1} \dots, \boldsymbol{v}_t)_i = \sum_{m=1}^{k} \beta_{im} \sum_{\ell=0}^{L-1} \overline{a}_p(\boldsymbol{\lambda}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \tau, \ell) v_{m,t-\ell},$$

with $\boldsymbol{\theta} = (\boldsymbol{\beta}_i, \boldsymbol{\lambda}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \tau)$ and

$$\overline{a}_p(\boldsymbol{\lambda}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \tau, \ell) = \frac{\exp\left(\frac{1}{\tau}\sum_{j=1}^{K} \lambda_j^2 \cos(2\pi\omega_j \ell + \gamma_j)\right)}{\sum_{\ell'=0}^{p-1} \exp\left(\frac{1}{\tau}\sum_{j=1}^{K} \lambda_j^2 \cos(2\pi\omega_j \ell' + \gamma_j)\right)}.$$

This aggregation scheme is motivated by self-attention models (we refer the reader to Bahdanau et al. (2014), Vaswani et al. (2017) for more details), but to retain linearity only considers a relative positional encoding with a Toeplitz structure. Observe that the aggregation parameters are shared across all n dimensions in contrast to the Almon lag scheme in Example 1.G.2.

Some authors (see for example Mariano and Murasawa (2003), Bańbura and Modugno (2014)) have imposed different restrictions on the form of the factor loadings matrices or aggregation function, particularly for one-dimensional mixed-frequency factor models of quarterly GDP growth rates and monthly covariates, which are motivated by approximations of growth rates. We do not pursue this additional restriction in this work.

Kalman filtering techniques have been used routinely for handling missing observations in multi-frequency DFMs, see Harvey et al. (1998). In this work, we leverage modern auto-differentiation libraries (Abadi et al., 2016, Dillon et al., 2017) to compute the gradient of the log-likelihood based on Kalman filtering formulae and estimate the static parameters $\boldsymbol{\theta}$ by gradient ascent of the log-likelihood. For alternative estimation approaches using EM that could be extended to this setting, we refer the reader to Bańbura and Modugno (2014). Nonlinear or non-Gaussian dynamic factor models in a mixed frequency setting have been considered in Gagliardini et al. (2017), Leippold and Yang (2019) that rely on particle filtering methods in conjunction with backward simulation algorithms as in Godsill et al. (2004), while Schorfheide et al. (2018) consider a Bayesian approach

using particle MCMC (see Andrieu et al. (2010)). Such approaches can become computationally expensive and are not considered for benchmarking purposes.

While previous mixed-frequency DFMs (see Mariano and Murasawa (2003), Bańbura and Modugno (2014) for a more thorough discussion) often consider time series which are sampled at two frequencies, we introduce here a flexible mixed-frequency DFM that describes $L \in \mathbb{N}^+$ collections of distinct time series sampled at frequencies $\{\kappa_1, \dots, \kappa_L\}$ and each consisting of $\{n_1, \dots, n_L\}$ series, respectively. In the same setting as in Section 1.3, each group of n_l, $l \in [L]$, time series sampled at frequency κ_l contains observations $(\boldsymbol{y}^{(l)}_{t,s|\kappa_l})$ with $\boldsymbol{y}^{(l)}_{t,s|\kappa_l} \in \mathbb{R}^{n_l}$ for all $t \in \mathbb{Z}$ and $s \in \{0, \dots, \kappa_l - 1\}$. Let $\kappa_{\max} = \max_l \kappa_l$. Suppose that the latent factor dynamics are updated at the highest sampling frequency based on the linear transition

$$\boldsymbol{f}_{t,s+1|\kappa_{\max}} = A_{\boldsymbol{\theta}} \boldsymbol{f}_{t,s|\kappa_{\max}} + R_{\boldsymbol{\theta}} \boldsymbol{u}_{t,s+1|\kappa_{\max}}, \tag{1.46}$$

where

$$\boldsymbol{f}_{t,s|\kappa_{\max}} = \left(\boldsymbol{v}^\top_{t,s|\kappa_{\max}}, \dots, \boldsymbol{v}^\top_{t,s-p+1|\kappa_{\max}}\right)^\top,$$

with $A_{\boldsymbol{\theta}}$ given in (1.45) for the special case where $A^{(\ell)}_{\boldsymbol{\theta}} = \mathbb{O}_k$ for $\ell \geq 2$, $p = \kappa_{\max}$ and

$$A^{(1)}_{\boldsymbol{\theta}} = \bar{A} \frac{\rho}{\max\left\{\rho, |\lambda_1(\bar{A})|\right\}}$$

with parameters $\rho \in (0,1)$, $\bar{A} \in \mathbb{M}_k$ and with $\lambda_1(\bar{A})$ denoting the largest eigenvalue of $\bar{A}$. In the simplified scenario of first-order autoregressive dynamics, we parameterize $R_{\boldsymbol{\theta}} \in \mathbb{M}_k$ to be positive definite and diagonal and $\boldsymbol{u}_{t,s+1|\kappa_{\max}}$ are a sequence of IID k-dimensional standard Gaussian variables.

Notice that Kalman filtering formulas yield the first moment

$$\widehat{\boldsymbol{f}}_{t,s|\kappa_{\max}} = \mathbb{E}\left[\boldsymbol{f}_{t,s|\kappa} | \boldsymbol{y}_{1,0|\kappa_{\max}}, \dots, \boldsymbol{y}_{t,s|\kappa_{\max}}\right]$$

recursively online, see for example Appendix 1.G.1 for details in the general time-inhomogeneous case. Due to the linearity in (1.46), for any $h \in \mathbb{N}$,

$$\widehat{\boldsymbol{f}}_{t,s+h|\kappa_{\max}} = \mathbb{E}\left[\boldsymbol{f}_{t,s+h|\kappa_{\max}} | \boldsymbol{y}_{1,0|\kappa_{\max}}, \dots, \boldsymbol{y}_{t,s|\kappa_{\max}}\right] = A^h_{\boldsymbol{\theta}} \widehat{\boldsymbol{f}}_{t,s|\kappa_{\max}}.$$

Furthermore, from the linearity of the aggregation scheme, we obtain the forecasts for any $s, h \in \mathbb{N}$,

$$\mathbb{E}\left[\boldsymbol{y}^{(l)}_{t,s+h|\kappa_l} | \boldsymbol{y}_{1,0|\kappa_l}, \dots, \boldsymbol{y}_{t,s|\kappa_l}\right] = \mathsf{Agg}_{\boldsymbol{\theta}^{(l)}}\left(\widehat{\boldsymbol{f}}_{t,(s+h)q_l|\kappa_{\max}}\right), \tag{1.47}$$

where $q_l = \kappa_{\max}/\kappa_l$ (c.f. Section 1.3.1) and $\mathsf{Agg}_{\boldsymbol{\theta}^{(l)}}$ is the aggregation scheme for frequency l. We observe that there is a single latent factor process that describes the observations at all frequencies, in contrast, for instance, to hierarchical Hidden Markov Models (HMM) (Hihi and Bengio, 1995) where the latent variables evolve a priori at different time-scales. This time evolution of states is similar to the SMFESN models also developed in this paper.

It is possible to write the following mixed-frequency DFM model in Example 1.G.4 as a general time-inhomogeneous state-space system (1.41)-(1.42) by suitably parameterizing the time dependencies in the aggregation matrices. We provide more details on implementing our mixed frequency

DFM in Appendix 1.G.1 below. The standard Kalman filtering recursions utilized therein for parameter estimation have a cubic complexity in the dimension d or n of the Markovian factor process $\boldsymbol{f}$ or the observation process $\boldsymbol{y}$, respectively, at every time step. The marginal log-likelihood is optimized based on stochastic gradient methods with adaptive step sizes (Kingma and Ba, 2014) and is generally not a concave function of the parameter values.[11]

Example 1.G.4 (Quarterly-Monthly-Daily DFM Model)**.** We consider $n_{(\mathtt{6d})}$ time series that result from averaging daily time series over 6 days, yieling 12 observations per quarter that are denoted as $\boldsymbol{y}^{(\mathtt{6d})}$. Furthermore, we consider $n_{(\mathtt{m})}$ monthly $\boldsymbol{y}^{(\mathtt{m})}$ as well as $n_{(\mathtt{q})}$ quarterly time series $\boldsymbol{y}^{(\mathtt{q})}$. We let $\kappa_{\max} = 72/6 = 12$ and update the k-dimensional latent factor process every 6 days in sync with $\boldsymbol{y}^{(\mathtt{6d})}$. We aggregate 6 days to significantly decrease the computational cost of the factor model. The latent factors are assumed to have the VAR(1) dynamics,[12]

$$\boldsymbol{v}_{t,s+1|12} = A^{(1)}\boldsymbol{v}_{t,s|12} + R\boldsymbol{u}_{t,s+1|12},$$

for any $s, t \in \mathbb{N}$, $A^{(1)} \in \mathbb{M}_{k,k}$ $R \in \mathbb{M}_k$ and IID k-dimensional standard Gaussian variables $\boldsymbol{u}_{t,s|12}$. The averaged daily data is described by

$$\boldsymbol{y}^{(\mathtt{6d})}_{t,s|12} = \beta^{(\mathtt{6d})}\boldsymbol{v}_{t,s|12} + S^{(\mathtt{6d})}\boldsymbol{w}^{(\mathtt{6d})}_{t,s|12}$$

for any $s, t \in \mathbb{N}$, $\beta^{(\mathtt{6d})} \in \mathbb{M}_{n_{(\mathtt{6d})},k}$, $S^{(\mathtt{6d})} \in \mathbb{M}_{n_{(\mathtt{6d})}}$ and IID $n_{(\mathtt{6d})}$-dimensional standard Gaussian variables $\boldsymbol{w}^{(\mathtt{6d})}_{t,s|12}$. The monthly data in the stock aggregation scheme is modeled as

$$\boldsymbol{y}^{(\mathtt{m})}_{t,s|3} = \beta^{(\mathtt{m})}\boldsymbol{v}_{t,4s|12} + S^{(\mathtt{m})}\boldsymbol{w}^{(\mathtt{m})}_{t,s|3},$$

with $\beta^{(\mathtt{m})} \in \mathbb{M}_{n_{(\mathtt{m})},k}$, $S^{(\mathtt{m})} \in \mathbb{M}_{n_{(\mathtt{m})}}$ and IID $n_{(\mathtt{m})}$-dimensional standard Gaussian variables $\boldsymbol{w}^{(\mathtt{m})}_{t,s|3}$. Alternatively, an Almon aggregation scheme yields the model

$$\boldsymbol{y}^{(\mathtt{m})}_{t,s|3} = \beta^{(\mathtt{m})}\sum_{\ell=0}^{3}\overline{\boldsymbol{\varphi}}(\boldsymbol{\psi}^{(\mathtt{m})}, \ell) \odot \boldsymbol{v}_{t,(4s-\ell)|12} + S^{(\mathtt{m})}\boldsymbol{w}^{(\mathtt{m})}_{t,s|3},$$

with $\beta^{(\mathtt{m})} \in \mathbb{M}_{n_{(\mathtt{m})},k}$, $S^{(\mathtt{m})} \in \mathbb{M}_{n_{(\mathtt{m})}}$, IID $n_{(\mathtt{m})}$-dimensional standard Gaussian variables $\boldsymbol{w}^{(\mathtt{m})}_{t,s|3}$ and $\overline{\boldsymbol{\varphi}}(\boldsymbol{\psi}^{(\mathtt{m})}, \ell) = \left(\overline{\varphi}(\psi^{(\mathtt{m})}{}_1, \ell), \ldots, \overline{\varphi}(\psi^{(\mathtt{m})}{}_k, \ell)\right)^\top \in \mathbb{R}^k$. The symbol $\odot$ stands for the Hadamard or componentwise matrix product.

The quarterly components can be analogously described as

$$\boldsymbol{y}^{(\mathtt{q})}_{t} = \beta^{(\mathtt{q})}\boldsymbol{v}_{t,0|12} + S^{(\mathtt{q})}\boldsymbol{w}^{(\mathtt{q})}_{t}$$

for a stock aggregation scheme, while the Almon scheme writes as

$$\boldsymbol{y}^{(\mathtt{q})}_{t} = \beta^{(\mathtt{q})}\sum_{\ell=0}^{11}\overline{\boldsymbol{\varphi}}(\boldsymbol{\psi}^{(\mathtt{q})}, \ell) \odot \boldsymbol{v}_{t,-\ell|12} + S^{(\mathtt{q})}\boldsymbol{w}^{(\mathtt{q})}_{t},$$

[11] We compute gradients of the marginal log-likelihood using a Kalman filter implementation for a time-inhomogeneous linear Gaussian state space model in TensorFlow Probability (Dillon et al., 2017).

[12] Because of the AR(1) dynamics, we do not write it in the companion form of the latent factor. However, unless one uses the stock aggregation scheme, one still needs to keep track of the past factor values for modeling monthly or quarterly observables.

with $\beta^{(\mathfrak{q})} \in \mathbb{M}_{n_{(\mathfrak{q})},k}$, $S^{(\mathfrak{m})} \in \mathbb{M}_{n_{(\mathfrak{q})}}$, IID $n_{(\mathfrak{q})}$-dimensional standard Gaussian variables $\boldsymbol{w}_t^{(\mathfrak{m})}$ and $\overline{\boldsymbol{\varphi}}(\boldsymbol{\psi}^{(\mathfrak{q})}, \ell) = \left(\overline{\varphi}(\psi^{(\mathfrak{q})}{}_1, \ell), \ldots, \overline{\varphi}(\psi^{(\mathfrak{q})}{}_k, \ell)\right)^\top \in \mathbb{R}^k$.

1.G.1 Mixed-frequency DFM Implementation

This section gives additional details on implementing non-homogeneous dynamic factor models, such as the mixed frequency model introduced in the main text. We notice that the conditioning notation in this section should not be confused with our temporal notation in Definition 1.1.1.

Kalman filtering and computational complexity. The sufficient statistics of the posterior distribution of the latent factor $\boldsymbol{f}_t|\boldsymbol{y}_{0:t}$ can be updated recursively by the Kalman filter updates in the linear Gaussian setting. First, propagate the prior

$$\begin{aligned}\widehat{\boldsymbol{f}}_{t+1|t,\boldsymbol{\theta}} &= A_{\boldsymbol{\theta}} \widehat{\boldsymbol{f}}_{t|t,\boldsymbol{\theta}} \\ \widehat{\Sigma}_{t+1|t,\boldsymbol{\theta}} &= A_{\boldsymbol{\theta}} \widehat{\Sigma}_{t+1|t,\boldsymbol{\theta}} A_{\boldsymbol{\theta}}^\top + S_{t+1,\boldsymbol{\theta}} S_{t+1,\boldsymbol{\theta}}^\top.\end{aligned}$$

Compute the innovation covariance

$$\Gamma_{t+1,\boldsymbol{\theta}} = \Lambda_{t+1} \widehat{\Sigma}_{t+1|t,\boldsymbol{\theta}} \Lambda_{t+1}^\top + R_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}^\top$$

and the Kalman gain

$$K_{t+1,\boldsymbol{\theta}} = \widehat{\Sigma}_{t+1|t,\boldsymbol{\theta}} \Lambda_{t+1,\boldsymbol{\theta}}^\top \Gamma_{t+1,\boldsymbol{\theta}}^{-1}.$$

Then, update the statistics with the new information y_{t+1},

$$\begin{aligned}\widehat{\boldsymbol{f}}_{t+1|t+1,\boldsymbol{\theta}} &= \widehat{\boldsymbol{f}}_{t+1|t,\boldsymbol{\theta}} - K_{t+1,\boldsymbol{\theta}} \left(y_{t+1} - \Lambda_{t+1,\boldsymbol{\theta}} \widehat{\boldsymbol{f}}_{t+1|t,\boldsymbol{\theta}}\right) \\ \widehat{\Sigma}_{t+1|t+1,\boldsymbol{\theta}} &= \left(\mathbb{I} - K_{t+1,\boldsymbol{\theta}} \Lambda_{t+1,\boldsymbol{\theta}}\right) \widehat{\Sigma}_{t+1|t,\boldsymbol{\theta}}.\end{aligned}$$

Notice that the inverse of the log-determinant of the innovation matrices $\Gamma_{t,\boldsymbol{\theta}}$ are required for computing the Kalman gains and the marginal log-likelihood, respectively, which yield a cubic computational complexity in the dimension of the observation process. Alternatively, one can apply matrix inversion or determinant lemmas to obtain a computational complexity that is cubic in the dimension of the Markovian factor process $\boldsymbol{f}_t$. For an alternative approach in high-dimensions that imposes a dynamic factor structure after a projection of the observations onto a low-dimensional space, see Jungbacker and Koopman (2015), and Bräuning and Koopman (2014) for a collapsed mixed-frequency DFM.

Model selection. The model parameters $\boldsymbol{\theta}$ are learned to jointly maximize the log-likelihood of the observed data for all frequencies. This is in contrast to the parameter estimation approach for MIDAS, which minimizes the MSE of low-frequency predictions conditional on observing the high-frequency series. We remark that a different log-likelihood weighting for the different frequencies in DFMs has been suggested in Blasques et al. (2016), but requires cross-validation to optimize such weightings. Nevertheless, the introduced DFM contains several hyperparameters that need to be chosen, such as the latent state space dimension k or the order p of the latent Markov process.

One possibility is to select such hyperparameters by evaluating the low-frequency predictions on a validation set. Approaches for choosing the dimensions of the latent factor process have been under-explored in the mixed-frequency setting, but see Bai and Ng (2007), Hallin and Liška (2007) for possible criteria in general dynamic factor models. In our implementation, we choose $p = 1$, as this allows for a differentiable model parametrization with stationary factor dynamics. We set $k = 5$ for the small dataset and $k = 10$ for the medium dataset.

Parameter estimation and forecasting. Based on the results from the Kalman filtering recursions, the model parameters $\boldsymbol{\theta}$ are learned by maximizing the marginal log-likelihood using $\ell_t(\boldsymbol{\theta}) = -\log p_{\boldsymbol{\theta}}(\boldsymbol{y}_{0:t}) = -\sum_{s=0}^{t} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_s|\boldsymbol{y}_{0:s-1})$ where $p_{\boldsymbol{\theta}}(\boldsymbol{y}_s|\boldsymbol{y}_{0:s-1})$ is Gaussian with mean $\Lambda_{s,\boldsymbol{\theta}}\widehat{\boldsymbol{f}}_{s|s-1,\boldsymbol{\theta}}$ and covariance $\Gamma_{s,\boldsymbol{\theta}}$. Gradients of $\ell_t(\boldsymbol{\theta})$ can be computed using algorithmic differentiation.

For fixed $\boldsymbol{\theta} \in \Theta$ and $h \in \mathbb{N}$, let

$$\mu_{t+h|t,\boldsymbol{\theta}}(\boldsymbol{y}_{t+h}|\boldsymbol{y}_{0:t}) = \int g_{t+h,\boldsymbol{\theta}}(\boldsymbol{y}_{t+h}|\boldsymbol{f}_{t+h}) \prod_{\ell=1}^{h} h_{t+\ell,\boldsymbol{\theta}}(\boldsymbol{f}_{t+\ell}|\boldsymbol{f}_{t+\ell-1})\mathrm{d}\boldsymbol{f}_{t+\ell}\pi_{t|t,\boldsymbol{\theta}}(\boldsymbol{f}_t|\boldsymbol{y}_{0:t})\mathrm{d}\boldsymbol{f}_t$$

be the h-step predictive distribution of the data, while $\pi_{t|t,\boldsymbol{\theta}}(\boldsymbol{f}_t|\boldsymbol{y}_{0:t})$ is the filtering distribution of the latent state $\boldsymbol{f}_t|\boldsymbol{y}_{0:t}$. The mean of $\mu_{t+h|t,\boldsymbol{\theta}}(\cdot|\boldsymbol{y}_{0:t})$ is $\widehat{\boldsymbol{y}}_{t+h|t,\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{y}_{t+h}|\boldsymbol{y}_{0:t}]$. For some $t, \tau \geq 0$, let us write $\widehat{\boldsymbol{f}}_{t+\tau|t,\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{f}_{t+\tau}|\boldsymbol{y}_{0:t}]$ and $\Sigma_{t+\tau|t,\boldsymbol{\theta}} = \mathrm{Cov}_{\boldsymbol{\theta}}[\boldsymbol{f}_{t+\tau} - \widehat{\boldsymbol{f}}_{t+\tau|t,\boldsymbol{\theta}}|\boldsymbol{y}_{0:t}]$ for the mean and covariance of the latent process, respectively. For linear Gaussian dynamics, Kalman filtering allows for computing the filtered mean $\widehat{\boldsymbol{f}}_{t|t,\boldsymbol{\theta}}$ and covariance matrices $\widehat{\Sigma}_{t|t,\boldsymbol{\theta}}$ analytically.

For fixed $\boldsymbol{\theta}$, the τ-step ahead prediction function $H^{\tau}_{t,\boldsymbol{\theta}}(\boldsymbol{y}_{0:t}) = \widehat{\boldsymbol{y}}_{t+\tau|t,\boldsymbol{\theta}} = \Lambda_{t+\tau,\boldsymbol{\theta}}\widehat{\boldsymbol{f}}_{t+\tau|t,\boldsymbol{\theta}}$ is linear due to the Kalman filtering recursion. For $s \leq t$, consider also the prediction $H^{\star\tau}_{s,t}(\boldsymbol{y}_{0:t}) = \mathbb{E}_{\boldsymbol{\theta}^\star(\boldsymbol{y}_{0:s})}[\boldsymbol{y}_{t+\tau}|\boldsymbol{y}_{0:t}]$ that is based on the sample $\boldsymbol{y}_{0:t}$, but where $\boldsymbol{\theta}^\star(\boldsymbol{y}_{0:s}) = \arg\min_{\boldsymbol{\theta}} \ell_s(\boldsymbol{\theta})$ maximizes the marginal likelihood of data $\boldsymbol{y}_{0:s}$ only. This setting allows to implement different parameter estimation setups from Section 1.C.1. For instance, the fixed parameter setup corresponds to fixing s, which yields a fixed training set $\boldsymbol{y}_{0:s}$ to estimate $\boldsymbol{\theta}$. In the expanding window setup, both s and t are expanded, while a rolling window setting updates the dataset $\boldsymbol{y}_{0:s}$ by rolling over the data.

1.H High-Frequency Forecasts

To better understand how the use of high-frequency data impacts forecasting, as an additional empirical experiment we investigate high-frequency (HF) forecasts of all models in the Small-MD dataset. We restrict our analysis to this dataset because the computational burden to construct HF forecasts can be high: when using daily data and using our suggested 24 days-per-month interpolation, one quarter amounts to 72 daily frequency observations, which means HF forecasts can involve thousands of data points, and for DFM and M-MFESN models this setup can be quite computationally onerous.

Constructing HF forecasts with MIDAS is trivial once the aggregation weights have been estimated, even though a practical implementation requires care in constructing the appropriate lag matrices. Recall for Section 1.F that the MIDAS equation with L regressors $(z^{(l)}_{t,s|\kappa_l})_{t,s}$ with

$z^{(l)}_{t,s|\kappa_l} \in \mathbb{R}$, $l \in [L]$ for all $t \in \mathbb{Z}$ and $s \in \{0, \dots, \kappa_l - 1\}$ can be written as

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{l=1}^{L} \beta_l \sum_{k=0}^{K_l} \varphi(\boldsymbol{\theta}_l, k) z^{(l)}_{t,-k|\kappa_l} + \epsilon_t.$$

For clarity, we suppress the dynamic autoregressive component, as it has the same frequency as the target. Now assume that we include $n_{(\mathtt{m})}$ monthly and $n_{(\mathtt{d})}$ daily frequency regressors in the model that are sampled $\kappa_{(\mathtt{m})} = 3$ and $\kappa_{(\mathtt{d})} = 72$ times per quarter and hence $\kappa_{\max} = 72$. Therefore we can partition the regression above in the following way

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{l=1}^{n_{(\mathtt{m})}} \beta_l \sum_{k=0}^{K_l} \varphi(\boldsymbol{\theta}_l, k) z^{(l)}_{t,-k|3} + \sum_{l=n_{(\mathtt{m})}+1}^{L} \beta_l \sum_{k=0}^{K_l} \varphi(\boldsymbol{\theta}_l, k) z^{(l)}_{t,-k|72} + \epsilon_t$$

with $L = n_{(\mathtt{m})} + n_{(\mathtt{d})}$.

Assuming parameter estimates $\widehat{\alpha}_0, \widehat{\alpha}_1, \dots, \widehat{\alpha}_p$ and $\{(\widehat{\beta}_l, \widehat{\boldsymbol{\theta}}_l)\}_{l=1}^{L}$ are available, the HF forecast $\widehat{y}_{t+1,s|72}$ is given by

$$\widehat{y}_{t+1,s|72} = \widehat{\alpha}_0 + \sum_{i=1}^{p} \widehat{\alpha}_i y_{t-i} + \sum_{l=1}^{n_{(\mathtt{m})}} \widehat{\beta}_l \sum_{k=0}^{K_l} \varphi(\widehat{\boldsymbol{\theta}}_l, k) z^{(l)}_{t,\lfloor s/24 \rfloor - k|3} + \sum_{l=n_{(\mathtt{m})}+1}^{L} \widehat{\beta}_l \sum_{k=0}^{K_l} \varphi(\widehat{\boldsymbol{\theta}}_l, k) z^{(l)}_{t,s-k|72}.$$

For DFMs, high-frequency forecasts can be constructed using (1.46) and (1.47) to iterate factors forward in time and then aggregate them according to estimated loadings or a weighting scheme.

Multi-frequency ESN models are also able to yield high-frequency forecasts in a straightforward manner. For simplicity, let us consider the case as in Example 1.3.1 of an aligned S-MFESN model that has been fit to a quarterly target with monthly and daily input data. The reservoir is run in high-frequency, $\kappa_{\max}$ steps per quarter, according to state equation

$$\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s|72} = \alpha \boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s-1|72} + (1-\alpha)\sigma(A\boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s-1|72} + C\boldsymbol{z}^{(\mathtt{m},\mathtt{d})}_{t,s|72} + \boldsymbol{\zeta}).$$

Suppose a coefficient matrix $\widehat{W}$ has been estimated. Then, as states between low-frequency periods t and $t+1$ are collected, we can immediately construct the high-frequency one-step-ahead forecasts

$$\widetilde{y}_{t+1,s|72} = \widehat{W}^{\top} \boldsymbol{x}^{(\mathtt{m},\mathtt{d})}_{t,s|72}.$$

For M-MFESN models HF forecasts require slightly more care. For example, when dealing with the multi-reservoir MFESN model of Example 1.3.2, we must repeat the most recent monthly state at daily frequency correctly.

1.I Robustness Analysis

1.I.1 MIDAS

As we discuss briefly in the main text, parameter optimization is a principal problem in implementing any MIDAS model. Even though explicit formulas exist for both gradient and Hessian of the MIDAS loss objective when an Almon weighting scheme is used (see Kostrov (2021)), there is

no known guarantee that the loss itself is convex or even locally convex. In practice, for a given starting point (or point set) a numerical optimizer might only converge to a local minimum.

We observe this in practice, and we explore its effects on the robustness of MIDAS forecasts. We report summary results for our simulations in Figure 1.20. Our proposal is, given a MIDAS model specification and a set of starting points for evaluating the loss, to run an optimizer (for example, L-BFGS-B with explicit gradient) and select the smallest local minimum. By repeating this procedure multiple times, we collect a set of MIDAS parameters and study both the variation between the parameter vectors and the implied one-step ahead forecasts.

To be precise, our procedure is as follows:

1. For a total of B repetitions:

   (a) Choose M initialization points for the optimizer. We draw 64 points inside the hypercube of edge length 0.025 using a low-discrepancy Sobol sequence. The choice of a down-scaled hypercube as a domain comes from the empirical fact that the Almon exponential scheme may produce extremely large values even for relatively small coefficients. A straightforward way to see this is to notice that given any arbitrary small value for θ_1 and θ_2 in (1.38), for lag index k sufficiently large weight $\exp(\theta_1 k + \theta_2 k^2)$ will overflow at any given numerical precision. This means that one should adjust the MIDAS optimization domain based on the number of lags in the model.

   (b) For each initialization point, run the optimizer of choice.

   (c) Among the resulting M (local) loss minimizers, select and store the one with the lowest loss value.

2. With the resulting B stored minimizer:

   - Construct a low-dimensional projection of the high-dimensional minima to see their relative location in the parameter space and to compare their gradient and loss values, see Figure 1.20 (a)-(b).
   - Use each minimizer to produce MIDAS one-step ahead forecasts and plot quantile frequency plots of the forecast variation due to initialization; see Figure 1.20 (c).

Figure 1.20 shows that the best minimizers among initial Sobol sets are clustered together. To construct this 2D projection of the high-dimensional Almon coefficient space (including autoregressive lags and intercept), we use the well-known t-SNE procedure developed in van der Maaten (2009), which is an unsupervised dimensionality reduction algorithm capable of preserving the latent high-dimensional structure. This approach naturally implies that the Euclidean distances in the plot are suggestive of “clustering" rather than the actual latent distance between points. In Figures 1.20 (a)-(b), we see that the L-BFGS-B optimizer with explicit gradient achieves good convergence in terms of gradient norm and also that the resulting cluster of minimizers has close loss values. However, one can see that there is no single loss minimum: Figure 1.20 instead suggests that the local structure of the MIDAS loss function is very uneven, and therefore many distinct local minima can be achieved even when choosing a large number of initialization values for the

optimizer. This means that the “multi-start" strategy suggested in Kostrov (2021) to alleviate issues in MIDAS model estimation is insufficient.

The effects of non-negligible variation in parameter values on forecasts appear to be significant. Looking at Figure 1.20 (c), we can see wide frequency bands for the one-step ahead forecasts constructed using the Small-MD dataset and fixed parameter values. In particular, the Financial Crisis period seems to induce larger deviations in forecasts, consistent with the intuition that data with larger variation causes stronger model sensitivity when making forecasts.

1.I.2 MFESN

Since ESN models, and thus MFESN models, require random sampling of parameter matrices, the size of which is often large, there is inherent variability in any ESN model forecast. In theory, because all MFESN state parameters $(\widetilde{A}, \widetilde{C}, \widetilde{\boldsymbol{\zeta}})$ are drawn independently of each other, one could try to decompose the variance of any MFESN into the share due to parameter sampling and the share due to data sampling. Unfortunately, in practice, such decomposition is hard to derive. Cross-validation of ridge penalties and rolling and expanding window estimation are non-trivial data-dependent operations that complicate inference. In this work, we limit ourselves to numerically evaluating the effect of reservoir coefficient sampling on MFESN forecast variance.

Our approach is straightforward: given an MFESN model specification, c.f. Table 1.1, and a forecasting setup (fixed parameters, expanding or rolling window), we resample the reservoir state matrix parameters, perform cross-validation and possibly train-test sample windowing, and finally construct pointwise forecasts. Once a sufficiently large set of resampling forecasts has been computed, we plot frequency intervals in Figures 1.22 and 1.24. From Figure 1.22, we can see that the single-reservoir MFESN model with reservoir size 30 produces forecasts with a meaningful amount of variability induced by parameter resampling. Forecasts exhibit more variation when using an expanding or rolling window estimation strategy, even though the overall forecasts align with the GDP realizations. A similar discussion to that of MIDAS applies here: forecast sensitivity increases with underlying data variation, exacerbated in periods of systemic economic crisis.

Figure 1.24 suggests that larger MFESN models produce significantly more stable forecasts regarding model resampling. Note that the M-MFESN model [A] has a monthly frequency reservoir that is approximately 3 times the size of the S-MFESN model [A]. This stability is preserved even in expanding or rolling window settings, even though a slightly higher variation is apparent at the height of the 2008 Financial Crisis. We hypothesize that this reduction in variance due to model parameter sampling is due to the concentration of measure phenomena that prevail in high-dimensional spaces. Figure 1.24 suggests that larger MFESN models produce significantly more stable forecasts regarding model resampling. Note that the M-MFESN model [A] has a monthly frequency reservoir that is approximately 3 times the size of the S-MFESN model [A]. This stability is preserved even in expanding or rolling window settings, even though a slightly higher variation is apparent at the height of the 2008 Financial Crisis. We hypothesize that this reduction in variance due to model parameter sampling is due to the concentration of measure phenomena that prevail in high-dimensional spaces.

1.J Additional Figures

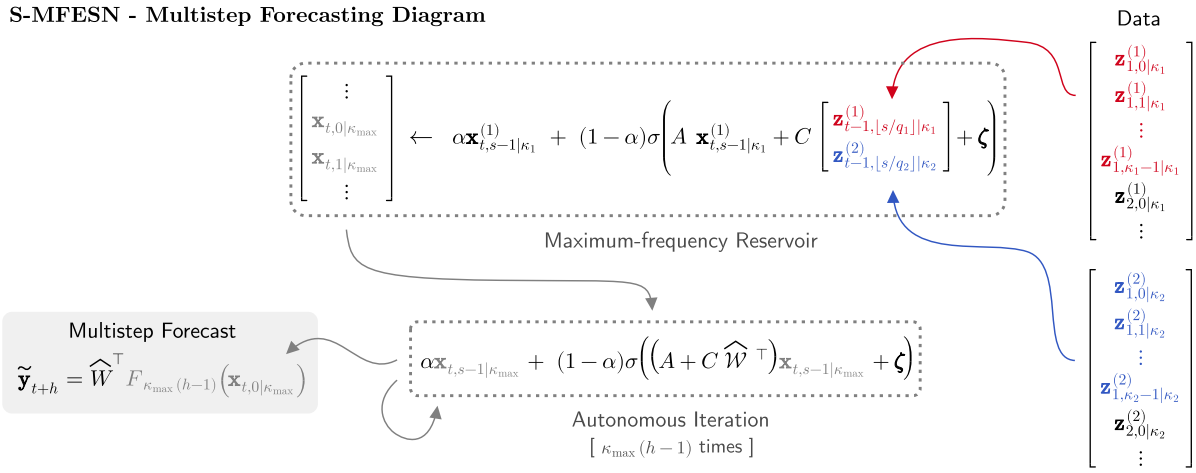


Figure 1.12

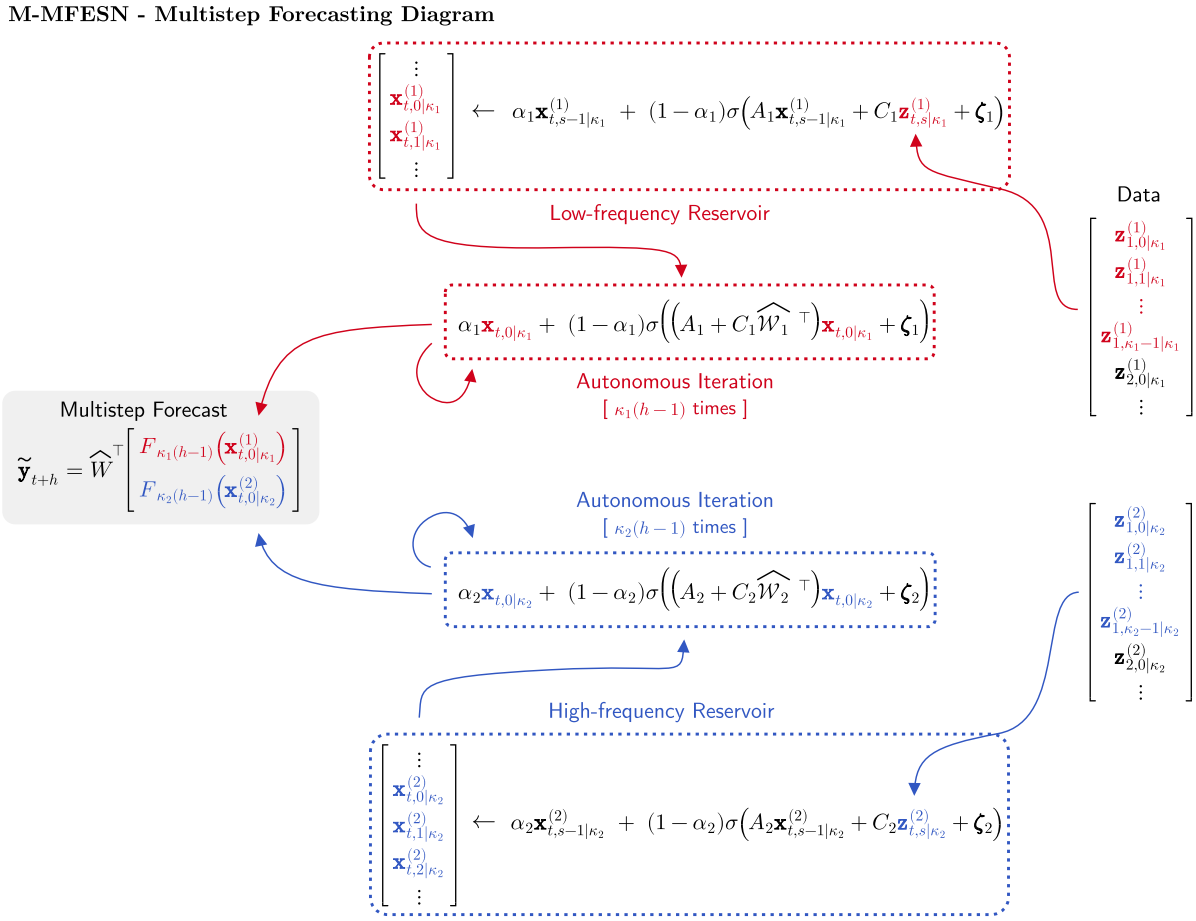
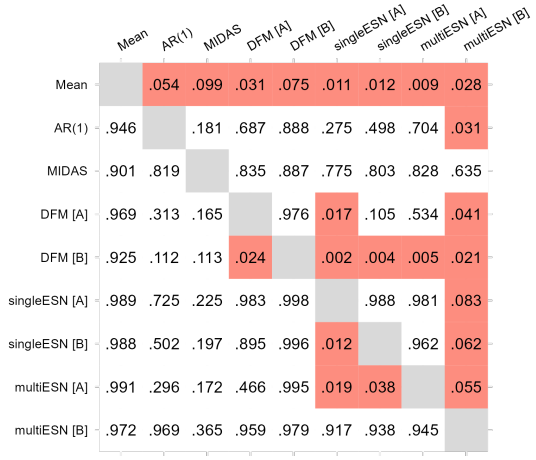


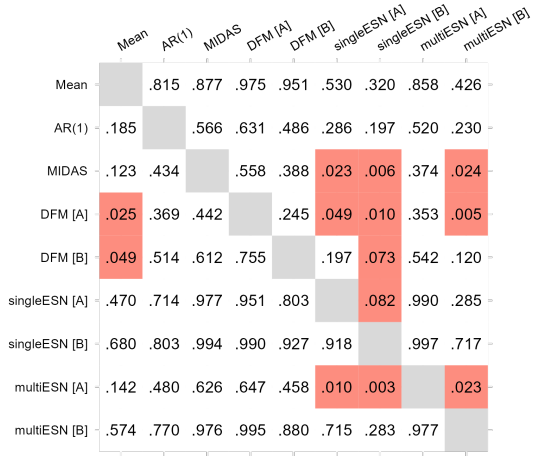
Figure 1.13

Figure 1.14: 1-Step-ahead GDP Forecasting – Modified Diebold-Mariano – Small-MD Dataset

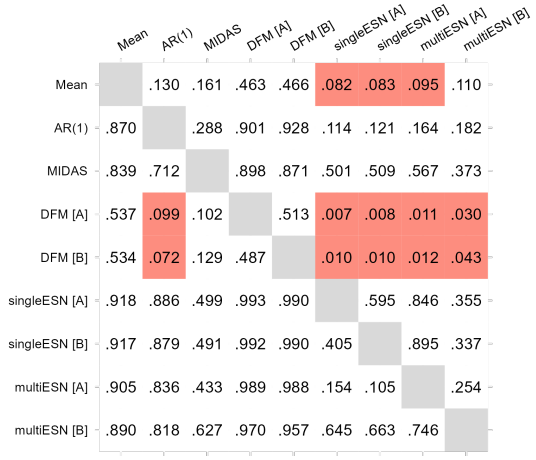
(a) Fixed 2007



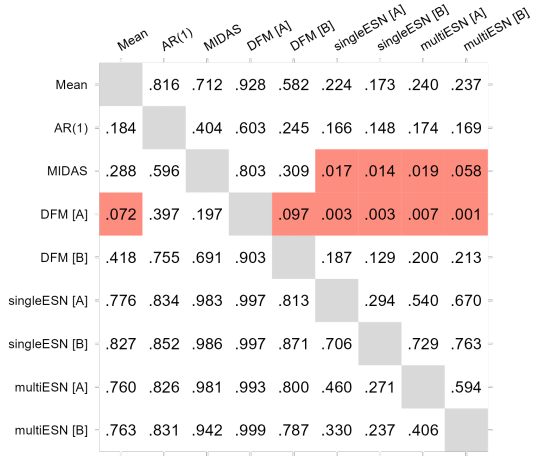
(b) Fixed 2011



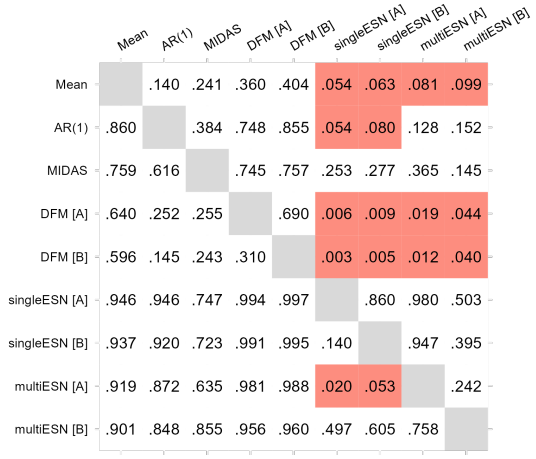
(c) Expanding 2007



(d) Expanding 2011



(e) Rolling 2007



(f) Rolling 2011

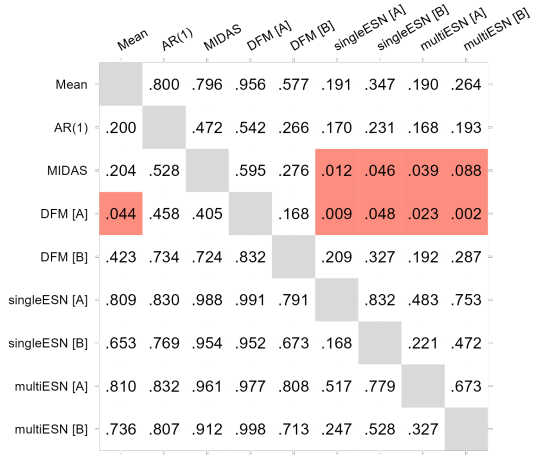
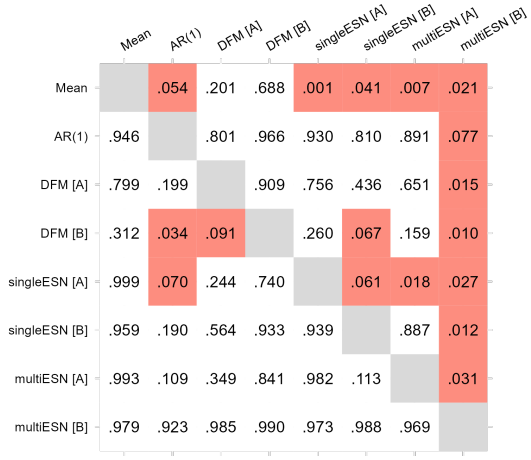


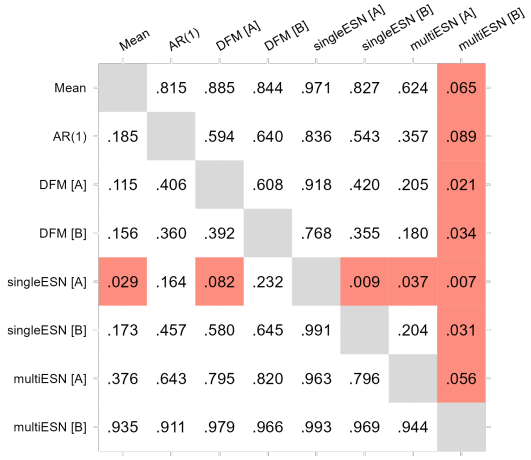
Figure 1.15: p-values of the pairwise Modified Diebold-Mariano tests between models of Table 1.3. Tests are one-sided and carried out row-wise: the null hypothesis for row i and column j reads as "the ith-row model forecasts *more accurately* than the jth-column model". Rejections at the 10% level are highlighted in red.

Figure 1.16: 1-Step-ahead GDP Forecasting – Modified Diebold-Mariano – Medium-MD Dataset

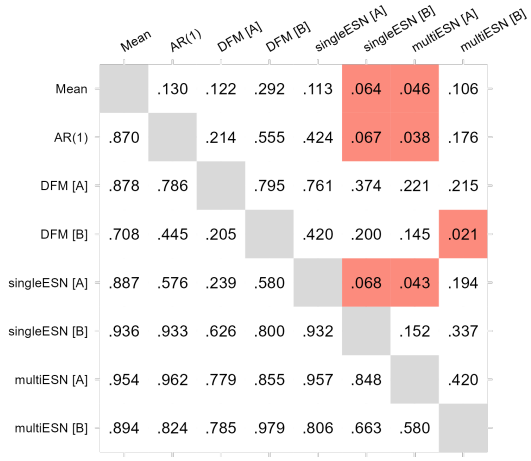
(a) Fixed 2007



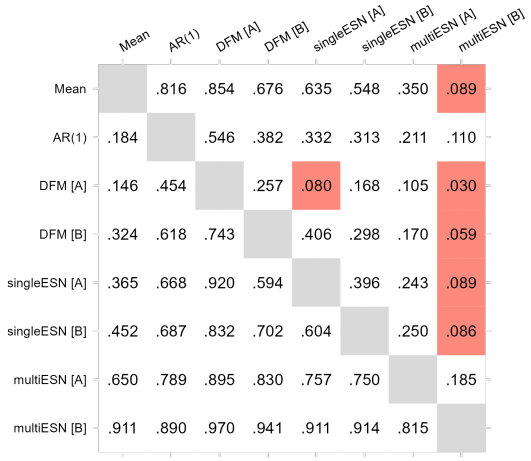
(b) Fixed 2011



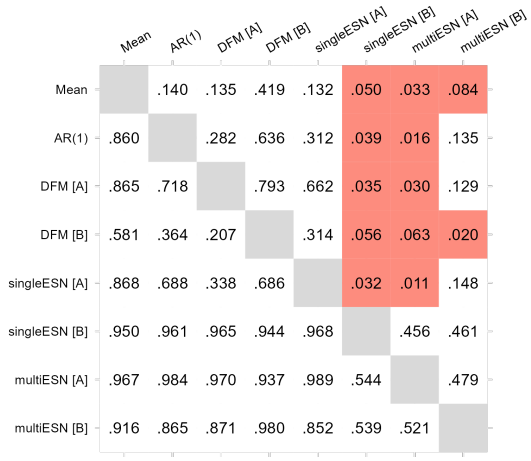
(c) Expanding 2007



(d) Expanding 2011



(e) Rolling 2007



(f) Rolling 2011

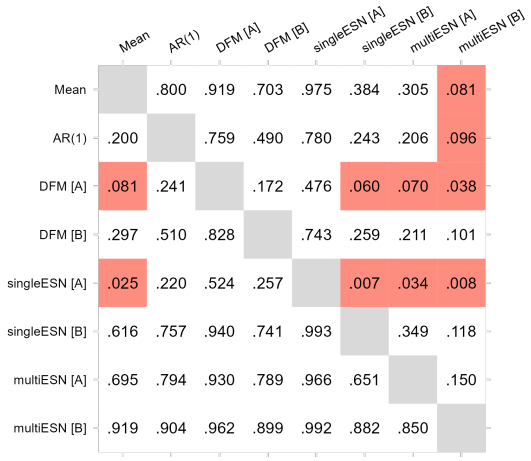
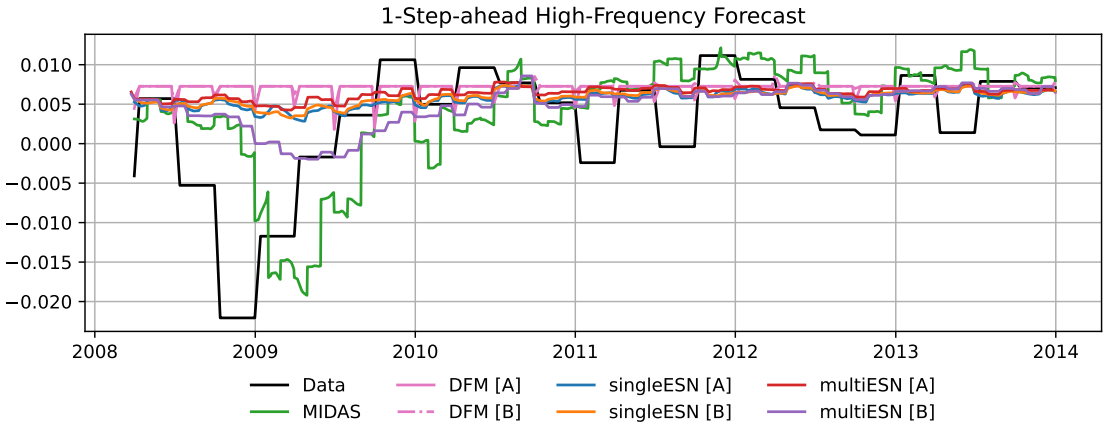


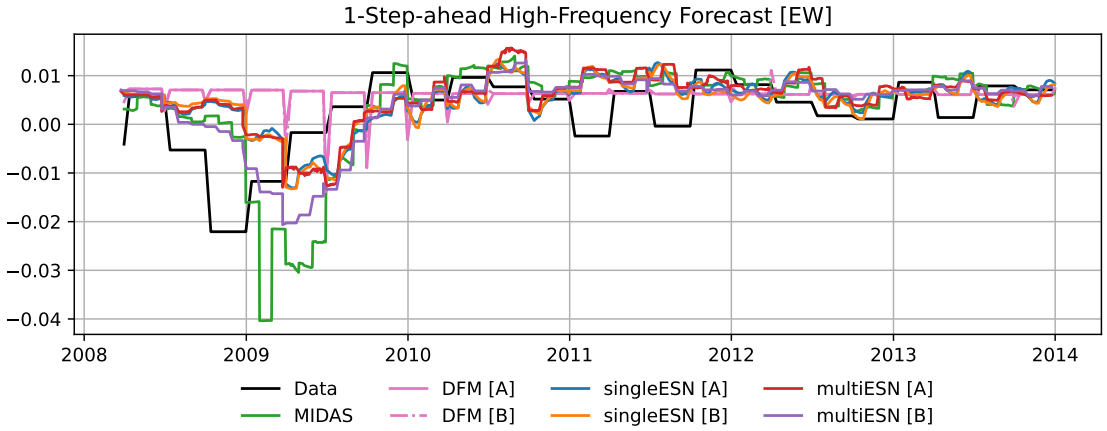
Figure 1.17: p-values of pairwise Modified Diebold-Mariano tests between models of Table 1.3. Tests are one-sided and carried out row-wise: the null hypothesis for row i and column j reads as "the ith-row model forecasts *more accurately* than the jth-column model". Rejections at the 10% level are highlighted in red.

Figure 1.18: 1-Step-ahead High-Frequency GDP Forecasting – 2007 Sample – Small-MD Dataset

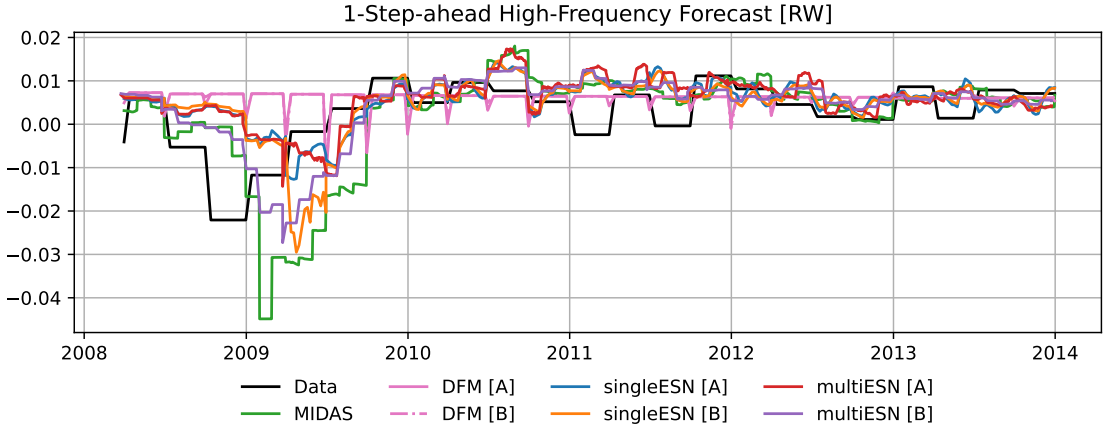
(a) Fixed



(b) Expanding



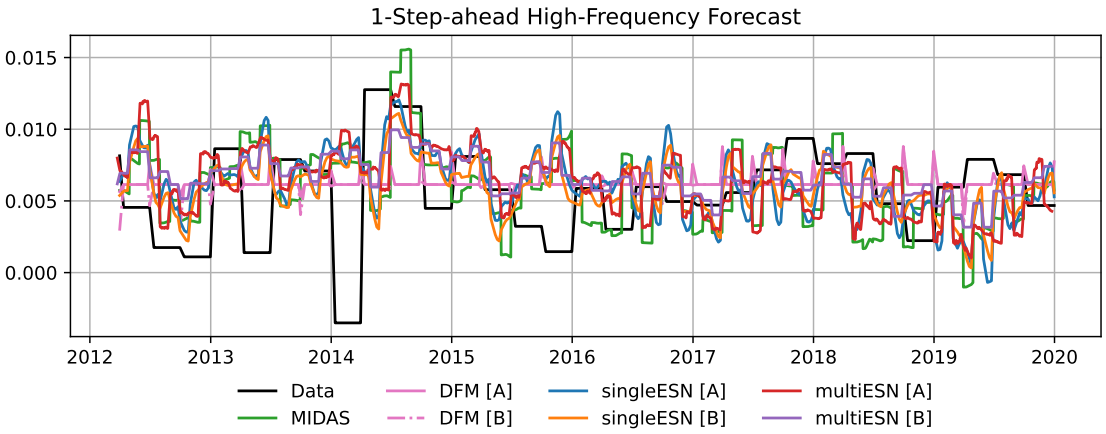
(c) Rolling



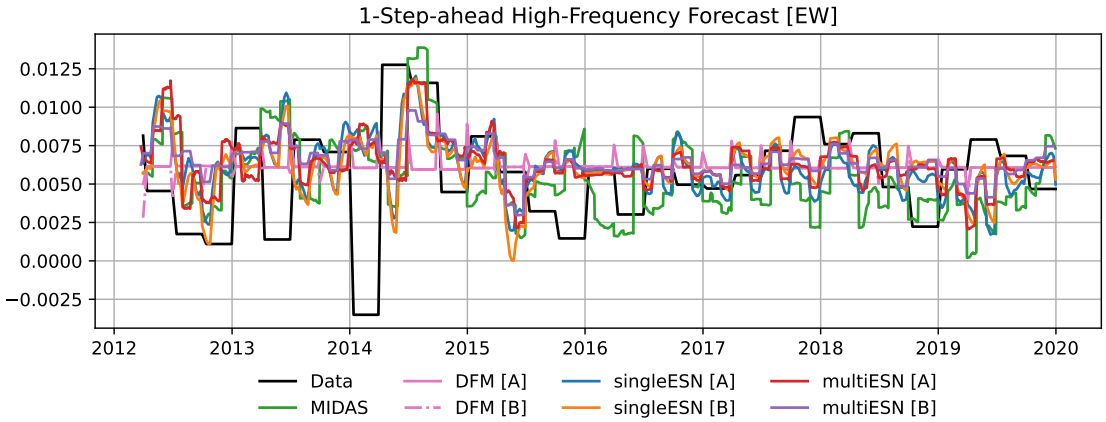
Note: Forecasts for the 2007 sample are presented up to Q4 2013 to better display the high-frequency behavior of models during the Financial Crisis period.

Figure 1.19: 1-Step-ahead High-Frequency GDP Forecasting – 2011 Sample – Small-MD Dataset

(a) Fixed



(b) Expanding



(c) Rolling

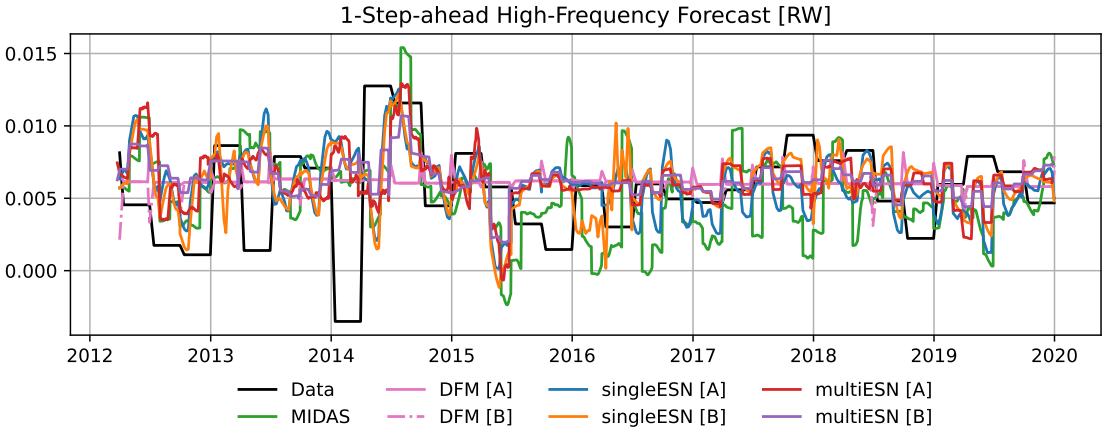
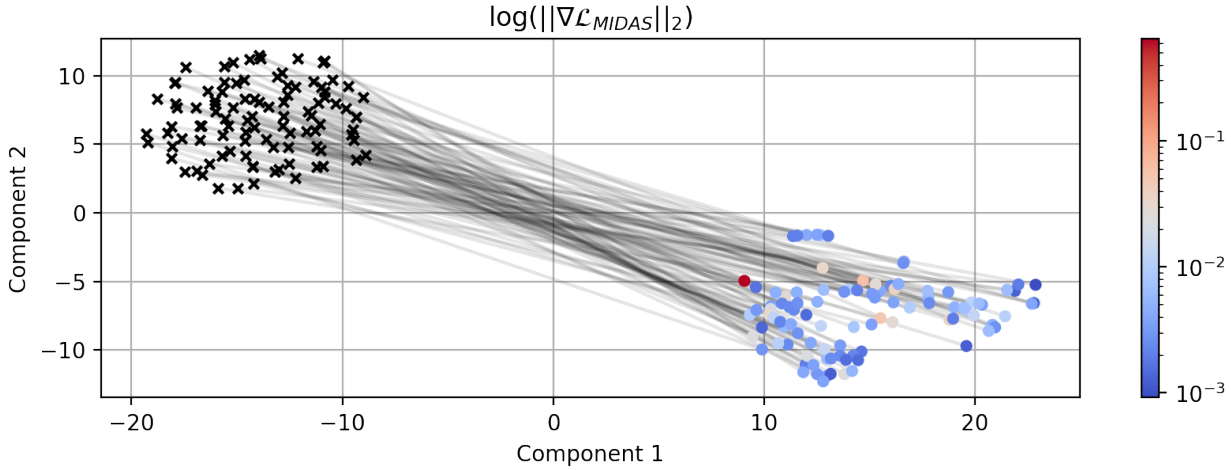


Figure 1.20: MIDAS Robustness Plots – 2007 Sample – Small-MD Dataset

(a) MIDAS Loss t-SNE Embedding: Gradient Norm



(b) MIDAS Loss t-SNE Embedding: Loss Norm

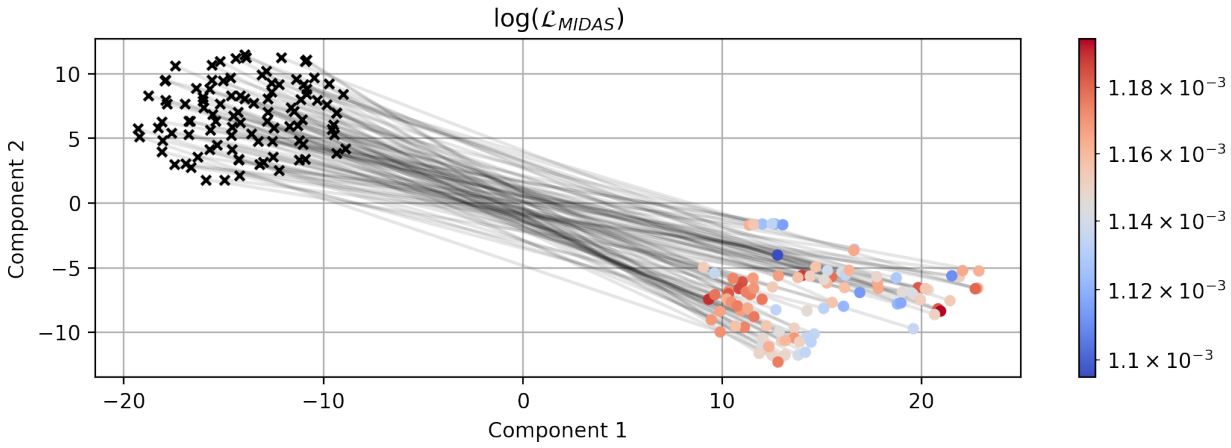
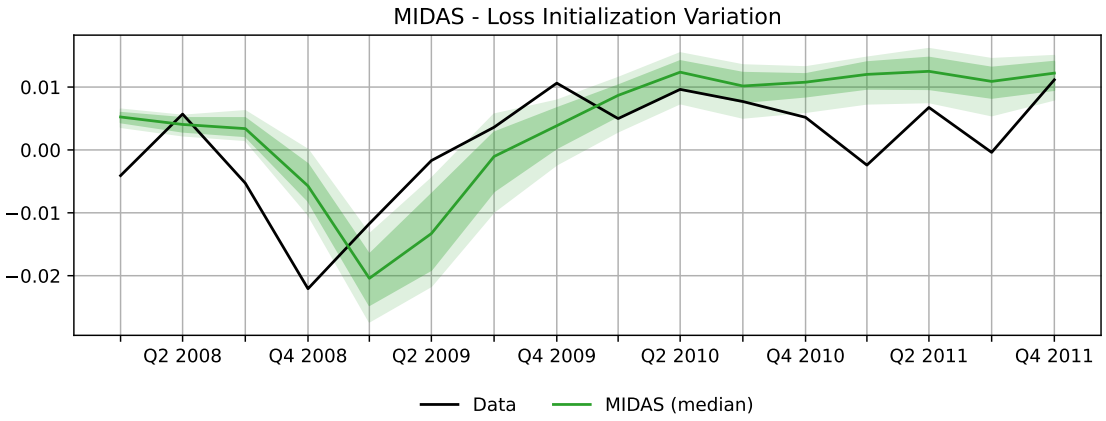
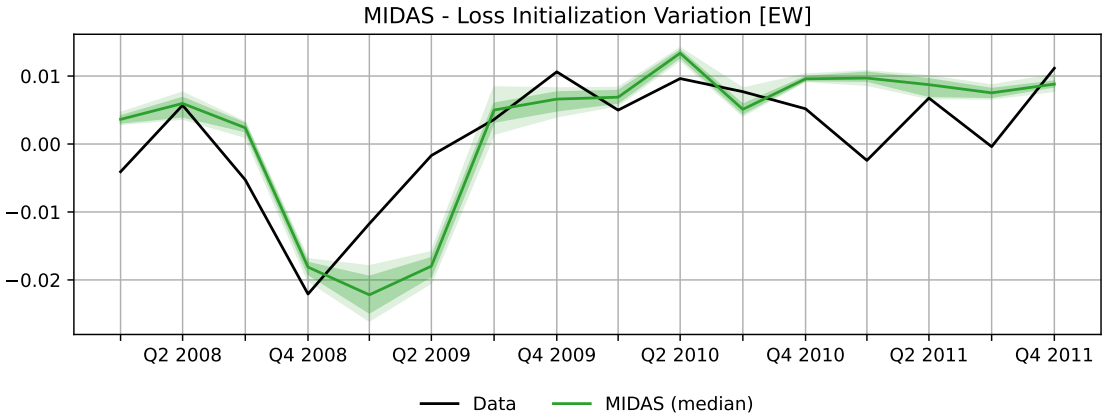


Figure 1.21: MIDAS Robustness Plots – 2007 Sample – Small-MD Dataset

(a) Fixed Parameters



(b) Expanding



(c) Rolling

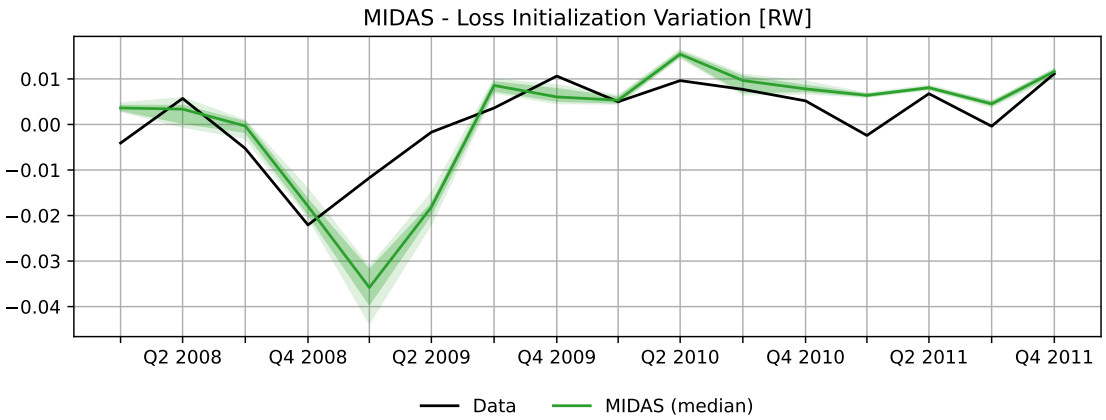
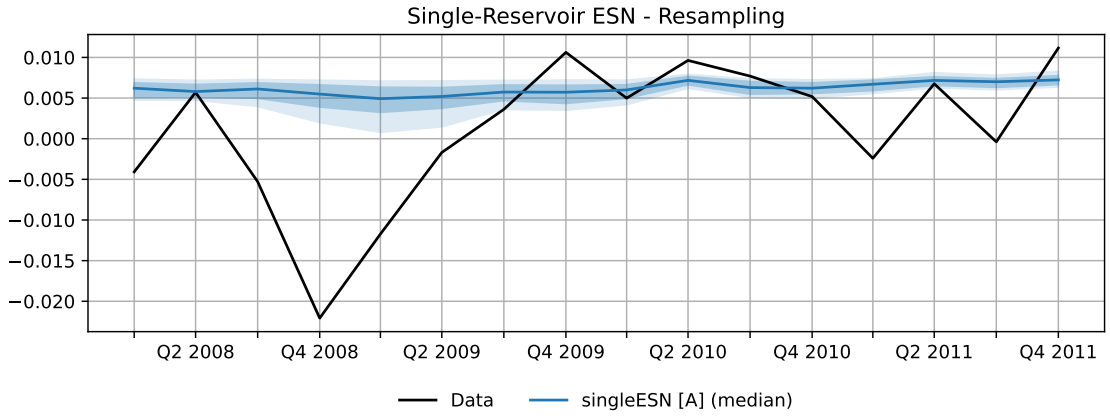
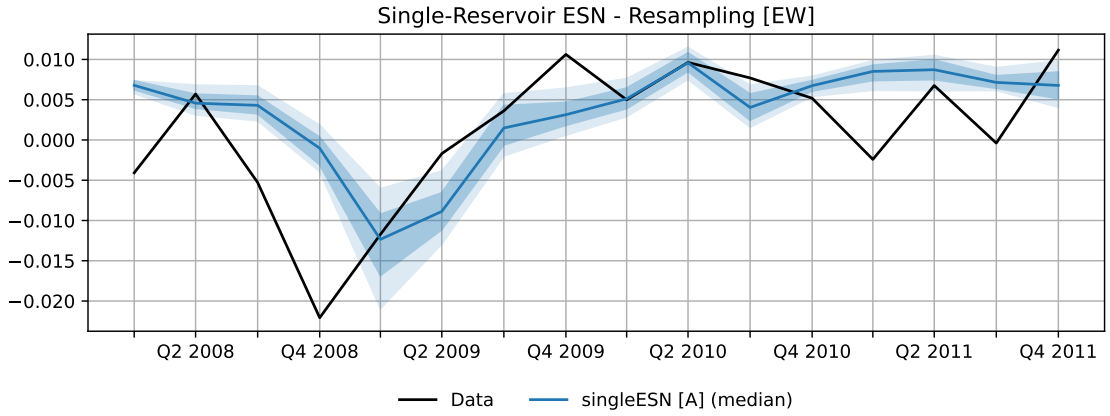


Figure 1.22: ESN Robustness Plots – 2007 Sample – Small-MD Dataset

(a) Fixed Parameters



(b) Expanding Window



(c) Rolling Window

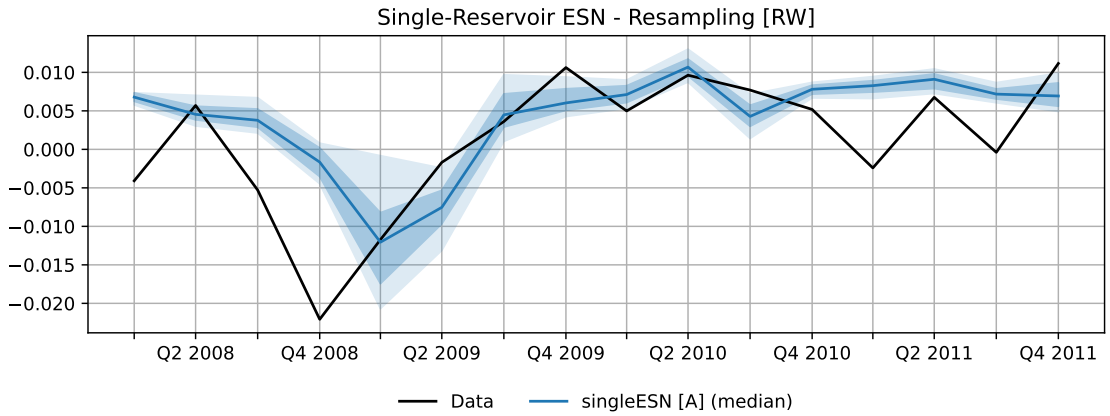
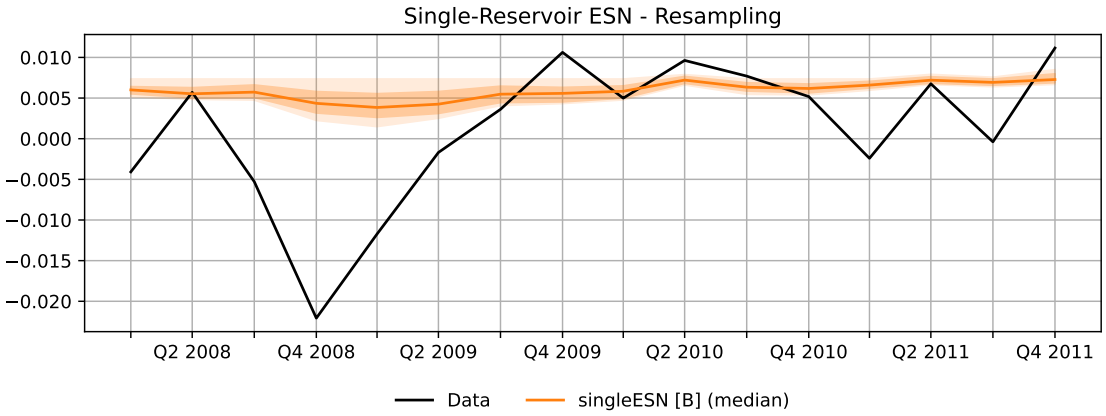
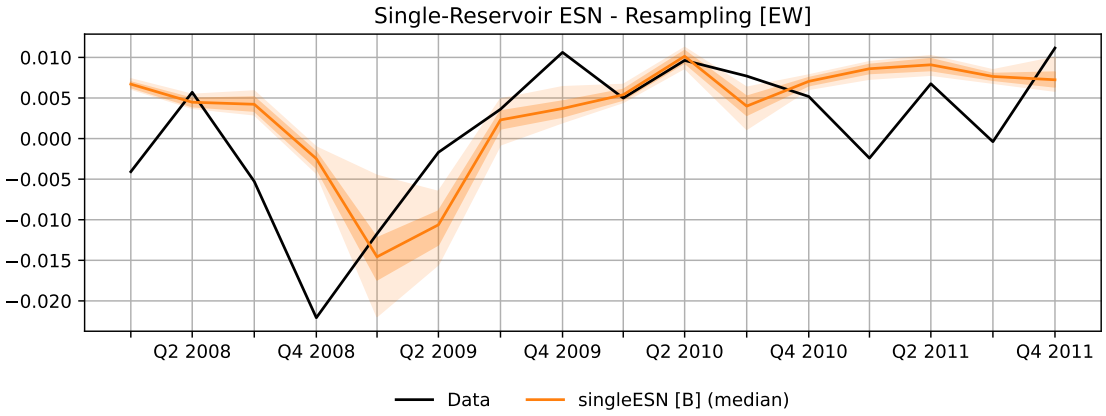


Figure 1.23: ESN Robustness Plots – 2007 Sample – Small-MD Dataset

(a) Fixed Parameters



(b) Expanding Window



(c) Rolling Window

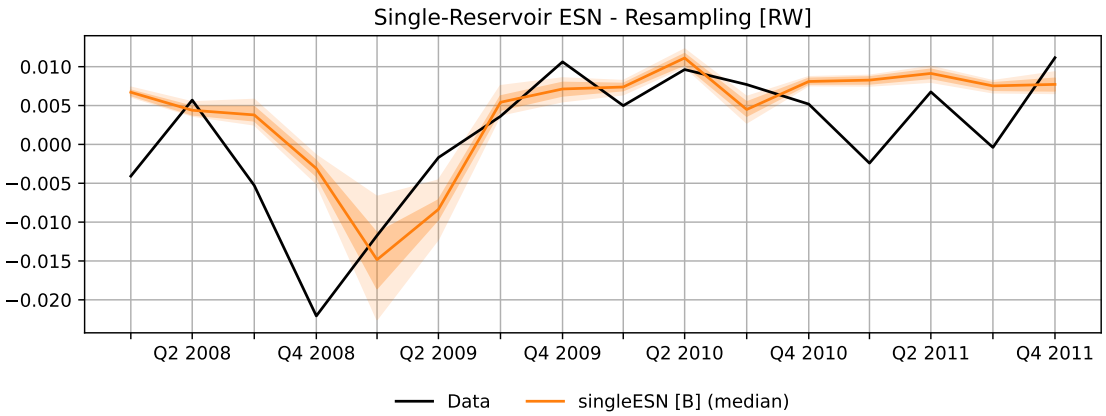
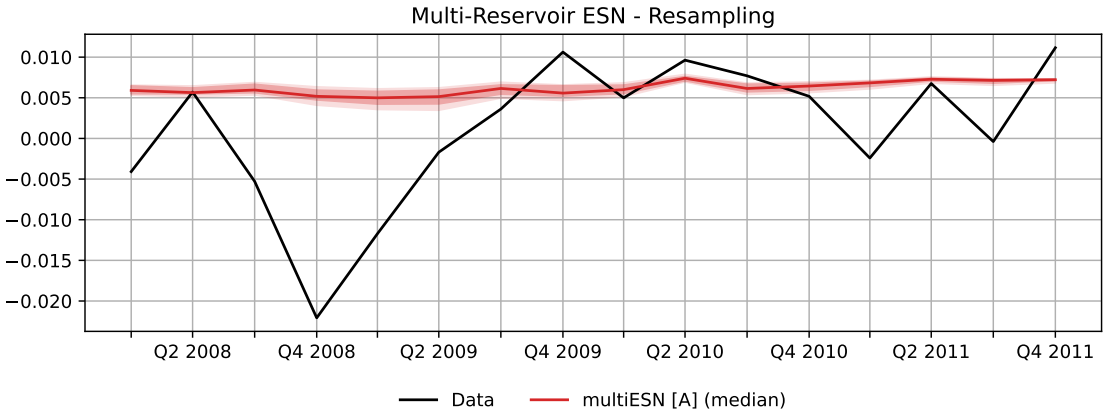
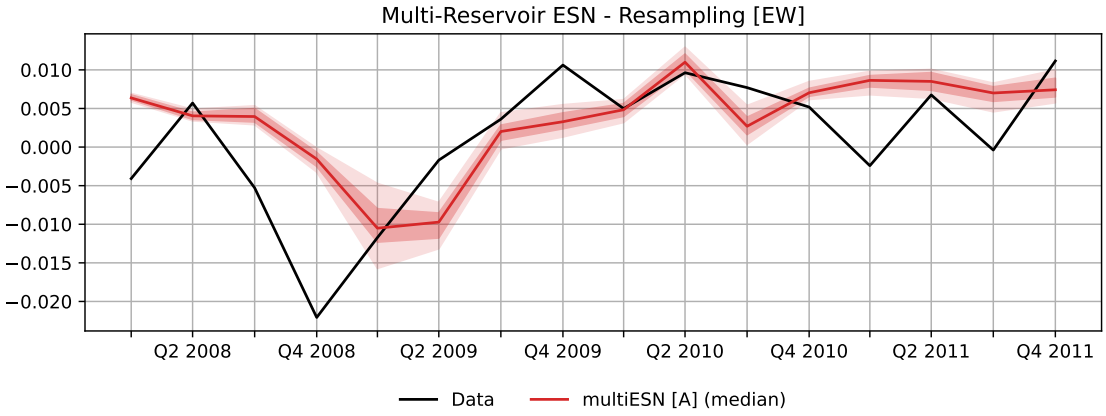


Figure 1.24: ESN Robustness Plots – 2007 Sample – Small-MD Dataset

(a) Fixed Parameters



(b) Expanding Window



(c) Rolling Window

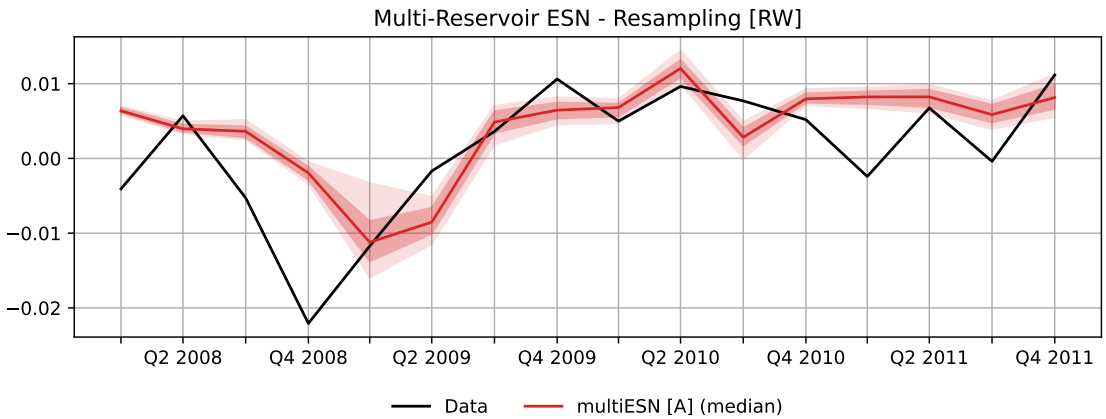
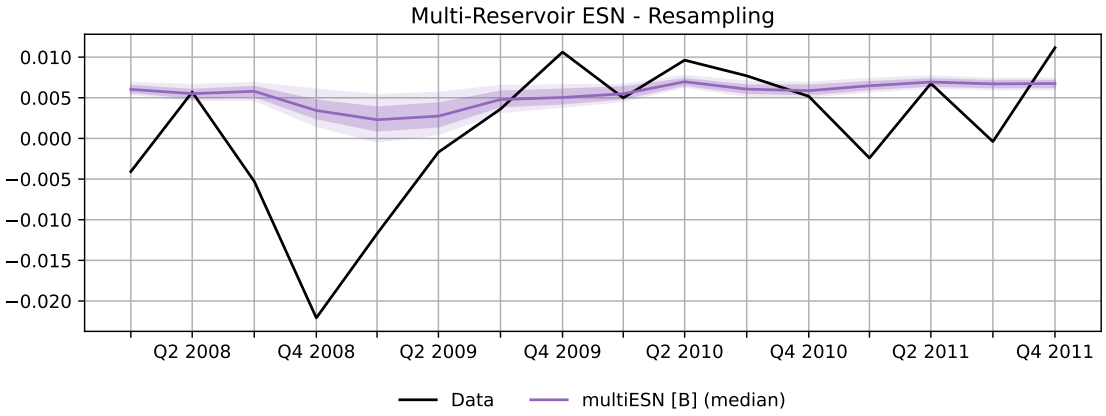
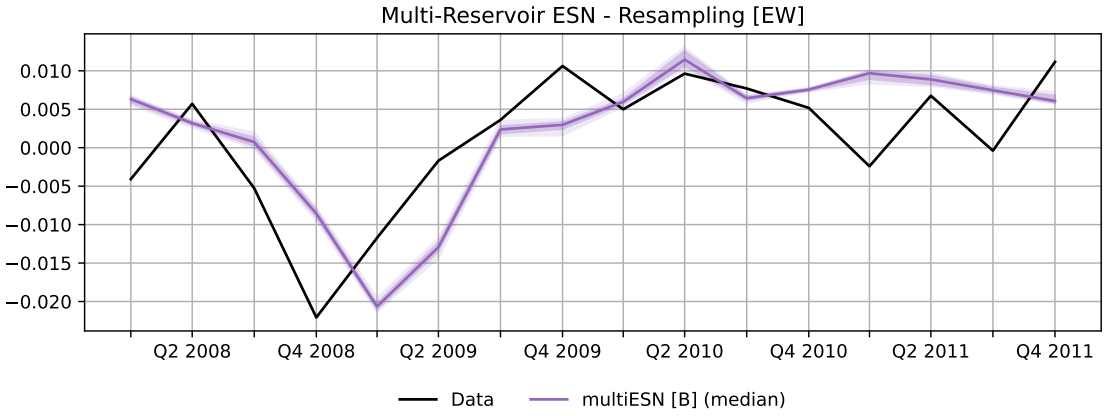


Figure 1.25: ESN Robustness Plots – 2007 Sample – Small-MD Dataset

(a) Fixed Parameters



(b) Expanding Window



(c) Rolling Window

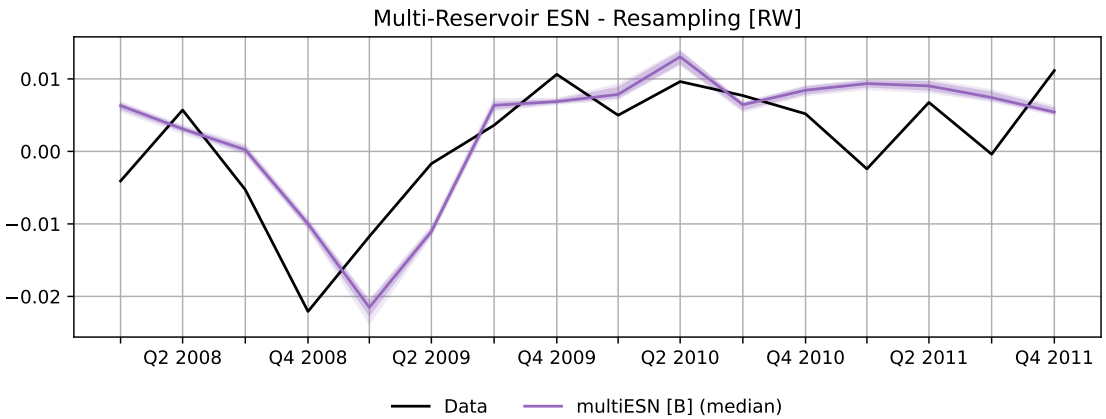
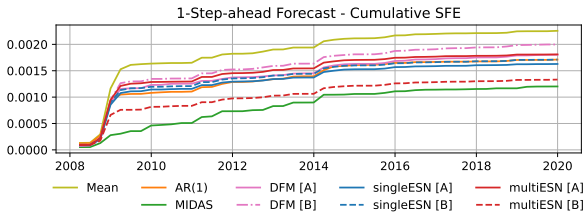
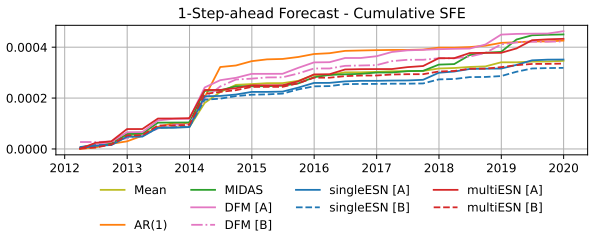


Figure 1.26: 1-Step-ahead GDP Forecasting, Fixed Parameters - Small-MD Dataset

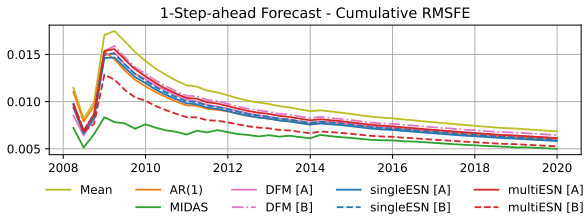
(a) Pre-crisis model



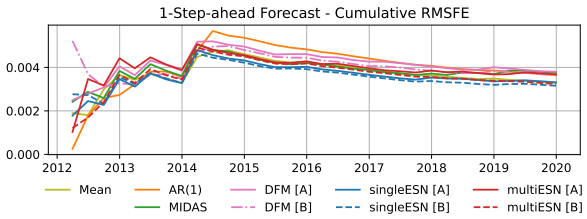
(b) Post-crisis model



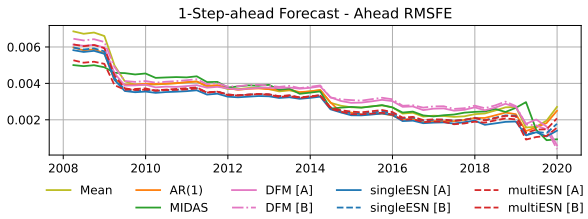
(c)



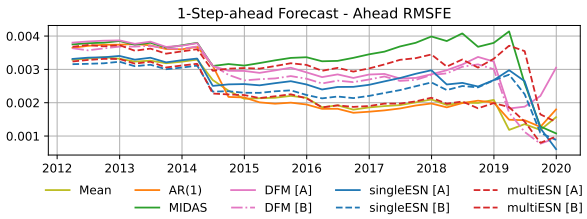
(d)



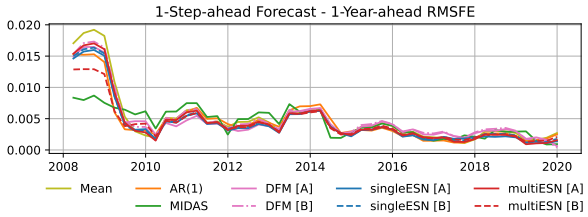
(e)



(f)



(g)



(h)

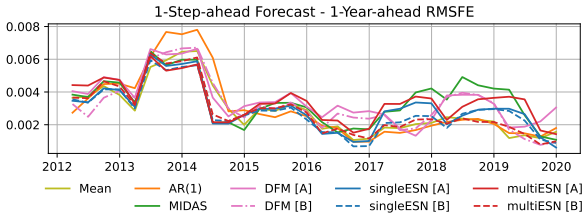
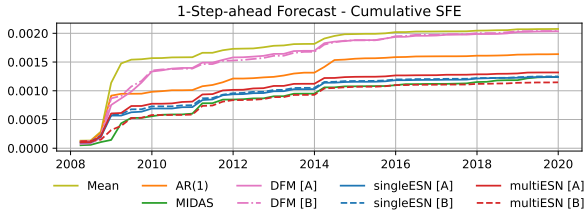
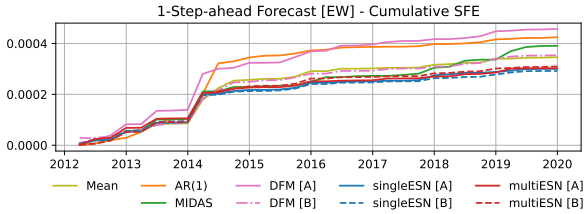


Figure 1.27: 1-Step-ahead GDP Forecasting, Expanding Window - Small-MD Dataset

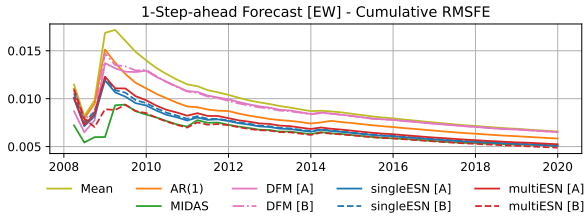
(a) Pre-crisis model



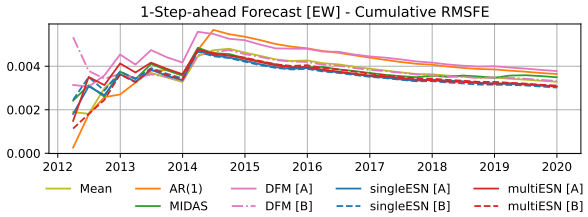
(b) Post-crisis model



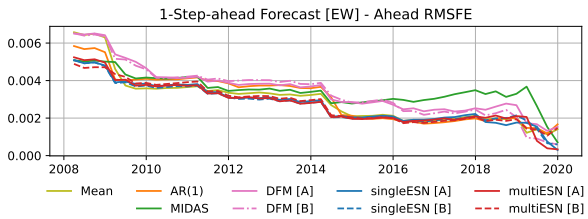
(c)



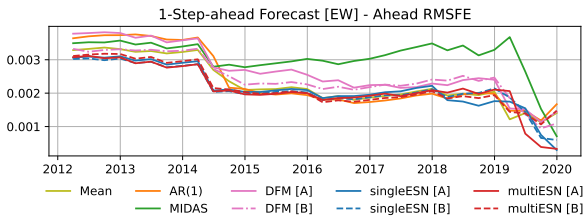
(d)



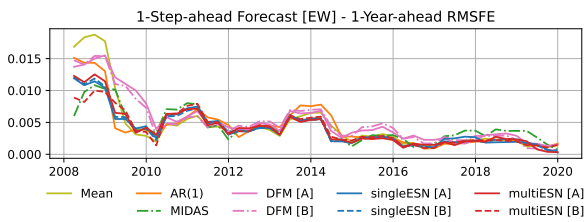
(e)



(f)



(g)



(h)

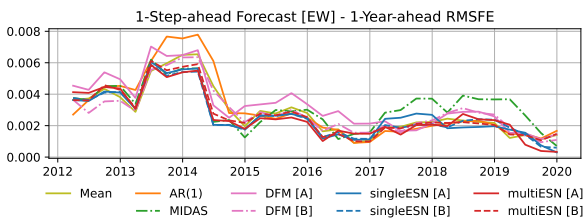
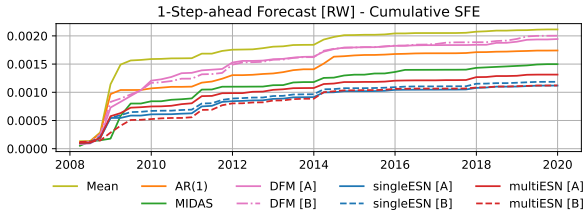
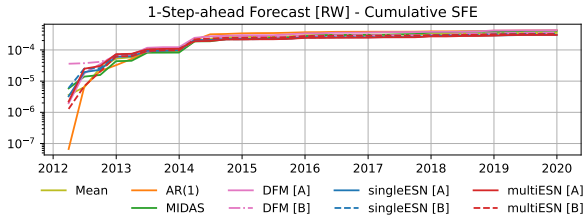


Figure 1.28: 1-Step-ahead GDP Forecasting, Rolling Window - Small-MD Dataset

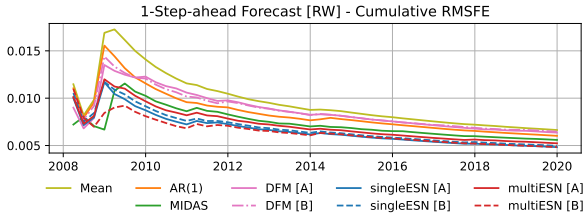
(a) Pre-crisis model



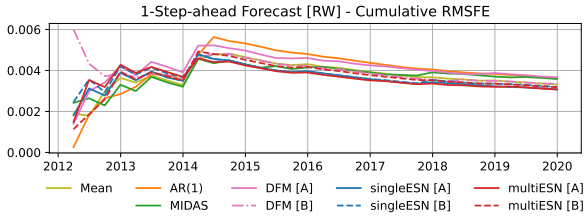
(b) Post-crisis model



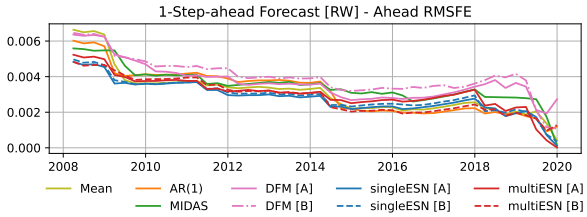
(c)



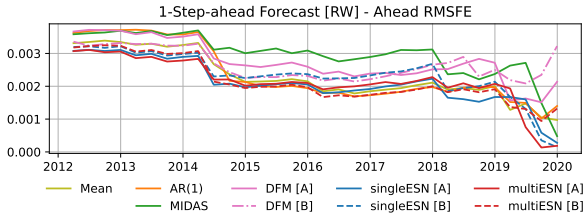
(d)



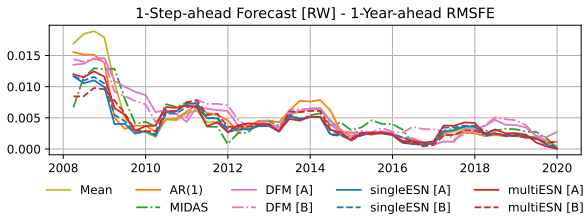
(e)



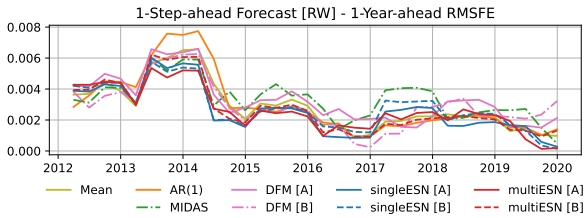
(f)



(g)



(h)



Chapter 2

Ridge Regularized Estimation of VAR Models for Inference

2.1 Introduction

While the idea of using ridge regression for vector autoregressive model estimation dates back to Hamilton (1994b), there seems to be no complete analysis of its properties and asymptotic theory in the literature. This paper fills this gap by analyzing the geometric and distributional properties of ridge in a VAR estimation framework, discussing its comparison to well-known Bayesian approaches and deriving the validity of cross-validation as a selection procedure for the ridge penalty.

First, I show that the shrinkage induced by the ridge estimator, while intuitive in the setting of an isotropic penalty, produces complex effects when estimating a VAR model with a more flexible penalization scheme. This implies that the benefits of the bias-variance trade-off (Hastie, 2020) may be hard to gauge a priori. I provide a tractable example where ridge can yield estimates that have higher autoregressive dependence than the least squares solution. To better understand how different ridge penalization strategies can be designed, I also make a comparison with Bayesian VAR estimators commonly used in macroeconometric practice.

Second, I generalize the analysis of Fu and Knight (2000) and prove the consistency and asymptotic normality of the ridge estimator, a result that seems to be missing in the literature. For standard inference, the ridge penalty should either be negligible in the limit or its centering converge in probability to the true parameter vector. In both these cases, there is no asymptotic bias and no reduction in variance. Alternatively, in settings where a researcher is willing to assume that a subset of the VAR parameters features small coefficients, one can achieve an asymptotic reduction of variance by correctly tuning the ridge penalty matrix. I further derive the properties of cross-validation, which is a popular approach in practical applications to tune penalized estimators (Hastie et al., 2009, Bergmeir et al., 2018b). More specifically, I show that cross-validation is able to select penalties that are asymptotically valid for inference. In passing, I also prove that, in an autoregressive setup, the time dependence of regressors has an exponentially small effect on in-sample prediction error evaluation.

Lastly, I use Monte Carlo simulations to study the performance of the different ridge approaches discussed, focusing on impulse response inference. I consider two exercises: one is based on a three-variable VARMA(1,1) data generating process from Kilian and Kim (2011); the other is a VAR(5) model estimated in levels from a set of seven macroeconomic series, following Giannone et al. (2015). The finding is that ridge can lead to improvements over unregularized methods in impulse

response confidence interval length, while Bayesian estimators show the best overall performance due to the underlying flexibility of their priors.

Related Literature. This paper does not discuss the high-dimensional setting, where the number of regressors grows together with the sample size. Some important work has been done in this direction already. Dobriban and Wager (2018) derive an explicit expression for the predictive risk of ridge regression assuming a high-dimensional random effects model. Other works in this vein are Liu and Dobriban (2020), Patil et al. (2021) and Hastie et al. (2022b), which are mostly focused on penalty selection by cross-validation, as well as structural features of ridge. Generally speaking, the complexity of analyzing ridge regression in high dimensions is a challenge to precisely understanding its practical implications. As I show below, in the context of finite-dimensional VARs, asymptotic inference demands that the ridge penalty becomes asymptotically negligible at appropriate rates. Thus, a challenge is understanding in what way high-dimensional time series problems would benefit from the use of ridge. This question is beyond the scope of this paper.

In the time series forecasting literature, ridge regression is commonly used for prediction. I provide a partial list of contributions in this direction. Inoue and Kilian (2008) use ridge regularization for forecasting U.S. consumer price inflation and argue that it compares favorably with bagging techniques; De Mol et al. (2008) use a Bayesian VAR with posterior mean equivalent to a ridge regression in forecasting; Ghosh et al. (2019) again study the Bayesian ridge, this time however in the high-dimensional context; Goulet Coulombe et al. (2022), Fuleky (2020a), Babii et al. (2021), and Medeiros et al. (2021) compare LASSO, ridge and other machine learning techniques for forecasting with large economic datasets. Fuleky (2020a) gives a textbook treatment of penalized time series estimation, including ridge, but does not discuss inference. The ridge penalty is considered within a more general mixed ℓ_1-ℓ_2 penalization setting in Smeekes and Wijler (2018), who study the performance and robustness of penalized estimates for constructing forecasts.

Regarding inference, Li et al. (2023) provided a general exploration of shrinkage procedures in the context of structural impulse response estimation. Very recently, Cavaliere et al. (2022) suggested a methodology for inference on ridge-type estimators that relies on bootstrapping. Finally, shrinkage of autoregressive models to constrained sub-models was discussed by Hansen (2016b) in a more general setting.

Finally, various estimation problems can either be cast as or augmented with ridge-type regressions. Goulet Coulombe (2023) shows that the estimation of VARs with time-varying parameters can be written as ridge regression. Plagborg-Møller (2016) and Barnichon and Brownlees (2019) both use ridge to derive smoothed local projection impulse response functions.

Outline. Section 2.2 provides a discussion of the ridge penalty and the ridge VAR estimator. In Section 2.3 I deal with the properties of ridge-induced shrinkage in the autoregressive coefficients. I discuss the connections between frequentist and Bayesian ridge for VAR models within Section 2.4. Section 2.5 develops the asymptotic theory and inference result in the case where there is no asymptotic shrinkage. This includes studying the property of cross-validation under dependence. Section 2.6 provides inference and CV results in a setting where some shrinkage of a subset of parameters is possible. Section 2.7 presents Monte Carlo simulations focused on impulse response

estimation. Section 2.8 concludes. Finally, the appendices contain more detailed derivations, as well as all proofs, additional tables and further information on simulations.

NOTATION. Define $\mathbb{R}_+$ to be the set of strictly positive real numbers. Vectors $v \in \mathbb{R}^N$ and matrices $A \in \mathbb{R}^{N\times M}$ are always denoted with lower and upper-case letters, respectively. Throughout, I will use I_M to represent the identity matrix of dimension M. For any vector $v \in \mathbb{R}^N$, $\|v\|$ is the Euclidean norm. For any matrix $A \in \mathbb{R}^{N\times M}$, $\|A\|$ is the spectral norm unless stated otherwise; $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ is the maximal entry norm; $\|A\|_F = (\text{tr}\{A'A\})^{-1/2}$ is the Frobenius norm; $\text{vec}(\cdot)$ is the vectorization operator and $\otimes$ is the Kronecker product (Lütkepohl, 2005). If a vector represents a vectorized matrix, then it will be written in bold, that is, for $A \in \mathbb{R}^{N\times M}$ I write $\text{vec}(A) = \boldsymbol{a} \in \mathbb{R}^{NM}$. Let $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_{K^2p}\}$, $\lambda_i > 0$ for all $i = 1, \ldots, K^2p$. To give the partial ordering of diagonal positive semi-definite penalization matrices, let $\Lambda_1 = \text{diag}\{\lambda_{1,j}\}_{j=1}^{K^2p}$ and $\Lambda_2 = \text{diag}\{\lambda_{2,j}\}_{j=1}^{K^2p}$. I write $\Lambda_1 \prec \Lambda_2$ if $\lambda_{1,i} < \lambda_{2,i}$ for all $i = 1, \ldots, K^2p$; $\Lambda_1 \preceq \Lambda_2$ if $\lambda_{1,i} \leq \lambda_{2,i}$ for all i and $\exists\ j \in 1, \ldots, K^2p$ such that $\lambda_{1,j} < \lambda_{2,j}$. Symbols $\xrightarrow{P}$ and $\xrightarrow{d}$ are used to indicate convergence in probability and convergence in distribution, respectively.

2.2 Ridge Regularized VAR Estimation

Let $y_t = (y_{1t}, \ldots, y_{Kt})'$ be a K-dimensional vector autoregressive process with lag length $p \geq 1$ and parametrization

$$y_t = \nu_t + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + u_t, \tag{2.1}$$

where $u_t = (u_{1t}, \ldots, u_{Kt})'$ is additive noise such that u_t are identically, independently distributed with $\mathbb{E}[u_{it}] = 0$ and $\text{Var}[u_t] = \Sigma_u$, and ν_t is a deterministic trend. For simplicity, in the remainder I shall assume that $\nu_t = 0$ so that y_t has no trend component – equivalently, y_t is a de-trended series.

For a given sample size T define $Y = (y_1, \ldots, y_T) \in \mathbb{R}^{K\times T}$, $z_t = (y_t', y_{t-1}', \ldots, y_{t-p+1}')' \in \mathbb{R}^{Kp}$, $Z = (z_0, \ldots, z_{T-1}) \in \mathbb{R}^{Kp\times T}$, $B = (A_1, \ldots, A_p) \in \mathbb{R}^{K\times Kp}$, $U = (u_1, \ldots, u_T) \in \mathbb{R}^{K\times T}$, and vectorized counterparts $\boldsymbol{y} = \text{vec}(Y)$, $\boldsymbol{\beta} = \text{vec}(B)$ and $\boldsymbol{u} = \text{vec}(U)$. Accordingly, $Y = BZ + U$ and $\boldsymbol{y} = (Z' \otimes I_K)\boldsymbol{\beta} + \boldsymbol{u}$, where $\Sigma_{\boldsymbol{u}} = I_K \otimes \Sigma_u$. Importantly, throughout this work, I will assume that the cross-sectional dimension K remains fixed.

Ridge regularization is a modification of the least squares objective by the addition of a term dependent on the Euclidean norm of the coefficient vector. The *isotropic* Ridge-regularized Least Squares (RLS) estimator is therefore defined as

$$\hat{\boldsymbol{\beta}}^R(\lambda) := \arg\min_{\boldsymbol{\beta}} \frac{1}{T} \left\|\boldsymbol{y} - (Z' \otimes I_K)\boldsymbol{\beta}\right\|^2 + \lambda \left\|\boldsymbol{\beta}\right\|^2,$$

where $\lambda > 0$ is the scalar regularization parameter or regularizer. When $\lambda \left\|\boldsymbol{\beta}\right\|^2$ is replaced with quadratic form $\boldsymbol{\beta}'\Lambda\boldsymbol{\beta}$ for a positive definite matrix Λ the above is often termed Tikhonov regularization. To avoid confusion, I shall also refer to it as "ridge", since in what follows Λ will always be assumed diagonal. As Λ does not, in general, penalize all coefficients equally, it can be used to construct an *anisotropic* ridge estimator. By solving the normal equations (see Appendix 2.A.1),

the RLS estimator with positive semi-definite regularization matrix $\Lambda \in \mathbb{R}^{K^2p \times K^2p}$ is shown to be

$$\hat{\boldsymbol{\beta}}^R(\Lambda) = \left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)^{-1} \frac{(Z \otimes I_K)\boldsymbol{y}}{T}.$$

When a centering vector $\boldsymbol{\beta}_0 \neq 0$ is included in penalty $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\Lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, the RLS estimator becomes

$$\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) = \left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)^{-1} \left(\frac{(Z \otimes I_K)\boldsymbol{y}}{T} + \Lambda\boldsymbol{\beta}_0\right). \tag{2.2}$$

In the context of multivariate estimation, one has to make a further distinction between two related types of ridge estimators. I let $\hat{B}^R(\Lambda, \boldsymbol{\beta}_0)$ be the de-vectorized coefficient estimator obtained from reshaping $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ to a $K \times Kp$ matrix. But one can also consider the *matrix RLS estimator* $\hat{B}^R_{\text{mat}}(\Lambda_{Kp}, B_0)$ given by

$$\hat{B}^R_{\text{mat}}(\Lambda_{Kp}, B_0) = T^{-1}(Y + B_0\,\Lambda_{Kp})Z'\left(T^{-1}ZZ' + \Lambda_{Kp}\right)^{-1},$$

where $\Lambda_{Kp} = \text{diag}\{\lambda_1, \ldots, \lambda_{Kp}\}$ and B_0 is a centering matrix. The distinction is important because the vectorized and matrix RLS estimators in general need not coincide. As discussed in Supplementary Appendix 2.A.2, $\hat{B}^R(\Lambda, \boldsymbol{\beta}_0)$ allows for more general penalty structures compared to $\hat{B}^R_{\text{mat}}(\Lambda_{Kp}, B_0)$. I, therefore, focus on the former rather than the latter.

Remark 2.2.1. Equation (2.2) implies that $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ and, therefore, $\hat{\boldsymbol{\beta}}^R(\Lambda)$ provide *simultaneous* estimation of all the coefficients in $\boldsymbol{\beta}$. However, by analogy with ordinary least squares VAR estimation, one may also consider an *equation-by-equation* ridge regression (ebe-RLS) scheme. For $k = 1, \ldots, K$, let $\boldsymbol{y}_k = (Z' \otimes I_K)\boldsymbol{\beta}_k + \boldsymbol{u}_k$ be the autoregressive equation for the kth series of y_t. Then, we can define the kth equation RLS estimator to be

$$\hat{\boldsymbol{\beta}}^R_k(\Lambda, \boldsymbol{\beta}_0) = \left(\frac{ZZ'}{T} \otimes I_K + \Lambda_k\right)^{-1} \left(\frac{(Z \otimes I_K)\boldsymbol{y}_k}{T} + \Lambda\boldsymbol{\beta}_{0k}\right),$$

where $0 \preceq \Lambda_k$ and $\boldsymbol{\beta}_{0k}$ are the kth equation regularizer and centering, respectively. Notice that the ebe-RLS approach allows, by construction, to penalize the estimates for one component differently than for another, and the two can be independently chosen. This provides a higher degree of freedom than the one afforded by, for example, the anisotropic lag-adapted scheme proposed in Section 2.3.2 or the Bayesian schemes of Section 2.4. However, implementing ebe-RLS in applications inherently implies that data-driven tuning of Λ_k will be significantly more computationally intensive – with costs growing linearly in K. Due to this higher complexity, in both theoretical derivation and simulations below, I will focus on studying the properties of the simultaneous RLS estimator.

Remark 2.2.2. Further regarding ebe-RLS, another way to approach estimation is through the *recursive form* of the VAR model. Let $\Sigma_u = P^{-1}DP^{-1\prime}$, where P^{-1} is an unitriangular matrix and D a diagonal matrix, so that we may write

$$Y = GZ - \widetilde{P}Y + D^{-1/2}E,$$

where $G = PB$, $\widetilde{P} = P - I_K$ and noise term E has identity covariance matrix. Estimation can now be performed in ebe-RLS fashion, and matrices P, B and D are recovered (Hausman, 1983). Notice,

however, that in this framework the *ordering* of variables plays a role, since it also determines the decomposition of Σ_u. Indeed, even if a penalization scheme is fixed, permuting the entries may yield different penalized estimates for P, B and D, so that both slope and covariance parameter estimates are different, implying (structural) IRF estimates will also differ. However, note that this issue is somewhat mirrored in a recursive shock identification approach: after estimation, $\widehat{\Sigma}_u$ is Cholesky decomposed to identify the shocks' rotation, and the ordering of variables is key and must be economically justified.

2.3 Shrinkage

In this section, I discuss both the isotropic ridge penalty, i.e. the "standard" ridge approach, and an anisotropic penalty that is better adapted to the VAR setting. An important result is that, even in simple setups with only two variables, the shrinkage induced by ridge can either increase or reduce bias, as well as the stability of autoregressive estimates.

Throughout this section, I consider *fixed* design matrices and the focus will be on the geometric properties of ridge.

2.3.1 Isotropic Penalty

The most common way to perform a ridge regression is through isotropic regularization, that is, $\Lambda = \lambda I$ for some scalar $\lambda \geq 0$. Isotropic ridge has been extensively studied, see for example the comprehensive review of Hastie (2020). With regard to shrinkage, an isotropic ridge penalty can be readily studied.

Proposition 2.3.1. *Let $Z \in \mathbb{R}^{M\times T}$, $Y \in \mathbb{R}^T$ for $T > M$ be regression matrices. For $\lambda_\bullet > \lambda > 0$ and isotropic RLS estimator $\hat{\beta}^R(\lambda) := (T^{-1}ZZ' + \lambda I_M)^{-1}(T^{-1}ZY)$ it holds*

$$\left\|\hat{\beta}^R(\lambda_\bullet)\right\| < \left\|\hat{\beta}^R(\lambda)\right\|.$$

Proof. Using the full singular-value decomposition (SVD), decompose $T^{-1/2}Z = UDV' \in \mathbb{R}^{M\times T}$ where U is $M \times M$ orthogonal, D is $M \times T$ diagonal and V is $T \times T$ orthogonal. Write

$$\begin{aligned}
\hat{\beta}^R(\lambda_\bullet) &= (T^{-1}ZZ' + \lambda_\bullet I_M)^{-1}(T^{-1}ZY) \\
&= (UDV'VDU' + \lambda_\bullet I_M)^{-1}UDV'(T^{-1/2}Y) \\
&= U(D^2 + \lambda_\bullet I_M)^{-1}DV'(T^{-1/2}Y) \\
&= U(D^2 + \lambda_\bullet I_M)^{-1}(D^2 + \lambda I_M)(D^2 + \lambda I_M)^{-1}DV'(T^{-1/2}Y) \\
&= \left[U(D^2 + \lambda_\bullet I_M)^{-1}(D^2 + \lambda I_M)U'\right]\hat{\beta}^R(\lambda).
\end{aligned}$$

Since $D^2 = \operatorname{diag}\{\sigma_j^2\}_{j=1}^M$, the term within brackets is $U\operatorname{diag}\{(\sigma_j^2 + \lambda)/(\sigma_j^2 + \lambda_\bullet)\}_{j=1}^M U'$. Moreover, because the spectral norm is unitary invariant and $\lambda_1 > \lambda_2$, it follows that

$$\left\|U(D^2 + \lambda_\bullet I_M)^{-1}(D^2 + \lambda I_M)U'\right\| = \left\|\operatorname{diag}\{(\sigma_j^2 + \lambda)/(\sigma_j^2 + \lambda_\bullet)\}_{j=1}^M\right\| < 1.$$

Finally, by the sub-multiplicative property it holds

$$\left\|\hat{\beta}^R(\lambda_\bullet)\right\| \leq \left\|U(D^2+\lambda_1 I_M)^{-1}(D^2+\lambda I_M)U'\right\| \cdot \left\|\hat{\beta}^R(\lambda)\right\| < \left\|\hat{\beta}^R(\lambda)\right\|$$

as claimed. □

Proposition 2.3.1 and its proof expose the main ingredients of ridge regression. From the SVD of $T^{-1/2}Z$ used above, it is clear that ridge regularization acts uniformly along the orthogonal directions that are the columns of V. The improvement in the conditioning of the inverse comes from all diagonal factors $[(D^2+\lambda_\bullet I_M)^{-1}D]_j = \sigma_j/(\sigma_j^2+\lambda_\bullet)$ being well-defined, even when $\sigma_j = 0$ (as is the case in systems with collinear regressors).

However, directly applying isotropic ridge to vector autoregressive models is not necessarily the most effective estimation approach. Stable VAR models show decay in the absolute size of coefficients over lags. Thus, it is reasonable to choose a more general ridge penalty that can accommodate lag decay.

2.3.2 Lag-Adapted Penalty

I now consider a different form for Λ that is of interest when applying ridge specifically to a VAR model. Define family $\mathcal{F}^{(p)}$ of *lag-adapted* ridge penalty matrices as

$$\mathcal{F}^{(p)} = \{\text{diag}\{\lambda_1,\dots,\lambda_p\} \otimes I_{K^2} \mid \lambda_i \in \mathbb{R}_+,\, i = 1,\dots,p\},$$

where each λ_i intuitively implies a different penalty for the elements of each coefficient matrix A_i, $i = 1,\dots,p$.[1] The family $\mathcal{F}^{(p)}$ allows imposing a ridge penalty that is coherent with the lag dimension of an autoregressive model. It is parametrized by p distinct penalty factors, meaning that the penalization is *anisotropic*.

Proposition 2.3.2. *Let $Z \in \mathbb{R}^{Kp\times T}$, $\boldsymbol{y} \in \mathbb{R}^{KT}$ for $T > Kp$ be multivariate VAR regression matrices. Given subset $\mathcal{S} \subseteq \{1,\dots,p\}$ of cardinality $s = |\mathcal{S}|$, for $\Lambda^{(p)} \in \mathcal{F}^{(p)}$ define $\hat{\boldsymbol{\beta}}^R(\Lambda^{(p)})_{[\mathcal{S}]}$ as the vector of sK^2 coefficient estimates located at indexes $1+K^2(j-1),\dots,K^2 j$ for $j \in \mathcal{S}$. Let $\mathcal{S}^c = \{1,\dots,p\} \setminus \mathcal{S}$ be the complement of $\mathcal{S}$.*

(a) If $\lambda_1 \geq \lambda_2$, then $\left\|\hat{\boldsymbol{\beta}}^R(\lambda_1 I_{K^2p})_{[\mathcal{U}]}\right\| \leq \left\|\hat{\boldsymbol{\beta}}^R(\lambda_2 I_{K^2p})_{[\mathcal{U}]}\right\|$ for any $\mathcal{U} \subset \{1,\dots,K^2p\}$. The inequality is strict when $\lambda_1 > \lambda_2$.

(b) Let $\hat{\boldsymbol{\beta}}^{LS}_{[\mathcal{S}]}$ be the least squares estimator of the autoregressive model with only the lags indexed by $\mathcal{S}$ included and zeros as coefficients for the lags indexed by $\mathcal{S}^c$. Similarly, let $\Lambda^{(p)}_{[\mathcal{S}]}$ be the

[1] Note that with a lag-adapted penalty it is also possible to directly use the matrix ridge estimator since the penalty for $\hat{\boldsymbol{\beta}}^R$ is given by $\text{diag}\{\lambda_1,\dots,\lambda_p\} \otimes I_{K^2} = (\text{diag}\{\lambda_1,\dots,\lambda_p\} \otimes I_K) \otimes I_K$, see Supplementary Appendix 2.A.2. Importantly, this kind of structure is minimal in terms of modeling the relative size of coefficients *within* each coefficient matrix A_i. If economic theory or intuition provides information about the effects of one specific variable and lag on another – say, the contemporaneous effect of the first series on the second series is zero – more structure can be integrated into the ridge penalty matrix. This would mean, however, that different ridge estimator forms are not equivalent.

subset of diagonal elements in $\Lambda^{(p)}$ *penalizing the lags in* $\mathcal{S}$*. Then*

$$\lim_{\substack{\Lambda^{(p)}_{[\mathcal{S}]}\to 0 \\ \Lambda^{(p)}_{[\mathcal{S}^c]}\to\infty}} \hat{\boldsymbol{\beta}}^R(\Lambda^{(p)}) = \hat{\boldsymbol{\beta}}^{LS}_{[\mathcal{S}]},$$

where $\Lambda^{(p)}_{[\mathcal{S}]} \to 0$ *and* $\Lambda^{(p)}_{[\mathcal{S}^c]} \to \infty$ *are to be intended as the element-wise convergence.*

Proposition 2.3.2 shows that the limiting geometry of a lag-adapted ridge estimator is thus identical to that of a least squares regression run on the subset specified by $\mathcal{S}$. By controlling the size of coefficients $\{\lambda_1, \dots, \lambda_p\}$ it is therefore possible to obtain pseudo-model-selection. However, in the next section, I show that anisotropic penalization produces complex effects on the model's coefficient estimates.

2.3.3 Illustration of Anisotropic Penalization

In this section, I aim to illustrate the effects of a lag-adapted ridge penalty on VAR coefficients estimates using a particular example. This further helps motivate and contextualize the results of the simulation exercises provided in Section 2.7. More generally, before moving on to the discussion of more sophisticated forms of ridge regression, it is important to gain some intuition regrading the properties of anisotropic penalization, which I highlight with the help of a simple bivariate VAR(2) model.

Note that, since ridge operates along principal components, there is no immediate relationship between a specific subset of the estimated coefficients and a given diagonal block of $\Lambda^{(p)}$. With regard to autoregressive modeling, three effects are of interest: the shrinkage of coefficient matrices A_i relative to the choice of λ_i; the entity of the bias introduced by shrinkage, and the impact of shrinkage on the persistence of the estimated model.

In order to showcase these effects, I consider the VAR(2) model

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + u_t, \quad u_t \sim \text{i.i.d. } \mathcal{N}(0, \Sigma_u),$$

where

$$A_1 = \begin{bmatrix} 0.8 & 0.1 \\ -0.1 & 0.7 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0.1 & -0.2 \\ -0.1 & 0.1 \end{bmatrix}, \quad \Sigma_u = \begin{bmatrix} 0.3 & 0 \\ 0 & 5 \end{bmatrix}.$$

A single sample of length $T = 200$ is drawn, demeaned and used to estimate coefficients A_1 and A_2. The VAR(2) model is fitted using the lag-adapted ridge estimator $\hat{B}^R(\Lambda^{(2)})$, where $\Lambda^{(2)} = \text{diag}\{\lambda_1, \lambda_2\} \otimes I_2$. Note that $\hat{B}^R(\Lambda^{(2)})$ can be partitioned into estimates $\hat{A}^R_1(\Lambda^{(2)})$ and $\hat{A}^R_2(\Lambda^{(2)})$ for the respective parameter matrices.

SHRINKAGE. To illustrate shrinkage, I consider the restricted case of $\lambda_1 \in [10^{-2}, 10^6]$ and $\lambda_2 = 0$. The ridge estimator is computed for varying λ_1 over a logarithmically spaced grid. Figure 2.1a shows that $\left\|\hat{B}^R(\Lambda^{(2)})\right\|_F \approx \left\|\hat{B}^{LS}\right\|_F$ for $\lambda_1 \approx 0$, but as the penalty increases $\left\|\hat{A}^R_1(\Lambda^{(2)})\right\|_F$ decreases while $\left\|\hat{A}^R_2(\Lambda^{(2)})\right\|_F$ grows. The resulting behavior of $\left\|\hat{B}^R(\Lambda^{(2)})\right\|_F$ is non-monotonic in λ_1, although indeed $\left\|\hat{B}^R(\Lambda^{(2)})\right\|_F < \left\|\hat{B}^{LS}\right\|_F$ in the limit $\lambda_1 \to \infty$. This effect is due to the model selection

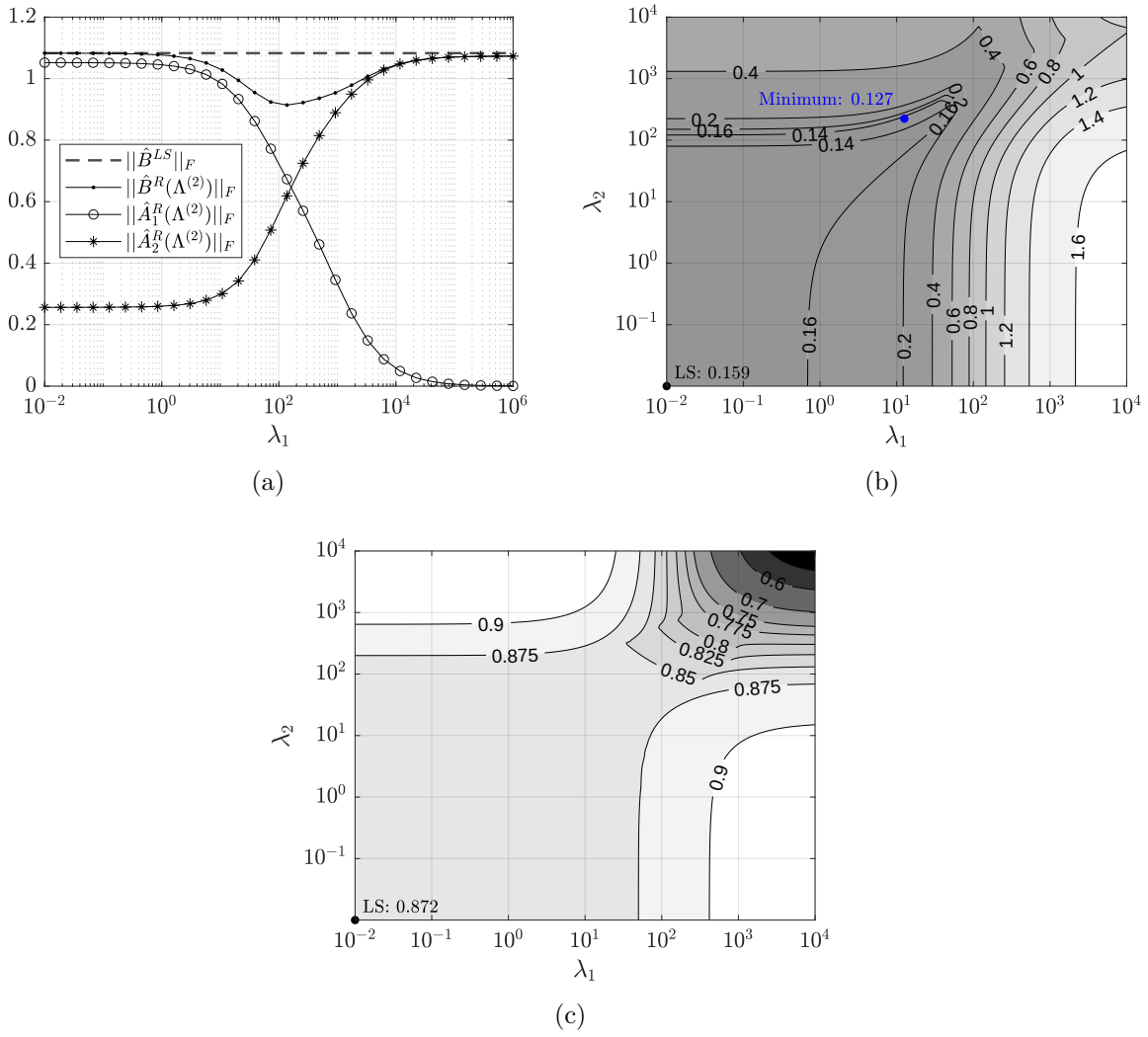


Figure 2.1: Shrinkage of coefficients estimate in Frobenius norm (a); bias induced by shrinkage (b); change in stability of estimated VAR model at different levels of penalization, measured by the absolute value of the largest companion form eigenvalue (c).

properties of lag-adapted ridge, and the resulting omitted variable bias. Therefore, in practice, it is not generally true that anisotropic ridge induces monotonic shrinkage of estimates.

Bias. Since ridge bias is hard to study theoretically, I use a simulation with the same setup of Figure 2.1a, this time with $\lambda_1, \lambda_2 \in [10^{-2}, 10^4]$. The grid is logarithmic with 150 points. Figure 2.1b presents a level plot of the sup-norm ridge bias $\left\|\hat{B}^R(\Lambda^{(2)}) - B\right\|_\infty$ given multiple combinations of λ_1 and λ_2. While there can be gains compared to the least squares estimator $\hat{B}^{LS}$, they are modest. Moreover, level curves of the bias surface show that gains concentrate in a very thin region of the parameter space. Consequently, one may imagine that, in practice, any (data-driven) ridge penalty selection criterion is unlikely to yield bias improvement over least squares. Yet, in large VAR models with many lags, the reduction in variance of the ridge estimator often yields improvements over un-regularized procedures (Li et al., 2023). However, the bias-variance trade-off in ridge is not a free-lunch when performing inference. Pratt (1961) showed that it is not possible

to produce a test (equivalently, a CI procedure) which is valid uniformly over the parameter space and yields meaningfully smaller confidence intervals than any other valid method.

STABILITY. To study the stability of ridge VAR estimates, I reuse the results of the bias simulation above. Let $\mathbb{A}$ be the companion matrix of $[A_1, A_2]$, and $\hat{\mathbb{A}}^R$ the companion matrix of estimates $[\hat{A}_1^R(\Lambda^{(2)}), \hat{A}_2^R(\Lambda^{(2)})]$. For all combinations (λ_1, λ_2), I compute the largest eigenvalue $\omega_1(\hat{\mathbb{A}}^R)$ of $\hat{\mathbb{A}}^R$. Note that if $|\omega_1(\hat{\mathbb{A}})| < 1$, then the estimated VAR(2) is stable (Lütkepohl, 2005). Figure 2.1c presents the level sets for the surface of maximal eigenvalue moduli, and for comparison $|\omega_1(\hat{B}^{LS})|$ is shown at the origin.[2] While along the main diagonal there is a clear decrease in $|\omega_1(\hat{\mathbb{A}}^R)|$ as isotropic penalization increases, when λ_1 is large and $\lambda_2 \ll 1$ (or vice versa) the maximal eigenvalue increases instead. Therefore, an estimate of a VAR model obtained with anisotropic ridge may be *closer* to unit root than the least squares estimate.

2.4 Bayesian and Frequentist Ridge

So far, I have discussed standard ridge penalization schemes. In this section, I study the posterior mean of Bayesian VAR (BVAR) priors commonly applied in the macroeconometrics literature. I show that such posteriors are in fact specific GLS formulations of the ridge estimator. This comparison highlights that ridge can be seen as a way to embed prior knowledge into the least squares estimation procedure by means of centering and rescaling coefficient estimates.

2.4.1 Litterman-Minnesota Priors

In Bayesian time series modeling, the so-called Minnesota or Litterman prior has found great success (Litterman, 1986). For stationary processes which one believes to have reasonably small dependence, a zero-mean normal prior can be put on the VAR parameters, with non-zero prior variance. Assuming that the covariance matrix of errors Σ_u is known, the Litterman-Minnesota has posterior mean

$$\overline{\boldsymbol{\beta}} \mid \Sigma_u = \left[\underline{V}_{\boldsymbol{\beta}}^{-1} + (ZZ' \otimes \Sigma_u^{-1})\right]^{-1} (Z \otimes \Sigma_u^{-1})\boldsymbol{y}, \tag{2.3}$$

where $\underline{V}_{\boldsymbol{\beta}} \succ 0$ is the prior covariance matrix of $\boldsymbol{\beta}$ (Lütkepohl, 2005). It is common to let $\underline{V}_{\boldsymbol{\beta}}$ be diagonal, and often the entries follow a simple pattern which depends on lag, individual components variances, and prior hyperparameters. For example, Bańbura et al. (2010) suggest the following structure for the diagonal

$$v_{i,jk} = \begin{cases} \dfrac{\lambda^2}{i^2} & \text{if } j = k, \\ \theta \dfrac{\lambda^2}{i^2} \dfrac{\sigma_j^2}{\sigma_k^2} & \text{if } j \neq k, \end{cases} \tag{2.4}$$

where $v_{i,jk}$ is the prior variance for coefficients $(A_i)_{jk}$ for $i = 1, \ldots, p$ and $j, k = 1, \ldots, K$. Here, σ_j is the j-th diagonal element of Σ_u, $\theta \in (0, 1)$ specifies beliefs on the explanatory importance of own lags relative to other variables' lags, while $\lambda \in [0, \infty]$ controls the overall tightness of the prior. The extreme $\lambda = 0$ yields a degenerate prior centered at $\overline{\boldsymbol{\beta}} = 0$, while $\lambda = \infty$ reduces the posterior

[2] If $\Lambda^{(2)} \to 0$, then by continuity of eigenvalues it follows that $|\omega_1(\hat{\mathbb{A}}^R)| \to |\omega_1(\hat{\mathbb{A}}^{LS})|$, see Supplementary Appendix 2.A.3.

mean to the OLS estimate $\hat{\boldsymbol{\beta}}^{LS}$. Factor $1/i^2$, which explicitly shrinks variance at higher lags, was originally introduced by De Mol et al. (2008), who formally developed the idea that coefficients at deeper lags should be coupled with more penalizing priors. Note that, in (2.4), assuming $\sigma_j^2 = \sigma_k^2$ for all $j, k = 1, \dots, K$ and setting $\theta = 1$, produces a $\underline{V}_{\boldsymbol{\beta}}$ that has a lag-adapted structure with quadratic lag decay.[3]

Equation (2.3) more generally demonstrates that the Minnesota posterior mean is equivalent to a ridge procedure. It is important to notice that, while with least squares the OLS and GLS estimators of VAR coefficients coincide, this is not the case with ridge regression. Regularizing a GLS regression will yield

$$\hat{\boldsymbol{\beta}}^{RGLS}(\Lambda) := \left[\Lambda + (ZZ' \otimes \Sigma_u^{-1})\right]^{-1} (Z \otimes \Sigma_u^{-1})\boldsymbol{y} \tag{2.5}$$

instead of $\hat{\boldsymbol{\beta}}^{R}$, which is equivalent to (2.3) under an appropriate choice of Λ. While I develop the asymptotic results for $\hat{\boldsymbol{\beta}}^{R}$ assuming a centering parameter $\underline{\boldsymbol{\beta}}_0 \neq 0$ in general, I do not directly study the properties $\hat{\boldsymbol{\beta}}^{RGLS}$. The generalization to GLS ridge employing the least squares error covariance estimator $\hat{\Sigma}_T^{LS}$ should follow from straightforward arguments. In Section 2.7, I focus on providing evidence on the application $\hat{\boldsymbol{\beta}}^{RGLS}$ in terms of its pointwise impulse response estimation mean-squared error.

Remark 2.4.1. In principle, ridge penalties can be designed to implement shrinkage towards nonstationary or long memory priors, too. Very recently, for example, Bauwens et al. (2023) have suggested a ridge type strategy to estimate a one-lag long memory model: their penalization scheme follows naturally from the assumption that an observed AR(1) series originates from an infinite-dimensional VAR(1) with an appropriate off-diagonal structure. One may also think of applying a unit-root-centered matrix RLS estimator $\hat{B}^R_{\text{mat}}(\Lambda_{Kp}, B_0^\dagger)$, where $B_0^\dagger := (I_K, 0_K, \dots, 0_K) \in \mathbb{R}^{K \times Kp}$. This is, in fact, exactly the centering of the Litterman-Minnesota prior (Bańbura et al., 2010). Notice, however, that this type of prior imposes very strict assumptions on the *form* of the unit-root – namely, each component is unaffected by any of the others.[4] Finally, shrinkage to subspaces associated with a factor model specification has also been explored (Huber and Koop, 2023).

2.4.2 Hierarchical Priors

Recent research on Bayesian vector autoregressions exploit more sophisticated priors compared to the Litterman-Minnesota design. Giannone et al. (2015) develop an advanced BVAR model by setting up hierarchical priors which entail not only model parameters, but also hyperparameters. They impose

$$\Sigma_u \;\sim\; \text{IW}(\underline{\Psi}, \underline{d}),$$

[3] Bańbura et al. (2010) also assume $\theta = 1$ in their BVAR estimation. They wish to relax the Litterman-Minnesota assumption that Σ_u is a fixed, diagonal matrix and implement estimation directly by augmenting their data with appropriately constructed dummy variables (Kadiyala and Karlsson, 1997). This approach, however, is selected primarily for computation reasons due to the size of their Bayesian model.

[4] While stationarity of autoregressive estimates can be easily enforced using the Yule-Walker estimator (Brockwell and Davis, 1991), exact unit-root behavior is inherently hard to encode via penalization due to the complex geometry of the stationary region, see the discussion by Heaps (2023).

$$\boldsymbol{\beta} \mid \Sigma_u \sim \mathcal{N}\left(\underline{\boldsymbol{\beta}}, \lambda(\Sigma_u \otimes \underline{\Omega})\right),$$

for hyperparameters $\underline{\boldsymbol{\beta}}$, $\underline{\Omega}$, $\underline{\Psi}$ and $\underline{d}$, where IW is the Inverse-Wishart distribution. Here, too, scalar $\lambda \in [0, \infty]$ controls prior tightness. Let $\underline{B}$ be the matrix form of the VAR coefficient prior mean, so that $\text{vec}(\underline{B}) = \underline{\boldsymbol{\beta}}$. The resulting (conditional) posterior mean $\overline{B}$ is given by

$$\overline{B} \mid \Sigma_u = \left[(\lambda\underline{\Omega})^{-1} + ZZ'\right]^{-1} \left[ZY + (\lambda\underline{\Omega})^{-1}\underline{B}\right]. \tag{2.6}$$

Observe that equation (2.6) is effectively equivalent to a centered ridge estimator, c.f. (2.2).

The introduction of a hierarchical prior leaves space to add informative hyperpriors on the model hyperparameters, allowing for a more flexible fit. Indeed, removing the zero centering constraint from the prior on $\boldsymbol{\beta}$ can improve estimation. It is often the case that economic time series show a high degree of correlation and temporal dependence, therefore imposing $\boldsymbol{\beta} = 0$ as in the Minnesota prior is inadequate. In fact, Giannone et al. (2015) show that their approach yields substantial improvements in forecasting exercises, even when hyperparameter priors are relatively flat and uninformative.

2.5 Standard Inference

In this section, I state the main asymptotic results for the RLS estimator $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ with general regularization matrix Λ. I shall allow Λ and non-zero centering coefficient $\boldsymbol{\beta}_0$ to be, under appropriate assumptions, random variables dependent on sample size T. In particular, $\boldsymbol{\beta}_0$ may be a consistent estimator of $\boldsymbol{\beta}$.

I will impose the following assumptions.

Assumptions

A. $\{u_t\}_{t=1}^T$ is a sequence of i.i.d. random variables with $\mathbb{E}[u_{it}] = 0$, covariance $\mathbb{E}[u_t u_t'] = \Sigma_u$ non-singular positive definite and $\mathbb{E}\left|u_{it}u_{jt}u_{mt}u_{nt}\right| < \infty$, $i, j, m, n = 1, \ldots, K$.

B. There exists $\rho > 1$ such that $\det(I_K - \sum_{i=1}^p A_i z^i) \neq 0$ for all complex z, $|z| \leq \rho$.

C. There exist $0 < \underline{m} \leq \overline{m} < \infty$ such that $\underline{m} \leq \omega_K(\Gamma) \leq \omega_1(\Gamma) \leq \overline{m}$, where $\Gamma = \mathbb{E}[z_t z_t']$ is the autocovariance matrix of z_t and $\omega_1(\Gamma)$, $\omega_K(\Gamma)$ are its largest and smallest eigenvalues, respectively.

Assumption A is standard and allows proving the main asymptotic results with well-known theoretical devices. Assuming u_t is white noise or assuming y_t respects strong mixing conditions (Davidson, 1994) would require more careful consideration in asymptotic arguments but is otherwise a simple generalization, although more involved in terms of notation, see e.g. Boubacar Mainassara and Francq (2011). Assumption B guarantees that y_t has no unit roots and is stable. Of course, many setups of interest do not satisfy this assumption, the most significant ones being unit roots, cointegrated VARs, and local-to-unity settings. Incorrect identification of unit roots does not invalidate the use of LS or ML estimators (Phillips, 1988, Park and Phillips, 1988, 1989, Sims et al., 1990), however inference is significantly impacted as a result (Pesavento and Rossi, 2006, Mikusheva, 2007, 2012). Assumption C is standard in the literature regarding penalized estimation

and does not imply significant additional constraints on the process y_t, c.f. Assumption A. It is sufficient to ensure that for large enough T the plug-in sample autocovariance estimator is invertible even under vanishing Λ.

Before stating the main theorems, let

$$\begin{aligned}\hat{\Gamma} &= T^{-1}ZZ', \\ \hat{U} &= Y - \hat{B}^R Z, \\ \hat{\Sigma}_u^R &= T^{-1}\hat{U}\hat{U}',\end{aligned}$$

be the regression covariance matrix, regression residuals and sample innovation covariance estimator, respectively.

Theorem 2.5.1. *Let Assumptions A-C hold and define $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ be the centered RLS estimator as in (2.2). If $\sqrt{T}\Lambda \xrightarrow{P} \Lambda_0$ and $\boldsymbol{\beta}_0 \xrightarrow{P} \underline{\boldsymbol{\beta}}_0$, where Λ_0 is a positive semi-definite diagonal matrix and $\underline{\boldsymbol{\beta}}_0$ is a constant vector, then*

(a) $\hat{\Gamma} \xrightarrow{P} \Gamma$,

(b) $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) \xrightarrow{P} \boldsymbol{\beta}$,

(c) $\hat{\Sigma}_u^R \xrightarrow{P} \Sigma_u$,

(d) $\sqrt{T}\left(\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\Gamma^{-1}\Lambda_0(\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}), \Gamma^{-1} \otimes \Sigma_u\right)$.

Theorem 2.5.1 considers the most general case, and, as previously mentioned, gives the asymptotic distribution of $\hat{\boldsymbol{\beta}}^R$ under rather weak conditions for the regularizer Λ. The resulting normal limit distribution is clearly dependent on the unknown model parameters $\boldsymbol{\beta}$, complicating inference.

However, it is possible – under strengthened assumptions for Λ or $\boldsymbol{\beta}_0$ – for $\hat{\boldsymbol{\beta}}^R$ to have a zero-mean Gaussian limit distribution.

Theorem 2.5.2. *In the setting of Theorem 2.5.1, results (a)-(c) hold and (d) simplifies to*

(d′) $\sqrt{T}\left(\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(0, \Gamma^{-1} \otimes \Sigma_u\right)$

if either

(i) $\Lambda = o_P\left(T^{-1/2}\right)$,

(ii) $\Lambda = O_P\left(T^{-1/2}\right)$ *and* $\boldsymbol{\beta}_0 - \boldsymbol{\beta} = o_p(1)$.

The following corollary is immediate.

Corollary 2.5.3. *Let $\hat{\boldsymbol{\beta}}_0$ be a consistent and asymptotically normal estimator of $\boldsymbol{\beta}$. Then, under condition* (i) *or* (ii) *of Theorem 2.5.2 results* (a)-(d′) *hold.*

2.5.1 Joint Inference

To handle smooth transformations of VAR coefficients, such as impulse responses (Lütkepohl, 1990), I also derive a standard joint limit result for both $\hat{\boldsymbol{\beta}}^R$ and the variance estimator $\hat{\Sigma}_u^R$.

Theorem 2.5.4. *Let $\hat{\boldsymbol{\sigma}}^R = \text{vec}(\hat{\Sigma}_u^R)$ and $\boldsymbol{\sigma} = \text{vec}(\Sigma_u)$. Under the assumptions of Theorem 2.5.1,*

$$\sqrt{T}\begin{bmatrix}\hat{\boldsymbol{\beta}}^R - \boldsymbol{\beta} \\ \hat{\boldsymbol{\sigma}}^R - \boldsymbol{\sigma}\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix}\Gamma^{-1}\Lambda_0(\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) \\ 0\end{bmatrix}, \begin{bmatrix}\Gamma^{-1}\otimes\Sigma_u & 0 \\ 0 & \Omega\end{bmatrix}\right).$$

Under assumption (1) or (2) of Theorem 2.5.2,

$$\sqrt{T}\begin{bmatrix}\hat{\boldsymbol{\beta}}^R - \boldsymbol{\beta} \\ \hat{\boldsymbol{\sigma}}^R - \boldsymbol{\sigma}\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix}\Gamma^{-1}\otimes\Sigma_u & 0 \\ 0 & \Omega\end{bmatrix}\right),$$

where $\Omega = \mathbb{E}\left[\text{vec}(u_t u_t')\,\text{vec}(u_t u_t')'\right] - \boldsymbol{\sigma}\boldsymbol{\sigma}'$.

This result is key as it allows, under the stated assumptions on the penalizer, to construct valid asymptotic confidence intervals and, specifically, perform impulse response inference, as done in the simulations of Section 2.7 using the Delta Method (Lütkepohl, 2005).

2.5.2 Cross-validation

In practice, the choice of ridge penalty is often data-driven, and cross-validation is a very popular approach to select Λ. I now turn to the properties of CV as applied to the RLS estimator $\hat{\boldsymbol{\beta}}^R(\Lambda)$.

For simplicity, assume that y_t is an AR(p) process, that is, $K = 1$. In this setting,

$$\hat{\boldsymbol{\beta}}^R(\Lambda) = \left(\frac{ZZ'}{T} + \Lambda\right)^{-1}\frac{Z\boldsymbol{y}}{T},$$

where $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$. Following Patil et al. (2021), the *prediction error* of ridge estimator $\hat{\boldsymbol{\beta}}^R(\Lambda)$ given penalty Λ is

$$\text{Err}\left(\hat{\boldsymbol{\beta}}^R(\Lambda)\right) := \mathbb{E}_{\tilde{y},\tilde{z}}\left[\left(\tilde{y} - \tilde{z}'\hat{\boldsymbol{\beta}}^R(\Lambda)\right)^2 | Z, \boldsymbol{y}\right],$$

where $\tilde{y}$ and $\tilde{z}$ are random variables from an independent copy of y_t. In particular, $\tilde{z}$ is the vector of p lags of $\tilde{y}$. Moreover, the error curve for Λ is given by

$$\text{err}(\Lambda) := \text{Err}\left(\hat{\boldsymbol{\beta}}^R(\Lambda)\right).$$

The prediction error is crucial because it allows to determine the oracle optimal penalization,

$$\Lambda^* := \arg\min_{\Lambda\succeq 0}\ \text{err}(\Lambda).$$

Clearly, $\text{err}(\Lambda)$ is unavailable in practice and Λ^* must be substituted with a feasible alternative. Cross-validation proposes to construct a collection of paired, non-overlapping subsets of the sample data such that the first subset of the pair (estimation set) is used to estimate the model, while the second (validation set) is used to provide an empirical estimate of the prediction error. The CV penalty is then selected to minimize the total error over validation sets. A very popular approach to

build cross-validation subsets is k-fold CV, wherein the sample is split into k blocks, so-called *folds*, of sequential observations (possibly after shuffling the data). Each fold determines a validation set, and is paired with its complement, which gives the estimation set. For more details, see e.g. Hastie et al. (2009).

Again with the intent of keeping complexity low – as this work is not focused on cross-validation – I will make the additional simplifying assumption that CV is implemented with two folds and one pair. Specifically, the first fold is the estimation set, where Z and $\boldsymbol{y}$ are constructed and $\hat{\boldsymbol{\beta}}^R(\Lambda)$ is estimated. The second fold is the validation set and yields $\widetilde{Z}$, $\widetilde{\boldsymbol{y}}$, where $\widetilde{Z} \in \mathbb{R}^{p\times\widetilde{T}}$ and $\widetilde{\boldsymbol{y}} \in \mathbb{R}^{\widetilde{T}}$. To account for dependence, a buffer of m observations between validation and estimation folds is introduced. The last observation of y_t in the estimation set is y_T, while the first observation in the validation set is $\widetilde{y}_1 := y_{T+m+1}$, that is, the total number of available observations is $T+m+\widetilde{T}+2p+1$. This is a stylized version of the CV setup of Burman et al. (1994) – also called *m-block* or *non-dependent cross-validation* in Bergmeir et al. (2018b) – and is effectively equivalent to an out-of-sample (OOS) validation scheme. Thus, the 2-fold m-buffered CV error curve is

$$\mathrm{cv2}_m(\Lambda) := \frac{1}{\widetilde{T}}\sum_{s=1}^{\widetilde{T}}\left(\widetilde{y}_s - \widetilde{z}_s'\hat{\boldsymbol{\beta}}^R(\Lambda)\right)^2. \tag{2.7}$$

Theorem 2.5.5. *Under Assumptions A-C, for every Λ in the cone of diagonal positive definite penalty matrices with diagonal entries in $(\lambda_{\min},\infty)$, $\lambda_{\min}\geq 0$, it holds that*

$$\mathrm{cv2}_m(\Lambda) - \mathrm{err}(\Lambda) \xrightarrow{a.s.} 0$$

as $T,\widetilde{T}\to\infty$. Furthermore, the convergence is uniform in Λ over compact subsets of penalty matrices.

In the current setup, the joint limit $T,\widetilde{T}\to\infty$ should be thought as $\widetilde{T}/T\to\gamma\in(0,1)$, where aspect ratio γ determines the balance of the cross-validation split.

Remark 2.5.1. Under Assumption C, $\omega_K(\hat{\Gamma})>0$ for T large. Therefore, the bounds derived in the proof of Theorem 2.5.5 are finite even if $\Lambda=0$. In fact, it is easily seen that the behavior of $\mathrm{err}(\Lambda)$ and $\mathrm{cv2}_m(\Lambda)$ is consistent at the endpoints $\Lambda=0$ and $\Lambda\to\infty$, see Patil et al. (2021). Observe that

$$\mathrm{cv2}_m(\Lambda)\to\Sigma_u \quad\text{and}\quad \mathrm{err}(\Lambda)\to\Sigma_u$$

as $\Lambda\to 0$, while

$$\mathrm{cv2}_m(\Lambda)\to\Gamma \quad\text{and}\quad \mathrm{err}(\Lambda)\to\Gamma$$

as $\Lambda\to\infty$, as needed.

Theorem 2.5.5 thus shows that $\mathrm{cv2}_m(\Lambda)$ gives an asymptotically valid way to evaluate the prediction error curve, and thus tune Λ, over any compact set of diagonal positive semi-definite penalization matrices. Moreover, in Theorem C.2.1, Supplementary Appendix 2.C.2, I show that the impact of dependence due to the VAR data generating process is exponentially small for m sufficiently large. This property of $\mathrm{cv2}_m(\Lambda)$ is desirable because it lets one choose m small also in applications with moderate sample sizes, and it theoretically justifies the prescription of Burman

et al. (1994).

2.5.3 Asymptotically Valid CV

So far, I have shown that a simple 2-fold CV – or, equivalently, out-of-sample validation – correctly estimates the predictive error of the ridge estimator, even under dependence. I turn now to the question of selecting an *asymptotically valid* penalty, that is, a Λ such that condition (1) of Theorem 2.5.2 is fulfilled. This enables inference, since one is in a setting where the bias is asymptotically negligible.

The idea is to scale the ridge penalty used at the estimation step of CV by a factor $\sqrt{T}$, so that the validated penalty converges to zero at an appropriate rate as both T and $\widetilde{T}$ grow. In other words, an over-smoothed ridge regression turns out to be key when studying cross-validation. To derive this result, first let

$$\hat{\boldsymbol{\beta}}^R_{\blacklozenge}(\Lambda) := \left(\frac{ZZ'}{T} + \sqrt{T}\Lambda\right)^{-1} \frac{Z\boldsymbol{y}}{T}$$

be the *over-smoothed ridge estimator*.

Theorem 2.5.6. *Under Assumptions A-C, let $\mathcal{I}_\lambda$ be the compact set of diagonal positive semidefinite penalization matrices Λ such that $\|\Lambda\|_{\max} \leq \lambda < \infty$. It holds*

$$\Lambda^*_{\blacklozenge} := \arg\min_{\Lambda \in \mathcal{I}_\lambda} \operatorname{Err}\left(\hat{\boldsymbol{\beta}}^R_{\blacklozenge}(\Lambda)\right) = o_p(T^{-1/2}).$$

Remark 2.5.2. The previous theorem is stated in terms of the oracle predictive error $\operatorname{Err}\left(\hat{\boldsymbol{\beta}}^R_{\blacklozenge}\left(\widetilde{\Lambda}\right)\right)$, which equals the 2-fold CV error curve up to a factor of order $O_P(\widetilde{T}^{-1/2})$. Therefore, assuming that the CV aspect ratio γ is strictly between zero and one, the result of Theorem 2.5.6 also directly generalizes to an empirically cross-validated penalty.

2.6 Inference with Shrinkage

Fu and Knight (2000) have argued that results such as Theorems 2.5.1 and 2.5.2 portray penalized estimators in a somewhat unfair light, because they result in asymptotic distributions showing no bias-variance trade-off. Indeed, they show that ridge shrinkage yields estimates with asymptotic variance no different from that of least squares. Of course, in finite samples shrinkage has an effect on $\Gamma^{-1} \otimes \Sigma_u$ since $\hat{\Sigma}^R_T$ is used in place of $\hat{\Sigma}^{LS}_T$ to estimate the error term variance matrix. To better understand the value of ridge penalization in practice, one should therefore consider the situation where shrinkage is *not* asymptotically negligible for at least a subset of coefficients. A motivating example would be that of a VAR(∞) model derived by inverting a stable VARMA(p, q) process: for i sufficiently large, coefficient matrices A_i decay exponentially to zero.[5] One should thus be able to exploit such structural information about the autoregressive coefficients to asymptotically improve on the bias-variance trade-off. Following this intuition and the discussion of lag-adapted penalty matrices in Section 2.3.2, I shall now consider the empirically relevant regression setup

[5] This result follows from a straightforward generalization of Lemma 2.C.1 in Supplementary Appendix 2.C. The choice of norm to measure such decay is not fundamental, as they are equivalent given that dimension K is fixed.

where one assumes that a subset of VAR coefficients are small (with respect to sample size), but not necessarily zero. Thus, to have inference reflect this type of shrinkage, an asymptotic framework with non-negligible penalization of higher-order lag coefficients is in fact more appropriate than that of Theorem 2.5.1.[6]

Formally, assume that for some $0 < n \leq p$ one can partition the VAR coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1 \in \mathbb{R}^{K^2(p-n)}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{K^2 n}$, and assume that $\boldsymbol{\beta}_2 = T^{-(1/2+\delta)}\, \boldsymbol{b}_2$ for $\delta > 0$ and $\boldsymbol{b}_2 \in \mathbb{R}^{K^2 n}$ is fixed. Such ordered partitioning of $\boldsymbol{\beta}$ is without loss of generality.[7] In this setup, it is clearly desirable to penalize $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ differently when constructing the ridge penalty. Let $\Lambda = \text{diag}\{(L_1', L_2')'\} \otimes I_K$ where $L_1 \in \mathbb{R}_+^{K^2(p-n)}$ and $L_2 \in \mathbb{R}_+^{K^2 n}$. Assume that

$$L_1 = o_P(T^{-1/2}) \quad \text{and} \quad L_2 \xrightarrow{P} \overline{L}_2$$

for a fixed vector $\overline{L}_2 \in \mathbb{R}_+^{K^2 n}$. In particular, letting $\Lambda_1 = \text{diag}\{L_1\}$ and $\Lambda_2 = \text{diag}\{L_2\}$,

$$\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \otimes I_K \xrightarrow{P} \overline{\Lambda} \otimes I_K \quad \text{where} \quad \overline{\Lambda} = \begin{bmatrix} 0 & 0 \\ 0 & \overline{\Lambda}_2 \end{bmatrix},\ \overline{\Lambda}_2 \succ 0. \tag{2.8}$$

One can now develop an asymptotic result which shows non-negligible shrinkage in the limit distribution of the ridge estimator. For simplicity of exposition, here I will assume that ridge centering $\boldsymbol{\beta}_0$ is chosen to be zero.

Theorem 2.6.1. *In the setting of Theorem 2.5.1, assume that, for $0 < n \leq p$,*

(i) $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ *where* $\boldsymbol{\beta}_1 \in \mathbb{R}^{K^2(p-n)}$ *and* $\boldsymbol{\beta}_2 = T^{-(1/2+\delta)}\, \boldsymbol{b}_2$ *for* $\delta > 0$, $\boldsymbol{b}_2 \in \mathbb{R}^{K^2 n}$ *fixed.*

(ii) $\Lambda = \text{diag}\{(L_1', L_2')'\}$ *where* $L_1 \in \mathbb{R}_+^{K^2(p-n)}$ *and* $L_2 \in \mathbb{R}_+^{K^2 n}$.

(iii) $L_1 = o_P(T^{-1/2})$ *and* $L_2 \xrightarrow{P} \overline{L}_2$ *as* $T \to \infty$.

(iv) $\boldsymbol{\beta}_0 = 0$.

Let $\Gamma_{\overline{\Lambda}} = \Gamma + \overline{\Lambda}$ *where* $\overline{\Lambda} \succeq 0$ *is given by* (2.8). *Then, results (a)-(c) hold and*

(d″) $\sqrt{T}\left(\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(0, \Gamma_{\overline{\Lambda}}^{-1}\, \Gamma\, \Gamma_{\overline{\Lambda}}^{-1} \otimes \Sigma_u\right)$

It is easy to see that indeed the term $\Gamma_{\overline{\Lambda}}^{-1}\, \Gamma\, \Gamma_{\overline{\Lambda}}^{-1}$ in Theorem 2.6.1 is weakly smaller than Γ^{-1} in the positive-definite sense. Note that

$$\begin{aligned} \Gamma_{\overline{\Lambda}}^{-1}\, \Gamma\, \Gamma_{\overline{\Lambda}}^{-1} \preceq \Gamma^{-1} &\iff (\Gamma + \overline{\Lambda})^{-1}\Gamma \preceq \Gamma^{-1}(\Gamma + \overline{\Lambda}) \\ &\iff I_{K^2 p} - (\Gamma + \overline{\Lambda})^{-1}\overline{\Lambda} \preceq I_{K^2 p} + \Gamma^{-1}\overline{\Lambda} \\ &\iff 0 \preceq ((\Gamma + \overline{\Lambda})^{-1} + \Gamma^{-1})\overline{\Lambda} \end{aligned}$$

The last inequality is true by definition of $\overline{\Lambda}$. Shrinkage gains are concentrated at the components that have non-zero asymptotic shrinkage, i.e. those penalized by $\overline{L}_2$.

[6] Such an approach to inference also follows De Mol et al. (2008), who argue for explicit lag penalization within BVAR priors on similar theoretical grounds. In the context of maximum-likelihood estimation, the use of appropriate and plausible model restrictions to improve efficiency by shrinkage, rather than perform hypothesis testing, has also been discussed by Hansen (2016a).

[7] The dimensions of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are chosen to be multiples of K^2 to better conform to the lag-adapted setting. This choice is also without loss of generality and simplifies exposition.

Remark 2.6.1. A key point in the application of Theorem 2.6.1 is identification of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. In practice, one may then proceed in two ways. As discussed in Section 2.4, one can see the ridge approach as a frequentist "counterpart" to implementing a Bayesian prior. Therefore, the researcher may split $\boldsymbol{\beta}$ into subsets of small and large parameters based on economic intuition, domain knowledge or preliminary information. Alternatively, in the following section, I show that cross-validation is able to automatically tune Λ appropriately.

Finally, it is immediate to generalize the argument of Theorem 2.6.1 to the case where $\boldsymbol{\beta}$ is not split into subsets based on the relative size of coefficients, but rather a non-zero, *partially consistent* centering sequence $\boldsymbol{\beta}_0$ is used.

Corollary 2.6.2. *Consider the setup of Theorem 2.6.1, where now assumptions (i) and (iv) are replaced by*

(i′) $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ *where* $\boldsymbol{\beta}_1 \in \mathbb{R}^{K^2(p-n)}$ *and* $\boldsymbol{\beta}_2 \in \mathbb{R}^{K^2 n}$ *are fixed.*

(iv′) $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}', \boldsymbol{\beta}_{02}')'$ *where* $\boldsymbol{\beta}_{01} \in \mathbb{R}^{K^2(p-n)}$ *is such that* $\boldsymbol{\beta}_{01} \neq \boldsymbol{\beta}_1$*, and* $\boldsymbol{\beta}_{02} = \boldsymbol{\beta}_2 + T^{-(1/2+\delta)}\boldsymbol{b}_2$ *for* $\delta > 0$*,* $\boldsymbol{b}_2 \in \mathbb{R}^{K^2 n}$ *fixed.*

Then, results (a)-(c) and (d″) still hold.

2.6.1 Cross-validation with Partitioned Coefficients

One can use the same approach applied to derive Theorem 2.5.6 to show that cross-validating the RLS estimator with $\text{Err}(\hat{\boldsymbol{\beta}}^R_{\blacklozenge}(\Lambda))$ is also asymptotically valid under partitioning.

Corollary 2.6.3. *Consider the setup of Theorem 2.6.1 and assume that the assumptions of Theorem 2.5.6 are met. It holds*

$$\begin{bmatrix} \Lambda_{1,\blacklozenge} & 0 \\ 0 & \Lambda_{2,\blacklozenge} \end{bmatrix} := \arg\min_{\Lambda \in \mathcal{I}_\lambda} \text{Err}\left(\hat{\boldsymbol{\beta}}^R_{\blacklozenge}(\Lambda)\right) = \begin{bmatrix} o_p(T^{-1/2}) & 0 \\ 0 & o_P(1) \end{bmatrix}$$

Moreover, any $\Lambda_{2,\blacklozenge}$ *such that* $0 \preceq \Lambda_{2,\blacklozenge} \preceq \lambda I$ *is asymptotically valid.*

In theory, one would like to be able to quantify the gains obtained in the asymptotic shrinkage setup of Theorem 2.6.1 compared to the standard setting of Theorems 2.5.1 and 2.5.2, particularly when using cross-validation. Unfortunately, it is in general hard to study the cross-validation error loss even in setups without dependence. Stephenson et al. (2021), in fact, show that the ridge leave-one-out CV loss is not generally convex. This suggests that studying the behavior of CV when penalizing with a diagonal anisotropic Λ can be a very complex task in a finite sample setup.

2.7 Simulations

To study the performance of ridge-regularized estimators, I now perform simulation exercises focused on impulse response functions (IRFs). Throughout the experiments I will consider structural impulse responses, and I assume that identification can be obtained in a recursive way (Kilian and

Table 2.1: List of Estimation Methods

Type	Name	Description
Frequentist	LS	Least squares estimator
	RIDGE	Ridge estimator, CV penalty
	RIDGE-GLS	GLS ridge estimator, CV penalty
	RIDGE-AS	Ridge estimator with asymptotic shrinkage, CV penalty
	LP	Local projections with Newey-West covariance estimate
Bayesian	BVAR-CV	Litterman-Minnesota Bayesian VAR, CV tightness prior
	H-BVAR	Hierarchical Bayesian VAR of Giannone et al. (2015)

Lütkepohl, 2017a), which is a widely used approach for structural shock identification in macroeconometrics.

I consider two setups:

1. The three-variable VARMA(1,1) design of Kilian and Kim (2011), representing a small-scale macro model. I term this setup "A".

2. A VAR(5) model in levels, using the model specification of Giannone et al. (2015) with the dataset of Hansen (2016b) consisting of $K = 7$ variables in levels.[8] I term this setup "B". For the ease of exposition, in the discussion I will tabulate results only for three variables – real GDP, investment and federal funds rate – but complete tables can be found in Supplementary Appendix 2.D.5.

The specification of Kilian and Kim (2011) has already been extensively used in the literature as a benchmark to gauge the basic properties of inference methods. On the other hand, the estimation task of Giannone et al. (2015) involves more variables and a higher degree of persistence. This setting is useful to evaluate the effects of ridge shrinkage when applied to realistic macroeconomic questions. It is also a suitable test bench to compare Bayesian methods with frequentist ridge.

Estimators. For frequentist methods, I include both $\hat{\boldsymbol{\beta}}^R$ and $\hat{\boldsymbol{\beta}}^{RGLS}$ ridge estimators as well as the local projection estimator of Jordà (2005). For Bayesian methods, I implement both the Minnesota prior approach of Bańbura et al. (2010) with stationary prior and the hierarchical prior BVAR of Giannone et al. (2015).[9] The full list of method I consider is given in Table 2.1. To make methods comparable, I have extended the ridge estimators to include an intercept in the regression. A precise discussion regarding the tuning of penalties and hyperparameters of all methods can be found in Appendix 2.D.

[8]The dataset is supplied by the author at `https://users.ssc.wisc.edu/~bhansen/progs/var.html`. While the data provided by Hansen (2016b) includes releases until 2016, I do not include more recent quarterly data since this is a simulation exercise. Moreover, due to the effects of the COVID-19 global pandemic, an extended sample would likely only add data released until Q4 2019 due to overwhelming concerns of a break point.

[9]To estimate hierarchical prior BVARs I rely on the original MATLAB implementation provided by Giannone et al. (2015) on the authors' website at `http://faculty.wcas.northwestern.edu/gep575/GLPreplicationWeb.zip`.

2.7.1 Pointwise MSE

The first two simulation designs explore the MSE performance of ridge-type estimators versus alternatives. Let $\theta_{km}(h)$ be the horizon h structural IRF for variable k given a unit shock from variable m. To compute the MSE for each k, define

$$\text{MSE}_k(h) := \sum_{m=1}^{K} \mathbb{E}\left[\left(\hat{\theta}_{km}(h) - \theta_{km}(h)\right)^2\right],$$

which is the total MSE for the kth variable over all possible structural shocks. In simulations, I use B replications to estimate the expectation. All MSEs are normalized by the mean squared error of the least squares estimator.

Setup A. A time series of length $T = 200$ is generated a number $B = 10\,000$ of times for replication. All VAR estimators are computed using $p = 10$ lags, while LPs include $q = 10$ regression lags. Table 2.2 shows relative MSEs for this design. It is important to notice that, in this situation, GLS ridge has remarkably low performance at horizon $h = 1$ compared to other methods. The primary issue is that Σ_u features strong correlation between components, and thus the diagonal lag-adapted structure does not shrink along the appropriate directions. This is much less prominent as the horizon increases due to the fact that impulse responses eventually decay to zero, since the underlying VARMA DGP is stationary. While there is no clear ranking, the MSE of the baseline ridge VAR estimator is in between those of the BVAR and hierarchical BVAR approaches. The degrading quality of local projection estimates are mainly due to the smaller samples available in regressions at each increasing horizon (Kilian and Kim, 2011). This behavior is one of the prime reasons behind the development of LP shrinkage estimators, like that proposed in Plagborg-Møller (2016) or the SLP estimator of Barnichon and Brownlees (2019).

Setup B. Using the data of Hansen (2016b), I estimate and simulate a stationary but highly persistent VAR(5) model using the same sample size and number of replications as Setup A. For all methods, $p = 5$ lags are used, so that VAR estimators are correctly specified. The results can be found in Table 2.3. In this setup, unlike in the previous experiment, one can clearly notice that impulse responses computed via cross-validated ridge show increasing MSE as horizon h grows. There are two main reasons behind this behavior. First, the chosen setup features a very persistent data generating process, as the largest root of the underlying VAR model is 0.9945. This means that the true IRFs revert to zero only over long horizons, while lag-adapted ridge estimates yields models with lower persistence and thus flatter impulse responses. Secondly, the dataset from Hansen (2016b) is not normalized, and the included series have markedly heterogenous variances. Since GLS ridge shrinks along covariance-rotated data, shrinkage is adjusted according to each series variance, unlike that baseline ridge estimator $\hat{\boldsymbol{\beta}}^R$. The MSE for the Fed Fund Rate impulse responses shows that the pointwise difference between baseline and GLS ridge can be severe for long horizon IRFs when the DGP is highly persistent. On short horizons, Bayesian estimators perform on par or better than baseline least squares estimates, while at longer horizons differences are less stark. It is, however, clear that the hierarchical prior BVAR of Giannone et al. (2015) shows the

Table 2.2: MSE Relative to OLS – Setup A

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Investment Growth	RIDGE	0.97	0.74	0.64	0.64	0.65	0.63	0.60
	RIDGE-GLS	5.16	0.89	0.55	0.47	0.44	0.41	0.38
	LP	1.00	1.05	1.13	1.52	2.15	3.20	4.87
	BVAR-CV	1.55	0.84	0.70	0.70	0.71	0.70	0.66
	H-BVAR	1.80	0.66	0.53	0.52	0.54	0.53	0.50
Deflator	RIDGE	0.93	0.78	0.69	0.68	0.67	0.64	0.59
	RIDGE-GLS	2.43	0.83	0.59	0.52	0.48	0.44	0.40
	LP	1.00	1.05	1.13	1.44	1.99	2.90	4.47
	BVAR-CV	1.03	0.89	0.74	0.73	0.73	0.70	0.66
	H-BVAR	1.01	0.70	0.58	0.56	0.55	0.53	0.50
Paper Rate	RIDGE	0.94	0.76	0.66	0.66	0.66	0.64	0.60
	RIDGE-GLS	1.80	0.87	0.59	0.52	0.47	0.43	0.39
	LP	1.00	1.05	1.13	1.46	1.99	2.86	4.31
	BVAR-CV	0.87	0.87	0.74	0.73	0.73	0.71	0.66
	H-BVAR	0.81	0.69	0.57	0.55	0.56	0.54	0.51

Table 2.3: MSE Relative to OLS – Setup B

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Real GDP	RIDGE	1.11	1.08	1.16	1.06	0.90	0.89	0.94
	RIDGE-GLS	1.16	1.00	0.99	1.00	0.93	0.93	0.95
	LP	1.00	1.14	1.37	1.52	1.72	1.98	2.24
	BVAR-CV	0.90	0.87	1.04	1.01	0.92	0.92	0.98
	H-BVAR	0.83	0.62	0.78	0.73	0.62	0.62	0.68
Investment	RIDGE	1.49	1.27	1.17	0.99	0.70	0.73	1.61
	RIDGE-GLS	1.34	1.14	1.02	1.02	0.86	0.82	0.86
	LP	1.00	1.15	1.40	1.63	2.03	2.76	3.59
	BVAR-CV	1.51	1.01	0.97	0.97	0.93	1.08	1.24
	H-BVAR	1.06	0.68	0.69	0.66	0.63	0.87	1.14
Fed Funds Rate	RIDGE	2.17	1.21	0.96	0.93	1.03	4.00	53.18
	RIDGE-GLS	1.21	1.04	0.90	0.93	0.90	0.88	0.91
	LP	1.00	1.18	1.51	1.71	1.97	2.44	2.99
	BVAR-CV	0.92	0.94	0.91	0.90	0.86	0.87	0.92
	H-BVAR	0.75	0.77	1.32	1.38	1.25	1.15	1.20

overall best results. As in the previous setup, local projections show degrading performance at larger horizons.

Remark 2.7.1. The comparison between methods in both Setup A and Setup B is largely consis-

Table 2.4: Impulse Response Inference – Setup A – CI Coverage

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Investment Growth	LS	0.88	0.88	0.87	0.88	0.91	0.93	0.94
	RIDGE	0.90	0.92	0.94	0.93	0.94	0.95	0.95
	RIDGE-AS	0.90	0.92	0.88	0.88	0.88	0.89	0.89
	LP	0.88	0.97	0.99	0.99	0.99	0.99	0.99
	BVAR-CV	0.77	0.88	0.88	0.90	0.92	0.94	0.96
	H-BVAR	0.72	0.89	0.89	0.92	0.93	0.95	0.96
Deflator	LS	0.88	0.87	0.86	0.88	0.91	0.92	0.94
	RIDGE	0.91	0.92	0.93	0.92	0.93	0.94	0.95
	RIDGE-AS	0.91	0.91	0.88	0.88	0.87	0.87	0.88
	LP	0.88	0.97	0.99	0.99	0.99	0.99	1.00
	BVAR-CV	0.80	0.86	0.88	0.91	0.93	0.94	0.96
	H-BVAR	0.84	0.88	0.90	0.92	0.94	0.95	0.97
Paper Rate	LS	0.87	0.86	0.86	0.88	0.90	0.92	0.94
	RIDGE	0.90	0.91	0.93	0.93	0.93	0.94	0.95
	RIDGE-AS	0.89	0.90	0.89	0.88	0.88	0.88	0.88
	LP	0.87	0.97	0.99	0.99	0.99	0.99	0.99
	BVAR-CV	0.82	0.84	0.87	0.90	0.92	0.93	0.95
	H-BVAR	0.88	0.88	0.90	0.92	0.93	0.95	0.96

tent with the findings of Li et al. (2023), who make extensive computational simulations by simulating from synthetic DGPs. They provide a comprehensive treatment of the question of which model – VAR or LP – is best suited for IRF inference in a given scenario in terms of bias-variance trade-off. They show that a key balance of bias versus variance exists between LP and VAR estimates of impulse responses: LPs tend to have low bias, due to their flexibility, but they also feature large variance at higher horizons. Their results allow one to better understand the trade-offs at play in Table 2.2 and Table 2.3. In particular, it is clear that ridge shrinkage is beneficial at short horizons only if the penalization scheme is well-adapted to the DGP at hand. Otherwise, as is the case for RIDGE and RIDGE-GLS methods, the induced bias can be such that ridge MSEs surpass that of OLS estimates. One also finds that the medium and long horizons MSE gains over LPs are more pronounced in cases of moderate dependence, but in the case of the Federal Funds Rate IRFs in Setup B zero-centered RIDGE estimates thoroughly mistake long-term behavior.

2.7.2 Confidence Intervals

I now try and evaluate whether ridge shrinkage has a negative impact on inference. There have also been recent contributions directly aimed at studying shrinkage effects. Using the same simulation setups as in the previous section, I investigate coverage and size properties of pointwise CIs constructed using the methods in Table 2.1. All confidence intervals are constructed with nominal 90% level coverage.

In this set of simulations, I swap GLS ridge for the asymptotic shrinkage ridge estimator, $\hat{\boldsymbol{\beta}}^R_{as}$, see Section 2.6, since the latter allows for a partially non-negligible penalization in the limit. To implement $\hat{\boldsymbol{\beta}}^R_{as}$, one needs to choose a partition of $\boldsymbol{\beta}$ which identifies asymptotically negligible

Table 2.5: Impulse Response Inference – Setup A – CI Length

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Investment Growth	LS	2.99	5.11	5.78	5.35	4.79	4.17	3.56
	RIDGE	3.13	5.20	5.82	5.17	4.48	3.78	3.09
	RIDGE-AS	3.11	5.15	4.84	4.33	3.70	3.06	2.48
	LP	2.99	7.50	10.97	12.89	13.99	14.55	14.70
	BVAR-CV	2.84	4.48	4.70	4.38	3.99	3.56	3.11
	H-BVAR	2.71	4.20	4.50	4.29	3.96	3.56	3.13
Deflator	LS	1.19	1.92	2.23	2.14	1.94	1.71	1.46
	RIDGE	1.24	1.97	2.25	2.09	1.84	1.54	1.26
	RIDGE-AS	1.24	1.95	1.95	1.78	1.52	1.25	1.01
	LP	1.19	3.03	4.56	5.42	5.90	6.14	6.21
	BVAR-CV	1.03	1.69	1.87	1.80	1.67	1.50	1.31
	H-BVAR	1.01	1.64	1.83	1.79	1.67	1.51	1.33
Paper Rate	LS	0.97	1.42	1.64	1.57	1.44	1.27	1.09
	RIDGE	1.01	1.44	1.65	1.53	1.36	1.16	0.95
	RIDGE-AS	1.01	1.43	1.42	1.31	1.13	0.94	0.77
	LP	0.97	2.19	3.28	3.90	4.26	4.43	4.48
	BVAR-CV	0.84	1.22	1.35	1.30	1.21	1.09	0.96
	H-BVAR	0.85	1.21	1.34	1.31	1.22	1.10	0.97

coefficient. To do this, I split $\boldsymbol{\beta}$ by lag and penalize all coefficients with lag orders greater than a given threshold $\overline{p}$, such that $1 < \overline{p} < p$. In setup A, I choose $\overline{p} = 6$, while in setup B I set $\overline{p} = 3$. In Bayesian methods, including the cross-validated Minnesota BVAR, I construct high-probability intervals by drawing from the posterior. Comparison between frequentist CIs and Bayesian posterior densities is not generally valid, because they are not analogous concepts. Therefore, the discussion below is intended to highlight differences in *structure* between ridge approaches.

Setup A. Simulations with the DGP of Kilian and Kim (2011), presented in Tables 2.4 and 2.5, highlight some of the advantages of applying ridge when performing inference. Focusing on estimator $\hat{\boldsymbol{\beta}}^R$, it is clear that CI coverage is in fact higher than the intervals obtained by least squares estimation in all situations. At impact, ridge CIs are larger than the LS baseline, but they shrink as horizons increase. Thus, is IRFs revert relatively quickly to zero, ridge can effectively reduce length while preserving coverage. As discussed in Section 2.3, these gains are inherently local to the DGP – shrinkage to zero at deep lags embodies correct prior knowledge of a weakly persistent process. For Bayesian estimators, one can note that quantile intervals at small horizons tend to be shorter compared to least squares and ridge methods.

Setup B. The effects of ridge shrinkage on a DGP with high persistence are much more severe, as shown in Tables 2.6 and 2.7. Focusing on frequentist ridge, one can observe that close to impact ($h = 1$) ridge has similar or even higher coverage than other methods for real GDP[10] However, as the IRF horizon grows, shrinkage often leads to severe undercoverage, with asymptotic shrinkage

[10]This also is the case with consumption and compensation, see also Tables 9 and 10 in Supplementary Appendix 2.D.5.

Table 2.6: Impulse Response Inference – Setup B: CI Coverage

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Real GDP	LS	0.87	0.81	0.75	0.72	0.71	0.72	0.73
	RIDGE	0.90	0.79	0.66	0.62	0.65	0.68	0.68
	RIDGE-AS	0.89	0.72	0.61	0.58	0.61	0.65	0.65
	LP	0.87	0.93	0.94	0.94	0.93	0.93	0.91
	BVAR-CV	0.70	0.71	0.63	0.64	0.71	0.75	0.76
	H-BVAR	0.84	0.86	0.76	0.76	0.83	0.88	0.88
Investment	LS	0.87	0.82	0.76	0.73	0.75	0.82	0.87
	RIDGE	0.85	0.79	0.65	0.62	0.73	0.80	0.81
	RIDGE-AS	0.82	0.69	0.59	0.57	0.68	0.77	0.77
	LP	0.87	0.94	0.94	0.95	0.94	0.94	0.94
	BVAR-CV	0.70	0.73	0.67	0.71	0.77	0.81	0.83
	H-BVAR	0.80	0.86	0.81	0.82	0.87	0.88	0.88
Fed Funds Rate	LS	0.85	0.83	0.80	0.78	0.77	0.79	0.80
	RIDGE	0.79	0.77	0.74	0.68	0.68	0.72	0.72
	RIDGE-AS	0.78	0.66	0.68	0.64	0.64	0.68	0.69
	LP	0.85	0.94	0.96	0.96	0.95	0.94	0.93
	BVAR-CV	0.76	0.72	0.76	0.77	0.77	0.81	0.83
	H-BVAR	0.87	0.86	0.74	0.73	0.78	0.84	0.87

Table 2.7: Impulse Response Inference – Setup B: CI Length (rescaled $\times 100$)

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Real GDP	LS	0.71	1.56	2.07	2.31	2.32	2.24	2.15
	RIDGE	0.79	1.56	1.85	1.95	1.92	1.85	1.77
	RIDGE-AS	0.74	1.31	1.65	1.76	1.75	1.70	1.64
	LP	0.71	2.42	4.21	5.40	5.90	5.91	5.70
	BVAR-CV	0.53	1.23	1.74	2.00	2.10	2.13	2.15
	H-BVAR	0.58	1.36	1.87	2.16	2.32	2.44	2.55
Investment	LS	3.38	6.65	7.89	7.89	7.31	6.69	6.18
	RIDGE	3.79	6.81	6.93	6.46	5.79	5.19	4.73
	RIDGE-AS	3.59	5.57	6.11	5.77	5.21	4.72	4.34
	LP	3.37	10.16	16.00	18.85	19.06	18.22	17.23
	BVAR-CV	2.64	5.26	6.59	6.91	6.78	6.57	6.38
	H-BVAR	2.89	5.74	7.08	7.54	7.63	7.60	7.58
Fed Funds Rate	LS	0.25	0.39	0.43	0.43	0.41	0.38	0.35
	RIDGE	0.29	0.39	0.37	0.36	0.33	0.30	0.29
	RIDGE-AS	0.27	0.31	0.33	0.32	0.30	0.28	0.27
	LP	0.25	0.59	0.88	1.01	1.05	1.03	0.98
	BVAR-CV	0.21	0.31	0.36	0.37	0.36	0.35	0.34
	H-BVAR	0.23	0.36	0.42	0.44	0.45	0.45	0.46

estimator $\hat{\boldsymbol{\beta}}_{as}^{R}$ giving the worst results. In comparison, Bayesian methods are much more reliable at all horizons, although the only estimator that can consistently improve upon the benchmark least squares VAR CIs is the hierarchical prior BVAR of Giannone et al. (2015). The reason behind this

is simple enough: the implementation of the Minnesota-prior BVAR I have used here has a white noise prior on all variables, which in this case is far from the truth. Indeed, Bańbura et al. (2010) implement the same BVAR by tuning the prior to a random walk for very persistent variables in their applications. In this sense, the cross-validated BVAR considered – which is assumed centered at zero – is really the flip-side of ridge estimators. Therefore, the addition of a prior on the mean of the autoregressive parameters as done by Giannone et al. (2015) is a key element to perform shrinkage in high persistence setups in a way that does not systematically undermine asymptotic inference on impulse responses.

2.8 Conclusion

In this paper, I have studied ridge regression and its application to vector autoregressive model estimation in detail. This appears to be the first work that provides a thorough analysis of ridge penalization in the context of time series data, including geometric as well as asymptotic properties. I have also derived results on the validity of cross-validation as a method to select the penalty intensity in practice, and I have shown that CV produces asymptotically valid penalization rates. Finally, I have compared both frequentist and Bayesian ridge formulation in simulations aimed at quantifying the applicability of ridge for impulse response inference.

The key takeaway of this work is that ridge penalization is a useful approach to VAR estimation as long as the chosen penalty structure is well-adapted to the model's structure. Bayesian ridge posteriors are especially flexible, with hierarchical priors also allowing shrinkage towards non-zero coefficient vectors. However, it is important to note that the Bayesian approach also permits the researcher to specify uninformative priors, so that the influence of the priors' hyperparameters is less pronounced. This is not the case in frequentist ridge, c.f. including an explicit non-zero centering vector. However, prior knowledge or a pre-estimation procedure may be available to the researcher, so that ridge can be effectively implemented without the need to implement a BVAR.

To conclude, there are still avenues of research regarding ridge which would be interesting to develop. First and foremost, the high-dimensional setup, for which, however, it seems non-trivial to find a domain of applicability, as discussed in the introduction. Secondly, a more in-depth analysis of cross-validation, especially in the multivariate case, would be extremely valuable. Moreover, both the latter and former topics should be jointly addressed in the context of mild cross-sectional dimension growth, i.e. $K \to \infty$ such that $K/T \to \rho \in (0, 1)$, which is comparable to factor model setups.

Appendix

2.A Basic Ridge Properties

2.A.1 LS and RLS Estimators.

Lütkepohl (2005), Chapter 3, shows that the multivariate least squares and GLS estimator of parameter vector $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = ((Z'Z)^{-1}Z \otimes I_K)\boldsymbol{y}$$

as the minimizer of $S(\boldsymbol{\beta}) = T^{-1}\text{tr}[(Y-BZ)'\Sigma_u(Y-BZ)]$. The multivariate Ridge-regularized Least Squares (RLS) estimator considered in this paper is defined to be the minimizer of the regularized problem,

$$\begin{aligned} S^R(\boldsymbol{\beta};\Lambda) &= T^{-1}\text{tr}[(Y-BZ)'(Y-BZ)] + \text{tr}[B'\Lambda B] \\ &= \frac{\boldsymbol{y}'\boldsymbol{y}}{T} + \boldsymbol{\beta}'\left(\frac{ZZ'}{T}\otimes I_K\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}'\frac{(Z\otimes I_K)\boldsymbol{y}}{T} + \boldsymbol{\beta}'\Lambda\boldsymbol{\beta} \end{aligned}$$

The first partial derivative,

$$\frac{\partial S^R(\boldsymbol{\beta};\Lambda)}{\partial\boldsymbol{\beta}} = 2\left(\frac{ZZ'}{T}\otimes I_K\right)\boldsymbol{\beta} - 2\frac{(Z\otimes I_K)\boldsymbol{y}}{T} + 2\Lambda\boldsymbol{\beta},$$

yields the normal equations $(T^{-1}ZZ'\otimes I_K + \Lambda)\boldsymbol{\beta} = T^{-1}(Z\otimes I_K)\boldsymbol{y}$. The Hessian $\partial^2 S^R(\boldsymbol{\beta})/\partial^2\boldsymbol{\beta} = 2(T^{-1}ZZ'\otimes I_K + \Lambda)\boldsymbol{\beta}$ is positive definite when $\Lambda > 0$, thus indeed the minimum is achieved by

$$\hat{\boldsymbol{\beta}}^R(\Lambda) = \left(\frac{ZZ'}{T}\otimes I_K + \Lambda\right)^{-1}\frac{(Z\otimes I_K)\boldsymbol{y}}{T}.$$

Identical derivations prove that re-centering the ridge penalty at $\boldsymbol{\beta}_0 \in \mathbb{R}^{K^2p}$ produces the estimator

$$\hat{\boldsymbol{\beta}}^R(\Lambda,\boldsymbol{\beta}_0) = \left(\frac{ZZ'}{T}\otimes I_K + \Lambda\right)^{-1}\left(\frac{(Z\otimes I_K)\boldsymbol{y}}{T} + \Lambda\boldsymbol{\beta}_0\right).$$

2.A.2 Structure of the Regularization Matrix

The vectorized RLS estimator $\hat{\boldsymbol{\beta}}^R(\Lambda)$ has maximal flexibility in terms of the regularization structure that matrix $\Lambda = \text{diag}\{\lambda_{1,1},\ldots,\lambda_{K,p}\}$ $(K^2p \times K^2p)$ imposes. Since $\boldsymbol{\beta}$ contains all the coefficients of $(A_1,\ldots,A_p)$ it is indeed possible to individually penalize each lag of each series differently. In fact, by relaxing the assumption that Λ be a diagonal matrix, even more general penalization structures are possible, although I do not consider them in this paper.

An interesting special case arises if the RLS estimator is instead written in its matrix form[11],

$$\hat{B}^R_{\text{mat}}(\Lambda_{Kp}) = \frac{YZ'}{T}\left(\frac{ZZ'}{T} + \Lambda_{Kp}\right)^{-1}$$

where here it is of note that $\Lambda_{Kp} > 0$ has size $(Kp \times Kp)$. The regularization structure imposed is

[11] For details in the least squares case, see again Lütkepohl (2005), Chapter 3. The derivations for the ridge estimator are identical.

different in general than that in $\hat{\boldsymbol{\beta}}^R(\Lambda)$: Λ_{Kp} induces *column-wise* ridge regularization, which penalizes coefficient estimates uniformly over each of the Kp columns of B. The associated vectorized estimator then simplifies:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^R(\Lambda_{Kp}) &= \left(\left(\frac{ZZ'}{T} + \Lambda_{Kp}\right) \otimes I_K\right)^{-1} (Z \otimes I_K)\boldsymbol{y} \\ &= \left(\left(\frac{ZZ'}{T} + \Lambda_{Kp}\right)^{-1} \frac{Z \otimes I_K}{T}\right) \boldsymbol{y}\end{aligned}$$

On the other hand, the *devectorized* RLS estimator is given by

$$\hat{B}^R(\Lambda_{K^2p}) = \text{reshape}(\boldsymbol{\beta}^R(\Lambda), K, Kp)$$

that is, $\hat{B}^R$ is simply a restructuring of the vectorized estimator into a matrix with identical dimensions to B. Importantly then, $\hat{B}^R(\Lambda_{K^2p})$ is equivalent to $\hat{B}^R(\Lambda_{Kp})$ if $\Lambda_{K^2p} = \Lambda_{Kp} \otimes I_K$. Because $\boldsymbol{\beta}^R(\Lambda_{K^2p})$ and $\hat{B}^R(\Lambda_{K^2p})$ allow for the most generality in penalization structure, I will consider them to be the reference RLS estimators, so the dimension subscript to Λ will be dropped unless explicitly required.

2.A.3 Autocovariance and Asymptotic Conditioning

The conditioning of the autocovariance $\Sigma_y = \mathbb{E}[y_t y_t']$ is an important measure for the role that the regularization in the RLS estimator should be playing. This in turn depends on the eigenvalues of $\hat{\Sigma}_y$ with respect to those of Σ_y. Hoerl and Kennard (1970) showed in the linear regression setting that, when the sample covariance matrix deviates significantly from the identity matrix, its small eigenvalues excessively inflate the variance of least squares estimates, even though the regression problem itself is well-posed. This fragility is inherently a byproduct of finite sampling, and partially due to numerical procedures. Nowadays, numerical precision is virtually not a concern anymore, as robust linear algebra procedures are implicitly implemented in most scientific languages and toolboxes. Yet estimation issues tied to small or unfavorable data samples remain extremely relevant from both theoretical and practical viewpoints.

In the spirit of ridge as a regularization procedure, the following Lemma establishes convergence in probability of the ordered eigenvalues of the sample autocovariance matrix.

Lemma 2.A.1. *If* $\hat{\Sigma}_y = T^{-1}\sum_{t=1}^{T-1} y_t y_t' \xrightarrow{P} \Sigma_y$ *where* $\Gamma \in \mathbb{R}^{K\times K}$ *is positive definite, then*

$$\omega_j\left(\hat{\Sigma}_y\right) \xrightarrow{P} \omega_j(\Sigma_y)$$

where $\omega_j(A)$ *is the* j *largest eigenvalue of* A.

Proof. First, recall that for all matrices $A \in \mathbb{R}^{K\times K}$, the determinant $\det(A)$ is clearly a continuous mapping[12]. Furthermore, for any polynomial $g(z) = z^n + a_1 z^{n-1} + \ldots + a_n$, $a_i \in \mathbb{C}$ factored as $g(z) = (z - w_1)\cdots(z - w_n)$, $w_i \in \mathbb{C}$, where the ordering of roots w_i is arbitrary, it holds that for any $\epsilon > 0$ there exists $\delta > 0$ such that for every polynomial $h(z) = z^n + b_1 z^{n-1} + \ldots + b_n$ with $|a_i - b_i| < \delta$ decomposed as $g(z) = (z - \overline{w}_1)\cdots(z - \overline{w}_n)$, $|w_i - \overline{w}_i| < \epsilon$, $i = 1, \ldots, n$, see Whitney

[12]This follows from $\det(A_{i,j}) = \sum_\varsigma \text{sign}(\varsigma)\prod_{i=1}^K A_{\varsigma(i),i}$ for permutation ς over $\{1, \ldots, K\}$

(1972), Appendix V.4. This in particular implies that the roots of the characteristic polynomial of matrix A are continuous functions of its coefficients.

Let $\varrho_{\hat{\Sigma}_y}(z) = z^K + a_1 z^{K-1} + \ldots + a_K = (z - \hat{\omega}_1) \cdots (z - \hat{\omega}_K)$ and $\varrho_{\Sigma_y}(z) = z^K + b_1 z^{K-1} + \ldots + b_K = (z - \omega_1) \cdots (z - \omega_K)$ be the (real) characteristic polynomials of $\hat{\Sigma}_y$ and Σ_y respectively. Because of the continuity arguments above, for every $\epsilon > 0$ there exist δ_1, $\delta_2 > 0$ such that

$$\begin{aligned}\mathbb{P}(|\hat{\omega}_i - \omega_i| > \epsilon) &\leq \mathbb{P}(|a_i - b_i| > \delta_1) \\ &\leq \mathbb{P}(\left\|\hat{\Sigma}_y - \Sigma_y\right\| > \delta_2)\end{aligned}$$

for $i \in \{1, \ldots, K\}$. Since by assumption $\hat{\Sigma}_y \xrightarrow{P} \Sigma_y$, the RHS of the above converges to zero as $T \to \infty$, thus $\hat{\omega}_i \xrightarrow{P} \omega_i$. □

2.B Proofs

2.B.1 Shrinkage

Proof of Proposition 2.3.2

Proof. Notice that, by introducing $\Lambda_p := \text{diag}\{\lambda_1, \ldots, \lambda_p\}$, any lag-adapted regularization matrix can be written as $\Lambda^{(p)} = \Lambda_p \otimes I_{K^2} = (\Lambda_p \otimes I_K) \otimes I_K$, so that

$$\begin{aligned}\hat{\boldsymbol{\beta}}^R(\Lambda_i^{(p)}) &= \left[(ZZ' + \Lambda_{p,i} \otimes I_K) \otimes I_K\right]^{-1} (Z \otimes I_K)\boldsymbol{y} \\ &= \left[(ZZ' + \Lambda_{p,i} \otimes I_K)^{-1} \otimes I_K\right] (Z \otimes I_K)\boldsymbol{y}\end{aligned}$$

by the properties of Kronecker product. It is now possible to derive the statements of the proposition as follows:

(a) The result regarding isotropic regularizer $\Lambda^{(p)} = \lambda I_{K^2 p}$ is trivial given Proposition 2.3.1.

(b) Without loss of generality due to the ordering of lags in Z, one may write the Gram matrix ZZ' in a block fashion,

$$ZZ' + \Lambda_p = \begin{bmatrix}(ZZ')_{[\mathcal{S}]} + \Lambda_{[\mathcal{S}]} & D \\ D' & (ZZ')_{[\mathcal{S}^c]} + \Lambda_{[\mathcal{S}^c]}\end{bmatrix}$$

where $(ZZ')_{[\mathcal{S}^c]}$ is the sub-matrix containing all the components *not* indexed by subset $\mathcal{S}$, and the subscript has been dropped from Λ_p for ease of notation.
Define $A_{[\mathcal{S}]} = (ZZ')_{[\mathcal{S}]} + \Lambda_{[\mathcal{S}]}$, $B_{[\mathcal{S}^c]} = (ZZ')_{[\mathcal{S}^c]} + \Lambda_{[\mathcal{S}^c]}$ and $\Delta = (B_{[\mathcal{S}^c]} - D' A_{[\mathcal{S}]}^{-1} D)$. The matrix block-inversion formula yields

$$(ZZ' + \Lambda_p)^{-1} = \begin{bmatrix}A_{[\mathcal{S}]}^{-1} + A_{[\mathcal{S}]}^{-1} D \Delta^{-1} D' A_{[\mathcal{S}]}^{-1} & A_{[\mathcal{S}]}^{-1} D \Delta^{-1} \\ -\Delta^{-1} D' A_{[\mathcal{S}]}^{-1} & \Delta^{-1}\end{bmatrix}.$$

If $\Lambda_{[\mathcal{S}]} \to 0$ and $\Lambda_{[\mathcal{S}^c]} \to \infty$, then $A_{[\mathcal{S}]} \to (ZZ')_{[\mathcal{S}]}$, $B_{[\mathcal{S}^c]} \to \infty$. Therefore $\Delta^{-1} \to 0$, since for $\Lambda_{[\mathcal{S}^c]}$ sufficiently large $\left\|B_{[\mathcal{S}^c]}^{-1} D' A_{[\mathcal{S}]}^{-1} D\right\| < 1$ and thus the Sherman-Morrison-Woodbury

formula implies

$$\left\|(B_{[\mathcal{S}^c]} - D'A_{[\mathcal{S}]}^{-1}D)^{-1}\right\| \leq \frac{\left\|B_{[\mathcal{S}]}^{-1}\right\|}{1 - \left\|B_{[\mathcal{S}^c]}^{-1}D'A_{[\mathcal{S}]}^{-1}D\right\|} \to 0.$$

The above results finally yield

$$\left[(ZZ' + \Lambda_p)^{-1}Z \otimes I_K\right]\boldsymbol{y} \to \begin{bmatrix}(ZZ')_{[\mathcal{S}]} & 0\\ 0 & 0\end{bmatrix}(Z \otimes I_K)\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{[\mathcal{S}]}^{LS}$$

as required.

□

2.B.2 Ridge Asymptotic Theory

Proof of Theorem 2.5.1

Proof. (a) Assumptions A-B imply directly that $\hat{\Gamma}_T$ is a consistent estimator for Γ: in particular, y_t is a stationary, stable and ergodic VAR process.

(b) Rewriting $\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0)$ yields

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) &= \left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)^{-1}\left[T^{-1}(Z \otimes I_K)((Z' \otimes I_K)\boldsymbol{\beta} + \boldsymbol{u}) + \Lambda\boldsymbol{\beta}_0\right]\\
&= \left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)^{-1}\left[\left(\frac{ZZ'}{T} \otimes I_K\right)\boldsymbol{\beta} + \frac{(Z \otimes I_K)\boldsymbol{u}}{T} + \Lambda\boldsymbol{\beta}_0\right]\\
&= \left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)^{-1}\left[\left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)\boldsymbol{\beta} + \frac{(Z \otimes I_K)\boldsymbol{u}}{T} + \Lambda(\boldsymbol{\beta}_0 - \boldsymbol{\beta})\right]\\
&= \boldsymbol{\beta} + \left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)^{-1}\Lambda(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \left(\frac{ZZ'}{T} \otimes I_K + \Lambda\right)^{-1}\frac{(Z \otimes I_K)\boldsymbol{u}}{T}.
\end{aligned}$$

I study the last two terms of the last equality separately. The first term is $o_p(1)$,

$$\left(\left(\frac{ZZ'}{T}\right) \otimes I_K + \Lambda\right)^{-1}\Lambda(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) = \left(\left(\frac{ZZ'}{T}\right) \otimes I_K + o_p(1)\right)^{-1}o_p(1)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \xrightarrow{P} 0$$

since $(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = (\boldsymbol{\beta} - \underline{\boldsymbol{\beta}}_0) + (\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) = (\boldsymbol{\beta} - \underline{\boldsymbol{\beta}}_0) + o_p(1)$. Considering the matrix sequence

$$\boldsymbol{\zeta}_T = \left[T^{-1}(ZZ'), T^{-1}\Lambda\right],$$

which under Assumptions B and D.(ii) converges in probability to $[\Gamma, 0]$, by the continuous mapping theorem (Davidson, 1994) the second term gives

$$\left(\left(\frac{ZZ'}{T}\right) \otimes I_K + o_p(1)\right)^{-1}\left(\frac{1}{T}(Z \otimes I_K)\boldsymbol{u}\right) \xrightarrow{P} \Gamma^{-1}\,\mathbb{E}[(Z \otimes I_K)\boldsymbol{u}] = 0$$

under Assumption A.

(c) The residuals $\hat{U}$ can be written as

$$\hat{U} = Y - \hat{B}^R Z = BZ + U - \hat{B}^R Z = U + (B - \hat{B}^R)Z$$

Thus

$$\frac{\hat{U}\hat{U}'}{T} = \frac{UU'}{T} + (B - \hat{B}^R)\left(\frac{ZZ'}{T}\right)(B - \hat{B}^R)' + (B - \hat{B}^R)\left(\frac{ZU'}{T}\right) + \left(\frac{UZ'}{T}\right)(B - \hat{B}^R)' \quad (2.9)$$

From (a) one finds that $\text{vec}(B) - \text{vec}(\hat{B}^R) = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R = o_p(1)$, so $(B - \hat{B}^R) \xrightarrow{P} 0$, while $T^{-1}(ZZ') \xrightarrow{P} \mathbb{E}[ZZ']$ and $T^{-1}(ZU') \xrightarrow{P} \mathbb{E}[ZU'] = 0$: the terms involving these quantities then vanish asymptotically. Lastly, the first term of the sum gives

$$\frac{UU'}{T} = \frac{1}{T}\sum_{t=1}^{T} u_t u_t' \xrightarrow{P} \mathbb{E}[u_t u_t'] = \Sigma_u$$

for $T \to \infty$ under Assumptions A and B.

(d) With the same expansion used in (b),

$$\sqrt{T}(\hat{\boldsymbol{\beta}}^R(\Lambda, \boldsymbol{\beta}_0) - \boldsymbol{\beta}) = Q_T^{-1}\left(\sqrt{T}\Lambda\right)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + Q_T^{-1}\left(\frac{1}{\sqrt{T}}(Z \otimes I_K)\boldsymbol{u}\right)$$

where $Q_T = (T^{-1}ZZ' + \Lambda) \xrightarrow{P} \Gamma$. Following the arguments above, the first term in the sum converges in probability,

$$Q_T^{-1}\left(\sqrt{T}\Lambda\right)\left(\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta} + o_p(1)\right) \xrightarrow{P} \Gamma^{-1}\Lambda_0(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}}_0)$$

The second term has normal limiting distribution,

$$Q_T^{-1}\left(\frac{1}{\sqrt{T}}(Z \otimes I_K)\boldsymbol{u}\right) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma_u)$$

see Lütkepohl (2005), Proposition 3.1. By Slutky's theorem claim (d) follows.

□

Proof of Theorem 2.5.2

Proof. (1) Since condition (i) implies that $\sqrt{T}\Lambda \xrightarrow{P} 0$, results (a)-(c) are unchanged, while (d) now involves the limit

$$Q_T^{-1}\left(\sqrt{T}\Lambda\right)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) = Q_T^{-1} \cdot o_P(1) \cdot \left(\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta} + o_P(1)\right) \xrightarrow{P} 0$$

yielding (d′).

(2) Assuming $\boldsymbol{\beta}_0 \xrightarrow{P} \boldsymbol{\beta}$ simplifies the terms in the proof of Theorem 2.5.1 since now $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = o_P(1)$. The weaker rate imposed on Λ does not influence results (a)-(c). Moreover,

$$Q_T^{-1} \cdot (\sqrt{T}\Lambda) \cdot (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) = Q_T^{-1} \cdot O_P(1) \cdot o_P(1) \xrightarrow{P} 0$$

so that (d′) follows.

□

Proof of Theorem 2.5.4

Proof. I make a straightforward adaptation of the proof found in Hamilton (1994b), Proposition 11.2. Define $\hat{\Sigma}_u^* = T^{-1}(UU')$, which is expanded to

$$\begin{aligned}\hat{\Sigma}_u^* &= \frac{1}{T}(Y - BZ)(Y - BZ)' \\ &= \frac{1}{T}\left(Y - \hat{B}^R Z + (\hat{B}^R - B)Z\right)\left(Y - \hat{B}^R Z + (\hat{B}^R - B)Z\right)' \\ &= \hat{\Sigma}_u^R + (\hat{B}^R - B)\left(\frac{ZZ'}{T}\right)(\hat{B}^R - B)' + \\ &\quad + \frac{1}{T}\left((Y - \hat{B}^R Z)Z'(\hat{B}^R - B)' + (\hat{B}^R - B)Z(Y - \hat{B}^R Z)'\right)\end{aligned}$$

Contrary to the least squares estimator, cross-terms do not cancel out since for $\Lambda \succeq 0$ the residuals $(Y - \hat{B}^R Z)$ are not in the orthogonal space of Z. From the consistency results of Theorem 2.5.1,

$$(\hat{B}^R - B)\left(\frac{ZZ'}{T}\right)(\hat{B}^R - B)' = o_p(1)\left(\frac{ZZ'}{T}\right)o_p(1) \xrightarrow{P} 0$$

and

$$\sqrt{T}(\hat{B}^R - B)\left(\frac{ZZ'}{T}\right)(\hat{B}^R - B)' = O_p(1)\left(\frac{ZZ'}{T}\right)o_p(1) \xrightarrow{P} 0$$

Further,

$$\sqrt{T}\left[\frac{1}{T}(Y - \hat{B}^R Z)Z'(\hat{B}^R - B)'\right] = \left(\frac{\hat{U}Z'}{T}\right)\sqrt{T}(\hat{B}^R - B)' \xrightarrow{P} 0$$

since again $\sqrt{T}\hat{B}^R$ is asymptotically normal, and $T^{-1}(\hat{U}Z') = T^{-1}(UZ') + (B - \hat{B}^R)\cdot T^{-1}(ZZ') = T^{-1}(UZ') + o_p(1) \xrightarrow{P} \mathbb{E}[UZ'] = 0$. The same holds for the remaining transpose term, too.

It is thus proven that $\sqrt{T}(\hat{\Sigma}_u^* - \hat{\Sigma}_u^R) \xrightarrow{P} 0$, meaning $\sqrt{T}(\hat{\Sigma}_u^* - \Sigma_u) \xrightarrow{P} \sqrt{T}(\hat{\Sigma}_u^R - \Sigma_u)$ so that the two terms may be exchanged in computing the joint asymptotic distribution. Theorem 2.5.1 accordingly yields

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_0^R - \boldsymbol{\beta} \\ \operatorname{vec}(\hat{\Sigma}_u^R) - \operatorname{vec}(\Sigma_u)\end{bmatrix} \xrightarrow{P} \begin{bmatrix}Q_T^{-1}\left(\sqrt{T}\Lambda\right)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + Q_T^{-1}\frac{1}{\sqrt{T}}\boldsymbol{\xi} \\ \frac{1}{\sqrt{T}}\boldsymbol{\eta}\end{bmatrix}$$

where $\boldsymbol{\xi} = (Z \otimes I_K)\boldsymbol{u}$ and $\boldsymbol{\eta} = \operatorname{vec}(UU' - \Sigma_u)$. As in Hamilton (1994b), Proof of Proposition 11.2, $(\boldsymbol{\xi}', \boldsymbol{\eta})'$ is a martingale difference sequence, thus the claim

$$\sqrt{T}\begin{bmatrix}\hat{\boldsymbol{\beta}}_0^R - \boldsymbol{\beta} \\ \operatorname{vec}(\hat{\Sigma}_u^R) - \operatorname{vec}(\Sigma_u)\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix}\Gamma^{-1}\Lambda_0(\underline{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) \\ 0\end{bmatrix}, \begin{bmatrix}\Gamma^{-1} \otimes \Sigma_u & 0 \\ 0 & \Omega\end{bmatrix}\right)$$

as $T \to \infty$ follows. When the strengthened assumptions (1) or (2) of Theorem 2.5.2 are used instead, the non-zero limiting mean vanishes

$$Q_T^{-1} \cdot \Lambda \cdot \sqrt{T}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \xrightarrow{P} 0$$

proving that the joint asymptotic distribution is mean-zero Gaussian.

Finally, to compute the explicit expression of the asymptotic variance Ω, one must take care and note that u_t is not assumed to be normally distributed, contrary to the remainder of the proof in Hamilton (1994b), pp. 342-343. A correct expression for i.i.d. non-Gaussian u_t can be found in

Remark 2.1, Brüggemann et al. (2016), yielding

$$\Omega = \mathrm{Var}[\mathrm{vec}(u_t u_t')] = \mathbb{E}[\mathrm{vec}(u_t u_t')\,\mathrm{vec}(u_t u_t')'] - \boldsymbol{\sigma}\boldsymbol{\sigma}'$$

where $\boldsymbol{\sigma} = \mathrm{vec}(\Sigma_u)$. □

Proof of Theorem 2.6.1

Proof. The stated results reduce to studying the behavior of two components used in the proof of Theorem 2.5.1 and Theorem 2.5.2, under the additional simplification of $\boldsymbol{\beta}_0 = 0$.

(a) Identical to result (a) in Theorem 2.5.1.

(b) Consistency follows immediately by the fact that $\Lambda\,\boldsymbol{\beta} \xrightarrow{P} 0$.

(c) Follows from (c), Theorem 2.5.1 and (b) above.

(d″) The bias term in the expression of $\sqrt{T}(\hat{\boldsymbol{\beta}}^R - \boldsymbol{\beta})$ is driven by

$$\sqrt{T}\Lambda\,\boldsymbol{\beta} = \begin{bmatrix} \sqrt{T}\Lambda_1 \cdot \boldsymbol{\beta}_1 \\ \sqrt{T}\Lambda_2 \cdot T^{-(1/2+\delta)}\boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} o_P(1) \cdot \boldsymbol{\beta}_1 \\ \Lambda_2 \cdot T^{-\delta}\boldsymbol{b}_2 \end{bmatrix} \xrightarrow{P} 0$$

meaning there is no asymptotic bias. On the other hand,

$$\Big(\Big(T^{-1}ZZ'\Big) \otimes I_K + \Lambda\Big)^{-1} \xrightarrow{P} (\Gamma + \overline{\Lambda})^{-1} \otimes I_K.$$

Setting $(\Gamma + \overline{\Lambda}) = \Gamma_{\overline{\Lambda}}$ yields the claim since there are no further simplifications in the asymptotic variance formula, cf. proof of (d), Theorem 2.5.1.

□

Proof of Corollary 2.6.2

Proof. For (a)-(c), it is again the case that Theorem 2.5.1 and Theorem 2.5.2 provide the needed results without need of adaptation. To prove that (d″) still holds, one just needs to notice that the asymptotic bias now involves the term

$$\sqrt{T}\Lambda\,(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \begin{bmatrix} \sqrt{T}\Lambda_1 \cdot (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}) \\ \sqrt{T}\Lambda_2 \cdot (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{02}) \end{bmatrix} = \begin{bmatrix} o_P(1) \cdot (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}) \\ -\Lambda_2 \cdot T^{-\delta}\boldsymbol{b}_2 \end{bmatrix},$$

since, by assumption, have that $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{02} = -T^{-(1/2+\delta)}\boldsymbol{b}_2$. Thus, as before, $\sqrt{T}\Lambda\,(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \xrightarrow{P} 0$. □

2.C Cross-validation

Later in this section, the following lemma will be useful.

Lemma 2.C.1. *Let y_t a stationary and stable mean-zero AR(p) process with companion form matrix $\mathbb{A} \in \mathbb{R}^{p\times p}$. Then, the associated MA(∞) coefficients, $\{\vartheta_k\}_{k\in\mathbb{N}_0}$, decay exponentially for k sufficiently large, that is,*

$$|\vartheta_k| = O\big(\exp(-C_{\mathbb{A}}\,k)\big)$$

for some constant $C_{\mathbb{A}} > 0$.

Proof. Recall from e.g. Lütkepohl (2005) that if $\mathbb{A}$ is the companion matrix of the AR(p) model, then $\vartheta_k = \iota \mathbb{A}^j \iota'$, where $\iota := (1, 0, \dots, 0)' \in \mathbb{R}^p$. Note that $\|\iota\| = 1$ by construction and that the spectral radius of $\mathbb{A}$ is less than one under the assumption of stability. Let $\mathbb{A} = VDV^{-1}$ be the Jordan canonical form of the companion matrix and ω_1 the dominant eigenvalue: stability thus implies that $|\omega_1| < 1$. Now observe that, supposing D has $l \leq p$ diagonal blocks, for $k \geq 0$

$$\frac{\left|\iota \mathbb{A}^k \iota'\right|}{|\omega_1^k|} = \left|\iota V \left(\frac{D^k}{\omega_1^k}\right) V^{-1} \iota'\right| = \left|\iota V \begin{bmatrix} [1] & & & \\ & \left[\frac{1}{\omega_1^k} D_2^k\right] & & \\ & & \ddots & \\ & & & \left[\frac{1}{\omega_1^k} D_l^k\right] \end{bmatrix} V^{-1} \iota'\right|,$$

where [1] is the dominant Jordan block, while $D_2, \dots, D_l$ are the non-dominant blocks. Then, letting $k \to \infty$, one gets

$$\frac{\left|\iota \mathbb{A}^k \iota'\right|}{|\omega_1^k|} \to \left|\iota V \begin{bmatrix} [1] & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} V^{-1} \iota'\right| = C_D$$

for some constant $C_D < \infty$, as for $\ell \in \{2, \dots, l\}$ one can see that $D_\ell^k / \omega_1^k \to 0$. Since for complex ω_1 it holds $\left|\omega_1^k\right| = |\omega_1|^k$, by letting $C_{\mathbb{A}} = -\log(|\omega_1|)$ the proof is complete. □

2.C.1 Two-fold CV

Proof of Theorem 2.5.5

Proof. Write $\text{cv2}_m(\Lambda)$ as

$$\begin{aligned} \text{cv2}_m(\Lambda) &= \widetilde{T}^{-1} \left(\widetilde{\boldsymbol{y}} - \widetilde{Z}' \hat{\boldsymbol{\beta}}^R(\Lambda)\right)' \left(\widetilde{\boldsymbol{y}} - \widetilde{Z}' \hat{\boldsymbol{\beta}}^R(\Lambda)\right) \\ &= \widetilde{T}^{-1} \left[\widetilde{\boldsymbol{u}} + \widetilde{Z}' \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right)\right]' \left[\widetilde{\boldsymbol{u}} + \widetilde{Z}' \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right)\right] \\ &= \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right)' \left(\frac{\widetilde{Z}\widetilde{Z}'}{\widetilde{T}}\right) \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right) + 2 \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right)' \left(\frac{\widetilde{Z}\widetilde{\boldsymbol{u}}}{\widetilde{T}}\right) + \frac{\widetilde{\boldsymbol{u}}'\widetilde{\boldsymbol{u}}}{\widetilde{T}}. \end{aligned}$$

By a strong LLN for weakly dependent processes (see, for example, Rio 2017), it holds that $\widetilde{T}^{-1} \widetilde{Z}\widetilde{Z}' \xrightarrow{a.s.} \Gamma_z$, $\widetilde{T}^{-1} \widetilde{Z}\widetilde{\boldsymbol{u}} \xrightarrow{a.s.} \mathbf{0}$ and $\widetilde{T}^{-1} \widetilde{\boldsymbol{u}}'\widetilde{\boldsymbol{u}} \xrightarrow{a.s.} \Sigma_u$. Since, by a similar decomposition, it holds that

$$\text{err}(\Lambda) = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right)' \Gamma \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right) + \Sigma_u,$$

where $\Gamma = \mathbb{E}[z_t z_t']$ is positive definite, almost sure convergence is proven.

To prove uniform convergence over compact subsets, I follow the proof of Patil et al. (2021), Theorem 4.1, which relies on verifying the conditions of the Arzelà-Ascoli theorem. That is, one must prove that function $\text{cv2}_m(\Lambda)$ as well as its first derivatives are bounded over compact sets. As

the Arzelà-Ascoli theorem readily generalizes to Euclidean spaces of arbitrary (fixed) dimension, I will directly consider the matrix derivative when checking boundedness.

Assume that $\Lambda \in \mathcal{I}$, where $\mathcal{I}$ is a compact set of positive semidefinite penalization matrices Λ such that $\|\Lambda\|_{\max} < \infty$ and $\Lambda \succ \lambda_{\min} I$. Note $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda) = (\hat{\Gamma} + \Lambda)^{-1}\Lambda\boldsymbol{\beta} - (\hat{\Gamma} + \Lambda)^{-1}(T^{-1}Z\boldsymbol{u})$ where $\hat{\Gamma} = T^{-1}ZZ'$. Using this decomposition, one gets first

$$\begin{aligned}
|\mathrm{err}(\Lambda)| &\leq \Sigma_u + \|\Gamma\|_2 \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right\|_2^2 \\
&\leq \Sigma_u + \|\Gamma\|_2 \left[\left\|(\hat{\Gamma} + \Lambda)^{-1}\Lambda\right\|_2^2 \|\boldsymbol{\beta}\|_2^2 + \left\|(\hat{\Gamma} + \Lambda)^{-1}\right\|_2^2 \left\|T^{-1}Z\boldsymbol{u}\right\|_2^2 \right. \\
&\qquad \left. + 2\left\|(\hat{\Gamma} + \Lambda)^{-1}\right\|_2^2 \left\|T^{-1}Z\boldsymbol{u}\right\|_2 \|\Lambda\boldsymbol{\beta}\|_2\right] \\
&\overset{a.s.}{\leq} \Sigma_u + \|\Gamma\|_2 \frac{\omega_{\max}(\Lambda)^2 \|\boldsymbol{\beta}\|_2^2 + C_{zu}^2 + 2C_{zu}\,\omega_{\max}(\Lambda)\,\|\boldsymbol{\beta}\|_2}{(\omega_{\min}(\hat{\Gamma}) + \omega_{\min}(\Lambda))^2}
\end{aligned}$$

where the last line follows from applying Weil's eigenvalue inequalities (Bhatia, 1997) to $(\hat{\Gamma} + \Lambda)^{-1}$ and the fact that $T^{-1}Z\boldsymbol{u} \xrightarrow{a.s.} \mathbf{0}$ by a strong LLN, so that there exists a constant $C_{zu} > 0$ bounding $\left\|T^{-1}Z\boldsymbol{u}\right\|_2$ for T large enough.

Additionally, the matrix derivative of $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)$ with respect to Λ is

$$\frac{\partial(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda))}{\partial\Lambda} = (\hat{\Gamma} + \Lambda)^{-1}\boldsymbol{\beta} - (\hat{\Gamma} + \Lambda)^{-2}\Lambda\boldsymbol{\beta} - (\hat{\Gamma} + \Lambda)^{-2}(T^{-1}Z\boldsymbol{u}),$$

so that, by using similar argument as the one used above, one gets

$$\begin{aligned}
\left|\frac{\partial\,\mathrm{err}(\Lambda)}{\partial\Lambda}\right| &\overset{a.s.}{\leq} \Sigma_u + \|\Gamma\|_2 \left[\frac{\|\boldsymbol{\beta}\|_2^2}{(\omega_{\min}(\hat{\Gamma}) + \omega_{\min}(\Lambda))^2} + \frac{2\,\omega_{\max}(\Lambda)\,\|\boldsymbol{\beta}\|_2^2 + 2C_{zu}\,\|\boldsymbol{\beta}\|_2}{(\omega_{\min}(\hat{\Gamma}) + \omega_{\min}(\Lambda))^3} \right. \\
&\qquad \left. + \frac{\omega_{\max}(\Lambda)^2 \|\boldsymbol{\beta}\|_2^2 + C_{zu}^2 + 2C_{zu}\,\omega_{\max}(\Lambda)\,\|\boldsymbol{\beta}\|_2}{(\omega_{\min}(\hat{\Gamma}) + \omega_{\min}(\Lambda))^4}\right].
\end{aligned}$$

The almost sure bound in the last display is also clearly finite for any $\Lambda \in \mathcal{I}_\lambda$, as required.

One can easily bound $\mathrm{cv2}_m(\Lambda)$ and its first derivative as $\mathrm{err}(\Lambda)$, with only addition of an extra term depending on $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda))'(\widetilde{T}^{-1}\widetilde{Z}\widetilde{\boldsymbol{u}})$. This means that $\mathrm{err}(\Lambda) - \mathrm{cv2}_m(\Lambda)$ forms an equicontinuous family of functions with respect to Λ over any $\mathcal{I}_\lambda$. Therefore, Arzelà-Ascoli yields uniform convergence of a subsequence, and since the difference converges to zero pointwise, too, the entire sequence converges uniformly. □

2.C.2 Cross-validation under Dependence

The result of Theorem 2.5.5 may be only partially informative in practice, as it does not give information on how dependence, in terms of the buffer block of size m, impacts $\mathrm{cv2}_m(\Lambda)$. Indeed, due to averaging, the effects of time dependence between the estimation and evaluation folds are washed out in the limit $\widetilde{T} \to \infty$ even when m is fixed. Therefore, Theorem 2.5.5 is not useful in finite samples, where one would preferably set m to be as small as possible.

To address dependence, in the same setup as above, consider an alternative predictive error

measure, the *m-dependence prediction error*,

$$\mathrm{Err}_m\left(\hat{\boldsymbol{\beta}}^R(\Lambda)\right) := \mathbb{E}_{y_{T+m+1},z_{T+m+1}}\left[\left(y_{T+m+1} - z'_{T+m+1}\,\hat{\boldsymbol{\beta}}^R(\Lambda)\right)^2 |Z,\boldsymbol{y}\right],$$

and the associated error curve, $\mathrm{err}_m(\Lambda) := \mathrm{Err}_m\left(\hat{\boldsymbol{\beta}}^R(\Lambda)\right)$. The empirical counterpart to this quantity is given by $\mathrm{cv2}_m(\Lambda)$ for $\widetilde{T} = 1$. The next theorem shows that in the case of a purely autoregressive data generating process, the error one commits by choosing a finite buffer size is exponentially small for sufficiently large m.

Theorem 2.C.2. *Under Assumptions A-C, for every Λ in the cone of diagonal positive definite penalty matrices with diagonal entries in $(\lambda_{\min},\infty)$, for $m\to\infty$ it holds that*

$$\mathrm{err}_m(\Lambda) - \mathrm{err}(\Lambda) = O(\exp(-C_{\boldsymbol{\beta}}\, m))$$

where $C_{\boldsymbol{\beta}}$ is a constant that does not depend on Λ.

Proof. In line with the definition of $\mathrm{cv2}_m(\Lambda)$, I set $\widetilde{y}_1 = y_{T+m+1}$, $\widetilde{z} = z_{T+m+1}$ and $\widetilde{u}_1 = y_{T+m+1} - z'_{T+m+1}\boldsymbol{\beta}$. With the same approach as in the proof of Theorem 2.5.5, here one finds

$$\begin{aligned}\mathrm{err}_m(\Lambda) = &\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right)' \mathbb{E}\left[\widetilde{z}_1\widetilde{z}'_1|Z,\boldsymbol{y}\right]\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right)\\ &+ 2\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^R(\Lambda)\right)' \mathbb{E}\left[\widetilde{z}_1\widetilde{u}_1|Z,\boldsymbol{y}\right] + \mathbb{E}\left[\widetilde{u}_1^2|Z,\boldsymbol{y}\right],\end{aligned}$$

where I have removed the subscript from expectation $\mathbb{E}$ to make notation clearer. Since $\widetilde{u}_1$ is independent of $\widetilde{z}_1$, the cross term reduces to zero, while $\mathbb{E}\left[\widetilde{u}_1^2|Z,\boldsymbol{y}\right] = \Sigma_u$. Thus, it is the first term in the last display that is effected by dependence.

To see this, let $\widetilde{z}_{i1}$ for $1\le i\le p$ be the ith entry of $\widetilde{z}_1$. Then, using the MA(∞) decomposition of y_t, i.e. $y_t = \sum_{\ell=0}^{\infty}\phi_\ell u_{t-\ell}$, one can write

$$\begin{aligned}\widetilde{z}_{i1} = z_{T+m+1-i} &= \sum_{\ell=0}^{m-i}\phi_\ell u_{T+m+1-i-\ell} + \sum_{\ell=M+1-i}^{\infty}\phi_\ell u_{T+m+1-i-\ell}\\ &= \sum_{\ell=0}^{m-i}\phi_\ell u_{T+m+1-i-\ell} + \sum_{s=0}^{\infty}\phi_{m+1-i+s}u_{T-s}\\ &= \eta_i + \zeta_i.\end{aligned}$$

Note that η_i is independent of ζ_i, and ζ_i is belongs with the σ-algebra generated by Z and $\boldsymbol{y}$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\widetilde{z}_1\widetilde{z}'_1|Z,\boldsymbol{y}\right] &= \mathbb{E}\left[(\eta_i+\zeta_i)(\eta_i+\zeta_i)'|Z,\boldsymbol{y}\right]\\ &= \Gamma_\eta + \zeta_i\zeta_i',\end{aligned}$$

as $\mathbb{E}\left[\eta_i\zeta_i'|Z,\boldsymbol{y}\right] = \mathbb{E}\left[\eta_i|Z,\boldsymbol{y}\right]\zeta_i' = 0$.

Now, I prove that $\Gamma_\eta \to \Gamma_z$ and $\zeta_i \zeta_i' \to 0$ at an exponential rate. First, let

$$H_\phi := \begin{bmatrix} \sum_{s=0}^{\infty} \phi_{m+1+s}^2 & \sum_{s=0}^{\infty} \phi_{m+1+s}\phi_{m+1+s-1} & & \\ \sum_{s=0}^{\infty} \phi_{m+1+s-1}\phi_{m+1+s} & \sum_{s=1}^{\infty} \phi_{m+1+s}^2 & & \\ & & \ddots & \\ & & & \sum_{s=p}^{\infty} \phi_{m+1+s}^2 \end{bmatrix}$$

and observe that

$$\|\Gamma_\eta - \Gamma_z\|_2 \leq \Sigma_u \, p \, \|H_\phi\|_{\max} \leq C \exp(-C_\eta \, m),$$

since $\Gamma_\eta = \Sigma_u \otimes H_\phi$, lag order p is fixed and the maximal entry of H_ϕ decays exponentially for m sufficiently large following Lemma 2.C.1. Secondly, much in the same vein

$$\left\|\zeta_i \zeta_i'\right\|_2 = \zeta_i' \zeta_i \leq p \left(\sum_{s=0}^{\infty} \phi_{m+1-i+s} u_{T-s} \right)^2 \leq C' \exp(-C_\zeta \, m).$$

The proof concludes by setting $C_{\boldsymbol{\beta}} = \max(C_\eta, C_\zeta)$. □

Remark 2.C.1. Theorem 2.C.2 is reassuring because it suggests that, in practice, if the AR(p) model is correctly specified, one may keep m small and still get a valid prediction error estimate in sense of Theorem 2.5.5. In simulations, I set $m = 0$, which is a common simplification to more effectively exploit the entire sample and does not, as discussed above, effect consistency (Bergmeir et al., 2018b). Moreover, note that Theorem 2.C.2 intuitively gives a *worst-case* rate: the dependence of between $\widetilde{z}_t$ and data in the estimation set gets milder, on average, as $\widetilde{T}$ grows. Thus, if CV aspect ratio $\widetilde{T}/T$ is balanced, dependence only plays a negligible role.

2.C.3 Asymptotically Valid CV

Proof of Theorem 2.5.6

Proof. First, recall that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\blacklozenge}^R(\Lambda) &= \left(\frac{ZZ'}{T} + \sqrt{T}\Lambda \right)^{-1} \frac{Z\boldsymbol{y}}{T} \\ &= \boldsymbol{\beta} - \left(\frac{ZZ'}{T} + \sqrt{T}\Lambda \right)^{-1} \left(\sqrt{T}\Lambda \right) \boldsymbol{\beta} + \left(\frac{ZZ'}{T} + \sqrt{T}\Lambda \right)^{-1} \frac{Z\boldsymbol{u}}{T}. \end{aligned}$$

It also holds $\operatorname{Err}\left(\hat{\boldsymbol{\beta}}_{\blacklozenge}^R(\Lambda)\right) = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\blacklozenge}^R(\Lambda)\right)' \Gamma \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\blacklozenge}^R(\Lambda)\right) + \Sigma_u$. Now, notice that

$$\left(T^{-1} ZZ' + \sqrt{T}\Lambda\right)^{-1} = O_P(1) \quad \text{and} \quad \left(T^{-1} ZZ' + \sqrt{T}\Lambda\right)^{-1} \left(\sqrt{T}\Lambda\right) = O_P(1),$$

since $\Lambda \in \mathcal{I}_\lambda$. It follows

$$\operatorname{Err}\left(\hat{\boldsymbol{\beta}}_{\blacklozenge}^R(\Lambda)\right) = \boldsymbol{\beta}' \left(\sqrt{T}\Lambda\right) \left(\frac{ZZ'}{T} + \sqrt{T}\Lambda \right)^{-1} \Gamma \left(\frac{ZZ'}{T} + \sqrt{T}\Lambda \right)^{-1} \left(\sqrt{T}\Lambda\right) \boldsymbol{\beta} + \Sigma_u + O_P(\sqrt{T}).$$

One can now consider a sequence $\widetilde{\Lambda} = O_p(T^{-1/2})$ of regularizers in $\mathcal{I}_\lambda$. By taking the limit, one

gets that

$$\begin{aligned}\lim_{T\to\infty} \operatorname{Err}\left(\hat{\boldsymbol{\beta}}^R_{\blacklozenge}\left(\widetilde{\Lambda}\right)\right) &= \boldsymbol{\beta}' \cdot O_p(1) \cdot (\Gamma + O_p(1))^{-1} \Gamma (\Gamma + O_p(1))^{-1} \cdot O_p(1) \cdot \boldsymbol{\beta} + \Sigma_u \\ &\geq \Sigma_u = \lim_{T\to\infty} \operatorname{Err}\left(\hat{\boldsymbol{\beta}}^R_{\blacklozenge}(0)\right),\end{aligned}$$

meaning that $\widetilde{\Lambda}$ can not be optimal asymptotically, since the least squares solution at $\Lambda = 0$ achieves a lower predictive error. Additionally, any sequence with lower convergence order is also asymptotically invalid. Therefore, by contradiction, it must hold that $\Lambda_{\blacklozenge} = o_p(T^{-1/2})$. □

Proof of Corollary 2.6.3

Proof. The first results follows directly from Theorem 2.5.6. Further, the fact that $\Lambda_{2,\blacklozenge} = O_P(1)$ is trivial because it is assumed that $\Lambda \in \mathcal{I}_\lambda$.

To prove the second part of the theorem, one can simply notice that, given the assumption on the coefficients $\boldsymbol{\beta}$ in Theorem 2.6.1,

$$\left(\frac{ZZ'}{T} + \sqrt{T}\Lambda\right)^{-1}\left(\sqrt{T}\Lambda\right)\boldsymbol{\beta} = \left(\frac{ZZ'}{T} + \sqrt{T}\Lambda\right)^{-1}\begin{bmatrix}\sqrt{T}\Lambda_1 \cdot \boldsymbol{\beta}_1 \\ \Lambda_2 \cdot T^{-\delta}\boldsymbol{b}_2\end{bmatrix} \xrightarrow{P} 0,$$

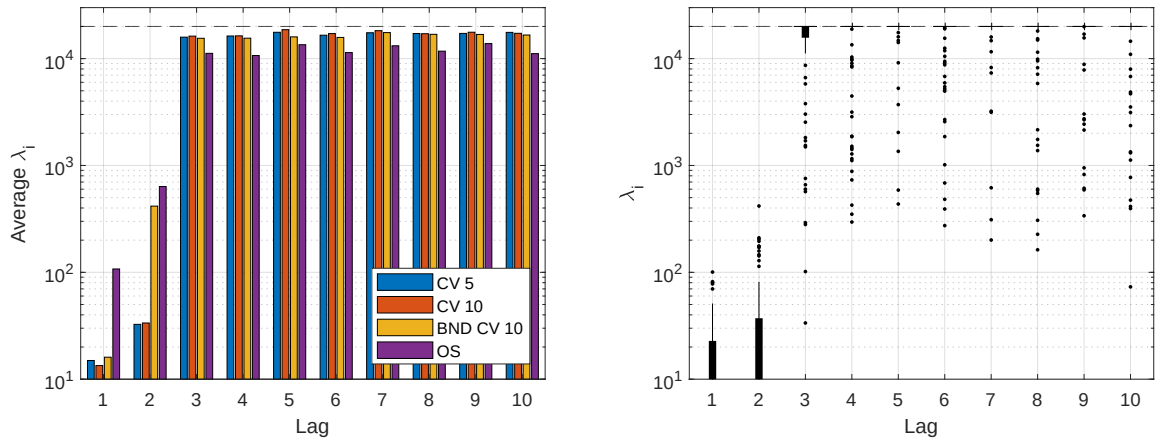
where Λ was partitioned into two diagonal blocks, Λ_1 and Λ_2, as done previously. Block Λ_1 must be $O_P(T^{-1/2})$ following the proof of Theorem 2.5.6. Finally, the fact that $T^{-\delta}\boldsymbol{b}_2 \to 0$ as $T \to \infty$ means that, in the limit, a nonzero Λ_2 does not yield a sub-optimal cross-validation loss value. □

2.D Monte Carlo Simulations

2.D.1 Cross-validation Details

To select the ridge penalty, I implement the lag-adapted structure and choose the relevant λ_i's using block non-dependent cross-validation (Burman et al., 1994, Bergmeir et al., 2018b). I constraint the optimization domain of λ_i to be $[0, 10^2]$, without discretization. An issue with cross-validation regards the GLS ridge estimator: the matrices involved can quickly become prohibitively large due to Kronecker products, making CV optimization impractical. To avoid this, I set the penalty for lag-adapted $\hat{\boldsymbol{\beta}}^{RGLS}(\Lambda)$ to be the same as that obtained for $\hat{\boldsymbol{\beta}}^R$ via CV, which means the regularizer is tuned sub-optimally. Nonetheless, an identical choice of Λ for both methods can help shed light on the difference in structure between the two estimators.

In contrast to Bańbura et al. (2010), I do not tune the shrinkage parameter of the Minnesota BVAR using a mean squared forecasting error (MSFE) criterion: instead, I again use block CV. Since a Minnesota prior can be easily implemented with the use of augmented regression matrices, cross-validation can be much more efficiently implemented than for $\hat{\boldsymbol{\beta}}^{RGLS}$. The resulting choice of prior tightness λ^2 is reasonable because CV, too, estimates the (one step ahead) forecasting risk. Since in this context only the mean of the posterior is used to compute pointwise impulse responses, one can even directly interpret the cross-validated Minnesota BVAR estimator as a refinement of GLS ridge.



(a) Average penalty selected by lag.

(b) Distribution of 10-fold CV penalties.

Figure 2.2: Comparison of penalty selection methods in Setup A, 1000 replications.

2.D.2 Penalty Selection in Simulations

In the simulations of Section 2.7, an interesting aspect to study is how data-driven penalty selection methods behave. Both average and individual behavior are important, because the former generally gives intuition for the kind of regularization structure that is selected for the model, while the latter is relevant for empirical modeling where estimation can be done only on one sample.

I use Setup A from the Monte Carlo experiments (with 1000 replications) and instead just focus on the behavior of a number of penalty validation techniques. The rational behind the choice is straightforward: Setup A involves models with more lags. Figure 2.2a shows the mean selected penalty parameter λ_i for $i \in \{1, \dots, 10\}$. The methods I compare are: out-of-sample validation (OOS) with a split of 80% of sample for estimation and 20% for testing; block cross-validation with 5 (CV 5) or 10 (CV 10) folds; block non-dependent cross validation with 10 folds (BND CV 10). The differences between 5 and 10-fold block CV methods are small, and both largely agree with BND cross-validation apart at lag 2. In contrast, out-of-sample validation appears to select on average much higher penalties at early lags and slightly lower ones at higher lags.

Regarding the distribution of selected λ_i over all replications, one can notice from Figure 2.2b that there is indeed important variation in the individual penalty choices. The evidence is for the specific case of 10-fold block CV, but it appears as a common pattern with other techniques, too. The implications of such variability in lag-adapted penalties are hard to gauge because in any given sample it is not possible to say whether the choice of $\{\lambda_i\}$ is good or bad outside of speculation. A guiding principle might be to compare CV with any "hyperpriors" one might have on Λ itself – like in Bayesian paradigm – but then parameteric penalty matrices like the one used with the Minnesota prior should be preferred. Indeed, the question of whether a more robust but still general method other than cross–validation can be applied to the time series context is highly relevant.

2.D.3 Penalty Selection with Many Lags

The minimization problem involved in OOS and CV penalty selection in its most general form suffers from the curse of dimensionality. This is somehow mitigated when using a lag-adapted regularizer since the loss with Λ^ℓ depends only on p non-negative parameters, rather then K^2p with non-block-diagonal Λ. But whenever p is chosen large (e.g. $p > 20$) the problem resurfaces.

I suggest a basic shortcut to make computation easier. Such simplification stems from the following observation. If one is willing to believe the assumptions of Section 2.6, then, because deep lag coefficients are small, after the first few lags penalization can be equally strong on all remaining lags with negligible additional bias. The shortcut, then, is to estimate only $\{\lambda_1, \dots, \lambda_r\}$ for $r < p$, then extrapolate and use $\{\lambda_1, \dots, \lambda_{r-1}, \lambda_r, \dots, \lambda_r\}$, where λ_r is repeated $p - r$ times, as lag-adapted penalty parameters. The idea is supported by the results in Figure 2.2a. However, this strategy is not generally appropriate, because it could be that even at relatively deep lags some coefficients are large, while on the other hand the early coefficients are small. Therefore, in applications where the ridge penalty needs to be estimated only once or a handful of times I would suggest to avoid this shortcut altogether.

2.D.4 Numerical Optimization

For a VAR(p) and a lag-adapted Λ^ℓ, a collection $\{\lambda_1, \dots, \lambda_p\}$ must be chosen. To implement OS and CV for the ridge estimators, I rely on MATLAB optimization routines, in both cases using the optimization function `patternsearch` from the MATLAB Optimization Toolbox. The domain of optimization is chosen to be the hypercube $[0, 10^2]^p$, where T is the sample size. The choice of a bounded domain is asymptotically valid, cf. Theorem 2.5.6 and Corollary 2.6.3.

In applications, since the CV loss needs not be convex (Stephenson et al., 2021), it appropriate to employ advanced optimization routines, e.g. genetic or pattern-based optimizers like `patternsearch`, if possible. When one only requires to estimate the VAR model once, then the selection of Λ is a one-time cost. The gains of better optimization solutions therefore are often superior to the higher computational costs one incurs in when using more sophisticated routines.

2.D.5 Additional Tables

Table 2.8: MSE Relative to OLS – Setup B

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Real GDP	RIDGE	1.11	1.08	1.16	1.06	0.90	0.89	0.94
	RIDGE-GLS	1.16	1.00	0.99	1.00	0.93	0.93	0.95
	LP	1.00	1.14	1.37	1.52	1.72	1.98	2.24
	BVAR-CV	0.90	0.87	1.04	1.01	0.92	0.92	0.98
	H-BVAR	0.83	0.62	0.78	0.73	0.62	0.62	0.68
GDP Deflator	RIDGE	2.25	2.10	1.81	1.54	1.39	1.47	3.37
	RIDGE-GLS	1.17	1.09	1.10	1.10	1.08	1.05	1.04
	LP	1.00	1.13	1.25	1.33	1.40	1.47	1.54
	BVAR-CV	1.06	1.03	1.00	0.95	0.92	0.92	0.93
	H-BVAR	0.81	0.91	1.10	1.11	1.04	0.98	0.94
Consumption	RIDGE	0.99	1.29	1.31	1.08	0.96	1.05	2.31
	RIDGE-GLS	0.94	0.96	1.04	0.99	0.95	0.96	0.99
	LP	1.00	1.13	1.32	1.44	1.63	1.83	2.03
	BVAR-CV	1.04	1.07	1.15	1.00	0.92	0.94	0.99
	H-BVAR	0.94	0.83	0.98	0.83	0.78	0.83	0.91
Investment	RIDGE	1.49	1.27	1.17	0.99	0.70	0.73	1.61
	RIDGE-GLS	1.34	1.14	1.02	1.02	0.86	0.82	0.86
	LP	1.00	1.15	1.40	1.63	2.03	2.76	3.59
	BVAR-CV	1.51	1.01	0.97	0.97	0.93	1.08	1.24
	H-BVAR	1.06	0.68	0.69	0.66	0.63	0.87	1.14
Hours	RIDGE	1.22	1.24	1.18	1.03	0.77	0.76	1.27
	RIDGE-GLS	1.07	1.05	1.01	1.03	0.90	0.85	0.90
	LP	1.00	1.14	1.33	1.53	1.81	2.35	2.92
	BVAR-CV	0.89	0.88	1.03	1.02	0.91	0.95	1.05
	H-BVAR	0.77	0.71	0.97	0.96	0.81	0.85	0.98
Compensation	RIDGE	0.85	0.99	0.85	0.94	1.13	1.40	4.70
	RIDGE-GLS	0.93	0.97	0.89	0.94	1.04	1.04	1.00
	LP	1.00	1.18	1.52	1.78	1.90	1.93	1.99
	BVAR-CV	1.07	0.92	0.94	0.93	0.99	1.01	0.98
	H-BVAR	0.86	0.80	1.12	1.31	1.35	1.27	1.22
Fed Funds Rate	RIDGE	2.17	1.21	0.96	0.93	1.03	4.00	53.18
	RIDGE-GLS	1.21	1.04	0.90	0.93	0.90	0.88	0.91
	LP	1.00	1.18	1.51	1.71	1.97	2.44	2.99
	BVAR-CV	0.92	0.94	0.91	0.90	0.86	0.87	0.92
	H-BVAR	0.75	0.77	1.32	1.38	1.25	1.15	1.20

Table 2.9: Impulse Response Inference – Setup B: CI Coverage

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Real GDP	LS	0.87	0.81	0.75	0.72	0.71	0.72	0.73
	RIDGE	0.90	0.79	0.66	0.62	0.65	0.68	0.68
	RIDGE-AS	0.89	0.72	0.61	0.58	0.61	0.65	0.65
	LP	0.87	0.93	0.94	0.94	0.93	0.93	0.91
	BVAR-CV	0.70	0.71	0.63	0.64	0.71	0.75	0.76
	H-BVAR	0.84	0.86	0.76	0.76	0.83	0.88	0.88
GDP Deflator	LS	0.86	0.83	0.80	0.76	0.73	0.72	0.70
	RIDGE	0.85	0.76	0.66	0.62	0.61	0.60	0.58
	RIDGE-AS	0.83	0.70	0.61	0.58	0.57	0.57	0.55
	LP	0.86	0.93	0.94	0.94	0.93	0.91	0.89
	BVAR-CV	0.76	0.72	0.70	0.72	0.72	0.71	0.70
	H-BVAR	0.84	0.83	0.79	0.78	0.78	0.78	0.77
Consumption	LS	0.87	0.80	0.75	0.72	0.70	0.70	0.70
	RIDGE	0.90	0.74	0.60	0.60	0.64	0.66	0.65
	RIDGE-AS	0.89	0.67	0.55	0.55	0.60	0.62	0.62
	LP	0.86	0.93	0.94	0.94	0.94	0.92	0.90
	BVAR-CV	0.73	0.66	0.60	0.63	0.70	0.73	0.74
	H-BVAR	0.84	0.79	0.70	0.74	0.79	0.82	0.84
Investment	LS	0.87	0.82	0.76	0.73	0.75	0.82	0.87
	RIDGE	0.85	0.79	0.65	0.62	0.73	0.80	0.81
	RIDGE-AS	0.82	0.69	0.59	0.57	0.68	0.77	0.77
	LP	0.87	0.94	0.94	0.95	0.94	0.94	0.94
	BVAR-CV	0.70	0.73	0.67	0.71	0.77	0.81	0.83
	H-BVAR	0.80	0.86	0.81	0.82	0.87	0.88	0.88
Hours	LS	0.86	0.81	0.76	0.74	0.74	0.79	0.81
	RIDGE	0.88	0.80	0.68	0.64	0.70	0.72	0.66
	RIDGE-AS	0.87	0.74	0.62	0.59	0.65	0.67	0.62
	LP	0.86	0.93	0.94	0.94	0.94	0.94	0.93
	BVAR-CV	0.73	0.72	0.64	0.66	0.74	0.78	0.77
	H-BVAR	0.88	0.86	0.73	0.72	0.80	0.85	0.86
Compensation	LS	0.86	0.82	0.76	0.75	0.72	0.68	0.67
	RIDGE	0.93	0.82	0.75	0.72	0.66	0.60	0.58
	RIDGE-AS	0.91	0.71	0.69	0.67	0.61	0.56	0.55
	LP	0.86	0.93	0.95	0.95	0.94	0.92	0.90
	BVAR-CV	0.78	0.71	0.69	0.71	0.69	0.67	0.69
	H-BVAR	0.85	0.83	0.80	0.81	0.82	0.82	0.83
Fed Funds Rate	LS	0.85	0.83	0.80	0.78	0.77	0.79	0.80
	RIDGE	0.79	0.77	0.74	0.68	0.68	0.72	0.72
	RIDGE-AS	0.78	0.66	0.68	0.64	0.64	0.68	0.69
	LP	0.85	0.94	0.96	0.96	0.95	0.94	0.93
	BVAR-CV	0.76	0.72	0.76	0.77	0.77	0.81	0.83
	H-BVAR	0.87	0.86	0.74	0.73	0.78	0.84	0.87

Table 2.10: Impulse Response Inference – Setup B: CI Length (rescaled ×100)

Variable	Method	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 16$	$h = 20$	$h = 24$
Real GDP	LS	0.71	1.56	2.07	2.31	2.32	2.24	2.15
	RIDGE	0.79	1.56	1.85	1.95	1.92	1.85	1.77
	RIDGE-AS	0.74	1.31	1.65	1.76	1.75	1.70	1.64
	LP	0.71	2.42	4.21	5.40	5.90	5.91	5.70
	BVAR-CV	0.53	1.23	1.74	2.00	2.10	2.13	2.15
	H-BVAR	0.58	1.36	1.87	2.16	2.32	2.44	2.55
GDP Deflator	LS	0.26	0.74	1.47	2.14	2.69	3.12	3.46
	RIDGE	0.32	0.81	1.41	1.96	2.43	2.82	3.13
	RIDGE-AS	0.31	0.71	1.28	1.79	2.23	2.59	2.88
	LP	0.26	1.09	2.68	4.39	5.95	7.23	8.13
	BVAR-CV	0.21	0.61	1.26	1.88	2.43	2.91	3.32
	H-BVAR	0.23	0.71	1.42	2.09	2.71	3.27	3.80
Consumption	LS	0.63	1.35	1.97	2.29	2.34	2.30	2.24
	RIDGE	0.71	1.35	1.74	1.93	1.96	1.93	1.88
	RIDGE-AS	0.67	1.15	1.57	1.75	1.79	1.77	1.74
	LP	0.63	2.04	3.92	5.27	5.89	6.00	5.88
	BVAR-CV	0.49	1.09	1.65	1.97	2.12	2.20	2.24
	H-BVAR	0.53	1.21	1.79	2.16	2.41	2.60	2.77
Investment	LS	3.38	6.65	7.89	7.89	7.31	6.69	6.18
	RIDGE	3.79	6.81	6.93	6.46	5.79	5.19	4.73
	RIDGE-AS	3.59	5.57	6.11	5.77	5.21	4.72	4.34
	LP	3.37	10.16	16.00	18.85	19.06	18.22	17.23
	BVAR-CV	2.64	5.26	6.59	6.91	6.78	6.57	6.38
	H-BVAR	2.89	5.74	7.08	7.54	7.63	7.60	7.58
Hours	LS	0.70	1.64	2.27	2.42	2.29	2.11	1.99
	RIDGE	0.82	1.73	2.10	2.06	1.86	1.66	1.52
	RIDGE-AS	0.79	1.49	1.87	1.85	1.68	1.51	1.40
	LP	0.70	2.49	4.52	5.62	5.83	5.57	5.26
	BVAR-CV	0.57	1.30	1.90	2.07	2.03	1.94	1.88
	H-BVAR	0.62	1.49	2.15	2.35	2.38	2.37	2.37
Compensation	LS	0.86	1.17	1.18	1.18	1.21	1.24	1.25
	RIDGE	0.97	1.21	1.07	1.05	1.06	1.06	1.06
	RIDGE-AS	0.93	0.95	0.95	0.96	0.97	0.98	0.98
	LP	0.86	1.80	2.53	2.87	3.11	3.26	3.32
	BVAR-CV	0.69	0.94	1.00	1.05	1.11	1.18	1.23
	H-BVAR	0.78	1.10	1.26	1.40	1.54	1.67	1.78
Fed Funds Rate	LS	0.25	0.39	0.43	0.43	0.41	0.38	0.35
	RIDGE	0.29	0.39	0.37	0.36	0.33	0.30	0.29
	RIDGE-AS	0.27	0.31	0.33	0.32	0.30	0.28	0.27
	LP	0.25	0.59	0.88	1.01	1.05	1.03	0.98
	BVAR-CV	0.21	0.31	0.36	0.37	0.36	0.35	0.34
	H-BVAR	0.23	0.36	0.42	0.44	0.45	0.45	0.46

Chapter 3

Impulse Response Analysis of Structural Nonlinear Time Series Models

3.1 Introduction

This paper presents a semi-nonparametric method to study the structural dynamic effects of unpredictable shocks in a class of nonlinear time series models.

Linear models are the foundation of economic structural time series modeling. The nature of linear models makes them especially tractable and apt at describing fundamental interactions and processes. For example, large classes of macroeconomic models in modern New Keynesian theory can be reduced to linear VARMA form via linearization techniques. This often justifies the application of the linear time series toolbox from a theoretical point of view. Concurrently, the work of Sims (1980) on VARs reinvigorated the strain of macroeconometric literature that seeks to study dynamic economic relationships. Brockwell and Davis (1991), Hamilton (1994b) and Lütkepohl (2005) provide detailed overviews of linear time series modeling and its developments. When the objects of interest are solely dynamic effects, the local projection (LP) approach of Jordà (2005) has also gained popularity as an alternative thanks to its flexibility and ease of implementation. LPs do not directly impose a linear model on the conditional distribution of the time series, but rather consist of linear lag regressions. Throughout this paper, the key dynamic effect under discussion will be the impulse response function (IRF), which is the common inference object of both linear VARMA and LP analyses.

Nonlinear methods seek to flexibly study the dependence structure between variables of interest by accommodating a potentially complex model structure. In recent years, research in nonlinear and asymmetric effects has grown, partly due to the increasing availability of data, making it feasible to estimate more elaborate models (Fuleky, 2020b). From a macroeconomic perspective, one can imagine at least three broad categories of nonlinearities that may be important to study. Sign-dependence of impulse responses is a potential key factor in the evaluation of monetary policy, as the specific effects of an interest rate change might be mitigated if the central bank implements a rate drop rather than a rate hike, while some others might be enhanced (Debortoli et al., 2020). If impulse responses are size-sensitive, large shocks and small shocks can have vastly different economic impacts, meaning that the policymaker must account for nonlinear scaling in the intensity of an intervention (Tenreyro and Thwaites, 2016). Finally, if the researcher's objective lies in studying exogenous changes impacting a variable that is nonlinear by definition, such as volatility indexes, any valid structural model should account for this feature.

The main contribution of this work is the development of an approach that allows estimating structural IRFs which can account for general nonlinear effects. This goal entails solving two related issues: first, structural identification of shocks, so that it is possible to give a valid economic interpretation to impulse responses; second, estimation of nonlinear functions in the setting of dependent data. In a linear setup, identification and estimation can be considered as distinct problems, but when working with nonlinear models these questions become intertwined. Without specific assumptions, nonlinear model classes are much too vast in terms of complexity: there are too many channels for any variable to affect any other. Disentangling such channels thus becomes impossible, and one cannot structurally interpret IRFs and dynamic effects such as multipliers. This problem can be solved by being more precise about the classes of models one is willing to entertain. I consider the structural nonlinear framework originally proposed by Gonçalves et al. (2021), which involves selecting one variable to identify the structural shocks of interest, X_t, and treating it separately from all other series, a vector Y_t, included in the model. By imposing a few additional assumptions on the dependence structure of innovations, one is able to include general nonlinear effects of X_t and its lags onto Y_t. By further allowing the lags of Y_t to influence X_t, this setup permits nonlinear dynamics to propagate to all variables over time. The significant upside of this paradigm is that structural identification is built-in, instead of being treated as a separate step. The latter path is most often taken in the literature by implementing the generalized impulse response function (GIRF) proposed by Koop et al. (1996). Kilian and Lütkepohl (2017b) have, however, highlighted that common linear identification strategies such as long-term and sign restrictions are generally impossible to impose in general nonlinear models, since closed-form expressions are not available but in a handful of special cases.

A weakness of the framework in Gonçalves et al. (2021) is that it requires choosing a specific functional form for the nonlinear components of the model, such as the negative-censoring map or a cubic map. These are used to tease out the sign and size effects of shocks.[1] Yet, correct prior knowledge of such terms is often unreasonable, especially in multivariate, multi-lag models. The natural way to avoid selecting a parametric nonlinear specification is to resort to semi-nonparametric techniques. Nonparametric time series methods have a long history in econometrics (Härdle et al., 1997), but until recently not much progress has been made in applying them to studying dynamic effects. Impulse response functions are objects that depend on the global properties of the model and, to be more precise, defining an IRF requires iterating shock perturbations over time. In a nonlinear model, the perturbation depends on the variables' state, so that one must consider the shock's effects across possible states. That is, different features of the nonlinear model such as level, slope, curvature must be evaluated over a range of values. Therefore, in this setting, an econometrician must provide error guarantees that are uniform over the variables' domain. In this work, I combine the uniform inference framework of Chen and Christensen (2015) with the structural nonlinear time series scheme discussed above. The general idea is to resort to semi-nonparametric series estimation and work in a physical dependence setup (Wu, 2005). On the one hand, I argue that physical dependence is a natural way of imposing assumptions that lead to estimable models, being more transparent than standard mixing conditions. On the other hand, the series approach

[1] The negative-censoring map applied to variable a is $a \mapsto \max(a, 0)$.

makes it easy to estimate models with linear and nonlinear components of the type considered in this paper. It also provides well-developed theoretical results to study uncertainty. Under appropriate regularity assumptions, I show that a two-step semi-nonparametric series estimation procedure is able to consistently recover the structural model in a uniform sense. This result encompasses the generated regressors' problem, which arises in the second step due to the structural identification strategy. Lastly, I prove that the nonlinear impulse response function estimates obtained are themselves asymptotically consistent and, thanks to an iterative algorithm, straightforward to compute in practice.

To validate the proposed methodology, I provide simulation evidence. The first set of results shows that, with realistic sample sizes, the efficiency costs of the semi-nonparametric procedure are small compared to correctly-specified parametric estimates. A second set of simulations demonstrates that whenever the nonlinear parametric model is mildly misspecified the large-sample bias is large, while for semi-nonparametric estimates it is negligible. Finally, I study how the IRFs computed with the new method compare with the ones from two previous empirical exercises. In a small, quarterly model of the US macroeconomy, I find that the parametric nonlinear and nonlinear appear to underestimate by intensity the GDP responses by 13% and 16%, respectively, after a large exogenous monetary policy shock. Moreover, sieve responses achieve maximum impact a year before their linear counterparts. Then, I evaluate the effects of interest rate uncertainty on US output, prices, and unemployment following Istrefi and Mouabbi (2018). In this exercise, the impact on industrial production of a one-deviation increase in uncertainty is approximately 54% stronger according to semi-nonparametric IRFs than the comparable linear specification. These findings suggest that structural impulse responses predicated on linear specifications might be appreciably underestimating shock effects.

Related Literature. Nonlinear models for dependent data have been extensively developed with the aim of analyzing diverse types of series, see e.g. the monographs of Tong (1990), Fan and Yao (2003), Gao (2007), Tsay and Chen (2018). Teräsvirta et al. (2010) provide a thorough discussion of nonlinear economic time series modeling, but, by only presenting the generalized IRF (GIRF) approach proposed by Koop et al. (1996), Potter (2000) and Gourieroux and Jasiak (2005), they do not explicitly address *structural* analysis.

Parametric nonlinear specifications are common prescriptions, for example, in time-varying models (Auerbach and Gorodnichenko, 2012, Caggiano et al., 2015) and state-depend models (Ramey and Zubairy, 2018). They have been and are commonly used in time-homogeneous models. Kilian and Vega (2011) provide a structural analysis of the effects of GDP on oil price shocks and, in contrast to previous literature, find that asymmetries play a negligible role: they do this by including a negative-censoring transformation of the structural variable and testing for significance. Caggiano et al. (2017), Pellegrino (2021) and Caggiano et al. (2021) use interacted VAR models to estimate effects of uncertainty and monetary policy shocks. From a finance perspective, Forni et al. (2023a,b) study the economic effects of financial shocks. Their generalized VMA specification, which is based on that of Debortoli et al. (2020), sets that innovations be transformed with the quadratic map.[2] Gambetti et al. (2022) study news shocks asymmetries by imposing that news

[2]I will discuss how their nonlinear model setup compares to the one I consider below.

changes enter their autoregressive model with a pre-specified threshold function.

Extension of nonparametric methods to nonlinear time series have already been discussed in the recent literature. For example, Kanazawa (2020) proposed to use radial basis function neural networks to estimate a nonlinear time series model of the US macroeconomy. This work focuses on estimating the GIRF of Koop et al. (1996), with its structural limitations: productivity is assumed to be a fully exogenous variable. Gourieroux and Lee (2023) provide a framework for nonparametric kernel estimation and inference of IRFs via local projections. Yet, they primarily work in the one-dimensional case and only mention economic identification in multivariate setups from the perspective of linear VARs. The work possibly closest to the present paper seems to be that of Lanne and Nyberg (2023), who develop a nearest-neighbor approach to impulse responses estimation that builds on the local projection idea and the GIRF concept. These papers, save for Gourieroux and Lee (2023), do not fully develop an asymptotic theory for their estimators, which makes it hard to judge the econometric assumptions under which they are applicable.

Outline. The remainder of this paper is organized as follows. Section 3.2 provides the general framework for the structural model. Section 3.3 describes the two-step semi-nonparametric estimation strategy, provides a thorough treatment of physical dependence assumptions and reports the key uniform consistency guarantees. Section 3.4 is devoted to the discussion of nonlinear impulse response function computation, validity and consistency. In Section 3.5, I report simulation results that show the performance of the proposed method, while in Section 3.6 I discuss empirical applications. Finally, Section 3.7 concludes. All proofs and additional technical results, as well as secondary plots, can be found in Appendices 3.A and 3.B, respectively.

Notation. A (vector) random variable will be denoted in capital or Greek letters, e.g. Y_t or ϵ_t, while its realization will be in lowercase Latin letters, that is y_t. For a process $\{Y_t\}_{t\in\mathbb{Z}}$, we write $Y_{t:s} = (Y_t, Y_{t+1}, \ldots, Y_{s-1}, Y_s)$, as well as $Y_{*:t} = (\ldots, Y_{t-2}, Y_{t-1}, Y_t)$ for the left-infinite history and $Y_{t:*} = (Y_t, Y_{t+1}, Y_{t+2}, \ldots)$ for its right-infinite history. The same notation is also used for random variable realizations. For a matrix $A \in \mathbb{R}^{d\times d}$ where $d \geq 1$, $\|A\|$ is the spectral norm, $\|A\|_\infty$ is the supremum norm and $\|A\|_r$ for $0 < r < \infty$ is the r-operator norm. For a random vector or matrix, I will use $\|\cdot\|_{L^r}$ to denote the associated L^r norm.

3.2 Model Framework

In this section, I introduce the nonlinear time series model that will be considered throughout the paper. This model setup will be a generalization of the one developed in Gonçalves et al. (2021) by letting the form of nonlinear components to remain unspecified until estimation. The idea behind the partial structural identification scheme is simple: if Z_t is the full vector of time series of interest, one must choose one series, call it X_t, as the structural variable, and add specific assumption on its dynamic effects on the remaining series, vector Y_t. The central goal will be the estimation of the impulse responses of Y_t due to a shock in X_t.

3.2.1 A Simple Nonlinear Monetary Policy Model

To begin with, it is useful to present a basic modeling setup with a straightforward economic interpretation, which may also serve as a concrete empirical example for the setting I will develop. To this end, I consider first a simple nonlinear monetary policy (MP) model which, however, captures all of the key ingredients of the general framework discussed in the next subsection. Consider the following hypothetical model of US macroeconomic time series:

$$\begin{aligned} X_t &= \rho X_{t-1} + \epsilon_{1t}, \\ \text{FFR}_t &= \alpha_{11}\text{FFR}_{t-1} + \alpha_{12}X_{t-1} + \beta_0^1\epsilon_{1t} + \epsilon_{21t}, \\ \text{GDP}_t &= \alpha_{21}\text{GDP}_{t-1} + \alpha_{22}\text{FFR}_{t-1} + G(X_t) + \beta_0^2\epsilon_{1t} + \epsilon_{22t}, \end{aligned}$$

where X_t is a structural monetary policy variable (for example, a credibility exogenous sequence of autocorrelated MP shocks), FFR_t is the Federal Funds Rate and GDP_t is the US Gross Domestic Product. Moreover, ϵ_{1t} and $\epsilon_{2t} := (\epsilon_{21t}, \epsilon_{22t})'$ are reciprocally independent sequences of shocks. Coefficients α_{11}, α_{12}, α_{21} and α_{22} induce a linear autoregressive structure for the endogenous variables FFR_t and GDP_t, while β_0^1 and β_0^2 determine the structural effects of ϵ_{1t} on FFR and GDP. Moreover, notice that the (sufficiently smooth) nonlinear function $G : \mathbb{R} \to \mathbb{R}$ implies that shocks ϵ_{1t} not only effect GDP contemporaneously in a linear fashion, but also nonlinearly through the level of X_t. To first aid conceptualization, one could think of setting G to be a known transformation which can tease out a specific effect of interest. For example, $G(X_t) = \max(0, X_t)$ to incorporate an asymmetry which depends on the sign of X_t.[3] Yet, a choice of G that is made a priori is hard to justify in general, and so the objective is rather to estimate G jointly with all other parameters. This will be the core issue at hand in the remainder of this paper.

To formally and effectively analyze this simple MP model and discuss its estimation, I separate the linear, nonlinear and structural parts. In vector form,

$$\begin{bmatrix} X_t \\ \text{FFR}_t \\ \text{GDP}_t \end{bmatrix} = \begin{bmatrix} \rho & 0 & 0 \\ \alpha_{12} & \alpha_{11} & 0 \\ 0 & \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} X_{t-1} \\ \text{FFR}_{t-1} \\ \text{GDP}_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ G(X_t) \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ \beta_0^1 & 1 & 0 \\ \beta_0^2 & 0 & 1 \end{bmatrix} \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{21t} \\ \epsilon_{22t} \end{bmatrix}$$

Now, by setting $Y_t := (\text{FFR}_t, \text{GDP}_t)'$ and $Z_t := (X_t, \text{FFR}_t, \text{GDP}_t)' \equiv (X_t, Y_t')'$, we obtain the equation

$$Z_t = A_1 Z_{t-1} + G_1(X_t) + B_0^{-1}\epsilon_t,$$

where A_1 is a matix function of $(\rho, \alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22})$, B_0^{-1} is a matrix function of (β_0^1, β_0^2) and $G_1(X_t) = (0, 0, G(X_t))'$. Throughout this paper, I will call the form in the above display the semi-reduced form, for reasons that will be made clear when presenting the general model. Finally, a key insight is that one can, with a mild abuse of notation, write G_1 too in a "functional matrix"

[3] See also the wider class of threshold autoregressive models (TAR) discussed by Fan and Yao (2003) and Teräsvirta et al. (2010).

form, that is

$$G_1(X_t) = \begin{bmatrix} 0 \\ 0 \\ G(X_t) \end{bmatrix} X_t \equiv \begin{bmatrix} 0 \\ 0 \\ G \end{bmatrix} X_t = G_1 X_t.$$

Below, this will formalism will prove very useful in terms of streamlining notation.

3.2.2 General Model

Let $Z_t := (X_t, Y_t')'$ where $X_t \in \mathcal{X} \subseteq \mathbb{R}$ and $Y_t \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$, and let $d = 1 + d_Y$ be the dimension of Z_t. I assume that the structural nonlinear data generating process has the form

$$B_0 Z_t = b + B(L) Z_{t-1} + F(L) X_t + \epsilon_t, \tag{3.1}$$

where $b = (b_1, b_2')' \in \mathbb{R}^d$ and $\epsilon_t = (\epsilon_1, \epsilon_2')' \in \mathcal{E} \subseteq \mathbb{R}^d$ are partitioned accordingly. Moreover, I assume that model (3.1) imposes a *linear* dependence of observables on Y_t and its lags, while series X_t can enter *nonlinearly*. That is, $B(L) = B_1 + B_2 L + \ldots + B_p L^{p-1}$ and $F(L) = F_0 + F_1 L + \ldots + F_p L^p$ are linear and functional lag polynomials, respectively.[4]

Matrices $(F_0, \ldots, F_p)$ are functional in the sense that their entries consist of real univariate functions, and the product between $F(L)$ and X_t is to be interpreted as functional evaluation, c.f. the example discussed above. That is,

$$F(L)X_t = \begin{bmatrix} f_{0,1}(X_t) \\ \vdots \\ f_{0,d}(X_t) \end{bmatrix} + \begin{bmatrix} f_{1,1}(X_{t-1}) \\ \vdots \\ f_{1,d}(X_{t-1}) \end{bmatrix} + \ldots + \begin{bmatrix} f_{p,1}(X_{t-p}) \\ \vdots \\ f_{p,d}(X_{t-p}) \end{bmatrix},$$

where $\{f_{j,l}\} \in \Lambda$ for $j = 0, \ldots, p$, $l = 1, \ldots, d$, and Λ is a sufficiently regular function class.[5] The modeling choice to remain within the autoregressive time series class with additive lag structure has two core advantages. First, it yields a straightforward generalization to classical linear models (Lütkepohl, 2005, Kilian and Lütkepohl, 2017b). Second, it keeps semi-nonparametric estimation of nonlinear components feasible. Additivity in variables and lags means that the curse of dimensionality involved with multivariate nonparametric estimation is effectively mitigated (Fan and Yao, 2003).

Let the lag polynomials be given by

$$B(L) = \begin{bmatrix} B_{11}(L) & B_{12}(L) \\ B_{21}(L) & B_{22}(L) \end{bmatrix}, \qquad F(L) = \begin{bmatrix} 0 \\ F_{21}(L) \end{bmatrix}.$$

This structural formulation means that the model equation for X_t is restricted to be linear in all regressors. It also implies that X_t does not depend contemporaneously on itself. Note that as long

[4] This is a minor abuse of notation compared to e.g. Lütkepohl (2005). The choice to use a matrix notation is due to the ease and clarity of writing a (multivariate) additive nonlinear model such as (3.1) in a manner consistent with standard linear VAR models. In cases where a real matrix $A \in \mathbb{R}^{d \times d}$ is multiplied with a conformable functional matrix F, I simply assume the natural product of a scalar times a function, e.g. $A_{ij} F_{k\ell}$, where $F_{k\ell}$ is a function, returning a new real function.

[5] To fix ideas, one may think of $\Lambda^q(M)$, the Hölder function class of smoothness $q > 0$ and domain $M \subseteq \mathbb{R}$. We shall make more precise assumptions regarding Λ in Section 3.3 when discussing model estimation.

as $B_{12}(L) \neq 0$, X_t still depends upon nonlinear functions of its own lags, which enter via lags of Y_t. Next, I impose that $B_0 \in \mathbb{R}^{d_Y \times d_Y}$ has the form

$$B_0 = \begin{bmatrix} 1 & 0 \\ -B_{0,12} & B_{0,22} \end{bmatrix},$$

where $B_{0,22}$ is non-singular and normalized to have unit diagonal. The structural model is thus given by

$$\begin{aligned} X_t &= b_1 + B_{12}(L)Y_{t-1} + B_{11}(L)X_{t-1} + \epsilon_{1t}, \\ B_{0,22}\, Y_t &= b_2 + B_{22}(L)Y_{t-1} + B_{21}(L)X_{t-1} + B_{0,12}X_t + F_{21}(L)X_t + \epsilon_{2t}. \end{aligned}$$

Moreover, it follows that B_0^{-1} exists and has form

$$B_0^{-1} = \begin{bmatrix} 1 & 0 \\ B_0^{21} & B_0^{22} \end{bmatrix}.$$

The constraints on B_0 yield a structural identification assumption and require that X_t be predetermined with respect to Y_t (Gonçalves et al., 2021). By introducing

$$\mu := B_0^{-1}b, \quad A(L) := B_0^{-1}B(L) \quad \text{and} \quad G(L) := B_0^{-1}F(L),$$

one thus obtains

$$\begin{aligned} X_t &= \mu_1 + A_{12}(L)Y_{t-1} + A_{11}(L)X_{t-1} + \epsilon_{1t}, \\ Y_t &= \mu_2 + A_{22}(L)Y_{t-1} + A_{21}(L)X_{t-1} + G_{21}(L)X_t + B_0^{21}\epsilon_{1t} + B_0^{22}\epsilon_{2t}, \end{aligned} \tag{3.2}$$

or, equivalently,

$$Z_t = \mu + A(L)Y_{t-1} + G(L)X_t + u_t, \tag{3.3}$$

where $u_t = [u_{1t}, u_{2t}]'$, $u_{1t} \equiv \epsilon_{1t}$ and $u_{2t} := B_0^{21}\epsilon_{1t} + B_0^{22}\epsilon_{2t}$. Given the structure of B_0^{-1}, one can see that $A_{12}(L) \equiv B_{12}(L)$, $A_{11}(L) \equiv B_{11}(L)$ and $G_{11}(L) = 0$. Importantly, one must also notice that $A_{12}(L)$ and $G_{21}(L) = B_0^{22}F_{21}(L)$ might now be not properly identified without further assumptions. Since $A_{21}(L)$ is not necessarily zero, linear effects of lags of X_t on Y_t can enter by means of both lag polynomials. To resolve this issue, I therefore assume that the functional polynomial $G_{21}(L)$ contains, at lags greater than zero, only *nonlinear* components.[6]

Example 3.2.1. (Bivariate Model with Exogenous Shocks). To give a concrete example of (3.2), assume that one wants to model the effects of monetary policy shocks on U.S. GDP growth following Romer and Romer (2004). Then, let

$$\begin{aligned} X_t &= \epsilon_{1t}, \\ Y_t &= \mu_2 + A_2 Y_{t-1} + G(X_t) + B_0^{21}\epsilon_{1t} + \epsilon_{2t}, \end{aligned}$$

where X_t are the policy shocks, which are assumed to be i.i.d., while Y_t is a macroeconomic

[6] When using a semi-nonparametric estimation strategy with B-splines, this will be feasible to implement numerically. When using wavelets, this also is a natural approach. In practice, however, some care must be taken to avoid constructing collinear regression matrices.

variable whose responses the researcher is interested in, e.g. GDP growth or PCE inflation. This setup is very minimal, and I assume here, for the sake of simplicity, that endogeneity of ϵ_{2t} does not pose a problem. Then, the term $G(X_t) + B_0^{21}\epsilon_{1t} \equiv G(\epsilon_{1t}) + B_0^{21}\epsilon_{1t} =: H(\epsilon_{1t})$ fully captures any contemporaneous effect of monetary policy shocks on Y_t. When $G(\epsilon_{1t}) = 0$, $H(\epsilon_{1t})$ and the model are linear. If $G(\epsilon_{1t}) = \beta_0 \max(0, \epsilon_{1t})$ for some $\beta_0 \neq 0$, function H is piece-wise linear: contractionary and expansionary shocks have, in general, different effects on Y_t, but shocks with the same sign have proportional impact. As a final example, if $G(\epsilon_{1t}) = \beta_0\, \epsilon_{1t}^3$ then $H(\epsilon_{1t})$ is a third-degree polynomial, so that both sign and size of monetary policy shocks are fundamental determinants of Y_t's impulse response function. In principle, to correctly quantify the repercussions of a specific monetary intervention a researcher must model all of these effects, unless they have a strong prior belief that either or both can be safely ignored. More complex nonlinear and asymmetric relations are also possible. A more robust strategy - as proposed in the present work - is to avoid choosing G (or H) as part of the model's specification, but rather to empirically estimate it jointly with all other coefficients.

Remark 3.2.1. (Constrained Models). The general approach of leaving $F(L)$ unconstrained is appealing when no precise economic intuition or information is available. However, there might be cases where the functional form of the nonlinear component is either partially known, or can be restricted. A simple restriction is that of a uniform functional over lags,

$$F(L) = F + FL + FL^2 + \ldots + FL^p.$$

This is a constraint effectively imposed by e.g. Gonçalves et al. (2021), Kilian and Vega (2011) and other references. They do this by fully specifying F, but nonparametric constraints may be desired, e.g. monotonicity. Constrained estimation of $F(L)$ is addressed in Remark 3.3.2 below.

The system of equations in (3.2) provides the so-called *pseudo-reduced form* model. By design, one does not need to identify the model fully, meaning that fewer assumptions on Z_t and ϵ_t are needed to estimate the structural effects of ϵ_{1t} on Y_t. This comes at the cost of not being able to simultaneously study structural effects with respect to ϵ_{2t}. An associated problem is that, in general, $G_{21}(L)X_t$ is correlated with innovation u_{2t} through $B_0^{21}\epsilon_{1t}$. The main challenge to structural shock identification of ϵ_{1t} thus lies in the fact that if $B_0^{21} \neq 0$ and $G_{21}(0) \neq 0$, there is endogeneity in the equations for Y_t since X_t depends linearly on ϵ_{1t}. Gonçalves et al. (2021) address the issue by proposing a two-step estimation procedure wherein one explicitly controls for ϵ_{1t} by using regression residuals $\widehat{\epsilon}_t$. In Section 3.3 below, I show that this approach also allows for consistent semi-nonparametric estimation of structural impulse responses.

Remark 3.2.2. (Identification Schemes). Forni et al. (2023a,b) provide an alternative nonlinear structural identification framework to that of Gonçalves et al. (2021). Their approach was originally introduced in Debortoli et al. (2020) and is based on the VMA form of the time series. Using the current notation, suppose that the structural representation of Z_t is given by

$$Z_t = b + Q(L)F(\epsilon_{1t}) + B(L)\epsilon_t$$

where ϵ_t are independent structural shocks with zero mean and identity covariance, while ϵ_{1t} iden-

tifies, e.g., financial innovations and shocks. $Q(L)$ and $B(L)$ are both linear lag polynomials and F is a nonlinear function to be specified by the researcher. Imposing some additional assumptions, the reduced-form assumed by Forni et al. (2023a) is

$$Z_t = \mu + A(L)Z_t + Q_0 F(\epsilon_{1t}) + B_0\epsilon_t, \tag{3.4}$$

where $F(x) = x^2$ in their baseline specification. Forni et al. (2023b) use an analogous model, while Debortoli et al. (2020) also consider more general setups where Q_0 is replaced by a general lag polynomial $D(L)$. These kinds of structural assumptions are similar but not identical to the ones imposed in Gonçalves et al. (2021) and this paper. For (3.4) to overlap with (3.2), one must assume that X_t is exogenous and independently distributed, so that its level does not affect the mapping of ϵ_{1t} through F. That is, (3.4) requires that only the *shocks* have nonlinear effects, not the structural variable itself. The upside of this approach is that one can directly and explicitly model asymmetry in the innovation process. The drawbacks are that, without a clear identification of a structural variable, one must fully identify B_0. Moreover, function F remains to be specified a priori. Note, however, that if innovation sequence ϵ_{1t} is observable, a generalization of the semi-nonparametric estimation results of this paper to the framework of Debortoli et al. (2020) would be straightforward.

I now state some preliminary assumptions for the model.

Assumption 1. $\{\epsilon_{1t}\}_{t\in\mathbb{Z}}$ and $\{\epsilon_{2t}\}_{t\in\mathbb{Z}}$ are mutually independent time series such that

$$\begin{bmatrix}\epsilon_{1t}\\ \epsilon_{2t}\end{bmatrix} \overset{\text{i.i.d.}}{\sim} \left(0, \begin{bmatrix}\sigma_1^2 & 0\\ 0 & \Sigma_2\end{bmatrix}\right)$$

where Σ_2 is a diagonal positive definite matrix.

Assumption 2. $\{Z_t\}_{t\in\mathbb{Z}}$ is strictly stationary, ergodic and such that $\sup_t \mathbb{E}[|Z_t|] < \infty$.

Assumption 3. The roots of equation $\det(I_d - A(L)L) = 0$ are outside the complex unit circle.

Assumption 1 follows Gonçalves et al. (2021). Assumption 2 is a high-level assumption on the properties of process $\{Z_t\}_{t\in\mathbb{Z}}$ and is common in the analysis of structural time series. Assumption 3 ensures that it is possible to invert lag polynomial $(I - A(L)L)$ in order to define impulse responses, as done below. However, Assumption 2 and 3 will not be sufficient to make sure that (3.2) is estimable from data, and in Section 3.3 additional constraints on $A(L)$ and $G(L)$ will be required in order to apply semi-nonparametric estimation. Moreover, Assumption 2 is not easily interpretable: functional lag polynomial $G(L)$ makes it impossible to reduce semi-structural equations (3.2) to an explicit infinite moving average form.

I will resolve both the former (sufficiency) and latter (interpretability) issue by using the non-linear dynamic model framework outlined by Pötscher and Prucha (1997). It will allow introducing regularity assumptions on the dependence of Z_t which enable the derivation of consistency of impulse response estimates.

3.2.3 Structural Nonlinear Impulse Responses

Starting from pseudo-reduced equations (3.2), by letting $\Psi(L) = (I_d - A(L)L)^{-1}$ one can further derive that

$$Z_t = \eta + \Theta(L)\epsilon_t + \Gamma(L)X_t, \tag{3.5}$$

where

$$\mu := \Psi(1)\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \quad \Theta(L) := \Psi(L)B_0^{-1}, \quad \text{and} \quad \Gamma(L) := \Psi(L)\begin{bmatrix}0\\ G_{21}(L)\end{bmatrix}.$$

To formally define impulse responses, it is useful to partition the polynomial $\Theta(L)$ according to

$$\Theta(L) := \begin{bmatrix}\Theta_{\cdot 1}(L) & \Theta_{\cdot 2}(L)\end{bmatrix},$$

where $\Theta_{\cdot 1}(L)$ represents the first column of matrices in $\Theta(L)$, and $\Theta_{\cdot 2}(L)$ the remaining d_Y columns.

Given impulse $\delta \in \mathbb{R}$ at time t, define the shocked innovation process as $\epsilon_{1s}(\delta) = \epsilon_s$ for $s \neq t$ and $\epsilon_{1t}(\delta) = \epsilon_t + \delta$, as well as the shocked structural variable as $X_s(\delta) = X_t$ for $s < t$ and $X_s(\delta) = X_s(Z_{t-1}, \epsilon_t + \delta, \epsilon_{t+1} \ldots, \epsilon_s)$ for $s \geq t$. Further, let

$$\begin{aligned} Z_{t+h} &:= \eta + \Theta_{\cdot 1}(L)\epsilon_{1t+h} + \Theta_{\cdot 2}(L)\epsilon_{2t+h} + \Gamma(L)X_t, \\ Z_{t+h}(\delta) &:= \eta + \Theta_{\cdot 1}(L)\epsilon_{1t+h}(\delta) + \Theta_{\cdot 2}(L)\epsilon_{2t+h} + \Gamma(L)X_t(\delta), \end{aligned}$$

be the time-t baseline and shocked series, respectively. The unconditional impulse response is given by

$$\text{IRF}_h(\delta) = \mathbb{E}\left[Z_{t+h}(\delta) - Z_{t+h}\right]. \tag{3.6}$$

The difference between shock and baseline is clearly

$$\begin{aligned} Z_{t+h}(\delta) - Z_{t+h} &= \Theta_{h,\cdot 1}\delta + \Gamma(L)X_t(\delta) - \Gamma(L)X_t \\ &= \Theta_{h,\cdot 1}\delta + (\Gamma_0 X_{t+h}(\delta) - \Gamma_0 X_{t+h}) + \ldots + (\Gamma_h X_t(\delta) - \Gamma_h X_t), \end{aligned}$$

therefore the unconditional IRF reduces to

$$\text{IRF}_h(\delta) = \Theta_{h,\cdot 1}\delta + \mathbb{E}\left[\Gamma_0 X_{t+h}(\delta) - \Gamma_0 X_{t+h}\right] + \ldots + \mathbb{E}\left[\Gamma_h X_t(\delta) - \Gamma_h X_t\right]. \tag{3.7}$$

Notice that, in (3.7), while one can linearly separate expectations in the impulse response formula, terms $\mathbb{E}\left[\Gamma_j X_{t+j}(\delta) - \Gamma_j X_{t+j}\right]$ for $0 \leq j \leq h$ cannot be meaningfully simplified. Coefficients Γ_j are functional, therefore it is not possible to collect them across $X_{t+j}(\delta)$ and X_{t+j}. Moreover, these expectations involve nonlinear functions of lags of X_t and cannot be computed explicitly. To address this issue, Section 3.4 provides an iterative procedure that makes computation of nonlinear impulse responses in (3.7) straightforward.

Remark 3.2.3. (Local Projection Approaches). As mentioned in the introduction, in recent years there has been growing interest in nonlinear IRF estimation procedures, and, accordingly, ways to generalize the LP framework. Jordà (2005) already suggested that nonlinear impulse responses can, in principle, be directly estimated with local projections via the so-called *flexible local projection* approach. The flexible LP method relies on the Volterra expansion of time series to account for

nonlinearities. There are multiple issues with this method. First, Jordà (2005) does not directly state how the validity of Volterra series implies the autoregressive form used in the LP regression. Second, the flexible LP proposal is fundamentally equivalent to adding polynomial factors to the linear regression specification. Thus, it is effectively a semi-nonparametric method, yet Jordà (2005) does not provide a theoretical analysis from this viewpoint. Moreover, no criterion or empirical rule-of-thumb for selecting the truncation order of the Volterra expansion are suggested, which becomes a key issue in practice. Due to these concerns, application of flexible LPs seems hard to justify from an econometric perspective.[7] Lanne and Nyberg (2023) propose to nonparametrically recover the conditional mean function with a nearest-neighbor (k-NN) regression estimator. Their method is very flexible, but requires appropriately choosing the neighborhood size k and a distance measure for histories of realizations, and the authors do not theoretically address these issues. Very recently, Gourieroux and Lee (2023) have considered nonlinear IRF estimation with kernel-based methods by means of a novel conditional quantile representation of the process. They prove kernel LP estimators based on such representation are consistent, and that the direct estimator is asymptotically normal. The theory is developed only for the univariate case, with an autoregressive structure of lag order one, limiting the applicability of their procedure.

3.3 Estimation

Pseudo-reduced form model (3.2) can be compactly rewritten as

$$\begin{aligned} X_t &= \Pi_1' W_{1t} + \epsilon_{1t}, \\ Y_t &= \Pi_2' W_{2t} + u_{2t}, \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} \Pi_1 &:= \Big(\eta_1,\ A_{1,11},\ \cdots,\ A_{p,11},\ A_{1,12}',\ \cdots,\ A_{p,12}'\Big)' \in \mathbb{R}^{1+pd}, \\ \Pi_2 &:= \Big[\eta_2 \quad G_{1,21} \quad \cdots \quad G_{p,21} \quad A_{1,22} \quad \cdots \quad A_{p,22} \quad B_0^{21}\Big]', \\ Z_{t-1:t-p} &:= \Big(X_{t-1},\ \ldots,\ X_{t-p}, Y_{t-1}',\ \ldots,\ Y_{t-p}'\Big)' \in \mathbb{R}^{pd}, \\ W_{1t} &:= \Big(1,\ Z_{t-1:t-p}'\Big)' \in \mathbb{R}^{1+pd}, \\ W_{2t} &:= \Big(1,\ X_t, Z_{t-1:t-p}',\ \epsilon_{1t}\Big)' \in \mathbb{R}^{3+pd}. \end{aligned}$$

Additionally, let $W_1 = (W_{11}, \ldots, W_{1n})'$ and $W_2 = (W_{21}, \ldots, W_{2n})'$ be the design matrices for X_t and Y_t, respectively.

Two-step Estimation Procedure. Since W_{2t} is an infeasible vector of regressors, to estimate Π_2 one can use $\widehat{W}_{2t} = (1, X_t, Z_{t-1:t-p}', \widehat{\epsilon}_{1t})'$, which now contains generated regressors in the form of residual $\widehat{\epsilon}_{1t}$. This approach is an adaptation of the two-step procedure put forth by Gonçalves et al. (2021), where I allow for semi-nonparametric estimation:

[7] Moreover, the complexity of estimating Volterra kernels grows exponentially with the kernel order, and thus more sophisticated approaches have been proposed to make estimation feasible, see e.g. Sirotko-Sibirskaya et al. (2020) and Movahedifar and Dickhaus (2023).

1. Regress X_t onto W_{1t} to get estimate $\widehat{\Pi}_1$ and compute residuals $\widehat{\epsilon}_{1t} = X_t - \widehat{\Pi}_1' W_{1t}$.

2. Fit Y_t using $\widehat{W}_{2t}$ to get estimate $\widehat{\Pi}_2$. Since $G_{1,21}, \ldots, G_{p,21}$ contain functional parameters, a semi-nonparametric estimation method is required.

3. Compute coefficients in $\widehat{\Theta}(L)$ and $\widehat{\Gamma}(L)$ from $\widehat{\Pi}_1$ and $\widehat{\Pi}_2$.

4. Consider the two paths with time t shocks $\epsilon_t + \delta$ versus ϵ_t: to construct the unconditional IRF, average over histories as well as future shocks by using the algorithm detailed in Proposition 3.4.1 or Proposition 3.4.2.

Gonçalves et al. (2021) only allow for pre-determined nonlinear transforms of X_t. The core contribution of this paper is allowing $G_{1,21}, \ldots, G_{p,21}$ to be estimated in a nonparametric way. I focus on series estimation in order to build on the extensive theory available in the setting of dependent data (Chen, 2013, Chen and Christensen, 2015). This further adds to the framework of Gonçalves et al. (2021), as their regularity assumptions are stated only as preconditions for a uniform LLN to hold and are not easy to interpret.

Remark 3.3.1. (Alternative Estimation Approaches). One does not need to limit estimation of the nonlinear functional parameters $G_{1,21}, \ldots, G_{p,21}$ to series-type estimators. The literature on nonparametric regression is mature, and thus kernel (Tsybakov, 2009), nearest-neighbor (Li and Racine, 2009), partitioning (Cattaneo et al., 2020) and deep neural network (Farrell et al., 2021) estimators are all potentially valid alternatives. For example, Huang et al. (2014) use kernel regression to perform density estimation and regression under physical dependence. However, thanks to both availability of uniform inference results (see also Belloni et al. 2015b) and ease of implementation, series methods stand out as a choice for semi-nonparametric time series estimation and nonlinear impulse response computation.

In the reminder of this section, I first introduce the semi-nonparametric series estimation strategy in detail. Then, I outline the core assumptions of the sieve setup. Special focus is put on the dependence structure of the data: rather than directly assuming β-mixing as in Chen and Christensen (2015), I shall consider physical dependence assumptions (Wu, 2005). to provide transparent conditions on the model itself that, if satisfied, ensure consistency. I prove that the proposed two-step semi-nonparametric procedure is uniformly consistent under physical dependence assumptions. These assumptions can be imposed directly on the model, and, as such, may be empirically checked, if necessary. The uniform asymptotic guarantees are first stated for the infeasible estimator involving true innovations ϵ_{1t} and later extended to encompass feasible estimator $\widehat{\Pi}_2$.

3.3.1 Semi-nonparametric Series Estimation

Starting from (3.8), one can introduce the ith-row coefficient matrices

$$G_i^{21} = \begin{bmatrix} G_{1,21} & \cdots & G_{p,21} \end{bmatrix}_i,$$
$$A_i^{22} = \begin{bmatrix} A_{1,22} & \cdots & A_{p,22} \end{bmatrix}_i,$$

and B_{0i}^{21} accordingly. Consider now the regression problem for each individual component of Y_t,

$$Y_{it} = G_i^{21} X_{t:t-p} + A_i^{22} Y_{t-1:t-p} + B_{0i}^{21} \epsilon_{1t} + u_{2it},$$

where $X_{t:t-p} := (X_t, \ldots, X_{t-p})'$ and $i = 1, \ldots, d_Y$. For simplicity of notation, I suppress intercept η_{2i}, but this is without loss of generality. Since G_i^{21} consists of $1+p$ functional coefficients and A_i^{22} can be segmented into p row vectors of length d_Y, it is possible to rewrite the above as

$$Y_{it} = \sum_{j=0}^{p} g_{ij}^{21}(X_{t-j}) + \sum_{j=1}^{p} A_{ij}^{22} Y_{t-j} + B_{0i}^{21} \epsilon_{1t} + u_{2it}. \tag{3.9}$$

I will use $\pi_{2,i} := [G_i^{21} \; A_i^{22} \; B_{0i}^{21}]'$ to identify the vector of coefficients in the equation for the ith component of Y_t. From (3.9), Π_2' can be decomposed in d_Y rows of coefficients, i.e.

$$\begin{bmatrix} Y_{1t} \\ \vdots \\ Y_{d_Y t} \end{bmatrix} = \begin{bmatrix} \pi_{2,1} \\ \vdots \\ \pi_{2,d_Y} \end{bmatrix} W_{2t} + u_{2t}$$

and one can treat each equation separately.

A semi-nonparametric series estimator for (3.9) is built on the idea that, if functions g_{ij}^{21} belong to an appropriate functional space, one can construct a growing collection of sets of basis functions – called a *sieve* – which, linearly combined, progressively approximate g_{ij}^{21}. That is, one can reduce the infinite dimensional problem of estimating the functional coefficients in $\pi_{2,i}$ to a linear regression problem. Although (3.9) features a sum of possibly nonlinear functions in $\{X_{t-j}\}_{j=0}^{p}$, as well as linear functions of $\{Y_{t-j}\}_{j=1}^{p}$ and ϵ_{1t}, constructing a sieve is straightforward.[8]

Assume that $g_{ij}^{21} \in \Lambda$, where Λ is a sufficiently regular function class to be specified in the following, and let $\mathbf{B}_\Lambda$ be a sieve for Λ. Let $b_{1\kappa}, \ldots, b_{\kappa\kappa}$ be the collection of $\kappa \geq 1$ sieve basis functions from $\mathbf{B}_\Lambda$ and define

$$\begin{aligned} b^\kappa(x) &:= (b_{1\kappa}(x), \ldots, b_{\kappa\kappa}(x))', \\ B_\kappa &:= (b^\kappa(X_{1:1-p}), \ldots, b^\kappa(X_{n:n-p}))'. \end{aligned}$$

The sieve space for $\pi_{2,i}$ is $\mathbf{B}_\Lambda^{1+p} \times \mathbb{R}^{1+pd_Y}$, where here $\mathbb{R}$ identifies the space of linear functions. Since the nonparametric components of Π_2 are linearly separable in the lag dimension, I take $\mathbf{B}_\Lambda^{1+p}$ to be a direct product of sieve spaces.[9] Importantly, the same sieve can be used for all components of Y_t, as I assume the specification of the model does not change across i.

Let $b_{\pi,1K}, \ldots, b_{\pi,KK}$ be the sieve basis in $\mathbf{B}_\Lambda^{1+p} \times \mathbb{R}^{1+pd_Y}$ which, for $\kappa \geq 1$ and $K = p\kappa + (1+pd_Y)$, is given by

$$b_{\pi,1K}(W_{2t}) = b_{1\kappa}(X_t),$$

[8] See Chen (2007) for a comprehensive exposition of sieve estimation. Chen and Shen (1998) and Chen (2013) also provide additional examples of partially linear semi-nonparametric models under dependence.

[9] It is not necessary to consider the more general case of tensor products of 1D sieve functions, as it would be the case for a general $(1+d_Y)$-dimensional function $G_i^{21}(X_t, X_{t-1}, \ldots, X_{t-p})$. As previously discussed, the additive structure avoids the curse of dimensionality which in nonlinear time series modeling if often a primary concern when working with moderate sample sizes (Fan and Yao, 2003).

$$
\begin{aligned}
&\vdots \\
b_{\pi,(p\kappa)K}(W_{2t}) &= b_{\kappa\kappa}(X_{t-p}), \\
b_{\pi,(p\kappa+1)K}(W_{2t}) &= Y_{t-1,1}, \\
&\vdots \\
b_{\pi,(K-1)K}(W_{2t}) &= Y_{t-p,d_Y}, \\
b_{\pi,KK}(W_{2t}) &= \epsilon_{1t},
\end{aligned}
$$

where κ fixes the size of the nonparametric component of the sieve. Note that K, the overall size of the sieve, grows linearly in κ, which itself controls the effective dimension of the nonparametric component of the sieve, $b_{\pi,1\kappa}, \ldots, b_{\pi,\kappa\kappa}$. In all theoretical results, I will focus on the growth rate of K rather than κ, as asymptotically they differ at most by a constant multiplicative factor.

The regression equation for $\pi_{2,i}$ is

$$Y_i = \pi_{2,i}' W_2 + u_{2i},$$

where $Y_i = (Y_{i1}, \ldots, Y_{in})'$ and $u_{2i} = (u_{2i1}, \ldots, u_{2in})'$. The estimation target is the conditional expectation $\pi_{2,i}(w) = \mathbb{E}[Y_{it} \mid W_{2t} = w]$ under the assumption $\mathbb{E}[u_{2it} \mid W_{2t}] = 0$. By introducing

$$
\begin{aligned}
b_\pi^K(w) &:= (b_{\pi,1K}(w), \ldots, b_{\pi,KK}(w))', \\
B_\pi &:= \left(b_\pi^K(W_{21}), \ldots, b_\pi^K(W_{2n})\right)',
\end{aligned}
$$

the *infeasible least squares series estimator* $\widehat{\pi}_{2,i}^*(w)$ is given by

$$\widehat{\pi}_{2,i}^*(w) = b_\pi^K(w)'(B_\pi' B_\pi)^{-1} B_K' Y_i.$$

Similarly, consider the feasible series regression matrices

$$
\begin{aligned}
b_\pi^K(w) &:= (b_{\pi,1K}(w), \ldots, b_{\pi,KK}(w))', \\
\widehat{B}_\pi &:= \left(b_\pi^K(\widehat{W}_{21}), \ldots, b_\pi^K(\widehat{W}_{2n})\right)'.
\end{aligned}
$$

Thus, the *feasible least squares series estimator* is

$$\widehat{\pi}_{2,i}(w) = b_\pi^K(w)'(\widehat{B}_\pi' \widehat{B}_\pi)^{-1} \widehat{B}_K' Y_i.$$

Given that the semi-nonparametric estimation problem is the same across i, to further streamline notation, where it does not lead to confusion I will let π_2 be a generic coefficient vector belonging to $\{\pi_{2,i}\}_{i=1}^p$, as well as define $\widehat{\pi}_2$, Y and u_2 accordingly.

Remark 3.3.2. (Constrained Sieve Estimation). The idea of constrained estimation was only briefly touched upon in Remark 3.2.1. In fully parametric nonlinear models, constraints are often imposed out of necessity or simplicity. If, say, $G_{1,21}$ is constituted only of the negative-censoring map, it is unclear why $G_{2,21}$ would be constituted instead of quadratic or cubic functions, for example. That is, *specific* parametric assumptions can be either unreasonable or hard to justify in

practice.[10] Yet, constrained semi-nonparametric estimation might be desirable at times.

If the shape of the regression function is to be constrained to ensure e.g. non-negativity, monotonicity or convexity, Chen (2007) gives examples of shape-preserving sieves, like cardinal B-spline wavelets. Constraints on a generic sieve can also be imposed at estimation time. For example, for simplicity suppose $d_Y = 1$ and $p = 2$, and that one wants to impose $G_{1,21} = G_{2,21}$. The constrained sieve estimator then solves

$$\min_{\boldsymbol{\beta}} \sum_{t=p+1}^{n} \left(Y_t - \boldsymbol{\beta}' b_\pi^K(W_{2t})\right)^2 \quad \text{subject to} \quad [I_\kappa, -I_\kappa, 0_{\kappa\times(1+pd_Y)}]\boldsymbol{\beta} = 0.$$

Analysis of restricted or constrained estimators, however, is still a challenging problem in nonparametric theory, c.f. Horowitz and Lee (2017), Freyberger and Reeves (2018), Chetverikov et al. (2018). Misspecification in particular is complex to address. Accordingly, I will not be imposing any specific restrictions on the nonlinear functions in Π_2 outside the ones necessary to derive uniform asymptotic theory.

Spline Sieve. The B-spline sieve $\text{BSpl}(\kappa, [0,1]^{d_Y}, r)$ of degree $r \geq 1$ over $[0,1]^{d_Y}$ can be constructed using the Cox-de Boor recursion formula. Alternatively, an equivalent way of constructing the spline sieve is as follows. For simplicity, let $d_Y = 1$ and let $0 < m_1 < \ldots < m_{\kappa-r-1} < 1$ be a set of knots. Then

$$b_{\text{spline}}^\kappa(x) := \left(1, x, x^2, \ldots, x^r, \max(x - m_1, 0)^r, \ldots, \max(x - m_{\kappa-r-1}, 0)^r\right)'.$$

The resulting spline sieve is piece-wise polynomial of degree r. Moreover, notice that in practice the spline sieve already contains a linear and constant term, so care must be taken to avoid collinearity (for example, by not including an additional intercept and linear term in X_t in the series regression).

3.3.2 Distributional and Sieve Assumptions

To develop the asymptotic uniform consistency theory, I rely on the general theoretical framework established by Chen and Christensen (2015). Basic distributional and sieve assumptions can be carried over from their setup mostly unchanged.

Assumption 4. (i) $\{\epsilon_t\}_{t\in\mathbb{Z}}$ are such that $\epsilon_t \overset{\text{i.i.d.}}{\sim} (0, \Sigma)$, (ii) $\{\epsilon_{1t}\}_{t\in\mathbb{Z}}$ and $\{\epsilon_{2t}\}_{t\in\mathbb{Z}}$ are mutually independent, (iii) $\epsilon_t \in \mathcal{E}$ for all $t \in \mathbb{Z}$ where $\mathcal{E} \subset \mathbb{R}^{d_Y}$ is compact, convex and has nonempty interior.

Assumption 5. (i) $\{Z_t\}_{t\in\mathbb{Z}}$ is a strictly stationary and ergodic time series, (ii) $X_t \in \mathcal{X}$ for all $t \in \mathbb{Z}$ where $\mathcal{X} \subset \mathbb{R}$ is compact, convex and has nonempty interior, (iii) $Y_t \in \mathcal{Y}$ for all $t \in \mathbb{Z}$ where $\mathcal{Y} \subset \mathbb{R}^{d_Y}$ is compact, convex and has nonempty interior.

Assumptions 4(i)-(ii) are a repetition of Assumption 1. As W_{2t} depends only on $X_{t:t-p}$, $Y_{t-1:t-p}$, and ϵ_{1t}, Assumption 1 also implies that entries of u_{2t} are independent of W_{2t}, so that $\mathbb{E}[u_{2it} \,|\, W_{2t}] = 0$.[11] Assumption 5(i) also follows from Assumption 2. However, thanks to the results derived in

[10] For more precise examples and a more in-depth discussion, see Section 2.1 of Chen (2013).

[11] Moreover, for any given i, the sequence $\{u_{2it}\}_{t\in\mathbb{Z}}$ is i.i.d. over time index t.

Section 3.3.3, below I will impose more primitive conditions on the model for Z_t that allow to recover 5(i). Assumption 4(iii) and Assumptions 5(ii)-(iii) imply that X_t, Y_t, as well as ϵ_t are bounded random variables. In (semi-)nonparametric estimation, imposing that X_t be bounded almost surely is a standard assumption. Since lags of Y_t and innovations ϵ_t contribute linearly to all components of Z_t, it follows that they too must be bounded. Unbounded regressors are more complex to handle when working in the nonparametric setting. Generalization from bounded to unbounded domains under dependence has already been discussed by e.g. Fan and Yao (2003). Chen and Christensen (2015) also allow for an expanding support by using weighted sieves. I leave this extension for future work.

It is, however, important to highlight that bounded support assumptions are relatively uncommon in time series econometrics. This is clear when considering the extensive literature available on linear models such as, e.g., state-space, VARIMA and dynamic factor models (Hamilton, 1994a, Lütkepohl, 2005, Kilian and Lütkepohl, 2017b, Stock and Watson, 2016). Avoiding Assumptions 4(iii) and 5(iii) can possibly be achieved with a change in the model's equations – so that, for example, lags of Y_t only effect X_t either via bounded functions or not at all – so I do not discuss this approach here. In practice, Assumptions 4(ii) and 5(ii)-(iii) are not excessively restrictive, as most credibly stationary economic series often have reasonable implicit (e.g. inflation) or explicit bounds (e.g. employment rate).[12]

Let $\mathcal{F}_t = \sigma(\ldots, \epsilon_{1t-1}, u_{2t-1}, Y_{t-1}, \epsilon_{1t}, u_{2t}, Y_t)$ be the natural filtration defined up to time t. Thanks to Assumptions 4 and 5 the following moment requirements hold trivially.

Assumption 6. (i) $\mathbb{E}[u_{2it}^2|\mathcal{F}_{t-1}]$ is uniformly bounded for all $t \in \mathbb{Z}$ almost surely, (ii) $\mathbb{E}[|u_{2it}|^{2+\delta}] < \infty$ for some $\delta > 0$, (iii) $\mathbb{E}[|Y_{it}|^{2+\delta}]$ is uniformly bounded for all $t \in \mathbb{Z}$ almost surely, and (iv) $\mathbb{E}[Y_{it}^2|\mathcal{F}_{t-1}] < \infty$ for any $\delta > 0$.

Now let $\mathcal{W}_2 \subset \mathbb{R}^d$ be the domain of W_{2t}. By assumption, $\mathcal{W}_2$ is compact and convex and is given by the direct product

$$\mathcal{W}_2 = \mathcal{X}^{1+p} \times \mathcal{Y}^p \times \mathcal{E}_1,$$

where $\mathcal{E}_1$ is the domain of structural innovations ϵ_{1t} i.e. $\mathcal{E} \equiv \mathcal{E}_1 \times \mathcal{E}_2$.

Assumption 7. Define $\zeta_{K,n} := \sup_{w \in \mathcal{W}_2} \|b_\pi^K(w)\|$ and

$$\lambda_{K,n} := [\lambda_{\min}(\mathbb{E}[\, b_\pi^K(W_{2t}) b_\pi^K(W_{2t})' \,])]^{-1/2}.$$

It holds:

(i) There exist $\omega_1, \omega_2 \geq 0$ s.t. $\sup_{w \in \mathcal{W}_2} \|\nabla b_\pi^K(w)\| \lesssim n^{\omega_1} K^{\omega_2}$.

(ii) There exist $\overline{\omega}_1 \geq 0$, $\overline{\omega}_2 > 0$ s.t. $\zeta_{K,n} \lesssim n^{\overline{\omega}_1} K^{\overline{\omega}_2}$.

(iii) $\lambda_{\min}(\mathbb{E}[\, b^K(W_{2t}) b^K(W_{2t})' \,]) > 0$ for all K and n.

Assumption 7 provides mild regularity conditions on the families of sieves that can be used for the series estimator. More generally, letting $\mathcal{W}_2$ be compact and rectangular makes Assumption 7

[12] This is not true, of course, when modeling extreme events like natural disasters, wars or financial crises. To study these types of series, however, researchers often apply specialized models. Thinking in this direction, a future development could be to extend the framework presented here to allow for innovations with unbounded support.

hold for commonly used basis functions (Chen and Christensen, 2015).[13] In particular, Assumption 7(i) holds with $\omega_1 = 0$ since the domain is fixed over the sample size.

In the proofs, it is useful to consider the orthonormalized sieve basis. Let

$$\widetilde{b}_\pi^K(w) := \mathbb{E}\left[b_\pi^K(W_{2t}) b_\pi^K(W_{2t})' \right]^{-1/2} b_\pi^K(w),$$
$$\widetilde{B}_\pi := \left(\widetilde{b}_\pi^K(W_{21}), \ldots, \widetilde{b}_\pi^K(W_{2n}) \right)'$$

be the orthonormalized vector of basis functions and the orthonormalized regression matrix, respectively.

Assumption 8. It holds that $\|(\widetilde{B}_\pi' \widetilde{B}_\pi / n) - I_K\| = o_P(1)$.

Assumption 8 is the key assumption imposed by Chen and Christensen (2015) to derive uniform converges rates under dependence. They prove that if $\{W_{2t}\}_{t\in\mathbb{Z}}$ is strictly stationary and β-mixing – with either geometric or algebraic decay, depending on the sieve family of interest – then Assumption 8 holds. Let $(\Omega, \mathcal{Q}, \mathbb{P})$ be the underlying probability space and define

$$\beta(\mathcal{A}, \mathcal{B}) := \frac{1}{2} \sup \sum_{(i,j)\in I\times J} |\mathbb{P}(A_i \cap B_i) - \mathbb{P}(A_i)\mathbb{P}(B_i)|$$

where $\mathcal{A}, \mathcal{B}$ are two σ-algebras, $\{A_i\}_{i\in I} \subset \mathcal{A}$, $\{B_j\}_{j\in J} \subset \mathcal{B}$ and the supremum is taken over all finite partitions of Ω. The h-th β-mixing coefficient of process $\{W_{2t}\}_{t\in\mathbb{Z}}$ is defined as

$$\beta(h) = \sup_t \beta\left(\sigma(\ldots, W_{2t-1}, W_{2t}), , \sigma(W_{2t+h}, W_{2t+h+1}, \ldots)\right),$$

and W_{2t} is said to be *geometric* or *exponential* β*-mixing* if $\beta(h) \leq \gamma_1 \exp(-\gamma_2 h)$ for some $\gamma_1 > 0$ and $\gamma_2 > 0$. The main issue with mixing assumptions is that they are, in general, hard to compute and evaluate. Therefore, especially in nonlinear systems, assuming that $\beta(h)$ decays exponentially over h imposes very high-level assumptions on the model. There are, however, many setups in which it is known that β-mixing holds under primitive assumptions (see Chen (2013) for examples).

In the next subsection, I will argue that using a different concept of dependence - one rooted in a physical understanding of the underlying stochastic process - leads to imposing transparent assumptions on the model's structure.

3.3.3 Physical Dependence Conditions

Consider now a *non-structural model* of the form

$$Z_t = G(Z_{t-1}, \epsilon_t). \tag{3.10}$$

This is a generalization of semi-reduced model (3.3) where linear and nonlinear components are absorbed into one functional term and B_0 is the identity matrix.[14] Indeed, note that models of the form $Z_t = G(Z_{t-1}, \ldots, Z_{t-p}, \epsilon_t)$ can be rewritten as (3.10) using a companion formulation. If ϵ_t is stochastic, (3.10) defines a causal nonlinear stochastic process. More generally, it defines a

[13] See Chen (2007), Belloni et al. (2015b) for additional discussion and examples of sieve families.

[14] In this specific subsection, shock identification does not play a role and, as such, one can safely ignore B_0.

nonlinear difference equation and an associated dynamical system driven by ϵ_t. Throughout this subsection, I shall assume that $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ as well as $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$.

Relying on the framework of Pötscher and Prucha (1997), I now introduce explicit conditions that allow to control dependence in nonlinear models by using the toolbox of physical dependence measures developed by Wu (2005, 2011). The aim is to use a dynamical system perspective to address the question of imposing meaningful assumptions on nonlinear dynamic models. This makes it possible to give more primitive conditions under which one can actually estimate (3.8) in a semi-nonparametric way.

Stability. An important concept for dynamical system theory is that of stability. Stability turns out to play a key role in constructing valid asymptotic theory, as it is well understood in linear models. It is also fundamental in developing the approximation theory of nonlinear stochastic systems.

Example 3.3.1. (Linear System). As a motivating example, first consider the linear system

$$Z_t = BZ_{t-1} + \epsilon_t$$

where we may assume that $\{\epsilon_t\}_{t\in\mathbb{Z}}$, $\epsilon_t \in \mathbb{R}^{d_Z}$, is a sequence of i.i.d. innovations.[15] It is well-known that this system is stable if and only if the largest eigenvalue of B is strictly less than one in absolute value (Lütkepohl, 2005). For a higher order linear system, $Z_t = B(L)Z_{t-1} + \epsilon_t$ where $B(L) = B_1 + B_2L + \ldots + B_pL^{p-1}$, stability holds if and only if $|\lambda_{\max}(\mathbf{B})| < 1$ where

$$\mathbf{B} := \begin{bmatrix} B_1 & B_2 & \cdots & B_p \\ I_{d_Z} & 0 & \cdots & 0 \\ 0 & I_{d_Z} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & \cdots & I_{d_Z} & 0 \end{bmatrix}$$

is the companion matrix.

Extending the notion of stability from linear to nonlinear systems requires some care. Pötscher and Prucha (1997) derived generic conditions allowing to formally extend stability to nonlinear models by first analyzing *contractive* systems.

Definition 3.3.1 (Contractive System)**.** *Let $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$, $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$, where $\{Z_t\}_{t\in\mathbb{Z}}$ is generated according to*

$$Z_t = G(Z_{t-1}, \epsilon_t).$$

The system is contractive if for all $(z, z') \in \mathcal{Z} \times \mathcal{Z}$ and $(e, e') \in \mathcal{E} \times \mathcal{E}$

$$\|G(z, \epsilon) - G(z', \epsilon')\| \leq C_Z\|z - z'\| + C_\epsilon\|e - e'\|$$

holds with Lipschitz constants $0 \leq C_Z < 1$ and $0 \leq C_\epsilon < \infty$.

[15] One could alternatively think of the case of a deterministic input, setting $\epsilon_t \sim P_t(a_t)$ where $P_t(a_t)$ is a Dirac density on the deterministic sequence $\{a_t\}_{t\in\mathbb{Z}}$.

Sufficient conditions to establish contractivity are

$$\sup\left\{\left\|\operatorname{stack}_{i=1}^{d_Z}\left[\frac{\partial G}{\partial Z}(z^i,e^i)\right]_i\right\|\,\middle|\, z^i\in\mathcal{Z},e^i\in\mathcal{E}\right\}<1 \tag{3.11}$$

and

$$\left\|\frac{\partial G}{\partial \epsilon}\right\|<\infty, \tag{3.12}$$

where the stacking operator $\operatorname{stack}_{i=1}^{d_Z}[\,\cdot\,]_i$ progressively stacks the rows, indexed by i, of its argument (which can be changing with i) into a matrix. Values $(z^i,e^i)\in\mathcal{Z}\times\mathcal{E}$ change with index i as the above condition is derived using the mean value theorem, therefore it is necessary to consider a different set of values for each component of Z_t.

It is easy to see, as Pötscher and Prucha (1997) point out, that contractivity is often a too strong condition to be imposed. Indeed, even in the simple case of a scalar AR(2) model $Z_t = b_1Z_{t-1}+b_2Z_{t-2}+\epsilon_t$, regardless of the values of $b_1,b_2\in\mathbb{R}$ contractivity is violated. This is due to the fact that in a linear AR(2) model studying contractivity reduces to checking $\|\mathbf{B}\|<1$ instead of $|\lambda_{\max}(\mathbf{B})|<1$, and the former is a stronger condition than the latter.[16] One can weaken contractivity – which must hold for G as a map from Z_{t-1} to Z_t – to the idea of *eventual contractivity*. That is, intuitively, one can impose conditions on the dependence of Z_{t+h} on Z_t for $h>1$ sufficiently large. To do this formally, I first introduce the definition of system map iterates.

Definition 3.3.2 (System Map Iterates)**.** *Let $Z_t\in\mathcal{Z}\subseteq\mathbb{R}^{d_Z}$, $\epsilon_t\in\mathcal{E}\subseteq\mathbb{R}^{d_Z}$ where $\{Z_t\}_{t\in\mathbb{Z}}$ is generated from a sequence $\{\epsilon_t\}_{t\in\mathbb{Z}}$ according to*

$$Z_t=G(Z_{t-1},\epsilon_t).$$

The h-order system map iterate is defined to be

$$\begin{aligned}G^{(h)}(Z_t,\epsilon_{t+1},\epsilon_{t+2},\dots,\epsilon_{t+h}) &:= G(G(\cdots G(Z_t,\epsilon_{t+1})\cdots,\epsilon_{t+h-1}),\epsilon_{t+h})\\ &= G(\cdot,\epsilon_{t+h})\circ G(\cdot,\epsilon_{t+h-1})\circ\cdots\circ G(Z_t,\epsilon_{t+1}),\end{aligned}$$

where $\circ$ signifies function composition and $G^{(0)}(Z_t)=Z_t$.

To shorten notation, in place of $G^{(h)}(Z_t,\epsilon_{t+1},\epsilon_{t+2},\dots,\epsilon_{t+h})$ I shall use $G^{(h)}(Z_t,\epsilon_{t+1:t+h})$. Additionally, for $1\le j\le h$, the partial derivative

$$\frac{\partial G^{(h^*)}}{\partial\epsilon_j}$$

for some fixed h^* is to be intended with respect to ϵ_{t+j}, the j-th entry of the input sequence. This derivative does not dependent on the time index since by assumption G is time-invariant and so is $G^{(h)}$.

Taking again the linear autoregressive model as an example,

$$Z_{t+h}=G^{(h)}(Z_t,\epsilon_{t+1:t+h})=B_1^hZ_t+\sum_{i=0}^{h-1}B_1^i\epsilon_{t+h-i}$$

[16] See Pötscher and Prucha (1997), pp.68-69.

since $G(z,\epsilon) = B_1 z + \epsilon$. If B_1 determines a stable system, then $\|B_1^h\| \to 0$ as $h \to \infty$ since G^h converges to zero, and therefore $\|B_1^h\| \leq C_Z < 1$ for h sufficiently large. It is thus possible to use system map iterates to define stability for higher-order nonlinear systems.

Definition 3.3.3 (Stable System)**.** *Let $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$, $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$, where $\{Z_t\}_{t\in\mathbb{Z}}$ is generated according to the system*

$$Z_t = G(Z_{t-1}, \epsilon_t).$$

The system is stable if there exists $h^* \geq 1$ such that for all $(z, z') \in \mathcal{Z} \times \mathcal{Z}$ and $(e_1, e_2, \dots e_{h^*}, e'_1, e'_2, \dots, e'_{h^*}) \in \bigtimes_{i=1}^{2h^*} \mathcal{E}$

$$\|G^{(h^*)}(z, e_{1:h^*}) - G^{(h^*)}(z', e'_{1:h^*})\| \leq C_Z\|z - z'\| + C_\epsilon\|e_{1:h^*} - e'_{1:h^*}\|$$

holds with Lipschitz constants $0 \leq C_Z < 1$ and $0 \leq C_\epsilon < \infty$.

It is important to remember that this definition encompasses systems with an arbitrary finite autoregressive structure, i.e., $Z_t = G(Z_{t-p+1}, \dots, Z_{t-1}, \epsilon_t)$ for $p \geq 1$, thanks to the companion formulation of the process. An explicit stability condition, similar to that discussed above for contractivity, can be derived by means of the mean value theorem. Indeed, for a system to be stable it is sufficient that, at iterate h^*,

$$\sup\left\{\left\| \operatorname{stack}_{i=1}^{d_Z}\left[\frac{\partial G^{(h^*)}}{\partial Z}(z^i, e^i_{1:h^*})\right]_i \right\| \,\middle|\, z^i \in \mathcal{Z}, e^i_{1:h^*} \in \bigtimes_{i=1}^{h^*}\mathcal{E}\right\} < 1 \tag{3.13}$$

and

$$\sup\left\{\left\| \frac{\partial G^{(h^*)}}{\partial \epsilon_j}(z, e_{1:h^*}) \right\| \,\middle|\, z \in \mathcal{Z}, e_{1:h^*} \in \bigtimes_{i=1}^{h^*}\mathcal{E}\right\} < \infty, \qquad j = 1, \dots, h^*. \tag{3.14}$$

Pötscher and Prucha (1997) have used conditions (3.11)-(3.12) and (3.13)-(3.14) as basis for uniform laws of large numbers and central limit theorems for L^r-approximable and near epoch dependent processes.

Physical Dependence. Wu (2005) first proposed alternatives to mixing concepts by proposing dependence measures rooted in a dynamical system view of a stochastic process. Much work has been done to use such measures to derive approximation results and estimator properties, see for example Wu et al. (2010), Wu (2011), Chen et al. (2016), and references within.

Definition 3.3.4. *If for all $t \in \mathbb{Z}$, Z_t has finite rth moment, where $r \geq 1$, the functional physical dependence measure Δ_r is defined as*

$$\Delta_r(h) := \sup_t \left\| Z_{t+h} - G^{(h)}(Z'_t, \epsilon_{t+1:t+h}) \right\|_{L^r}$$

where $\|\cdot\|_{L^r} = (\mathbb{E}[\|\cdot\|_r^r])^{1/r}$, Z'_t is due to $\mathcal{F}'_t = (\dots, \epsilon'_{t-1}, \epsilon'_t)$ and $\{\epsilon'_t\}_{t\in\mathbb{Z}}$ is an independent copy of $\{\epsilon_t\}_{t\in\mathbb{Z}}$.

Chen et al. (2016), among others, show how one may replace the geometric β-mixing assumption with a physical dependence assumption.[17] They show that the key sufficient condition is for $\Delta_r(h)$

[17] I adapt the definitions of Chen et al. (2016) to work with a system of the form $Z_t = G(Z_{t-1}, \epsilon_t)$.

to decay sufficiently fast as h grows.

Definition 3.3.5 (Geometric Moment Contracting Process)**.** *$\{Z_t\}_{t\in\mathbb{Z}}$ is geometric moment contracting (GMC) in L^r norm if there exists $a_1 > 0$, $a_2 > 0$ and $\tau \in (0, 1]$ such that*

$$\Delta_r(h) \leq a_1 \exp(-a_2\, h^\tau).$$

GMC conditions can be considered more general than β-mixing, as they encompass well-known counterexamples, e.g., the known counterexample provided by $Z_t = (Z_{t-1} + \epsilon_t)/2$ for ϵ_t i.i.d. Bernoulli r.v.s (Chen et al., 2016). In the following proposition I prove that if contractivity or stability conditions as defined by Pötscher and Prucha (1997) hold for G and $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence, then process $\{Z_t\}_{t\in\mathbb{Z}}$ is GMC under weak moment assumptions.

Proposition 3.3.6. *Assume that $\{\epsilon_t\}_{t\in\mathbb{Z}}$, $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$ are i.i.d. and $\{Z_t\}_{t\in\mathbb{Z}}$ is generated according to*

$$Z_t = G(Z_{t-1}, \epsilon_t),$$

where $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ and G is a measurable function.

(a) If contractivity conditions (3.11)-(3.12) *hold, $\sup_{t\in\mathbb{Z}}\|\epsilon_t\|_{L^r} < \infty$ for $r \geq 2$ and $\|G(\overline{z}, \overline{\epsilon})\| < \infty$ for some $(\overline{z}, \overline{\epsilon}) \in \mathcal{Z} \times \mathcal{E}$, then $\{Z_t\}_{t\in\mathbb{Z}}$ is GMC with*

$$\Delta_r(k) \leq a \exp(-\gamma h)$$

where $\gamma = -\log(C_Z)$ and $a = 2\|Z_t\|_{L^r} < \infty$.

(b) If stability conditions (3.13)-(3.14) *hold, $\sup_{t\in\mathbb{Z}}\|\epsilon_t\|_{L^r} < \infty$ for $r \geq 2$ and $\|\partial G/\partial Z\| \leq M_Z < \infty$, then $\{Z_t\}_{t\in\mathbb{Z}}$ is GMC with*

$$\Delta_r(k) \leq \bar{a} \exp(-\gamma_{h^*} h)$$

where $\gamma_{h^*} = -\log(C_Z)/h^*$ and $\bar{a} = 2\|Z_t\|_{L^r} \max\{M_Z^{h-1}, 1\}/C_Z < \infty$.

Proposition 3.3.6 is important in that it links the GMC property to transparent conditions on the structure of the nonlinear model. It also immediately allows handling multivariate systems, while previous work has focused on scalar systems (c.f. Wu (2011) and Chen et al. (2016)).

Finally, it is now possible to show that if $\{W_{2t}\}_{t\in\mathbb{Z}}$ satisfies physical dependence assumptions, then Assumption 8 is fulfilled, c.f. Lemma 2.2 in Chen and Christensen (2015) for β-mixing assumptions.

Lemma 3.3.7. *If Assumption 7(iii) holds and $\{W_{2t}\}_{t\in\mathbb{Z}}$ is strictly stationary and GMC then one may choose an integer sequence $q = q(n) \leq n/2$ with $(n/q)^{r+1} q K^\rho \Delta_r(q) = o(1)$ for $\rho = 5/2 - (r/2 + 2/r) + \omega_2$ and $r > 2$ such that*

$$\|(\widetilde{B}_\pi' \widetilde{B}_\pi/n) - I_K\| = O_P\left(\zeta_{K,n}\lambda_{K,n}\sqrt{\frac{q\log K}{n}}\right) = o_P(1)$$

provided $\zeta_{K,n}\lambda_{K,n}\sqrt{(q\log K)/n} = o(1)$.

It can be seen that Lemma 3.3.7 holds by setting $\sqrt{K(\log(n))^2/n} = o(1)$ and choosing $q(n) = \gamma^{-1}\log(K^\rho n^{r+1})$, where γ is the GMC factor introduced in Proposition 3.3.6. Therefore, the rate is the same as the one derived by Chen and Christensen (2015) for exponentially β-mixing regressors. As shown in Proposition 3.3.6, system contractivity and stability conditions both imply geometric moment contractivity, meaning that in place of Assumption 8 one may require the following.

Assumption 9. For $r > 2$ it holds either:

(i) $\{Z_t\}_{t\in\mathbb{Z}}$ is GMC in L^r norm,

(ii) $\{Z_t\}_{t\in\mathbb{Z}}$ is generated according to $Z_t = \Phi(Z_{t-1}, \dots, Z_{t-p}; \epsilon_t)$ where $\sup_{t\in\mathbb{Z}}\|\epsilon_t\|_{L^r} < \infty$ and Φ is either contractive according to Definition 3.3.1 or stable according to Definition 3.3.3.

It is straightforward to prove that if GMC conditions are imposed on $\{Z_t\}_{t\in\mathbb{Z}}$, this implies that $\{W_{2t}\}_{t\in\mathbb{Z}}$ is also GMC.[18] Therefore, Lemma 3.3.7 applies and Assumption 8 as well as Assumption 5(i) are verified.

3.3.4 Uniform Convergence and Consistency

Since the key asymptotic condition of Chen and Christensen (2015) is upheld under GMC assumptions, their uniform convergence bound on the approximation error of the series estimator can be applied. In order to do so, one must also impose some regularity conditions on π_2.

Without loss of generality, let $\mathcal{X} = [0,1]$ and let $\|\pi_2\|_\infty := \sup_{w\in\mathcal{Y}}|\pi_2(w)|$ be the sup-norm of the conditional mean function $\pi_2(w)$.

Assumption 10. The unconditional density of X_t is uniformly bounded away from zero and infinity over $\mathcal{X}$.

Assumption 11. For all $1 \le i \le d_Y$ and $0 \le j \le p$, the restriction of g_{ij}^{21} to $[0,1]$ belongs to the Hölder class $\Lambda^s([0,1])$ of smoothness $s \ge 1$.

Assumptions 10 and 11 are standard in the nonparametric regression literature. One only needs to restrict the complexity of functions g_{ij}^{21} since, for any i, the remainder of $\pi_{2,i}$ consists of linear functions. More precisely, what is really needed is that the nonparametric components of the sieve given by $b_{\pi,1K}, \dots, b_{\pi,KK}$ are able to approximate g_{ij}^{21} well enough.

Assumption 12. Sieve B_κ belongs to $\mathrm{BSpl}(\kappa, [0,1]^{d_Y}, r)$, the B-spline sieve of degree r over $[0,1]^{d_Y}$, or $\mathrm{Wav}(\kappa, [0,1]^{d_Y}, r)$, the wavelet sieve of regularity r over $[0,1]^{d_Y}$, with $r > \max\{s, 1\}$.

In the remainder of the paper, I will consider the cubic spline sieve ($r = 3$), but theoretical results are stated in the more general setting. Moreover, d will be the effective dimension of the joint estimation domain for G_i^{21}.

Theorem 3.3.8 (Chen and Christensen (2015))**.** *Let Assumptions 4, 5, 6, 7, 9, 10, 11 and 12 hold. If*

$$K \asymp (n/\log(n))^{d/(2s+d)},$$

[18] A formal argument can be found in Appendix 3.A.

then

$$\|\widehat{\pi}_2^* - \pi_2\|_\infty = O_P\left((n/\log(n)^{-s/(2s+d)}\right)$$

provided that $\delta \geq 2/s$ *(in Assumption 6) and* $d < 2s$*.*

In Theorem 3.3.8 the sup-norm consistency rate generally depends on the dimension d and thus, in principle, the curse of dimensionality slows down convergence compared to parametric estimation. Fortunately, under the current strctural model assumptions, the nonlinear functional components in π_2 are linearly separable in the lag dimension, and thus one may take $d = 1$ as effective dimension. This also means that condition $d < 2s$ is trivially satisfied.

Two-step Consistency. The following theorem ensures that the two-step estimation procedure produces consistent estimates. Since for impulse response functions one needs to study the iteration of the entire structural model, this results is stated in terms of the full coefficient matrices.

Theorem 3.3.9. *Let* $\{Z_t\}_{t\in\mathbb{Z}}$ *be determined by structural model* (3.1)*. Under Assumptions 1, 4, 5, 6, 7, 9, 10, 11 and 12, let* $\widehat{\Pi}_1$ *and* $\widehat{\Pi}_2$ *be the least squares and semi-nonparametric series estimators for* Π_1 *and* Π_2*, respectively, based on the two-step procedure. Then,*

$$\|\widehat{\Pi}_1 - \Pi_1\|_\infty = O_P(n^{-1/2})$$

and

$$\|\widehat{\Pi}_2 - \Pi_2\|_\infty \leq O_P\left(\zeta_{K,n}\lambda_{K,n}\frac{K}{\sqrt{n}}\right) + \|\widehat{\Pi}_2^* - \Pi_2\|_\infty,$$

where $\widehat{\Pi}_2^*$ *is the infeasible series estimator involving* ϵ_{1t}*.*

Sup-norm bounds for $\|\widehat{\Pi}_2^* - \Pi_2\|_\infty$ follow immediately from Lemma 2.3 and Lemma 2.4 in Chen and Christensen (2015). In particular, choosing the optimal nonparametric rate $K \asymp (n/\log(n))^{d/(2s+d)}$ for the infeasible estimator would yield

$$\|\widehat{\Pi}_2^* - \Pi_2\|_\infty = O_P\left((n/\log(n))^{-s/(2s+d)}\right)$$

as per Theorem 3.3.8. The condition for consistency in Theorem 3.3.9 reduces to

$$\frac{K^{3/2}}{\sqrt{n}} = o(1),$$

since for B-spline and wavelet sieves $\lambda_{K,n} \lesssim 1$ and $\zeta_{K,n} \lesssim \sqrt{K}$. It simple to show that if for the feasible estimator $\widehat{\Pi}_2$ the same rate $(n/\log(n))^{d/(2s+d)}$ is chosen for K, the consistency condition in the above display is fulfilled assuming $s \geq 1$ and $d = 1$.[19]

Remark 3.3.3. (Hyperparameter Selection). An important practical question when applying any series or kernel-type methods is the selection of hyperparameters. For the former, this entails the choice of the sieve's size K. Although theory provides only asymptotic rates, a number of methods can be used to select K, such as cross-validation, generalized cross-validation and Mallow's criterion

[19]The rate for K may be optimized by balancing the uniform (infeasible) rate with the error due to residuals. Since this paper is not concerned with finding the optimal rate, I do not perform this exercise here.

(Li and Racine, 2009). In the case of piece-wise splines, once size is selected, knots can be chosen to be the K uniform quantiles of the data. This ensures knots are not located in regions of the domain with very few observations. In simulations and applications, for simplicity, I select sieve sizes manually and locate knots approximately following empirical quantiles. In unreported numerical experiments, I check that results are robust to moderate changes in the number and approximate locations of spline knots.

3.4 Impulse Response Analysis

Once the model's linear, functional and structural coefficient are consistently estimated, computation of nonlinear impulse responses must be addressed. As discussed in Section 3.2, nonlinear IRFs are generally hard to lay hands on, since the functional MA(∞) form of the process is highly non-trivial. In this section, I will provide an explicit, iterative algorithm to compute responses that is numerically straightforward and does not require the construction of moving average functional coefficients. Moreover, since to derive uniform bounds it is assumed that the data has compact support, I will introduce a novel yet familiar IRF definition, called the relaxed impulse response function, which is compatible with boundedness. Lastly, I prove that semi-nonparametric IRF estimates are consistent with respect to their population counterparts.

3.4.1 Computation

Recall from equation (3.7) in Section 3.2.2 that impulse responses involve two moving average lag polynomials, $\Theta(L)$ for the linear model component and $\Gamma(L)$ for the nonlinear component, respectively. As a first step, one can derive a semi-explicit recursive algorithm for computing $\text{IRF}_h(\delta)$ in a manner that does not involve simulations of the innovations process.

Proposition 3.4.1 (Gonçalves et al. (2021), Proposition 3.1)**.** *Under Assumptions 1, 2 and 3, for any $h = 0, 1, \ldots, H$, let*

$$V_j(\delta) := \mathbb{E}[\Gamma_j X_{t+j}(\delta)] - \mathbb{E}[\Gamma_j X_{t+j}].$$

To compute

$$\text{IRF}_h(\delta) = \Theta_{h,\cdot 1}\delta + \sum_{j=0}^{h} V_j(\delta),$$

the following steps can be used:

(i) For $j = 0$, set $X_t(\delta) = X_t + \delta$ and $V_0(\delta) = \mathbb{E}[\Gamma_0 X_t(\delta)] - \mathbb{E}[\Gamma_0 X_t]$.

(ii) For $j = 1, \ldots, h$, let

$$\begin{aligned} X_{t+j}(\delta) &= X_{t+j} + \Theta_{j,11}\delta + \sum_{k=1}^{j}(\Gamma_{k,11}X_{t+j-k}(\delta) - \Gamma_{j,11}X_{t+j-k}) \\ &= \gamma_j(X_{t+j:t}; \delta), \end{aligned}$$

where γ_j are implicitly defined and depend on $\Theta(L)$ and $\Gamma(L)$.

(iii) For $j = 1, \dots, h$, compute

$$V_j(\delta) = \mathbb{E}[\Gamma_j \gamma_j(X_{t+j:t}; \delta)] - \mathbb{E}[\Gamma_j X_{t+j}].$$

The proof of Proposition 3.4.1 is identical to that in Gonçalves et al. (2021), with the only variation being that in the current setup it is not possible to collect the nonlinear function across $X_{t+j-k}(\delta)$ and X_{t+j-k}. Computation of $X_{t+j}(\delta)$ in step (ii) involves recursive evaluations of nonlinear functions, which is why the algorithm is semi-explicit. For each horizon h, one needs to evaluate $h+1$ iterations of $X_t(\delta)$. Importantly, however, this approach dispenses from the need to simulate innovations $\{\epsilon_{t+j}\}_{j=1}^{h-1}$ as the joint distribution of $\{X_{t+h-1}, X_{t+j-1}, \dots, X_t\}$ already contains all relevant information. Gonçalves et al. (2021) naturally argue that the algorithm outlined in Proposition 3.4.1 is significantly more efficient than schemes involving Monte Carlo simulations like e.g. the one used by Kilian and Vigfusson (2011).

However, $\{\Gamma_j\}_{j=1}^{h}$ are combinations of real and functional matrices and closed-form derivation is numerically impractical. Note that, by the definition of IRFs, the following *explicit* iterative algorithm is also valid.

Proposition 3.4.2. *In the same setup of Proposition 3.4.1, to compute* $\mathrm{IRF}_h(\delta)$ *the following steps can be used:*

(i′) For $j = 0$, let $X_t(\delta) = X_t + \delta$ and

$$\mathrm{IRF}_0(\delta) = \begin{bmatrix} \delta \\ B_0^{21}\delta \end{bmatrix} + \mathbb{E}\begin{bmatrix} 0 \\ G_{21,0}X_t(\delta) \end{bmatrix} - \mathbb{E}\begin{bmatrix} 0 \\ G_{21,0}X_t \end{bmatrix}.$$

(ii′) For $j = 1, \dots, h$, let

$$\begin{aligned} X_{t+j}(\delta) &= \mu_1 + A_{12}(L)Y_{t+j-1}(\delta) + A_{11}(L)X_{t+j-1}(\delta) + \epsilon_{1t+j}, \\ Y_{t+j}(\delta) &= \mu_2 + A_{22}(L)Y_{t+j-1}(\delta) + H_{21}(L)X_{t+j}(\delta) + B_0^{21}\epsilon_{1t+j} + u_{2t+j}, \end{aligned}$$

where $H_{21}(L) := A_{21}(L)L + G_{21}(L)$ and $u_{2t+j} := B_0^{22}\epsilon_{2t+j}$. Setting $Z_{t+j}(\delta) = (X_t(\delta), Y_t(\delta))'$ it holds

$$\mathrm{IRF}_h(\delta) = \mathbb{E}[Z_{t+j}(\delta)] - \mathbb{E}[Z_{t+j}].$$

Proposition 3.4.2 follows directly from the definition of the unconditional impulse response (3.6) combined with explicit iteration of the semi-reduced form (3.2) and sidesteps the MA(∞) formulation in (3.7). Step (i′) is trivial in nature. Step (ii′) may not seem useful when compared to (ii), since, in practice, innovations ϵ_{1t} and u_{2t} are not available. However, let

$$\widehat{\mu},\ \widehat{A}_{11}(L),\ \widehat{A}_{12}(L),\ \widehat{A}_{21}(L),\ \widehat{H}_{11}(L),\ \widehat{B}_0^{21}$$

be estimates of the model's coefficients derived, for example, from series estimator $\widehat{\Pi}_1$ and $\widehat{\Pi}_2$. In sample, one can compute residuals $\widehat{\epsilon}_{1t}$ and $\widehat{u}_{2t}$, and by definition it holds

$$\begin{aligned} X_t &= \widehat{\mu}_1 + \widehat{A}_{12}(L)Y_{t-1} + \widehat{A}_{11}(L)X_{t-1} + \widehat{\epsilon}_{1t}, \\ Y_t &= \widehat{\mu}_2 + \widehat{A}_{22}(L)Y_{t-1} + \widehat{H}_{21}(L)X_t + \widehat{B}_0^{21}\widehat{\epsilon}_{1t} + \widehat{u}_{2t}. \end{aligned}$$

This means that one can readily construct the shocked sequence recursively as

$$\widehat{X}_{t+j}(\delta) = \widehat{\mu}_1 + \widehat{A}_{12}(L)\widehat{Y}_{t+j-1}(\delta) + \widehat{A}_{11}(L)\widehat{X}_{t+j-1}(\delta) + \widehat{\epsilon}_{1t+j},$$
$$\widehat{Y}_{t+j}(\delta) = \widehat{\mu}_2 + \widehat{A}_{22}(L)\widehat{Y}_{t+j-1}(\delta) + \widehat{H}_{21}(L)\widehat{X}_{t+j}(\delta) + \widehat{B}_0^{21}\widehat{\epsilon}_{1t+j} + \widehat{u}_{2t+j},$$

for $j = 1, \dots, h$ where $\widehat{X}_t(\delta) = X_t + \delta$, $\widehat{X}_{t-s} = X_{t-s}$ for all $s \geq 1$ and similarly for $\widehat{Y}_t(\delta)$. To evaluate a structural IRF, over a sample of size n one can compute

$$\widehat{\mathrm{IRF}}_h(\delta) = \frac{1}{n-j}\sum_{t=1}^{n-j}\left[\widehat{Y}_{t+j}(\delta) - Y_t\right],$$

which is still considerably less demanding than Monte Carlo simulations. Additionally, the advantage in implementing steps (i′)-(ii′) over the procedure in Proposition 3.4.1 is that, when $\widehat{H}_{21}(L)$ is a semi-nonparametric estimate, iterating model equations is numerically much more straightforward than handling functional MA matrices $\{\widehat{\Gamma}_j\}_{j=1}^h$.

3.4.2 Nonlinear Responses with Relaxed Shocks

Following Proposition 3.4.1, the sample impulse response would be

$$\widehat{\mathrm{IRF}}_h(\delta) := \widehat{\Theta}_{h,\cdot 1}\delta + \sum_{j=0}^{h}\bar{V}_j(\delta), \tag{3.15}$$

where

$$\bar{V}_j(\delta) := \frac{1}{n-j}\sum_{t=1}^{n-j}\left[\widehat{\Gamma}_j\widehat{\gamma}_j(X_{t+j:t};\delta) - \widehat{\Gamma}_j X_{t+j}\right]$$

and $\widehat{\Theta}$, $\widehat{\Gamma}$ and $\widehat{\gamma}_j$ are plug-in estimates of the respective quantities based on $\widehat{\Pi}_1$ and $\widehat{\Pi}_2$. However, under Assumptions 4 and 5, the construction of impulse response (3.15) is improper. This can be immediately seen by noticing that, at impact,

$$X_t(\delta) = \gamma_j(X_t;\delta) = X_t + \delta,$$

meaning that $\mathbb{P}(X_t(\delta) \notin \mathcal{X}) > 0$ since there is a translation of size δ in the support of X_t. The problem is rooted in the fact that the standard definition of IRF involves a translation of the distribution of time t structural innovations, which is incompatible with the assumptions imposed in Section 3.3 to derive semi-nonparametric consistency.

There are multiple ways to address this issue. One option, which would require substantial technical work, is to extend Theorem 3.3.9 to encompass regressors with unbounded or expanding domains. A potential direction could be coupling the weighted sieves of Chen and Christensen (2015) with appropriately defined shocks. Instead, I propose to take a more direct approach by changing the *type* of structural shock one studies in a way consistent with bounded domains for all variables.

Definition 3.4.3. *A mean-shift structural shock $\epsilon_{1t}(\delta)$ is a transformation of ϵ_{1t} such that*

$$\mathbb{P}(\epsilon_{1t}(\delta) \in \mathcal{E}_1) = 1 \quad \text{and} \quad \mathbb{E}[\epsilon_{1t}(\delta)] = \delta.$$

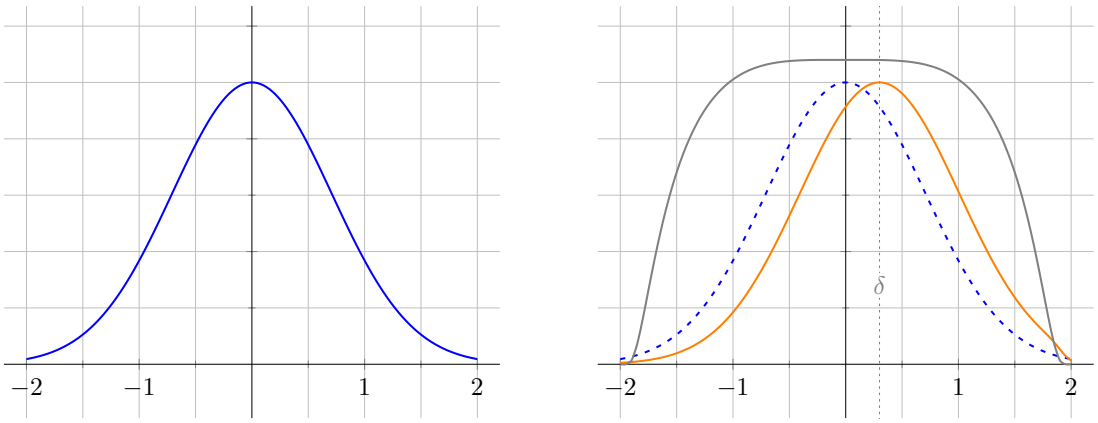


Figure 3.1: Example of symmetric shock relaxation. Unperturbed (left, blue) versus shocked (right, orange) densities of innovations ϵ_{1t}. The shock relaxation function (right, gray) and δ together determine the form of the relaxed shock used to compute the IRF.

A mean-shift shock is such that the distribution of time t innovations is shifted to have mean δ, while retaining support $\mathcal{E}$ almost surely. This definition is natural in that it makes evaluating the effect of the MA(∞) component of the unconditional IRF straightforward. With a mean-shift shock, at impact it holds

$$X_t(\delta) = X_t + \epsilon_{1t}(\delta) - \epsilon_{1t},$$

yet $\epsilon_{1t}(\delta) - \epsilon_{1t}$ is not known unless the transformation for the mean-shock is itself known. Unfortunately, the assumption that the mean of $\epsilon_{1t}(\delta)$ is exactly equal to δ requires that the distribution of ϵ_{1t} be known to properly choose a mean-shift transform. If instead one is willing to assume only that $\mathbb{E}[\epsilon_{1t}(\delta)] \approx \delta$, it is possible to sidestep this requirement by introducing a *shock relaxation function.*

Definition 3.4.4 (Shock Relaxation Function)**.** *A shock relaxation function is a map* $\rho : \mathcal{E}_1 \to [0, 1]$ *such that* $\rho(z) = 0$ *for all* $z \in \mathbb{R} \setminus \mathcal{E}_1$, $\rho(z) \geq 0$ *for all* $z \in \mathcal{E}_1$ *and there exists* $z_0 \in \mathcal{E}_1$ *for which* $\rho(z_0) = 1$.

In general, choosing a shock relaxation function without taking into account the shape of domain $\mathcal{E}_1$ does not necessarily imply that the relaxed shocks will not push the structural variable out-of-bounds. Therefore, I also introduce the notion of *compatibility.*

Definition 3.4.5 (Compatible Relaxation)**.** *Consider a shock* $\delta \in \mathbb{R}$ *and let* $\mathcal{E}_1 = [a, b]$.

(i) If $\delta > 0$, ρ *is said to be right-compatible with* δ *if*

$$\rho(z) \leq \frac{b - z}{|\delta|} \text{for all } z \in \mathcal{E}.$$

(ii) If $\delta < 0$, ρ *is said to be left-compatible with* δ *if*

$$\rho(z) \leq \frac{a + z}{|\delta|} \quad \text{for all } z \in \mathcal{E}.$$

(iii) Given shock size $|\delta| > 0$, ρ *is said to be compatible if it is both right- and left-compatible.*

By setting

$$\epsilon_{1t}(\delta) = \epsilon_{1t} + \delta\rho(\epsilon_{1t})$$

where ρ is compatible with δ, it follows that $X_t(\delta) = X_t + \delta\rho(\epsilon_{1t})$ and $|\mathbb{E}[\epsilon_{1t}(\delta)]| = |\delta\mathbb{E}[\rho(\epsilon_{1t})]| \leq |\delta|$ since $\mathbb{E}[\rho(\epsilon_{1t})] \in [0, 1)$ by definition of ρ. If ρ is a bump function, a relaxed shock is a structural shock that has been mitigated proportionally to the density of innovations at the edges of $\mathcal{E}_1$ and the squareness of ρ. For better intuition, Figure 3.1 provides a graphical rendition of shock relaxation of a symmetric error distribution with a bump function.

Remark 3.4.1. The definition of compatible relaxation function is *static*, as it considers only the impact effect of a shock. Nonetheless, the assumption that $X_t \in \mathcal{X}$ for all t must also hold for $X_t(\delta)$, the shocked structural variable. In theory, given δ, one can always either expand $\mathcal{X}$ or strengthen ρ so that compatibility is enforced at all horizons $1 \leq h \leq H$. For simulations, where one has access to the data generating process, the choice of domains and relaxation functions can be done transparently. In practice, some care is required. When working with empirical data, unless one is willing to assume X_t is wholly exogenous – as in Section 3.6.1 with monetary policy shocks – or strictly autoregressive, some scenarios are more amenable to analysis with the framework presented here than other. In Section 3.6.2, following Istrefi and Mouabbi (2018), I will let X_t be a non-negative uncertainty measure, so that negative shocks are harder to study without producing sequences that contain *negative* uncertainty values. Thus, I will focus on positive, contractionary shocks.

For a given X_t, transformation $X_t + \delta\rho(\epsilon_{1t})$ is not directly applicable since ϵ_{1t} is not observed. In practice, therefore, I will consider

$$\widehat{X}_t(\delta) := X_t + \delta\rho(\widehat{\epsilon}_{1t}).$$

For simplicity of notation, let $\widetilde{\delta}_t := \delta\rho(\epsilon_{1t})$. Similarly to Step (ii) of Proposition 3.4.1, given a path $X_{t+j:t}$ one finds

$$\begin{aligned} X_{t+j}(\widetilde{\delta}_t) &= X_{t+j} + \Theta_{j,11}\widetilde{\delta}_t + \sum_{k=1}^{j}(\Gamma_{k,11}X_{t+j-k}(\widetilde{\delta}_t) - \Gamma_{k,11}X_{t+j-k}) \\ &= \gamma_j(X_{t+j:t}; \widetilde{\delta}_t), \end{aligned}$$

The relaxed-shock impulse response is thus given by

$$\widetilde{\mathrm{IRF}}_h(\delta) := \mathbb{E}[Z_{t+j}(\widetilde{\delta}_t) - Z_{t+j}] = \Theta_{h,\cdot 1}\delta\,\mathbb{E}\left[\rho(\epsilon_{1t})\right] + \sum_{k=1}^{j}\mathbb{E}\left[\Gamma_k X_{t+j-k}(\widetilde{\delta}_t) - \Gamma_k X_{t+j-k}\right].$$

In what follows, I show that by replacing $\widetilde{\delta}_t$ with $\widehat{\widetilde{\delta}}_t = \delta\rho(\widehat{\epsilon}_{1t})$ it is possible to consistently estimate unconditional expectations involving $X_{t+j}(\widetilde{\delta}_t)$ as well as X_{t+j}, and thus $\widetilde{\mathrm{IRF}}_h(\delta)$, by averaging over sample realizations.

3.4.3 Relaxed Impulse Response Consistency

For a given $\delta \in \mathbb{R}$ and compatible shock relaxation function ρ, vector $V_j(\delta)$ is the nonlinear component of impulse responses. One can focus on a specific variable's response by introducing, for $1 \leq \ell \leq d$,

$$V_{j,\ell}(\delta) := \frac{1}{n-j} \sum_{t=1}^{n-j} \left[\Gamma_{j,\ell} \gamma_j(X_{t+j:t}; \widetilde{\delta}_t) - \Gamma_{j,\ell} X_{t+j} \right],$$

where $V_{j,\ell}(\delta)$ is the horizon j nonlinear effect on the ℓth variable and $\Gamma_{j,\ell}$ is the ℓth component of functional vector Γ_j. For the sake of notation I also define

$$v_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) := \Gamma_{j,\ell} \gamma_j(X_{t+j:t}; \widetilde{\delta}_t) - \Gamma_{j,\ell} X_{t+j}.$$

Let $\widehat{v}_{j,\ell}(X_{t+j:t}; \widehat{\widetilde{\delta}}_t)$ be its sample equivalent, so that

$$\widehat{v}_{j,\ell}(X_{t+j:t}; \widehat{\widetilde{\delta}}_t) = \widehat{\Gamma}_{j,\ell} \widehat{\gamma}_j(X_{t+j:t}; \widehat{\widetilde{\delta}}_t) - \widehat{\Gamma}_{j,\ell} X_{t+j},$$

$$\widehat{V}_{j,\ell}(\delta) = \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}(X_{t+j:t}; \widehat{\widetilde{\delta}}_t)$$

and

$$\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) = \Theta_{h,\cdot 1} \delta \, n^{-1} \sum_{t=1}^{n} \rho(\widehat{\epsilon}_{1t}) + \sum_{j=0}^{h} \widehat{V}_{j,\ell}(\delta)$$

for $1 \leq \ell \leq d$.

Theorem 3.4.6. *Let $\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta)$ be a semi-nonparametric estimate for the horizon h relaxed shock IRF of variable ℓ. Under the same assumptions as in Theorem 3.3.9*

$$\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) \xrightarrow{P} \widetilde{\mathrm{IRF}}_{h,\ell}(\delta)$$

for any fixed integers $0 \leq h < \infty$ and $1 \leq \ell \leq d$.

3.5 Simulations

This section is devoted to analyzing the empirical performance of the two-step semi-nonparametric estimation strategy discussed above. I will consider the two simulation setups employed by Gonçalves et al. (2021), with focus on bias and MSE of the estimated relaxed shocked impulse response functions. Additionally, I provide simulations under a modified design which highlight how in larger samples the non-parametric sieve estimator consistently recovers impulse responses, while a least-squares estimator constructed with a pre-specified nonlinear transform does not. In all simulations, I use a B-spline sieve of order 1.

3.5.1 Benchmark Bivariate Design

The first simulation setup involves a bivariate DGP where the structural shock does not directly affect other observables. This is a simple environment to check that indeed the two-step estimator

recover the nonlinear component of the model and impulse responses are consistently estimated, and that the MSE does not worsen excessively.

I consider three bivariate data generation processes. DGP 1 sets X_t to be a fully exogenous innovation process,

$$\begin{aligned} X_t &= \epsilon_{1t}, \\ Y_t &= 0.5Y_{t-1} + 0.5X_t + 0.3X_{t-1} - 0.4\max(0, X_t) + 0.3\max(0, X_{t-1}) + \epsilon_{2t}. \end{aligned} \tag{3.16}$$

DGP 2 adds an autoregressive component to X_t, but maintains exogeneity,

$$\begin{aligned} X_t &= 0.5X_{t-1} + \epsilon_{1t}, \\ Y_t &= 0.5Y_{t-1} + 0.5X_t + 0.3X_{t-1} - 0.4\max(0, X_t) + 0.3\max(0, X_{t-1}) + \epsilon_{2t}. \end{aligned} \tag{3.17}$$

Finally, DGP 3 add an endogenous effect of Y_{t-1} on the structural variable by setting

$$\begin{aligned} X_t &= 0.5X_{t-1} + 0.2Y_{t-1} + \epsilon_{1t}, \\ Y_t &= 0.5Y_{t-1} + 0.5X_t + 0.3X_{t-1} - 0.4\max(0, X_t) + 0.3\max(0, X_{t-1}) + \epsilon_{2t}. \end{aligned} \tag{3.18}$$

Following Assumption 1, innovations are mutually independent. To accommodate Assumptions 4 and 5, both ϵ_{1t} and ϵ_{2t} are drawn from a truncated standard Gaussian distribution over $[-3, 3]$.[20] All DGPs are centered to have zero intercept in population.

I evaluate bias and MSE plots using 1000 Monte Carlo simulation. For a chosen horizon H, the impact of a relaxed shock on ϵ_{1t} is evaluated on Y_{t+h} for $h = 1, \ldots, H$. To compute the population IRF, I employ a direct simulation strategy that replicates the shock's propagation through the model and I use 10 000 replications. To evaluate the estimated IRF, the two-step procedure is implemented: a sample of length n is drawn, the linear least squares and the semi-nonparametric series estimators of the model are used to estimate the model and the relaxed IRF is computed following Proposition 3.4.2. For the sake of brevity, I discuss the case of $\delta = 1$ and I set the shock relaxation function to be

$$\rho(z) = \exp\left(1 + \left[\left|\frac{z}{3}\right|^4 - 1\right]^{-1}\right)$$

over interval $[-3, 3]$ and zero everywhere else.[21] Choices of $\delta = -1$ and $\delta = \pm 0.5$ yield similar results in simulations, so I do not discuss them here.

Figure 3.2 contains the results for sample size $n = 240$. This choice is motivated by considering the average sample sizes found in most macroeconometric settings: it is equivalent to 20 years of monthly data or 60 yearly of quarterly data (Gonçalves et al., 2021). The benchmark method is an OLS regression that relies on a priori knowledge of the underlying DGP specification. Given the moderate sample size, to construct the cubic spline sieve estimator of the nonlinear component of the model I use a single knot, located at 0. The simulations in Figure 3.2 show that while the MSE is slighlty higher for the sieve model, the bias is comparable across methods. Note that for

[20] Let $e_{it} \sim \mathcal{N}(0, 1)$ for $i = 1, 2$, then the truncated Gaussian innovations used in simulation are set to be $\epsilon_{it} = \min(\max(-3, e_{it}), 3)$. The resulting r.v.s have a non-continuous density with two mass points at -3 and 3. However, in practice, since these masses are negligible, for the moderate sample sizes used this choice does not create issues.

[21] It can be easily checked that this choice of ρ is compatible with shocks of size $0 \leq |\delta| \leq 1$.

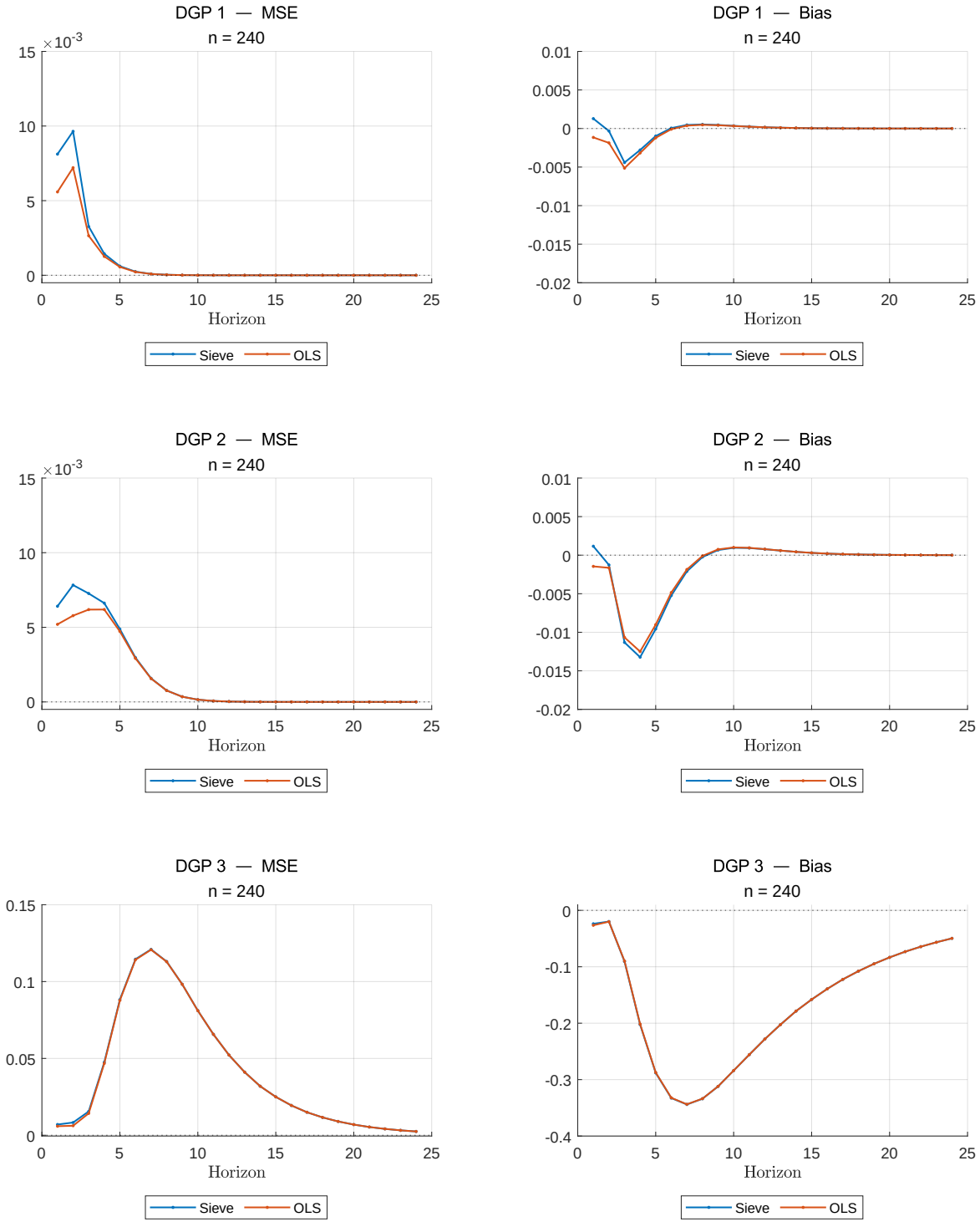


Figure 3.2: Simulations results for DGPs 1-3.

DGP 3, due to the dependence of the structural variable on non-structural series lags, the MSE and bias increase significantly, and there is no meaningful difference in performance between the two estimation approaches.

3.5.2 Structural Partial Identification Design

To showcase the validity of the proposed sieve estimator under the type of partial structural identification discussed in the paper, I again rely on the simulation design proposed by Gonçalves et al. (2021). All specifications are block-recursive, and require estimating the contemporaneous effects of a structural shock on non-structural variables, unlike in the previous section.

The form of the DGPs is

$$B_0 Z_t = B_1 Z_{t-1} + C_0 f(X_t) + C_1 f(X_{t-1}) + \epsilon_t,$$

where in all variations of the model

$$B_0 = \begin{bmatrix} 1 & 0 & 0 \\ -0.45 & 1 & -0.3 \\ -0.05 & 0.1 & 1 \end{bmatrix}, \quad C_0 = \begin{bmatrix} 0 \\ -0.2 \\ 0.08 \end{bmatrix}, \quad \text{and} \quad C_1 = \begin{bmatrix} 0 \\ -0.1 \\ 0.2 \end{bmatrix}.$$

I focus on the case $f(x) = \max(0, x)$, since this type of nonlinearity is simpler to study. DGP 4 treats X_t as an exogenous shock by setting

$$B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.15 & 0.17 & -0.18 \\ -0.08 & 0.03 & 0.6 \end{bmatrix};$$

DGP 5 add serial correlation to X_t,

$$B_1 = \begin{bmatrix} -0.13 & 0 & 0 \\ 0.15 & 0.17 & -0.18 \\ -0.08 & 0.03 & 0.6 \end{bmatrix};$$

and DGP 6 includes dependence on Y_{t-1},

$$B_1 = \begin{bmatrix} -0.13 & 0.05 & -0.01 \\ 0.15 & 0.17 & -0.18 \\ -0.08 & 0.03 & 0.6 \end{bmatrix}.$$

For these data generating processes, I employ the same setup of simulations with DGPs 1-3, including the number of replications as well as the type of relaxed shock. as well as the sieve grid. Here too I evaluate MSE and bias of both the sieve and the correct specification OLS estimators with as sample size of $n = 240$ observations. The results in Figure 3.3 show again that there is little difference in terms of performance between the semi-nonparametric sieve approach and a correctly-specified OLS regression.

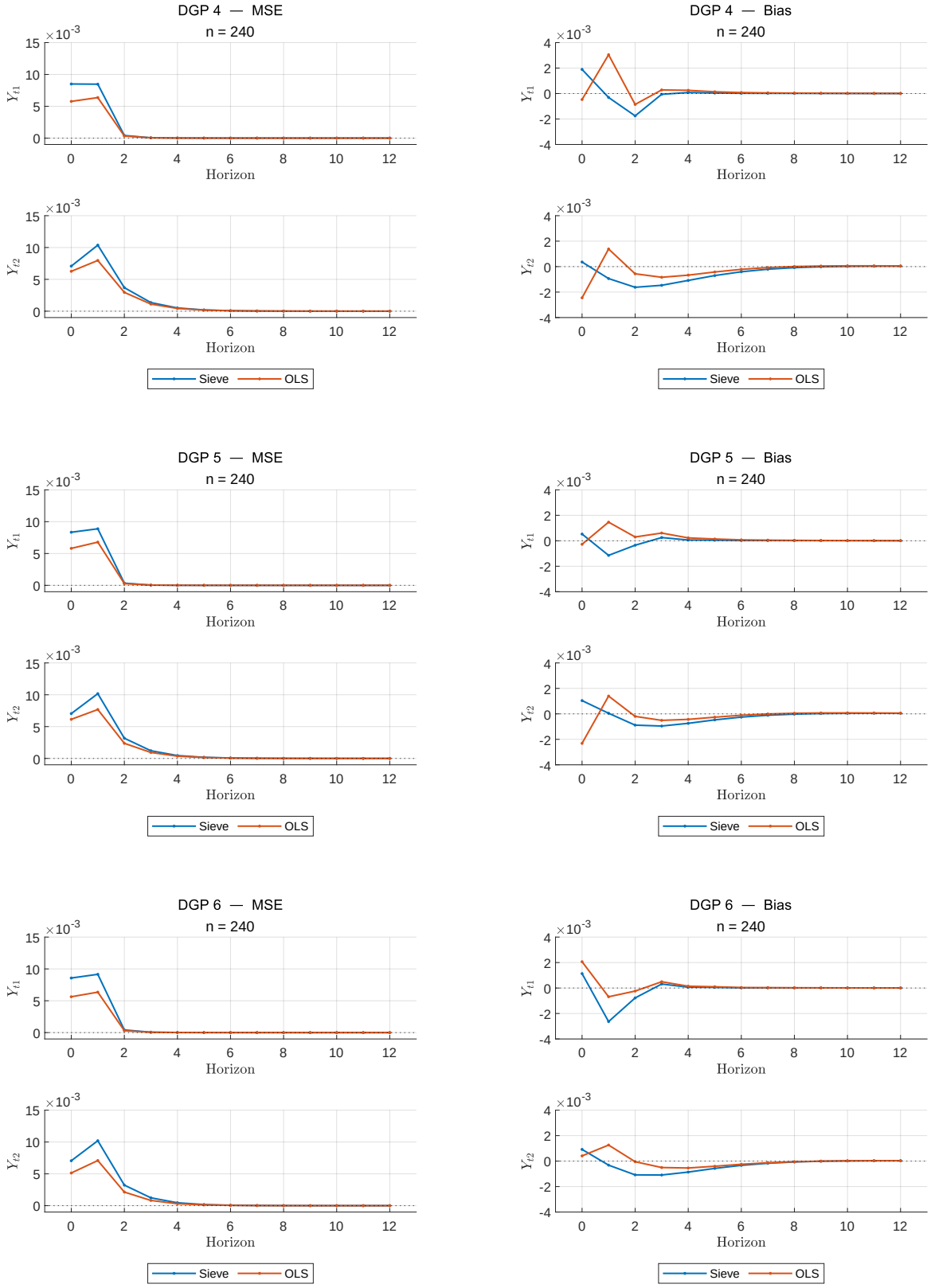


Figure 3.3: Simulations results for DGPs 4-6.

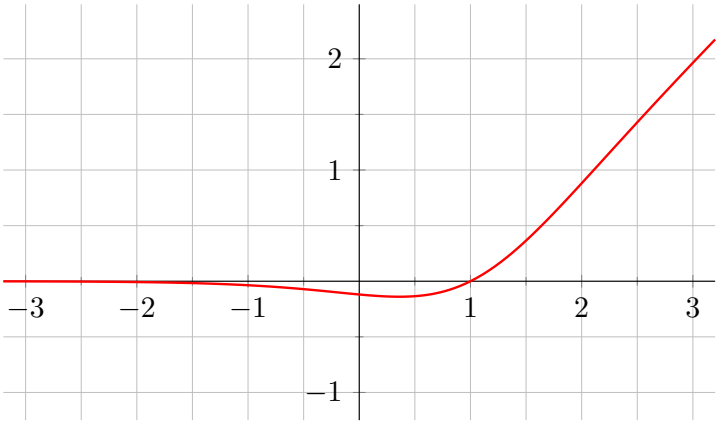


Figure 3.4: Plot of nonlinear function $\varphi(x)$ used in DGP 7.

3.5.3 Model Misspecification

The previous sections report results that support the use of the sieve IRF estimator in a sample of moderate size, since it performs comparably to a regression performed with a priori knowledge of the underlying DGP. I now show that the semi-nonparametric approach is also robust to model misspecification compared to simpler specifications involving fixed choices for nonlinear transformations.

To this end, I modify DGP 2 to use a smooth nonlinear transformation to define the effect of structural variable X_t on Y_t. That is, there is no compounding of linear and nonlinear effects. The autoregressive coefficient in the equation for X_t is also increased to make the shock more persistent. The new data generating process, DGP 2′, is, thus, given by

$$\begin{aligned} X_t &= 0.8X_{t-1} + \epsilon_{1t}, \\ Y_t &= 0.5Y_{t-1} + 0.9\varphi(X_t) + 0.5\varphi(X_{t-1}) + \epsilon_{2t}. \end{aligned} \tag{3.19}$$

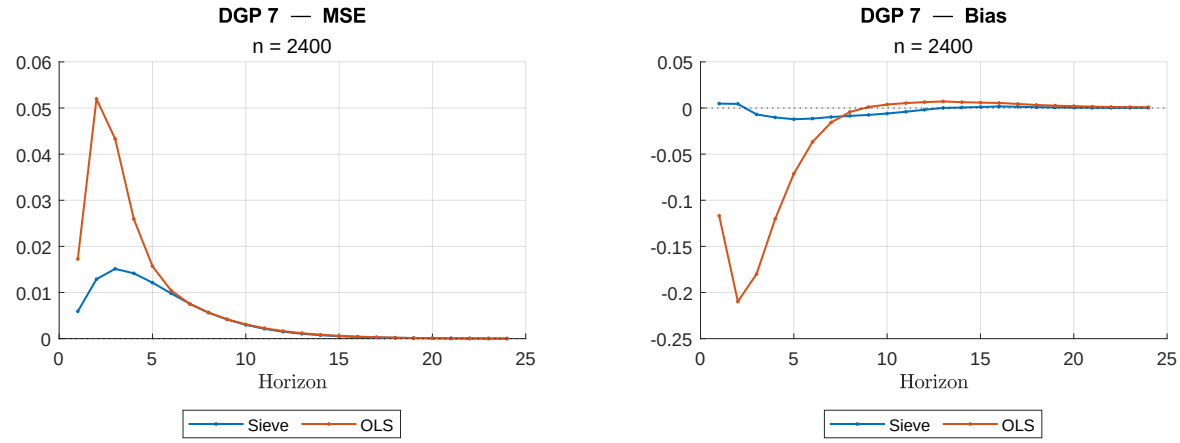
where $\varphi(x) := (x-1)(0.5 + \tanh(x-1)/2)$, which is plotted in Figure 3.4.

To emphasize the difference in estimated IRFs, in this setup I focus on $\delta = \pm 2$, which requires adapting the choice of innovations and shock relaxation function. In simulations of DGP 2′, ϵ_{1t} and ϵ_{2t} are both drawn from a truncated standard Gaussian distribution over $[-5, 5]$. The shock relaxation function of this setup is given by

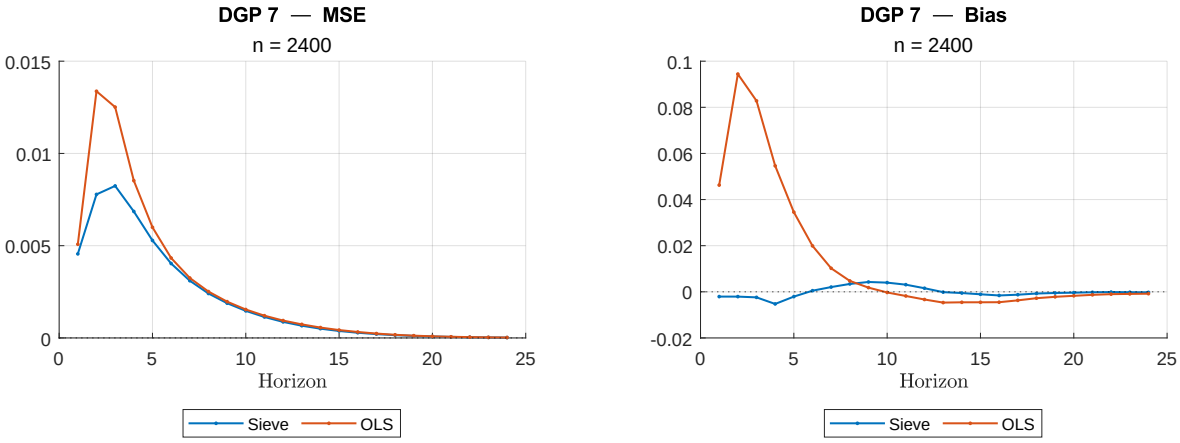
$$\rho(z) = \exp\left(1 + \left[\left|\frac{z}{5}\right|^{3.9} - 1\right]^{-1}\right)$$

over interval $[-5, 5]$ and zero everywhere else. This form of ρ is adapted to choices of δ such that $0 < |\delta| \leq 2$. The sieve grid now consists of 4 equidistant knots within $(-5, 5)$. I use the same numbers of replications as in the previous simulations. Finally, the regression design is identical to that used for DGP 2 under correct specification.

The results obtained with sample size $n = 2400$ are collected in Figure 3.5. I choose this larger sample size to clearly showcase the inconsistency of impulse responses under misspecification: as it can be observed, the simple OLS estimator involving the negative-censoring transform produces



(a) $\delta = +2$



(b) $\delta = -2$

Figure 3.5: Simulations results for DGP 7.

IRF estimates with consistently worse MSE and bias than those of the sieve estimator at almost all horizons. Similar results are also obtained for more moderate shocks $\delta = \pm 1$, but the differences are less pronounced. These simulations suggest that the semi-nonparametric sieve estimator can produce substantially better IRF estimates in large samples than methods involving nonlinear transformations selected a priori.

In this setup, it is also important to highlight the fact that the poor performance of OLS IRF estimates does not come from $\varphi(x)$ being "complex", and, thus, hard to approximate by combinations of simple functions. In fact, if in DGP 2′ function φ is replaced by $\widetilde{\varphi}(x) := \varphi(x+1)$, the differences between sieve and OLS impulse response estimates become minimal in simulations, with the bias of the latter decreasing by approximately an order of magnitude (see Figure 3.8 in Appendix 3.B). This is simply due the fact that $\widetilde{\varphi}(x)$ is well approximated by $\max(0, x)$ directly. However, one then requires either prior knowledge or sheer luck when constructing the nonlinear transforms of X_t for an OLS regression. The proposed series estimator, instead, just requires an appropriate choice of sieve. Many data-driven procedures to select sieves in applications have been proposed, see for example the discussion in Kang (2021).

3.6 Empirical Applications

In this section, I showcase the practical utility of the proposed semi-nonparametric sieve estimator by considering two applied exercises. First, I revisit the empirical analysis of Gonçalves et al. (2021), which is itself based on the work of Tenreyro and Thwaites (2016). This provides both linear and nonlinear benchmarks for the monetary policy responses within a compact econometric model. I find that, although the differences between approaches are mild, nonparametric IRFs in fact provide counter-evidence to the conclusions reported by Gonçalves et al. (2021). In the second application, I compare the linear and nonlinear impulse responses that are produced by uncertainty shocks in the setup studied by Istrefi and Mouabbi (2018). Here, sieve-estimated IRFs show differences in shape, timing and intensity, chiefly when the sign of the shock changes.

3.6.1 Monetary Policy Shocks

The objective of the empirical analysis in Gonçalves et al. (2021) is to analyze the effects of a monetary policy shock on a model of the US macroeconomy. Structural identification is achieved via a narrative approach, following the seminal work of Romer and Romer (2004).

The four-variable model is set up identically to the one of Gonçalves et al. (2021), Section 7. Let $Z_t = (X_t, \mathrm{FFR}_t, \mathrm{GDP}_t, \mathrm{PCE}_t)'$, where X_t is the series of narrative U.S. monetary policy shocks, FFR_t is the federal funds rate, GDP_t is log real GDP and PCE_t is PCE inflation.[22] As a pre-processing step, GDP is transformed to log GDP and then linearly detrended. The data is available quarterly and spans from 1969:Q1 to 2007:Q4. As in Tenreyro and Thwaites (2016), I

[22] In Gonçalves et al. (2021) p. 122, it is mentioned that CPI inflation is included in the model, but both in the replication package made available by one the authors (`https://sites.google.com/site/lkilian2019/research/code`) from which I source the data, and Tenreyro and Thwaites (2016), PCE inflation is used instead. Moreover, the authors say that both the FFR and PCE enter the model in first differences, yet in their code these variables are kept in levels. I keep their original formulation to allow for a proper comparison between estimation methods.

use a model with one lag, $p = 1$. Narrative shock X_t is considered to be an i.i.d. sequence, i.e. $X_t = \epsilon_{1t}$, therefore I assume no dependence on lagged variables when implementing pseudo-reduced form (3.2). Like in Gonçalves et al. (2021), I consider positive and negative shocks of size $|\delta| = 1$. As such, I choose

$$\rho(z) = \mathbb{I}\{|z| \leq 4\} \exp\left(1 + \left[\left|\frac{z}{4}\right|^6 - 1\right]^{-1}\right)$$

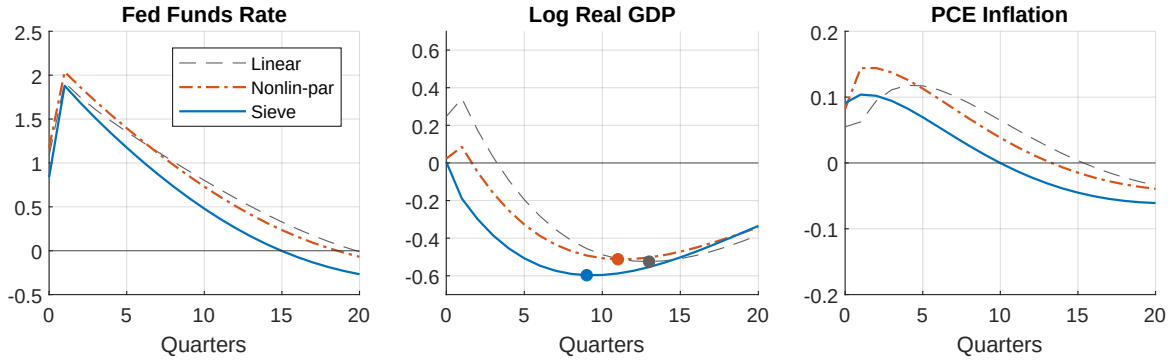
to be the shock relaxation functions. Figure 3.10 in Appendix 3.B provides a check for the validity of ρ given the sample distribution of X_t. Knots for sieve estimation are located at $\{-1, 0, 1\}$. The model is block-recursive, and the structural formulation of Section 3.2.2 allows identifying the U.S. monetary policy shocks without the need to impose additional assumptions on the remaining shocks. Gonçalves et al. (2021), following Tenreyro and Thwaites (2016), use two nonlinear transformations, $F(x) = \max(0, x)$ and $F(x) = x^3$, to try to gauge how negative versus positive and large versus small shocks, respectively, affect the U.S. macroeconomy. For clarity, below I refer to this approach as "parametric nonlinear method". Since the authors find that the inclusion of a cubic term does not meaningfully change impulse responses, I focus on comparing the IRFs estimated via sieve regression with the ones obtained by setting $F(x) = \max(0, x)$, as well as by not including nonlinear terms (i.e. linear IRFs).

Figure 3.6 shows the estimated impulse response to both a positive and negative unforeseen monetary policy shock. The impact on the federal funds rate is consistent across all three procedures, but there are important differences in GDP and inflation responses. In case of an exogenous monetary tightening change, the parametric nonlinear response for GDP, unlike in the case of linear and parametric nonlinear IRFs, is nearly zero at impact and has a monotonic decrease until around 10 quarters ahead. The change is shape is meaningful, as the procedure of Gonçalves et al. (2021) still yields a small short-term upward jump in GDP when a monetary tightening shock hits. Moreover, after the positive shock, the sieve GDP responses reaches its lowest value 4 and 2 quarters before the linear and parametric nonlinear responses, while its size is 13% and 16% larger, respectively.[23] Finally, the sieve PCE response is positive for a shorter interval, but looks to be more persistent once it turns negative also 10 months after impact.

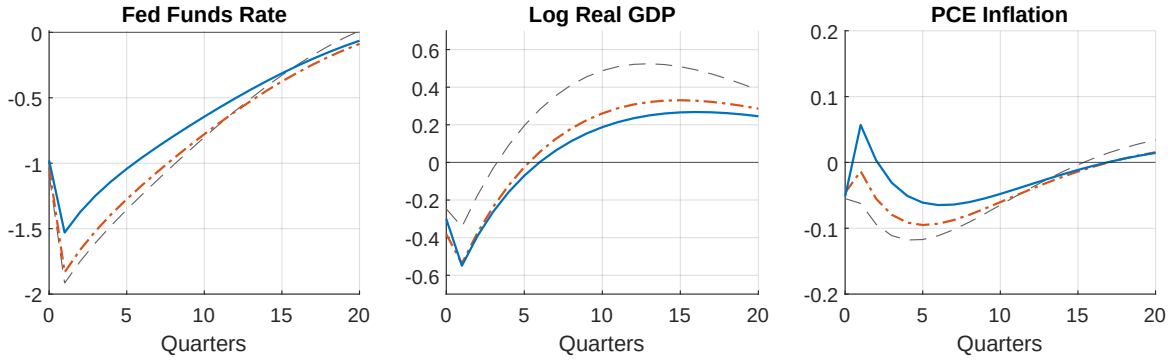
When the shock is expansionary, sieve IRFs show a pronounced asymmetry, even more than that of parametric nonlinear responses. One can observe that semi-nonparametric federal funds rate IRF is marginally mitigated compared to the alternative estimates. An important puzzle is due to the clearly negative impact on GDP. Indeed, both types of nonlinear responses show a drop in output in the first 5 quarters. Also note that the PCE inflation has a positive spike the first couple of quarters after impact. Such a quick change seems unrealistic, as one does not expect inflation to suddenly reverse sign, but, as Gonçalves et al. (2021) also remark, the overall impact on inflation of both shocks is small when compared to the change in federal funds rate.

This comparison between methods, and specifically the nature of nonparametric impulse responses, provides evidence that a small econometric model, such as the one studied by Tenreyro and Thwaites (2016), may be inadequate to fully capture the dynamic effects of monetary policy

[23] The strength of this effect changes across different shocks sizes, as Figure 3.12 in Appendix 3.B proves. As shocks sizes get smaller, nonlinear IRFs, both parametric and sieve, show decreasing negative effects.



(a) $\delta = +1$



(b) $\delta = -1$

Figure 3.6: Effect of an unexpected U.S. monetary policy shock on federal funds rate, GDP and inflation. Linear (gray, dashed), parametric nonlinear with $F(x) = \max(0, x)$ (red, point-dashed) and sieve (blue, solid) structural impulse responses. For $\delta = +1$, the lowest point of the GDP response is marked with a dot.

shocks. In both setups, however, impulse response interpretation is only suggestive, as confidence bands are missing and only pointwise IRFs are available. Whether the puzzles highlighted above would persist after accounting for estimation uncertainty is an important research question that I leave for future analysis.

3.6.2 Uncertainty Shocks

Uncertainty in interest rates appears to be a significant factor in recent economic history. Starting with the fundamental changes brought forth by the unprecedented measures of unconventional monetary policy after the 2007-2008 financial crisis, to the powerful economic stimuli during the COVID-19 pandemic, and finally the subsequent interest rate tightening and inflation phenomenon of 2022, central banks and institutional agents are often very concerned about uncertainty. Since traditional central bank policymaking is heavily guided by the principle that the central bank *can* and *should* influence expectations, controlling the (perceived) level of ambiguity in current and future commitments is key.

Istrefi and Mouabbi (2018) provide an analysis of the impact of unforeseen changes in the level of subjective interest rate uncertainty on the macroeconomy. They derive a collection of new indices based on short- and long-term profession forecasts. Their empirical study goes in depth into studying the different components that play a role in transmitting uncertainty shocks, but here I will focus on re-evaluating their structural impulse response estimates under the light of potentially-missing nonlinear effects. For the sake of simplicity, my evaluation will focus only on the 3-months-ahead uncertainty measure for short-term interest rate maturities (3M3M) and the US economy.[24]

Like in Istrefi and Mouabbi (2018), let $Z_t = (X_t, \text{IP}_t, \text{CPI}_t, \text{PPI}_t, \text{RT}_t, \text{UR}_t)'$ be a vector where X_t is the chosen uncertainty measure, IP_t is the (log) industrial production index, CPI_t is the CPI inflation rate, PPI_t is the producer price inflation rate, RT_t is (log) retail sales and UR_t is the unemployment rate. The nonlinear model specification is given by

$$Z_t = \mu + A_1 Z_{t-1} + A_2 Z_{t-1} + F_1(X_{t-1}) + F_2(X_{t-2}) + DW_t + u_t,$$

where W_t includes a linear time trend and oil price OIL_t.[25] The data has monthly frequency and spans the period between May 1993 and July 2015.[26] Note here that, following the identification strategy of Gonçalves et al. (2021), nonlinear functions F_1 and F_2 are to be understood as not effecting X_t, which is the structural variable. The linear VAR specification of Istrefi and Mouabbi (2018) is recovered by simply assuming $F_1 = F_2 = 0$ prior to estimation. Since they use recursive identification and order the uncertainty measure first, this model too is block-recursive.

I consider a positive shock with intensity $\delta = \sigma_{\epsilon,1}$, where $\sigma_{\epsilon,1}$ is the standard deviation of structural innovations. In this empirical exercise, the relaxation function is given by

$$\rho(z) = \mathbb{I}\left\{|z| \leq \frac{1}{4}\right\} \exp\left(1 + \left[|4x|^8 - 1\right]^{-1}\right)$$

and I set $\{0.1, 0.3\}$ to be the cubic spline knots. As 3M3M is a non-negative measure of uncertainty, some care must be taken to make sure that the shocked paths for X_t do not reach negative values. Figure 3.14 in Appendix 3.B shows that the relaxation function is compatible, and also that the shocked nonlinear paths of X_t with impulse δ and δ' all do not cross below zero.

Figure 3.7 presents both the linear and nonlinear structural impulse responses obtained. Importantly, even though Istrefi and Mouabbi (2018) estimate a Bayesian VAR model and here I consider a frequentist vector autoregressive benchmark, the shape of the IRFs is retained, c.f. the median response in the top row of their Figure 4. When uncertainty increases, industrial production drops, and the size and extent of this decrease is intensified in the nonlinear responses. In fact, the sieve IP response reaches a value that is 54% lower than that of the respective linear IRF.[27] A similar

[24] Istrefi and Mouabbi (2018) also provide comparisons with results obtained with the other uncertainty measures, which they comment are all very similar to the ones obtained with 3M3M. Their paper additionally evaluates a number of other highly developed countries.

[25] Inclusion of linear exogenous variables in the semi-nonparametric theoretical framework detail in Section 3.3 is straightforward as long as one can assume that they are stationary and weakly dependent. The choice of using $p = 2$ is identical to that of the original authors, based on BIC.

[26] I reuse the original data employed by the authors, who kindly shared it upon request, but rescale retail sales (RT_t) so that the level on January 2000 equals 100.

[27] Figure 3.15 in Appendix 3.B confirms that this difference is consistent over a range of shock sizes, too.

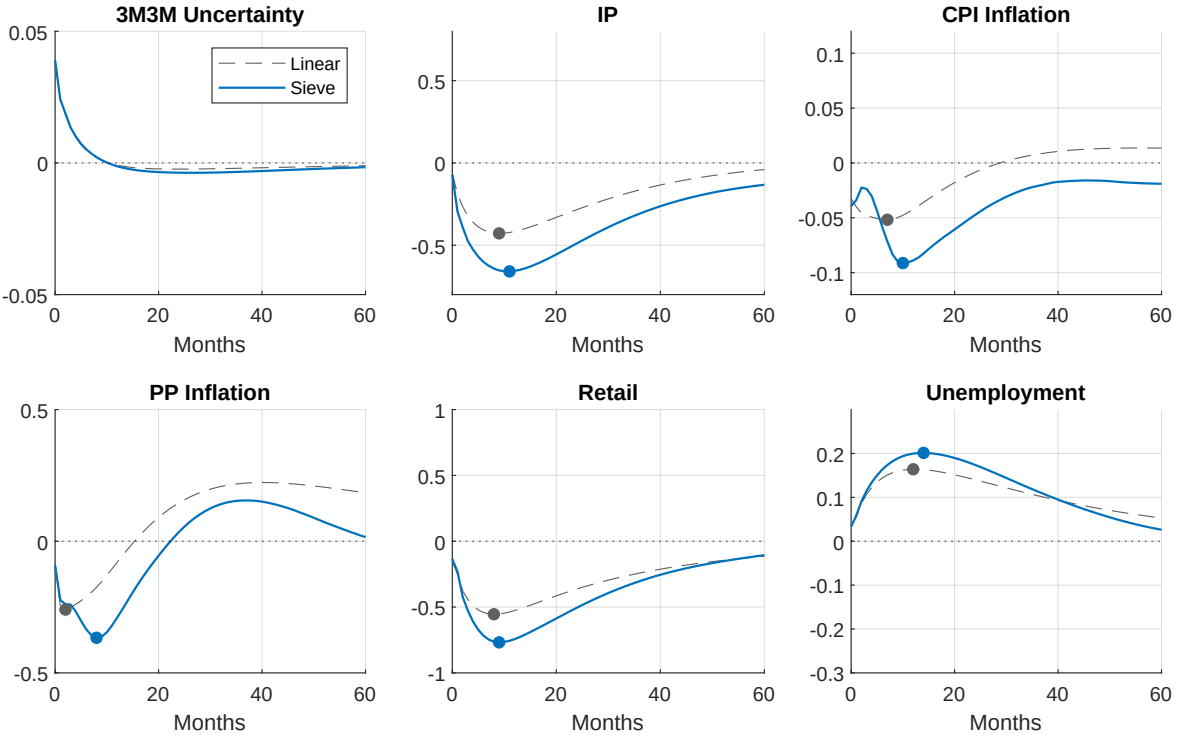


Figure 3.7: Effect of an unexpected, one-standard-deviation uncertainty shock to US macroeconomic variables. Linear (gray, dashed) and sieve (blue, solid) structural impulse responses. The extreme points of the responses are marked with a dot.

behavior holds true for retail sales (38% lower) and unemployment (23% higher), proving that this shock is more profoundly contractionary than suggested by the linear VAR model. Further, CPI and PP inflation both show short-term fluctuations which strengthen the short- and medium-term impact of the shock. CPI and PP nonlinear inflation responses are 76% and 41% stronger than their linear counterpart, respectively. These differences suggest that linear IRFs might be both under-estimating the short-term intensity and misrepresenting long-term persistence of inflation reactions. From another perspective, Nowzohour and Stracca (2020) presented evidence that consumer consumption growth, credit growth and unemployment do not co-move with the policy uncertainty index (EPU) of Baker et al. (2016), but are negatively correlated with financial volatility. Given the strength of nonlinear IRFs, this discrepancy may also suggest that the 3M3M uncertainty measure partially captures the financial channel, too.

The introduction of nonlinear terms in the structural VAR of Istrefi and Mouabbi (2018) thus provides evidence that fundamental impulse response features might otherwise be missed. Indeed, Figure 3.13 in Appendix 3.B - which plots regression functions of endogenous variables with respect to X_t - proves that high and low uncertainty levels may have significantly different effects on endogenous economic variables. In particular, at the second lag, tail effects appear to be milder, while at low levels changes in uncertainty have more pronounced impact.

3.7 Conclusion

This paper studies the application of semi-nonparametric series estimation to the problem of structural impulse response analysis for time series. After first discussing the partial identification model setup, I have used the conditions of system contractivity and stability to derive physical measures of the dependence for nonlinear systems. In turn, these allow to derive primitive conditions under which series estimation can be employed and structural IRFs are consistently estimated. The simulation results prove that this approach is valid in moderate samples and has the added benefit of being robust to misspecification of the nonlinear model components. Finally, two empirical applications showcase the utility in departing from both linear and parametric nonlinear specifications when estimating structural responses.

There are many possible avenues for extending the results I have presented here. A key aspect that I have not touched upon is inference in the form of confidence intervals: the theory of Chen and Christensen (2015) does not encompass uniform inference, and, as such, additional results have to be developed. Indeed, (uniform) confidence bands are necessary to properly quantify the uncertainty of IRF estimates. Belloni et al. (2015b) give a uniform asymptotic inference theory, but their derivations are limited to non-dependent data. Li and Liao (2020) and Cattaneo et al. (2022) provide theoretical coupling results that could be exploited in order to handle time series data. Chen and Christensen (2018) give a theory of uniform inference for panel IV setups, which could possibly be generalized to handle nonlinear IRFs. In the spirit of Kang (2021), it would be also important to derive inference results that are uniform in the selection of series terms, as, in practice, a data-driven procedure for selecting K should be used. Studying other sieve spaces, such as neural networks or shape-preserving sieves (Chen, 2007), would also be highly desirable. The latter can be especially useful in contexts where economic knowledge suggests that the nonlinear components of the model are e.g. strictly monotonic increasing or convex. Finally, sharpening of convergence rates used in the main proofs is of independent interest.

Appendix

Matrix Norms. Let

$$\|A\|_r := \max\{\|Ax\|_r \,|\, \|x\|_r \leq 1\}$$

be the r-operator norm of matrix $A \in \mathbb{C}^{d_1 \times d_2}$. The following Theorem establishes the equivalence between different operator norms as well as the compatibility constants.

Theorem 3..1 (Feng (2003))**.** *Let $1 \leq p, q \leq \infty$. Then for all $A \in \mathbb{C}^{d_1 \times d_2}$,*

$$\|A\|_p \leq \lambda_{p,q}(d_1)\lambda_{q,p}(d_2)\|A\|_q,$$

where

$$\lambda_{a,b}(d) := \begin{cases} 1 & \text{if } a \geq b, \\ d^{1/a-1/b} & \text{if } a < b. \end{cases}$$

This norm inequality is sharp.

In particular, if $p > q$ then it holds

$$\frac{1}{(d_2)^{1/q-1/p}}\|A\|_p \leq \|A\|_q \leq (d_1)^{1/q-1/p}\|A\|_p.$$

3.A Proofs

3.A.1 GMC Conditions and Proposition 3.3.6

Lemma 3.A.1. *Assume that $\{\epsilon_t\}_{t\in\mathbb{Z}}$, $\epsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_Z}$ are i.i.d., and $\{Z_t\}_{t\in\mathbb{Z}}$ is generated according to*

$$Z_t = G(Z_{t-1}, \epsilon_t),$$

where $Z_t \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ and G is a measurable function. If either

(a) Contractivity conditions (3.11)-(3.12) *hold,* $\sup_{t\in\mathbb{Z}}\|\epsilon_t\|_{L^r} < \infty$ *and* $\|G(\overline{z}, \overline{\epsilon})\| < \infty$ *for some* $(\overline{z}, \overline{\epsilon}) \in \mathcal{Z} \times \mathcal{E}$*;*

(b) Stability conditions (3.13)-(3.14) *hold,* $\sup_{t\in\mathbb{Z}}\|\epsilon_t\|_{L^r} < \infty$ *and* $\|\partial G/\partial Z\| \leq M_Z < \infty$*;*

then

$$\sup_t \|Z_t\|_{L^r} < \infty \quad w.p.1.$$

Proof.

(a) In a first step, we show that, given event $\omega \in \Omega$, realization $Z_t(\omega)$ is unique with probability one. To do this, introduce initial condition $z_\circ$ for $\ell > 1$ such that $z_\circ \in \mathcal{Z}$ and $\|z_\circ\| < \infty$. Define

$$Z_t^{(-\ell)}(\omega) = G^{(\ell)}(y_\circ, \epsilon_{t-\ell+1:t}(\omega)).$$

Further, let $Z_t'^{(-\ell)}$ be the realization with initial condition $z_\circ' \neq z_\circ$ and innovation realizations $\epsilon_{t-\ell+1:t}(\omega)$. Note that

$$\left\| Z_t^{(-\ell)}(\omega) - Z_t'^{(-\ell)}(\omega) \right\| \leq C_Z^\ell \left\| z_\circ - z_\circ' \right\|,$$

which goes to zero as $\ell \to \infty$. Therefore, if we set $Z_t(\omega) := \lim_{\ell\to\infty} Z_t^{(-\ell)}(\omega)$, $Z_t(\omega)$ is unique with respect to the choice of $z_\circ$ w.p.1. A similar recursion shows that

$$\left\| Z_t^{(-\ell)}(\omega) \right\| \leq C_Z^\ell \left\| z_\circ \right\| + \sum_{k=0}^{\ell-1} C_Z^k C_\epsilon \left\| \epsilon_{t-k}(\omega) \right\|.$$

By norm equivalence, this implies

$$\begin{aligned} \left\| Z_t^{(-\ell)} \right\|_{L^r} &\leq C_Z^\ell \left\| z_\circ \right\|_r + \sum_{k=0}^{\ell-1} C_Z^k C_\epsilon \left\| \epsilon_{t-k} \right\|_{L^r} \\ &\leq C_Z^\ell \left\| z_\circ \right\|_r + (1 - C_Z)^{-1} C_\epsilon \sup_{t\in\mathbb{Z}} \|\epsilon_t\|_{L^r} < \infty, \end{aligned}$$

and taking the limit $\ell \to \infty$ proves the claim.

(b) Consider again distinct initial conditions $z_\circ' \neq z_\circ$ and innovation realizations $\epsilon_{t-\ell+1:t}(\omega)$, yielding $Z_t'^{(-\ell)}(\omega)$ and $Z_t^{(-\ell)}(\omega)$, respectively. We may use the contraction bound derived in the proof of Proposition 3.3.6 (b) below, that is,

$$\left\| Z_t^{(-\ell)}(\omega) - Z_t'^{(-\ell)}(\omega) \right\|_r \leq C_Z^\ell C_2 \|z_\circ - z_\circ'\|_r,$$

where $C_2 > 0$ is a constant. With trivial adjustments, the uniqueness and limit arguments used for (a) above apply here too.

□

Proof of Proposition 3.3.6.

(a) By assumption it holds that for all $(z, z') \in \mathcal{Z} \times \mathcal{Z}$ and $(e, e') \in \mathcal{E} \times \mathcal{E}$

$$\|G(z, \epsilon) - G(z', \epsilon')\| \leq C_Z \|z - z'\| + C_\epsilon \|e - e'\|$$

holds, where $0 \leq C_Z < 1$ and $0 \leq C_\epsilon < \infty$. The equivalence of norms directly generalizes this inequality to any r-norm for $r > 2$. We study $\|Z_{t+h} - Z_{t+h}'\|_r$ where Z_{t+h}' is constructed with a time-t perturbation of the history of Z_{t+h}. Therefore, for any given t and $h \leq 1$ it holds that

$$\begin{aligned} \left\| Z_{t+h} - G^{(h)}(Z_t', \epsilon_{t+1:t+h}) \right\|_r &\leq C_Z \|G^{(h-1)}(Z_t, \epsilon_{t+1:t+h-1}) - G^{(h-1)}(Z_t', \epsilon_{t+1:t+h-1})\|_r \\ &\leq C_Z^h \|Z_t - Z_t'\|_r, \end{aligned}$$

since sequence $\epsilon_{t+1:t+h}$ is common between Z_{t+h} and Z_{t+h}'. Clearly then

$$\left\| Z_{t+h} - G^{(h)}(Z_t', \epsilon_{t+1:t+h}) \right\|_r \leq 2\|Z_t\|_r \exp(-\gamma h)$$

for $\gamma = -\log(C_Z)$. Letting $a = 2\|Z_t\|_r$ and shifting time index t backward by h, since $\sup_t \|Z_t\|_{L^r} < \infty$ w.p.1 from Lemma 3.A.1 the result for L^r follows with $\tau = 1$.

(b) Proceed similar to (a), but notice that now we must handle cases of steps $1 \leq h < h^*$. Consider iterate $h^* + 1$, for which

$$\begin{aligned}\left\| Z_{t+h+1} - G^{(h+1)}(Z'_t, \epsilon_{t+1:t+h+1}) \right\|_r &\leq C_Z \|G^{(h)}(G(Z_t, \epsilon_{t+1}), \epsilon_{t+2:t+h}) - G^{(h)}(G(Z'_t, \epsilon_{t+1}), \epsilon_{t+2:t+h})\|_r \\ &\leq C_Z^h \|G(Z_t, \epsilon_{t+1}) - G(Z'_t, \epsilon_{t+1})\|_r \\ &\leq C_Z^h M_Z \|Z_t - Z'_t\|_r\end{aligned}$$

by the mean value theorem. Here we may assume that $M_Z \geq 1$ otherwise we would fall under case (a), so that $M_Z \leq M_Z^2 \leq \ldots \leq M_Z^{h^*-1}$. More generally,

$$\left\| Z_{t+h+1} - G^{(h+1)}(Z'_t, \epsilon_{t+1:t+h+1}) \right\|_r \leq C_Z^{j(h)} \max\{M_Z^{h^*-1}, 1\} \|Z_t - Z'_t\|_r$$

for $j(h) := \lfloor h/h^* \rfloor$. Result (b) then follows by noting that $j(h) \geq h/h^* - 1$ and then proceeding as in (a) to derive GMC coefficients.

□

Companion and Lagged Vectors. The assumption of GMC for a process translates naturally to vectors that are composed of stacked lags of realizations. This, for example, is important in the discussion of Section 3.3 when imposing Assumption 9, since one needs that series regressors $\{W_{2t}\}_{t\in\mathbb{Z}}$ be GMC.

Recall that $W_{2t} = (X_t, X_{t-1}, \ldots, X_{t-p}, Y_{t-1}, \ldots, Y_{t-p}, \epsilon_{1t})$. Here we shall reorder this vector slightly to be

$$W_{2t} = (X_t, X_{t-1}, Y_{t-1}, \ldots, X_{t-p}, Y_{t-p}, \epsilon_{1t}).$$

For $h > 0$ and $1 \leq l \leq h$, let $Z'_{t+j} := \Phi^{(l)}(Z'_t, \ldots, Z'_{t-p}; \epsilon_{t+1:t+j})$ be the a perturbed version of Z_t, where $Z'_t, \ldots, Z'_{t-p}$ are taken from an independent copy of $\{Z_t\}_{t\in\mathbb{Z}}$. Define

$$W'_{2t} = (X'_t, X'_{t-1}, Y'_{t-1}, \ldots, X'_{t-p}, Y'_{t-p}, \epsilon_{1t}).$$

Using Minkowski's inequality

$$\begin{aligned}\|W_{2t+h} - W'_{2t+h}\|_{L^r} &\leq \|X_{t+h} - X'_{t+h}\|_{L^r} + \sum_{j=1}^{p} \|Z_{t+h-j} - Z'_{t+h-j}\|_{L^r} \\ &\leq \sum_{j=0}^{p} \|Z_{t+h-j} - Z'_{t+h-j}\|_{L^r},\end{aligned}$$

thus, since $p > 0$ is fixed finite,

$$\sup_t \|W_{2t+h} - W'_{2t+h}\|_{L^r} \leq \sum_{j=0}^{p} \Delta_r(h-j) \leq (p+1)\, a_{1Z} \exp(-a_{2Z} h).$$

Above, a_{1Z} and a_{2Z} are the GMC coefficients of $\{Z_t\}_{t\in\mathbb{Z}}$.

3.A.2 Lemma 3.3.7 and Matrix Inequalities under Dependence

In order to prove Lemma 3.3.7, the idea is to modify the approach of Chen and Christensen (2015), which relies on Berbee's Lemma and an interlaced coupling, to handle variables with physical dependence. Chen et al. (2016) provide an example on how to achieve this when working with self-normalized sums. In what follows I modify their ideas to work with random dependent matrices.

First of all, I recall below a Bernstein-type inequality for independent random matrices of Tropp (2012).

Theorem 3.A.2. *Let $\{\Xi_i\}_{i=1}^n$ be a finite sequence of independent random matrices with dimensions $d_1 \times d_2$. Assume $\mathbb{E}[\Xi_i] = 0$ for each i and $\max_{1\leq i\leq n}\|\Xi_i\| \leq R_n$ and define*

$$\varsigma_n^2 := \max\left\{\left\|\sum_{i=1}^n \mathbb{E}\left[\Xi_{i,n}\Xi_{j,n}'\right]\right\|, \left\|\sum_{i=1}^n \mathbb{E}\left[\Xi_{i,n}'\Xi_{j,n}\right]\right\|\right\}.$$

Then for all $z \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_i\right\| \geq z\right) \leq (d_1+d_2)\exp\left(\frac{-z^2/2}{nq\varsigma_n^2 + qR_n z/3}\right).$$

The main exponential matrix inequality due to Chen and Christensen (2015), Theorem 4.2 is as follows.

Theorem 3.A.3. *Let $\{X_i\}_{i\in\mathbb{Z}}$ where $X_i \in \mathcal{X}$ be a β-mixing sequence and let $\Xi_{i,n} = \Xi_n(X_i)$ for each i where $\Xi_n : \mathcal{X} \mapsto \mathbb{R}^{d_1\times d_2}$ be a sequence of measurable $d_1 \times d_2$ matrix-valued functions. Assume that $\mathbb{E}[\Xi_{i,n}] = 0$ and $\|\Xi_{i,n}\| \leq R_n$ for each i and define*

$$S_n^2 := \max\left\{\mathbb{E}\left[\|\Xi_{i,n}\Xi_{j,n}'\|\right], \mathbb{E}\left[\|\Xi_{i,n}'\Xi_{j,n}\|\right]\right\}.$$

Let $1 \leq q \leq n/2$ be an integer and let $I_\bullet = q\lfloor n/q\rfloor, \ldots, n$ when $q\lfloor n/q\rfloor < n$ and $I_\bullet = \emptyset$ otherwise. Then, for all $z \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_{i,n}\right\| \geq 6z\right) \leq \frac{n}{q}\beta(q) + \mathbb{P}\left(\left\|\sum_{i\in I_\bullet} \Xi_{i,n}\right\| \geq z\right) + 2(d_1+d_2)\exp\left(\frac{-z^2/2}{nqS_n^2 + qR_n z/3}\right),$$

where $\|\sum_{i\in I_\bullet}\Xi_{i,n}\| := 0$ whenever $I_\bullet = \emptyset$.

To fully extend Theorem 3.A.3 to physical dependence, I will proceed in steps. First, I derive a similar matrix inequality by directly assuming that random matrices $\Xi_{i,n}$ have physical dependence coefficient $\Delta_r^{\Xi}(h)$. In the derivations I will use that

$$\frac{1}{(d_2)^{1/2-1/r}}\|A\|_r \leq \|A\|_2 \leq (d_1)^{1/2-1/r}\|A\|_r.$$

for $r \geq 2$.

Theorem 3.A.4. *Let $\{\epsilon_j\}_{j\in\mathbb{Z}}$ be a sequence of i.i.d. variables and let $\{\Xi_{i,n}\}_{i=1}^n$,*

$$\Xi_{i,n} = G_n^{\Xi}(\ldots, \epsilon_{i-1}, \epsilon_i)$$

for each i, where $\Xi_n : \mathcal{X} \mapsto \mathbb{R}^{d_1 \times d_2}$, be a sequence of measurable $d_1 \times d_2$ matrix-valued functions. Assume that $\mathbb{E}[\Xi_{i,n}] = 0$ and $\|\Xi_{i,n}\| \leq R_n$ for each i and define

$$S_n^2 := \max\left\{\mathbb{E}\left[\|\Xi_{i,n}\Xi_{j,n}'\|\right], \mathbb{E}\left[\|\Xi_{i,n}'\Xi_{j,n}\|\right]\right\}.$$

Additionally assume that $\|\Xi_{i,n}\|_{L^r} < \infty$ for $r > 2$ and define the matrix physical dependence measure $\Delta_r^{\Xi}(h)$ as

$$\Delta_r^{\Xi}(h) := \max_{1\leq i\leq n} \left\| \Xi_{i,n} - \Xi_{i,n}^{h*} \right\|_{L^r},$$

where $\Xi_{i,n}^{h*} := G_n^{\Xi}(\ldots, \epsilon_{i-h-1}^*, \epsilon_{i-h}^*, \epsilon_{i-h+1}, \ldots, \epsilon_{i-1}, \epsilon_i)$ for independent copy $\{\epsilon_j^*\}_{j\in\mathbb{Z}}$. Let $1 \leq q \leq n/2$ be an integer and let $I_\bullet = q\lfloor n/q\rfloor, \ldots, n$ when $q\lfloor n/q\rfloor < n$ and $I_\bullet = \emptyset$ otherwise. Then, for all $z \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_{i,n}\right\| \geq 6z\right) \leq \frac{n^{r+1}}{q^r (d_2)^{r/2-1} z^r}\, \Delta_r^{\Xi}(q) + \mathbb{P}\left(\left\|\sum_{i\in I_\bullet} \Xi_{i,n}\right\| \geq z\right) + 2(d_1+d_2)\exp\left(\frac{-z^2/2}{nqS_n^2 + qR_n z/3}\right),$$

where $\|\sum_{i\in I_\bullet} \Xi_{i,n}\| := 0$ whenever $I_\bullet = \emptyset$.

Proof. To control dependence, we can adapt the interlacing block approach outlined by Chen et al. (2016). To interlace the sum, split it into

$$\sum_{i=1}^n \Xi_{i,n} = \sum_{j\in K_e} J_k + \sum_{j\in J_o} W_k + \sum_{i\in I_\bullet} \Xi_{i,n},$$

where $W_j := \sum_{i=q(j-1)+1}^{qj} \Xi_{i,n}$ for $j = 1, \ldots, \lfloor n/q\rfloor$ are the blocks, $I_\bullet := \{q\lfloor n/q\rfloor + 1, \ldots, n\}$ if $q\lfloor n/q\rfloor < n$ and J_e and J_o are the subsets of even and odd numbers of $\{1, \ldots, \lfloor n/q\rfloor\}$, respectively. For simplicity define $J = J_e \cup J_o$ as the set of block indices and let

$$W_j^\dagger := \mathbb{E}[\, W_j \,|\, \epsilon_\ell,\, q(j-2)+1 \leq \ell \leq qj\,].$$

Note that by construction $\{W_j^\dagger\}_{j\in J_e}$ are independent and also $\{W_j^\dagger\}_{j\in J_o}$ are independent. Using the triangle inequality we find

$$\begin{aligned}
\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_{i,n}\right\| \geq 6z\right) &\leq \mathbb{P}\left(\left\|\sum_{j\in J}(W_j - W_j^\dagger)\right\| + \left\|\sum_{j\in J} W_j^\dagger\right\| + \left\|\sum_{i\in I_\bullet} \Xi_{i,n}\right\| \geq 6z\right) \\
&\leq \mathbb{P}\left(\left\|\sum_{j\in J}(W_j - W_j^\dagger)\right\| \geq z\right) + \mathbb{P}\left(\left\|\sum_{j\in J_e} W_j^\dagger\right\| \geq z\right) \\
&\quad + \mathbb{P}\left(\left\|\sum_{j\in J_o} W_j^\dagger\right\| \geq z\right) + \mathbb{P}\left(\left\|\sum_{i\in I_\bullet} \Xi_{i,n}\right\| \geq z\right) \\
&= I + II + III + IV.
\end{aligned}$$

We keep term IV as is. As in the proof of Chen and Christensen (2015), terms II and III consist of sums of independent matrices, where each $W_j^\dagger$ satisfies $\|W_j^\dagger\| \leq qR_n$ and

$$\max\left\{\mathbb{E}\left[\|W_j^\dagger\, W_j^{\dagger\prime}\|\right], \mathbb{E}\left[\|W_j^{\dagger\prime}\, W_j^\dagger\|\right]\right\} \leq qS_n^2.$$

Then, using the exponential matrix inequality of Tropp (2012),

$$\mathbb{P}\left(\left\|\sum_{j\in J_e} W_k^\dagger\right\| \geq z\right) \leq (d_1+d_2)\exp\left(\frac{-z^2/2}{nqS_n^2+qR_nz/3}\right).$$

The same holds for the sum over J_o. Finally, we use the physical dependence measure Δ_r^Ξ to bound I. Start with the union bound to find

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{j\in J}(W_j-W_j^\dagger)\right\| \geq z\right) &\leq \mathbb{P}\left(\sum_{j\in J}\left\|W_j-W_j^\dagger\right\| \geq z\right)\\ &\leq \frac{n}{q}\,\mathbb{P}\left(\left\|W_j-W_j^\dagger\right\| \geq \frac{q}{n}\,z\right),\end{aligned}$$

where we have used that $\lfloor n/q\rfloor \leq n/q$. Since W_j and $W_j^\dagger$ differ only over a σ-algebra that is q steps in the past, by assumption

$$\left\|W_j-W_j^\dagger\right\|_{L^r} \leq q\,\Delta_r^\Xi(q),$$

which implies, by means of the rth moment inequality,

$$\mathbb{P}\left(\left\|W_j-W_j^\dagger\right\| \geq \frac{q}{n}\,z\right) \leq \mathbb{P}\left((d_2)^{1/r-1/2}\left\|W_j-W_j^\dagger\right\|_r \geq \frac{q}{n}\,z\right) \leq \frac{n^r}{q^{r-1}(d_2)^{r/2-1}z^r}\,\Delta_r^\Xi(q).$$

where $(d_2)^{1/r-1/2}$ is the operator norm equivalence constant such that $\|\cdot\| \geq (d_2)^{1/r-1/2}\|\cdot\|_r$ (Feng, 2003). Therefore,

$$\mathbb{P}\left(\left\|\sum_{j\in J}(W_j-W_j^\dagger)\right\| \geq z\right) \leq \frac{n^{r+1}}{q^r(d_2)^{r/2-1}z^r}\,\Delta_r^\Xi(q)$$

as claimed. □

Notice that the first term in the bound is weaker than that derived by Chen and Christensen (2015). The β-mixing assumption and Berbee's Lemma give strong control over the probability $\mathbb{P}(\|\sum_{j\in J}(W_j-W_j^\dagger)\| \geq z)$. In contrast, assuming physical dependence means we have to explicitly handle a moment condition. One might think of sharpening Theorem 3.A.4 by sidestepping the rth moment inequality (c.f. avoiding Chebyshev's inequality in concentration results), but I do not explore this approach here.

The second step is to map the physical dependence of a generic vector time series $\{X_i\}_{i\in\mathbb{Z}}$ to matrix functions.

Proposition 3.A.5. *Let $\{X_i\}_{i\in\mathbb{Z}}$ where $X_i = G(\ldots,\epsilon_{i-1},\epsilon_i) \in \mathcal{X}$ for $\{\epsilon_j\}_{j\in\mathbb{Z}}$ i.i.d. be a sequence with finite rth moment, where $r>0$, and functional physical dependence coefficients*

$$\Delta_r(h) = \sup_i \left\| X_{i+h} - G^{(h)}(X_i^*, \epsilon_{i+1:i+h}) \right\|_{L^r}$$

for $h \geq 1$. Let $\Xi_{i,n} = \Xi_n(X_i)$ for each i where $\Xi_n : \mathcal{X} \mapsto \mathbb{R}^{d_1\times d_2}$ be a sequence of measurable $d_1\times d_2$ matrix-valued functions such that $\Xi_n = (v_1,\ldots,v_{d_2})$ for $v_\ell \in \mathbb{R}^{d_1}$. If $\|\Xi_{i,n}\|_{L^r} < \infty$ and

$$C_{\Xi,\ell} := \sup_{x\in\mathcal{X}} \|\nabla v_\ell(x)\| \leq C_\Xi < \infty,$$

then matrices $\Xi_{i,n}$ have physical dependence coefficients

$$\Delta_r^{\Xi}(h) = \sup_i \left\| \Xi_{i,n} - \Xi_{i,n}^{h*} \right\|_{L^r} \leq \sqrt{d_1} \left(\frac{d_2}{d_1}\right)^{1/r} C_\Xi \, \Delta_r(h),$$

where $\Xi_{i,n}^{h*} = \Xi_n(G^{(h)}(X_i', \epsilon_{i+1:i+h}))$.

Proof. To derive the bound, we use $\Xi_n(X_i)$ and $\Xi_n(X_i^{h*})$ in place of $\Xi_{i,n}$ and $\Xi_{i,n}^{h*}$, respectively, where $X_i^{h*} = G^{(h)}(X_i^*, \epsilon_{i+1:i+h})$. First we move from studying the operator r-norm (recall, $r > 2$) to the Frobenius norm,

$$\left\|\Xi_n(X_i) - \Xi_n(X_i^{h*})\right\|_r \leq (d_2)^{1/2-1/r} \left\|\Xi_n(X_i) - \Xi_n(X_i^{h*})\right\|_F.$$

where as intermediate step we use the 2-norm. Let $\Xi_n = (v_1, \ldots, v_{d_2})$ for $v_\ell \in \mathbb{R}^{d_1}$ and $\ell \in 1, \ldots, d_2$, so that

$$\|\Xi_n\|_F = \sqrt{\sum_{\ell=1}^{d_2} \|v_\ell\|^2}$$

where $v_\ell = (v_{\ell 1}, \ldots, v_{\ell d_1})'$. Since $v_\ell : \mathcal{X} \mapsto \mathbb{R}^{d_1}$ are vector functions, the mean value theorem gives that

$$\left\|\Xi_n(X_i) - \Xi_n(X_i^{h*})\right\|_F \leq \sqrt{\sum_{\ell=1}^{d_2} C_{\Xi,\ell}^2 \, \|X_i - X_i^{h*}\|^2} \leq \sqrt{d_2}\, C_\Xi \, \|X_i - X_i^{h*}\|.$$

Combining results and moving from the vector r-norm to the 2-norm yields

$$\left\|\Xi_n(X_i) - \Xi_n(X_i^{h*})\right\|_r \leq (d_2)^{1-1/r} (d_1)^{1/2-1/r} \, C_\Xi \, \|X_i - X_i^{h*}\|_r.$$

The claim involving the L^r norm follows immediately. □

The following Corollary, which specifically handles matrix functions defined as outer products of vector functions, is immediate and covers the setups of series estimation.

Corollary 3.A.6. *Under the conditions of Proposition 3.A.5, if*

$$\Xi_n(X_i) = \xi_n(X_i)\xi_n(X_i)' + Q_n$$

where $\xi_n : \mathcal{X} \mapsto \mathbb{R}^d$ is a vector function and $Q_n \in \mathbb{R}^{d \times d}$ is nonrandom matrix, then

$$\Delta_r^{\Xi}(h) \leq d^{3/2-2/r} \, C_\xi \, \Delta_r(h),$$

where $C_\xi := \sup_{x \in \mathcal{X}} \|\nabla \xi_n(x)\| < \infty$.

Proof. Matrix Q_n cancels out since it is nonrandom and appears in both $\Xi_n(X_i)$ and $\Xi_n(X_i^{h*})$. Since $\Xi_n(X_i)$ is square, the ratio of row to column dimensions simplifies. □

The following Corollaries to Theorem 3.A.4 can now be derived in a straightforward manner.

Corollary 3.A.7. *Under the conditions of Theorem 3.A.4 and Proposition 3.A.5, for all $z \geq 0$*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \Xi_{i,n}\right\| \geq 6z\right) \leq \frac{n^{r+1}}{q^r z^r} (d_2)^{2-(r/2+1/r)} (d_1)^{1/2-1/r} C_\Xi \, \Delta_r(q) + \mathbb{P}\left(\left\|\sum_{i \in I_\bullet} \Xi_{i,n}\right\| \geq z\right)$$

$$+ 2(d_1 + d_2) \exp\left(\frac{-z^2/2}{nqS_n^2 + qR_n z/3}\right).$$

where $\Delta_r(\cdot)$ if the functional physical dependence coefficient of X_i.

Corollary 3.A.8. *Under the conditions of Theorem 3.A.4 and Proposition 3.A.5, if $q = q(n)$ is chosen such that*

$$\frac{n^{r+1}}{q^r}(d_2)^{2-(r/2+1/r)}(d_1)^{1/2-1/r} C_\Xi \, \Delta_r(q) = o(1)$$

and $R_n\sqrt{q\log(d_1+d_2)} = o(S_n\sqrt{n})$ then

$$\left\|\sum_{i=1}^n \Xi_{i,n}\right\| = O_P\left(S_n\sqrt{nq\log(d_1+d_2)}\right).$$

This result is almost identical to Corollary 4.2 in Chen and Christensen (2015), with the only adaptation of using Theorem 3.A.4 as a starting point. Condition $R_n\sqrt{q\log(d_1+d_2)} = o(S_n\sqrt{n})$ is simple to verify by assuming, e.g., $q = o(n/\log(n))$ since $\log(d_1+d_2) \lesssim \log(K)$ and $K = o(n)$.

Note that when $d_1 = d_2 \equiv K$, which is the case of interest in the series regression setup, the first condition in Corollary 3.A.8 reduces to

$$K^{5/2-(r/2+2/r)}\, C_\Xi \, \Delta_r(q) = o(1),$$

which also agrees with the rate of Corollary 3.A.6. Assumption 7(i) and a compact domain further allow to explicitly bound factor C_Ξ by

$$C_\Xi \lesssim K^{\omega_2},$$

so that the required rate becomes

$$K^\rho \, \Delta_r(q) = o(1), \quad \text{where} \quad \rho := \frac{3}{2} - \frac{r}{2} + \omega_2.$$

Proof of Lemma 3.3.7. The proof follows from Corollary 3.A.8 by the same steps of the proof of Lemma 2.2 in Chen and Christensen (2015). Simply take $\Xi_{i,n} = n^{-1}(\widetilde{(}b)_\pi^K(X_i)\widetilde{(}b)_\pi^K(X_i)' - I_K)$ and note that $R_n \le n^{-1}(1+\zeta_{K,n}^2\lambda_{K,n}^2)$ and $S_n \le n^{-2}(1+\zeta_{K,n}^2\lambda_{K,n}^2)$. □

For Lemma 3.3.7 to hold under GMC assumptions a valid choice for $q(n)$ is

$$q(n) = \gamma^{-1}\log(K^\rho n^{r+1})$$

where γ as in Proposition 3.3.6. This is due to

$$\begin{aligned}\left(\frac{n}{q}\right)^{r+1} qK^\rho\Delta_r(q) &\lesssim \frac{n^{r+1}}{q^r}K^\rho\exp(-\gamma q)\\ &\lesssim \frac{n^{r+1}K^\rho}{\log(K^\rho n^{r+1})^r}(K^\rho n^{r+1})^{-1}\\ &= \frac{1}{\log(K^\rho n^{r+1})^r} = o(1).\end{aligned}$$

Note then that, if $\lambda_{K,n} \lesssim 1$ and $\zeta_{K,n} \lesssim \sqrt{K}$, since

$$\zeta_{K,n}\lambda_{K,n}\sqrt{\frac{q\log K}{n}} \lesssim \sqrt{\frac{K\log(K^{\rho}n^{r+1})\log(K)}{n}} \lesssim \sqrt{\frac{K\log(n^{\rho+r+2})\log(n)}{n}} \lesssim \sqrt{\frac{K\log(n)^2}{n}},$$

to satisfy Assumption 8 we may assume $\sqrt{K\log(n)^2/n} = o(1)$ as in Remark 2.3 of Chen and Christensen (2015) for the case of exponential β-mixing regressors.

3.A.3 Theorem 3.3.9

Before delving into the proof of Theorem 3.3.9, note that we can decompose $\widehat{\Pi}_2 - \Pi_2$ as

$$\widehat{\Pi}_2 - \Pi_2 = (\widehat{\Pi}_2 - \widehat{\Pi}_2^*) + (\widehat{\Pi}_2^* - \widetilde{\Pi}_2) + (\widetilde{\Pi}_2 - \Pi_2),$$

where $\widetilde{\Pi}_2$ is the projection of Π_2 onto the linear space spanned by the sieve. The last two terms can be handled directly with the theory developed by Chen and Christensen (2015). Specifically, their Lemma 2.3 controls the second term (variance term), while Lemma 2.4 handles the third term (bias term). This means here we can focus on the first term, which is due to using generated regressors $\widehat{\epsilon}_{1t}$ in the second step.

Since $\widehat{\Pi}_2$ can be decomposed in d_Y rows of semi-nonparametric coefficients, i.e.,

$$Y_t = \begin{bmatrix} \pi_{2,1} \\ \vdots \\ \pi_{2,d_Y} \end{bmatrix} W_{2t} + \widetilde{u}_{2t},$$

we further reduce to the scalar case. Let π_2 be any row of Π_2 and, with a slight abuse of notation, Y the vector of observations of the component of Y_t of the same row, so that one may write

$$\begin{aligned} \widehat{\pi}_2(x) - \widehat{\pi}_2^*(x) &= \widetilde{b}_{\pi}^K(x)(\widehat{\widetilde{B}}_{\pi}'\widehat{\widetilde{B}}_{\pi})^{-}(\widehat{\widetilde{B}}_{\pi} - \widetilde{B}_{\pi})'Y + \widetilde{b}_{\pi}^K(x)\left[(\widehat{\widetilde{B}}_{\pi}'\widehat{\widetilde{B}}_{\pi})^{-} - (\widetilde{B}_{\pi}'\widetilde{B}_{\pi})^{-}\right]\widetilde{B}_{\pi}'Y \\ &= I + II \end{aligned}$$

where $\widetilde{b}_{\pi}^K(x) = \Gamma_{B,2}^{-1/2} b_{\pi}^K(x)$ is the orthonormalized sieve according to $\Gamma_{B,2} := \mathbb{E}[b_{\pi}^K(W_{2t})b_{\pi}^K(W_{2t})']$, $\widetilde{B}_{\pi}$ is the *infeasible* orthonormalized design matrix (involving ϵ_{1t}) and $\widehat{\widetilde{B}}_{\pi}$ is *feasible* orthonormalized design matrix (involving $\widehat{\epsilon}_{1t}$). In particular, note that

$$\widehat{B}_{\pi} = B_{\pi} + R_n, \quad \text{where} \quad R_n := \begin{bmatrix} 0 & & 0 & \widehat{\epsilon}_{11} - \epsilon_{11} \\ \vdots & \dots & \vdots & \vdots \\ 0 & & 0 & \widehat{\epsilon}_{1n} - \epsilon_{1n} \end{bmatrix} \in \mathbb{R}^{n\times K},$$

which implies $\widehat{\widetilde{B}}_{\pi} - \widetilde{B}_{\pi} = R_n\,\Gamma_{B,2}^{-1/2} =: \widetilde{R}_n$.

The next Lemma provides a bound for the difference $(\widehat{\widetilde{B}}_{\pi}'\widehat{\widetilde{B}}_{\pi}/n) - (\widetilde{B}_{\pi}'\widetilde{B}_{\pi}/n)$ that will be useful in the proof of Theorem 3.3.7 below.

Lemma 3.A.9. *Under the setup of Theorem 3.3.8, it holds*

$$\|(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n) - (\widetilde{B}_\pi'\widetilde{B}_\pi/n)\| = O_P(\sqrt{K/n}).$$

Proof. Using the expansion $\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi = \widetilde{B}_\pi'\widetilde{B}_\pi + (\widetilde{B}_\pi'\widetilde{R}_n + \widetilde{R}_n'\widetilde{B}_\pi) + \widetilde{R}_n'\widetilde{R}_n$, one immediately finds that

$$\|(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n) - (\widetilde{B}_\pi'\widetilde{B}_\pi/n)\| \leq 2\|\widetilde{B}_\pi'\widetilde{R}_n/n\| + \|\widetilde{R}_n'\widetilde{R}_n/n\|.$$

The second right-hand side factor satisfies $\|\widetilde{R}_n'\widetilde{R}_n/n\| \leq \lambda_{K,n}^2\|R_n'R_n/n\|$. Moreover,

$$\begin{aligned}
\|R_n'R_n/n\| &= \left\|\frac{1}{n}\sum_{t=1}^n(\widehat{\epsilon}_{1t} - \epsilon_{1t})^2\right\| \\
&= \left\|\frac{1}{n}\sum_{t=1}^n(\Pi_1 - \widehat{\Pi}_1)'W_{1t}W_{1t}'(\Pi_1 - \widehat{\Pi}_1)\right\| \\
&\leq \|\Pi_1 - \widehat{\Pi}_1\|^2\,\|W_1'W_1/n\| \\
&= O_P(n^{-1}),
\end{aligned}$$

since $\|W_1'W_1/n\| = O_P(1)$. Under Assumption 12, $\lambda_{K,n}^2/n = o_P(\sqrt{K/n})$ since B-splines and wavelets satisfy $\lambda_{K,n} \lesssim 1$. Consequently, $\|\widetilde{R}_n'\widetilde{R}_n/n\| = o_P(\sqrt{K/n})$.

Factor $\|\widetilde{B}_\pi'R_n/n\|$ is also straightforward, but depends on sieve dimension K,

$$\begin{aligned}
\|\widetilde{B}_\pi'R_n/n\| &\leq \left\|\frac{1}{n}\sum_{t=1}^n\widetilde{b}_\pi^K(W_{2t})(\widehat{\epsilon}_{1t} - \epsilon_{1t})\right\| \\
&= \left\|\frac{1}{n}\sum_{t=1}^n\widetilde{b}_\pi^K(W_{2t})W_{1t}'(\Pi_1 - \widehat{\Pi}_1)\right\| \\
&\leq \|\Pi_1 - \widehat{\Pi}_1\|\,\|\widetilde{B}_\pi'W_1/n\| \\
&= O_P(\sqrt{K/n}),
\end{aligned}$$

since $\|\widetilde{B}_\pi'W_1/n\| = O_P(\sqrt{K})$ as the column dimension of W_1 is fixed. The claim then follows by noting $O_P(\sqrt{K/n})$ is the dominating order of convergence. □

Proof of Theorem 3.3.9. Since $\widehat{\Pi}_1$ the least squares estimator of a linear equation, the rate of convergence is the parametric rate $n^{-1/2}$. The first result is therefore immediate.

For the second step, we consider

$$\|\widehat{\Pi}_2 - \Pi_2\|_\infty \leq \|\widehat{\Pi}_2 - \widehat{\Pi}_2^*\|_\infty + \|\widehat{\Pi}_2^* - \Pi_2\|_\infty,$$

and bound explicitly the first right-hand side term. For a given component of the regression function,

$$|\widehat{\pi}_2(x) - \widehat{\pi}_2^*(x)| \leq |I| + |II|.$$

We now control each term on the right side.

(1) It holds

$$|I| \leq \|\widetilde{b}_\pi^K(x)\|\,\|(\widehat{\widetilde{B}}_\pi'\widehat{\widetilde{B}}_\pi/n)^-\|\,\|(\widehat{\widetilde{B}}_\pi - \widetilde{B}_\pi)'Y/n\|$$

$$\leq \sup_{x\in\mathcal{W}_2} \|\widetilde{b}_\pi^K(x)\| \, \|(\widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n)^- \| \, \|(\widehat{\widetilde{B}}_\pi - \widetilde{B}_\pi)'Y/n\|$$
$$\leq \zeta_{K,n}\lambda_{K,n} \, \|(\widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n)^- \| \, \|(\widehat{\widetilde{B}}_\pi - \widetilde{B}_\pi)'Y/n\|.$$

Let $\mathcal{A}_n$ denote the event on which $\|\widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n - I_K\| \leq 1/2$, so that $\|(\widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n)^- \| \leq 2$ on $\mathcal{A}_n$. Notice that since $\|(\widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n) - (\widetilde{B}_\pi' \widetilde{B}_\pi/n)\| = o_P(1)$ (Lemma 3.A.9) and, by assumption, $\|\widetilde{B}_\pi' \widetilde{B}_\pi/n - I_K\| = o_P(1)$, then $\mathbb{P}(\mathcal{A}_n^{\mathrm{c}}) = o(1)$. On $\mathcal{A}_n$ then

$$|I| \lesssim \zeta_{K,n}\lambda_{K,n}^2 \, \|(\widehat{B}_\pi - B_\pi)'Y/n\| = \zeta_{K,n}\lambda_{K,n}^2 \, \|R_n'Y/n\|.$$

From $R_n'Y = \sum_{t=1}^n b_\pi^K(W_{2t})(\widehat{\epsilon}_{1t} - \epsilon_{1t})Y_t = (\Pi_1 - \widehat{\Pi}_1)'W_1'Y$ it follows that

$$\|R_n'Y/n\| \leq \|\Pi_1 - \widehat{\Pi}_1\| \, \|W_1'Y/n\|$$

on $\mathcal{A}_n$, meaning

$$|I| = O_P\left(\zeta_{K,n}\lambda_{K,n}^2/\sqrt{n}\right)$$

as $\|W_1'Y/n\| = O_P(1)$ and $\mathbb{P}(\mathcal{A}_n^{\mathrm{c}}) = o(1)$.

(2) Again we proceed by uniformly bounding II according to

$$|II| \leq \zeta_{K,n}\lambda_{K,n} \, \|(\widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n)^- - (\widetilde{B}_\pi' \widetilde{B}_\pi/n)^- \| \, \|\widetilde{B}_\pi'Y/n\|.$$

The last factor has order $\|\widetilde{B}_\pi'Y/n\| = O_P(\sqrt{K})$ since $\widetilde{B}_\pi$ is growing in row dimension with K. For the middle term, introduce

$$\Delta_B := \widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n - \widetilde{B}_\pi' \widetilde{B}_\pi/n$$

and event

$$\mathcal{B}_n := \left\{\|(\widetilde{B}_\pi' \widetilde{B}_\pi/n)^- \, \Delta_B\| \leq 1/2\right\} \cap \left\{\|\widetilde{B}_\pi' \widetilde{B}_\pi/n - I_K\| \leq 1/2\right\}.$$

On $\mathcal{B}_n$, we can apply the bound (Horn and Johnson, 2012)

$$\|(\widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n)^- - (\widetilde{B}_\pi' \widetilde{B}_\pi/n)^- \| \leq \frac{\|(\widetilde{B}_\pi' \widetilde{B}_\pi/n)^- \|^2 \, \|\Delta_B\|}{1 - \|(\widetilde{B}_\pi' \widetilde{B}_\pi/n)^- \, \Delta_B\|} \lesssim \|\widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n - \widetilde{B}_\pi' \widetilde{B}_\pi/n\|.$$

Since $\|\widehat{\widetilde{B}}_\pi' \widehat{\widetilde{B}}_\pi/n - \widetilde{B}_\pi' \widetilde{B}_\pi/n\| = O_P(\sqrt{K/n})$ by Lemma 3.A.9, we get

$$|II| = O_P\left(\zeta_{K,n}\lambda_{K,n}\frac{K}{\sqrt{n}}\right)$$

on $\mathcal{B}_n$. Finally, using $\mathbb{P}((A \cap B)^{\mathrm{c}}) \leq \mathbb{P}(A^{\mathrm{c}}) + \mathbb{P}(B^{\mathrm{c}})$ we note that $\mathbb{P}(\mathcal{B}_n^{\mathrm{c}}) = o(1)$ so that the bound asymptotically holds irrespective of event $\mathcal{B}_n$.

Thus, we have shown that

$$|\widehat{\pi}_2(x) - \widehat{\pi}_2^*(x)| \leq O_P\left(\zeta_{K,n}\lambda_{K,n}^2\frac{1}{\sqrt{n}}\right) + O_P\left(\zeta_{K,n}\lambda_{K,n}\frac{K}{\sqrt{n}}\right)$$

$$= O_P\left(\zeta_{K,n}\lambda_{K,n}\frac{K}{\sqrt{n}}\right)$$

as clearly $\sqrt{n}^{-1} = o(K/\sqrt{n})$ and, as discussed in the proof of Lemma 3.A.9, $\lambda_{K,n}^2/n = o_P(\sqrt{K/n})$. This bound is uniform in x and holds for each of the (finite number of) components of $\widehat{\Pi}_2$, therefore the proof is complete. □

3.A.4 Theorem 3.4.6

Before proving impulse response consistency, I show that compositions of the model's autoregressive nonlinear maps are also consistently estimated at any fixed horizon. This means that the "functional moving average" coefficient matrices Γ_j involved in Proposition 3.4.1 can be consistently estimated with $\widehat{\Pi}_1$ and $\widehat{\Pi}_2$.

Lemma 3.A.10. *Under the assumptions of Theorem 3.3.9 and for any fixed integer $j \geq 0$ it holds*

$$\|\widehat{\Gamma}_j - \Gamma_j\|_\infty = o_P(1).$$

Proof. By definition, recall that $\Gamma(L) = \Psi(L)G(L)$ where $\Psi = (I_d - A(L)L)^{-1}$. Since $\Psi(L)$ is an MA(∞) lag polynomial, we have that

$$\Gamma(L) = \left(\sum_{k=0}^{\infty}\Psi_k L^k\right)(G_0 + G_1 L + \ldots + G_p L^p),$$

where $\Psi_0 = I_d$, $\{\Psi_k\}_{k=1}^{\infty}$ are purely real matrices and G_0 is a functional vector that may also contain linear components (i.e. allow linear functions of X_t). This means that Γ_j is a convolution of real and functional matrices,

$$\Gamma_j = \sum_{k=1}^{\min\{j,\,p\}}\Psi_{j-k}G_k.$$

The linear coefficients of $A(L)$ can be consistently estimated by $\widehat{\Pi}_1$ and $\widehat{\Pi}_2$, and thus plug-in estimate $\widehat{\Psi}_j$ is consistent for Ψ_j (Lütkepohl, 2005). Therefore,

$$\begin{aligned}
\|\widehat{\Gamma}_j - \Gamma_j\|_\infty &\leq \sum_{k=1}^{\min\{j,p\}}\left\|\Psi_{j-k}G_k - \widehat{\Psi}_{j-k}\widehat{G}_k\right\|_\infty \\
&\leq \sum_{k=1}^{\min\{j,p\}}\left\|\Psi_{j-k} - \widehat{\Psi}_{j-k}\right\|_\infty\|G_k\|_\infty + \left\|\widehat{\Psi}_{j-k}\right\|_\infty\left\|G_k - \widehat{G}_k\right\|_\infty \\
&\leq \sum_{k=1}^{\min\{j,p\}} o_p(1)C_{G,k} + O_P(1)o_p(1) \\
&= o_p(1),
\end{aligned}$$

where $C_{G,k}$ is a constant and $\|G_k - \widehat{G}_k\|_\infty = o_p(1)$ as a direct consequence of Proposition 3.3.9. □

Note. Since we assume that the model respects either contractivity or stability conditions, the impulse responses must decay (eventually) exponentially fast to zero. This means that by "stitching" bounds appropriately, one should also be able to achieve convergence *uniformly* over $h = 0, 1, \ldots, \infty$.

Recall now that the sample estimate for the relaxed-shock impulse response is

$$\widehat{\overline{\mathrm{IRF}}}_{h,\ell}(\delta) = \Theta_{h,\cdot 1}\delta\, n^{-1}\sum_{t=1}^{n}\rho(\widehat{\epsilon}_{1t}) + \sum_{j=0}^{h}\widehat{V}_{j,\ell}(\delta)$$

where

$$\widehat{V}_{j,\ell}(\delta) = \frac{1}{n-j}\sum_{t=1}^{n-j}\widehat{v}_{j,\ell}(X_{t+j:t};\widehat{\widetilde{\delta}}_t) = \frac{1}{n-j}\sum_{t=1}^{n-j}\left[\widehat{\Gamma}_j\widehat{\gamma}_j(X_{t+j:t};\widehat{\widetilde{\delta}}_t) - \widehat{\Gamma}_j X_{t+j}\right].$$

Therefore, the estimated horizon h impulse response of the ℓth variable is

$$\widehat{\overline{\mathrm{IRF}}}_{h,\ell}(\delta) := \widehat{\Theta}_{h,\ell 1}\delta\, n^{-1}\sum_{t=1}^{n}\rho(\widehat{\epsilon}_{1t}) + \sum_{j=0}^{h}\left[\frac{1}{n-j}\sum_{t=1}^{n-j}\widehat{v}_{j,\ell}(X_{t+j:t};\widehat{\widetilde{\delta}}_t)\right].$$

Lemma 3.A.11. *Under the assumptions of Theorem 3.4.6 , let $x_{j:0} = (x_j,\ldots,x_0) \in \mathcal{X}^j$ and $\varepsilon \in \mathcal{E}_1$ be nonrandom quantities. Let $\widetilde{\delta}$ be the relaxed shock determined by δ, ρ and ε. Then*

(i) $\sup_{x_{j:0},\varepsilon}|\widehat{\gamma}_j(x_{j:0};\widetilde{\delta}) - \gamma_j(x_{j:0};\widetilde{\delta})| = o_P(1)$,

(ii) $\sup_{x_{j:0},\varepsilon}|\widehat{v}_{j,\ell}(x_{j:0};\widetilde{\delta}) - v_{j,\ell}(x_{j:0};\widetilde{\delta})| = o_P(1)$,

for any fixed integers $j \geq 0$ and $\ell \in \{1,\ldots,d\}$.

Proof.

(i) From Proposition 3.4.1, we have that

$$\widehat{\gamma}_j(x_{j:0};\delta) = x_j + \Theta_{j,11}\delta\rho(\varepsilon) + \sum_{k=1}^{j}(\Gamma_{k,11}x_{j-k}(\widetilde{\delta}) - \Gamma_{k,11}x_{j-k}),$$

thus

$$\begin{aligned}|\widehat{\gamma}_j(x_{j:0};\delta) - \gamma_j(x_{j:0};\delta)| &= \left|\sum_{k=1}^{j}\left[(\widehat{\Gamma}_{k,11}x_{j-k}(\widetilde{\delta}) - \widehat{\Gamma}_{k,11}x_{j-k}) - (\Gamma_{k,11}x_{j-k}(\widetilde{\delta}) - \Gamma_{k,11}x_{j-k})\right]\right| \\ &\leq \sum_{k=1}^{j}\left|\widehat{\Gamma}_{k,11}x_{j-k}(\widetilde{\delta}) - \Gamma_{k,11}x_{j-k}(\widetilde{\delta})\right| + \sum_{k=1}^{j}\left|\widehat{\Gamma}_{k,11}x_{j-k} - \Gamma_{k,11}x_{j-k}\right|.\end{aligned}$$

This yields

$$\sup_{x_{j:0},\varepsilon}|\widehat{\gamma}_j(x_{j:0};\widetilde{\delta}) - \gamma_j(x_{j:0};\widetilde{\delta})| \leq 2j\sup_{x\in\mathcal{X}}\left|\widehat{\Gamma}_{k,11}x - \Gamma_{k,11}x\right|.$$

Since j is finite and fixed and the uniform consistency bound of Lemma 3.A.10 holds, a fortiori $\sup_{x\in\mathcal{X}}\left|\widehat{\Gamma}_{k,11}x - \Gamma_{k,11}x\right| = o_P(1)$.

(ii) Similarly to above,

$$\begin{aligned}|\widehat{v}_{j,\ell}(x_{j:0};\widetilde{\delta}) - v_{j,\ell}(x_{j:0};\widetilde{\delta})| &= \left|\left(\widehat{\Gamma}_{j,\ell}\widehat{\gamma}_j(x_{j:0};\widetilde{\delta}) - \Gamma_{j,\ell}\gamma_j(x_{j:0};\widetilde{\delta})\right) - \left(\widehat{\Gamma}_{j,\ell}x_j - \Gamma_{j,\ell}x_j\right)\right| \\ &\leq \|\widehat{\Gamma}_{j,\ell} - \Gamma_{j,\ell}\|_\infty + \|\Gamma_{j,\ell}\|_\infty|\widehat{\gamma}_j(x_{j:0};\delta) - \gamma_j(x_{j:0};\delta)| \\ &\quad + |\widehat{\Gamma}_{j,\ell}x_j - \Gamma_{j,\ell}x_j| \\ &\leq 2\|\widehat{\Gamma}_{j,\ell} - \Gamma_{j,\ell}\|_\infty + C_{\Gamma,j,l}\,|\widehat{\gamma}_j(x_{j:0};\delta) - \gamma_j(x_{j:0};\delta)|,\end{aligned}$$

where we have used that $\gamma_j(x_{j:0}; \widetilde{\delta}) \in \mathcal{X}$ to derive the first term in the second line. In the last line, $C_{\Gamma,j,l}$ is a constant such that

$$\|\Gamma_{j,\ell}\|_\infty \leq \sum_{k=1}^{\min\{j,p\}} \|\Psi_{j-k}\|_\infty \|G_k\|_\infty \leq C_{\Gamma,j,l}.$$

The claim then follows thanks to Lemma 3.A.10 and (i).

□

In what follows, define $\widehat{v}_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t)$ to be a version of $v_{j,\ell}$ that is constructed using coefficient estimates from $\{\widehat{\Pi}_1, \widehat{\Pi}_2\}$ but evaluated on the true innovations ϵ_t.

Proof of Theorem 3.4.6. If we introduce

$$\widetilde{\mathrm{IRF}}_{h,\ell}(\delta)^* := \widehat{\Theta}_{h,\ell 1}\delta\, n^{-1}\sum_{t=1}^{n}\rho(\epsilon_{1t}) + \sum_{j=0}^{h}\left[\frac{1}{n-j}\sum_{t=1}^{n-j}\widehat{v}_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t)\right],$$

then clearly

$$\begin{aligned}\left|\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) - \widetilde{\mathrm{IRF}}_{h,\ell}(\delta)\right| &\leq \left|\widehat{\widetilde{\mathrm{IRF}}}_{h,\ell}(\delta) - \widetilde{\mathrm{IRF}}^*_{h,\ell}(\delta)\right| + \left|\widetilde{\mathrm{IRF}}^*_{h,\ell}(\delta) - \widetilde{\mathrm{IRF}}_{h,\ell}(\delta)\right| \\ &= I + II.\end{aligned}$$

To control II, we can observe

$$\begin{aligned}II &\leq \left|\widehat{\Theta}_{h,\ell 1}\delta\, n^{-1}\sum_{t=1}^{n}\rho(\epsilon_{1t}) - \Theta_{h,\ell 1}\delta\mathbb{E}[\rho(\epsilon_{1t})]\right| \\ &\quad + \sum_{j=0}^{h}\left|\frac{1}{n-j}\sum_{t=1}^{n-j}\widehat{v}_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) - \mathbb{E}[v_{j,\ell}(X_{t+j:t}; \widetilde{\delta})]\right| \\ &\leq \delta\left|\widehat{\Theta}_{h,\ell 1} - \Theta_{h,\ell 1}\right|\left|n^{-1}\sum_{t=1}^{n}\rho(\epsilon_{1t})\right| + \delta\left|\widehat{\Theta}_{h,\ell 1}\right|\left|n^{-1}\sum_{t=1}^{n}\rho(\epsilon_{1t}) - \mathbb{E}[\rho(\epsilon_{1t})]\right| \\ &\quad + \sum_{j=0}^{h}\left|\frac{1}{n-j}\sum_{t=1}^{n-j}\widehat{v}_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) - \mathbb{E}[v_{j,\ell}(X_{t+j:t}; \widetilde{\delta})]\right| \\ &\leq \delta\left|\widehat{\Theta}_{h,\ell 1} - \Theta_{h,\ell 1}\right|\left|n^{-1}\sum_{t=1}^{n}\rho(\epsilon_{1t})\right| + \delta\left|\widehat{\Theta}_{h,\ell 1}\right|\left|n^{-1}\sum_{t=1}^{n}\rho(\epsilon_{1t}) - \mathbb{E}[\rho(\epsilon_{1t})]\right| \\ &\quad + \sum_{j=0}^{h}\left|\frac{1}{n-j}\sum_{t=1}^{n-j}\widehat{v}_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) - v_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t)\right| \\ &\quad + \sum_{j=0}^{h}\left|\frac{1}{n-j}\sum_{t=1}^{n-j}v_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) - \mathbb{E}[v_{j,\ell}(X_{t+j:t}; \widetilde{\delta})]\right|.\end{aligned}$$

The first two terms in the last bound are $o_P(1)$ since $\left|\widehat{\Theta}_{h,\ell 1} - \Theta_{h,\ell 1}\right| = o_P(1)$, as discussed in Lemma 3.A.10, and $n^{-1}\sum_{t=1}^{n}\rho(\epsilon_{1t}) \xrightarrow{p} \mathbb{E}[\rho(\epsilon_{1t})]$ by a WLLN. For the other terms in the last sum above,

we similarly note that

$$\left| \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) - v_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) \right| = o_P(1)$$

from Lemma 3.A.11, while thanks again to a WLLN it holds

$$\left| \frac{1}{n-j} \sum_{t=1}^{n-j} v_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) - \mathbb{E}[v_{j,\ell}(X_{t+j:t}; \widetilde{\delta})] \right| = o_P(1).$$

Since h is fixed finite, this implies that $II = o_P(1)$.

Considering now I, we can write

$$\begin{aligned} I &\leq \delta \left| \widehat{\Theta}_{h,\ell 1} \right| \left| n^{-1} \sum_{t=1}^{n} \rho(\widehat{\epsilon}_{1t}) - \rho(\epsilon_{1t}) \right| + \sum_{j=0}^{h} \left| \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}(X_{t+j:t}; \widehat{\widetilde{\delta}}_t) - \widehat{v}_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) \right| \\ &= I' + I''. \end{aligned}$$

Since by assumption ρ is a bump function, thus continuously differentiable over the range of ϵ_t, by the mean value theorem

$$\left| n^{-1} \sum_{t=1}^{n} \rho(\widehat{\epsilon}_{1t}) - \rho(\epsilon_{1t}) \right| \leq n^{-1} \sum_{t=1}^{n} |\rho'_t| \, |\widehat{\epsilon}_{1t} - \epsilon_{1t}|$$

for a sequence $\{\rho'_t\}_{t=1}^n$ of evaluations of first-order derivative ρ' at values $\overline{\epsilon_t}$ in the interval with endpoint ϵ_t and $\widehat{\epsilon}_t$. One can use $|\rho'_t| \leq C_{\rho'}$ with a finite positive constant $C_{\rho'}$, and by recalling that $\widehat{\epsilon}_{1t} - \epsilon_{1t} = (\Pi_1 - \widehat{\Pi}_1)' W_{1t}$ one thus gets

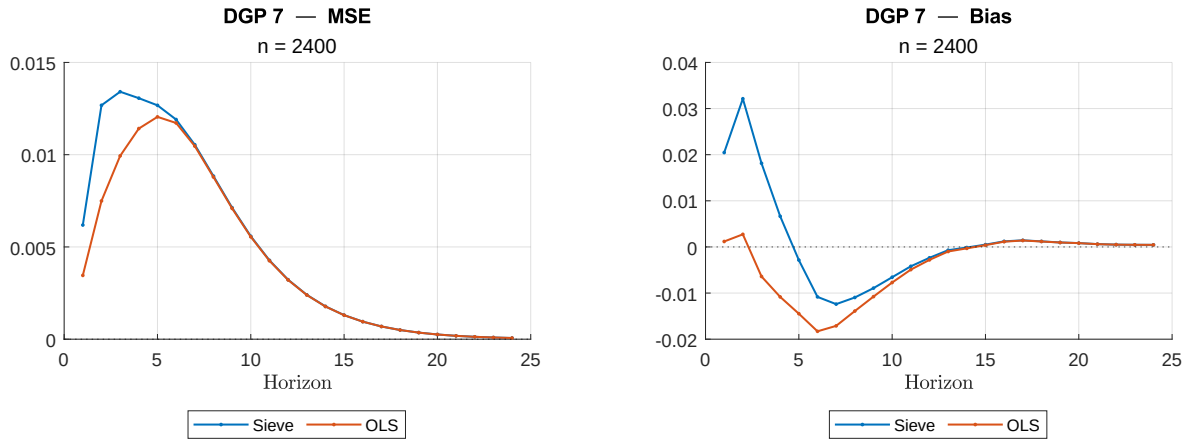
$$\left| n^{-1} \sum_{t=1}^{n} \rho(\widehat{\epsilon}_{1t}) - \rho(\epsilon_{1t}) \right| \leq C_{\rho'} \frac{1}{n} \sum_{t=1}^{n} \left| (\Pi_1 - \widehat{\Pi}_1)' W_{1t} \right| \leq C_{\rho'} \|\Pi_1 - \widehat{\Pi}_1\|_2 \, \frac{1}{n} \sum_{t=1}^{n} \|W_{1t}\|_2 = o_P(1).$$

This proves that term I' is itself $o_P(1)$. Finally, to control I'', we use that by construction estimator $\widehat{\Pi}_2$ is composed of sufficiently regular functional elements i.e. B-spline estimates of order 1 or greater. Thanks again to the mean value theorem

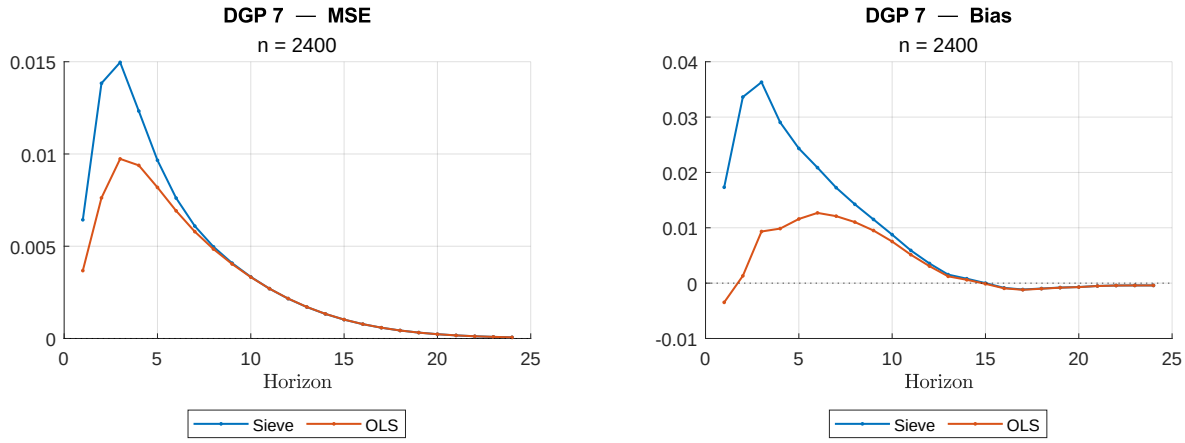
$$\begin{aligned} \left| \frac{1}{n-j} \sum_{t=1}^{n-j} \widehat{v}_{j,\ell}(X_{t+j:t}; \widehat{\widetilde{\delta}}_t) - \widehat{v}_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) \right| &\leq \frac{1}{n-j} \sum_{t=1}^{n-j} \left| \widehat{v}_{j,\ell}(X_{t+j:t}; \widehat{\widetilde{\delta}}_t) - \widehat{v}_{j,\ell}(X_{t+j:t}; \widetilde{\delta}_t) \right| \\ &\leq C_{\widehat{v}',j,\ell} \, \frac{1}{n-j} \sum_{t=1}^{n-j} |\widehat{\epsilon}_{1t} - \epsilon_{1t}| \end{aligned}$$

for any fixed j and some $C_{\widehat{v}',j,\ell} > 0$. This holds since $\widehat{v}_{j,\ell}$ is uniformly continuous by construction. Note that we have assumed that the nonlinear part of Π_2 belongs to a Hölder class with smoothness $s > 1$ (for simplicity, assume here that s is integer, otherwise a similar argument can be made). Then, even though $C_{\widehat{v}',j,\ell}$ depends on the sample, it is bounded above in probability for n sufficiently large. Following the discussion of term I', we deduce that the last line in the display above is $o_p(1)$. As h is finite and independent of n, it follows that also I'' is of order $o_P(1)$. □

3.B Additional Plots



(a) $\delta = +2$



(b) $\delta = -2$

Figure 3.8: Simulation results for DGP $2'$ when considering $\widetilde{\varphi}$ in place of φ.

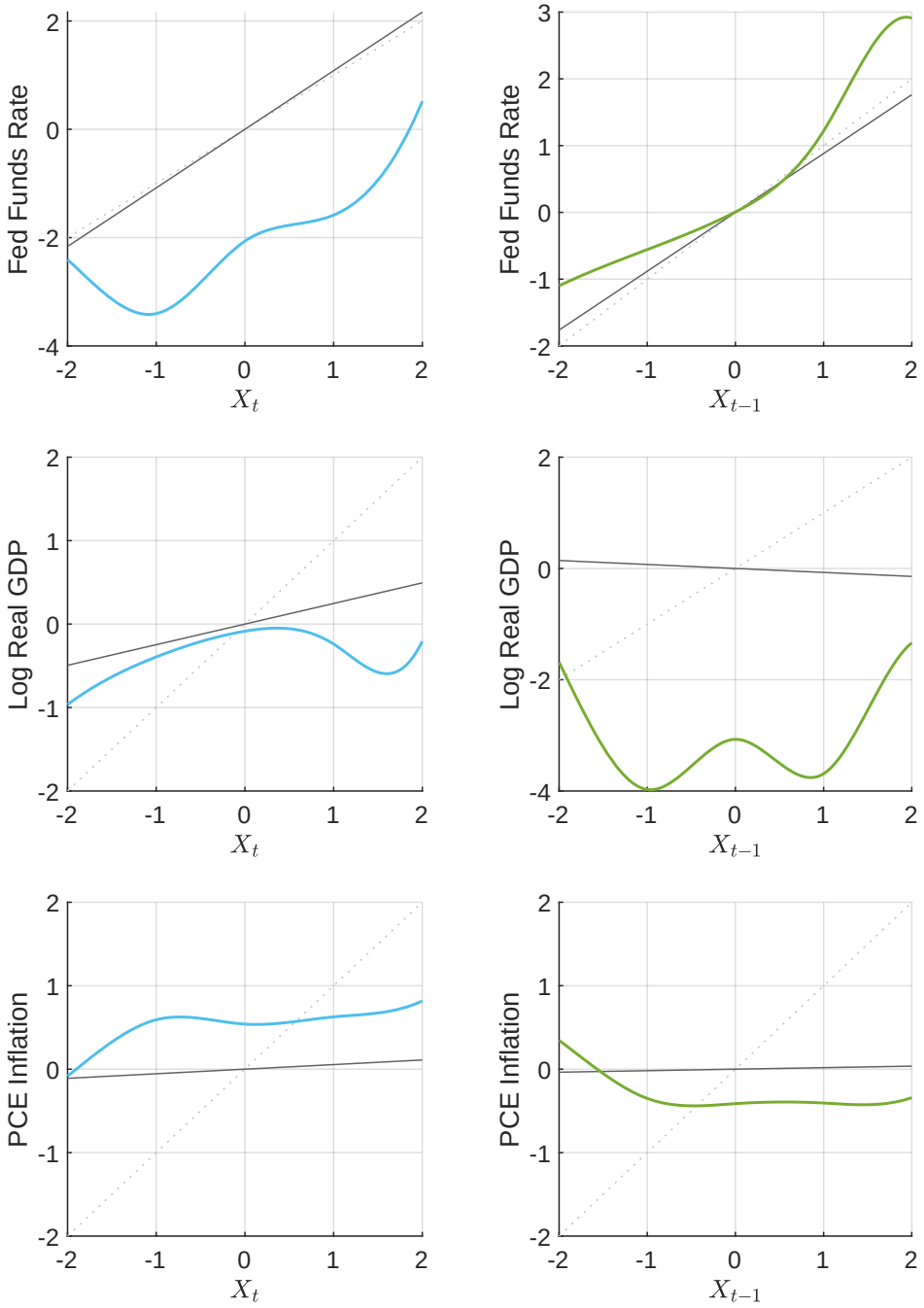
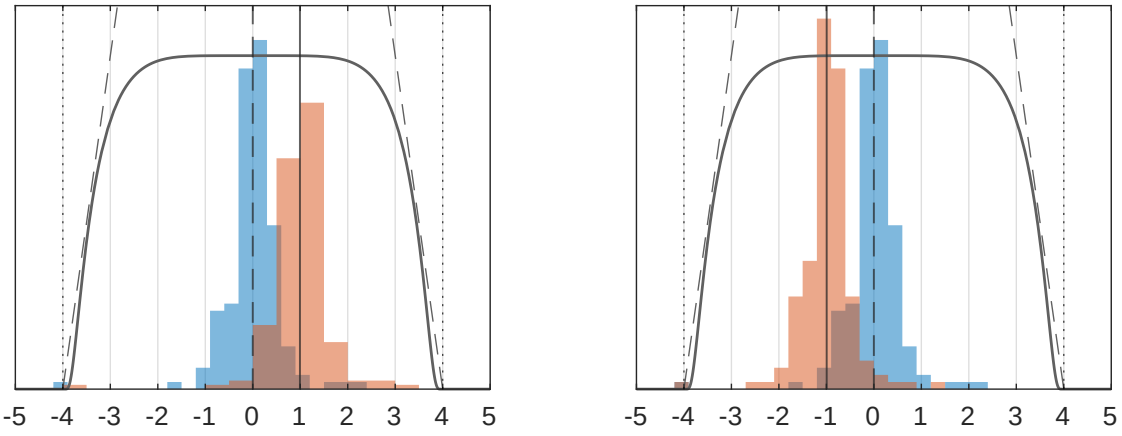
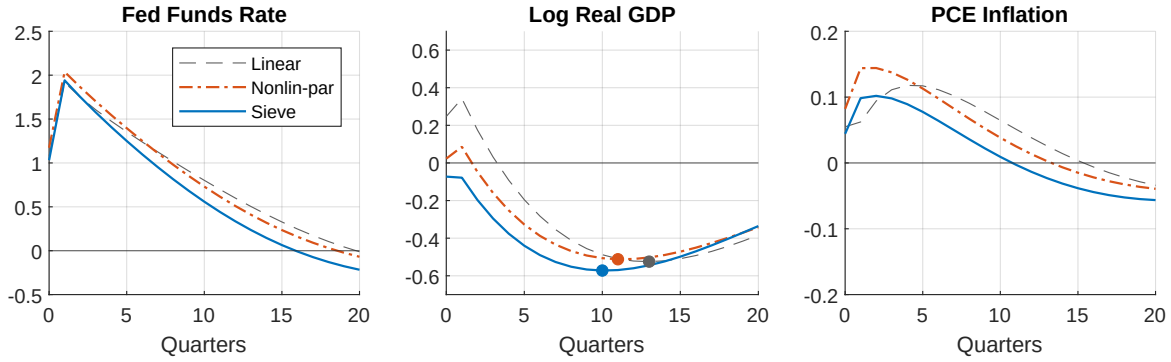


Figure 3.9: Estimated nonlinear regression functions for the narrative U.S. monetary policy variable. Contemporaneous (left side) and one-period lag (right side) effects are shown, linear and nonlinear functions. For comparison, linear VAR coefficients (dark gray) and the identity map (light gray, dashed) are shown as lines.

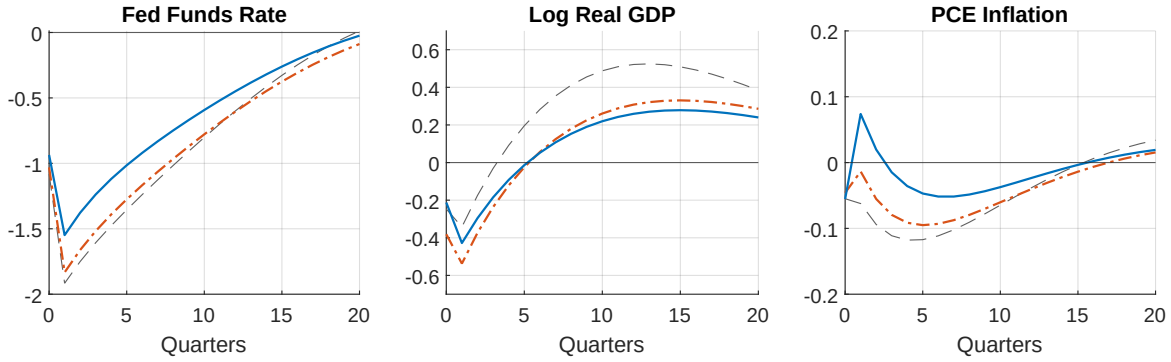


(a) $\delta = +1$ (b) $\delta = -1$

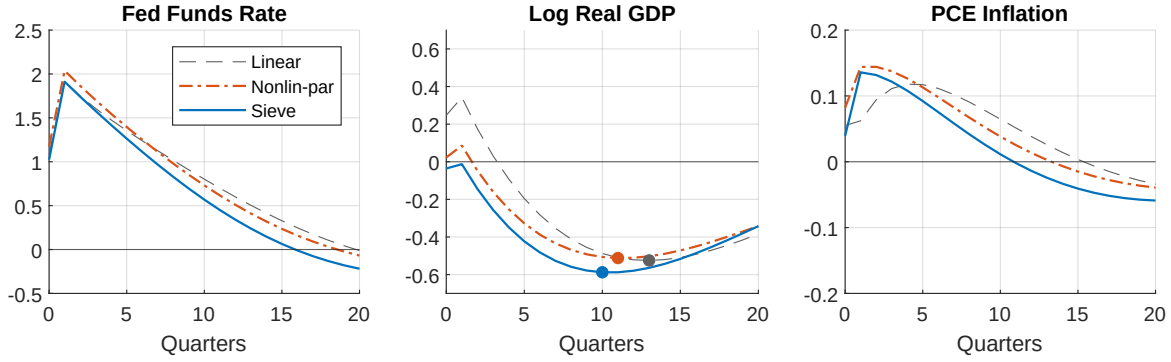
Figure 3.10: Comparison of histograms and shock relaxation function for a positive (left) and negative (right) shock in monetary policy. Original (blue) versus shocked (orange) distribution of the sample realization of ϵ_{1t}. The dashed vertical line is the mean of the original distribution, while the solid vertical line is the mean after the shock.



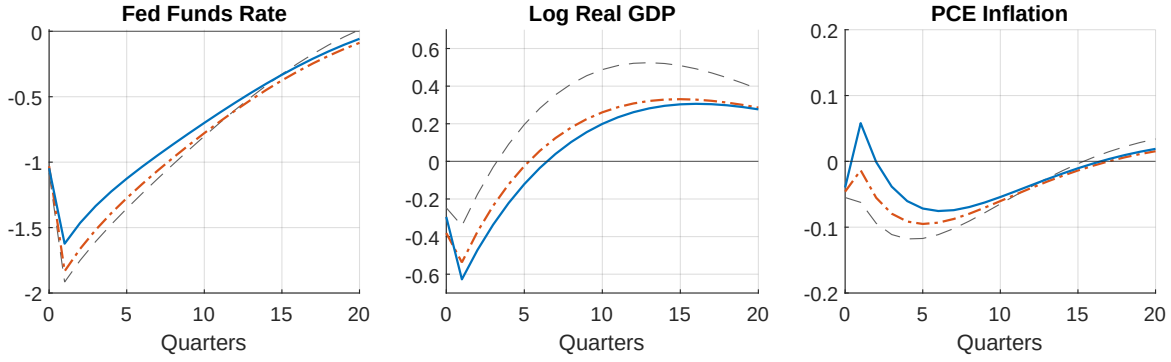
(a) $\delta = +1$, knots at $\{-1, 1\}$



(b) $\delta = -1$, knots at $\{-1, 1\}$



(c) $\delta = +1$, knot at $\{0\}$



(d) $\delta = -1$, knot at $\{0\}$

Figure 3.11: Robustness plots for U.S. monetary policy shock when changing knots compared to those used in Figure 3.6. Note that linear and parametric nonlinear responses do not change.

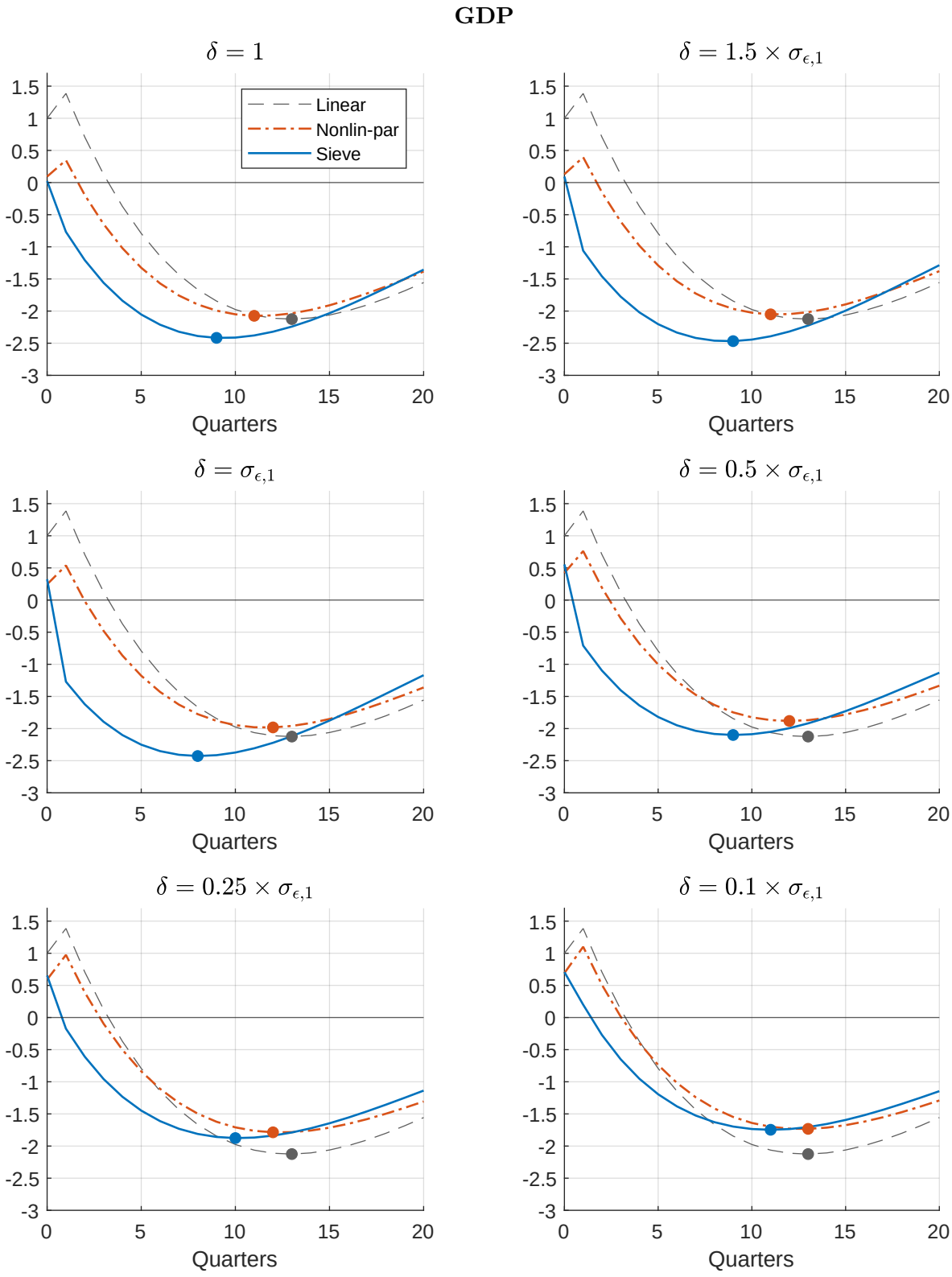


Figure 3.12: Relative changes in the GDP impulse responses function when the size of the shock is reduced from that used in Figure 3.6. The standard deviation of $X_t \equiv \epsilon_{1t}$ is $\sigma_{\epsilon,1} \approx 0.5972$. Linear IRFs are re-scaled such that for all values of δ the linear response at $h = 0$ is one in absolute value. Nonlinear IRFs are re-scaled by δ times the linear response scaling factor.

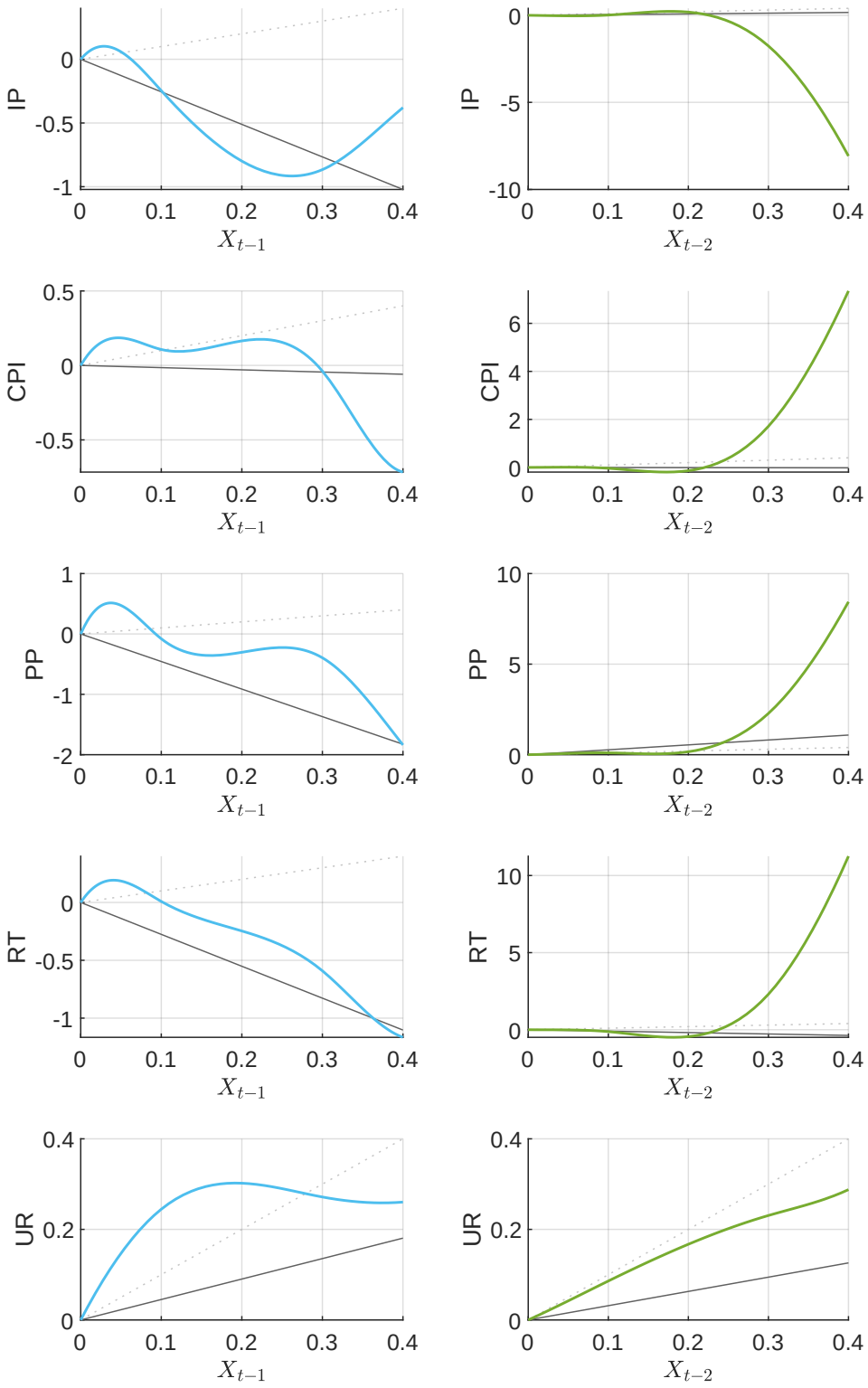
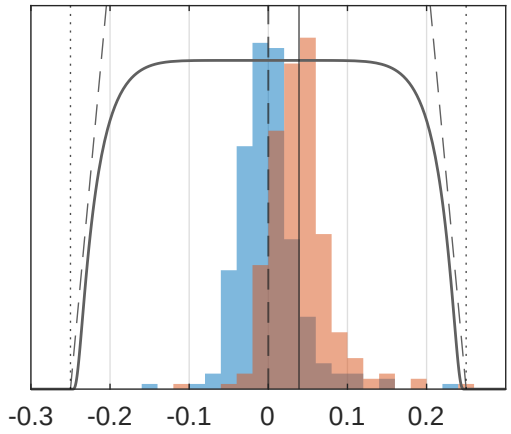
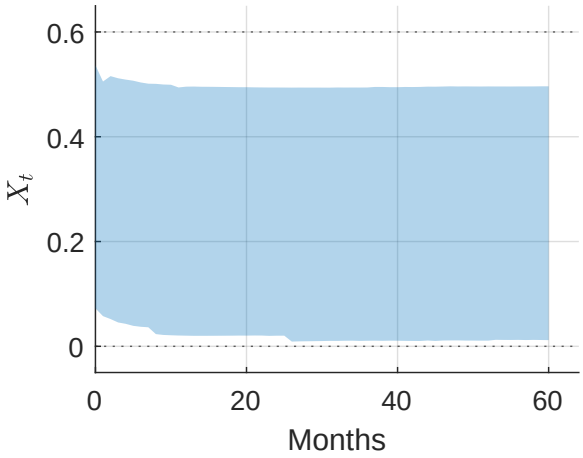


Figure 3.13: Estimated nonlinear regression functions for the 3M3M subjective interest rate uncertainty measure. One-period (left side) and two-period lag (right side) effects are shown, combining linear and nonlinear functions. For comparison, linear VAR coefficients (dark gray) and the identity map (light gray, dashed) are shown as lines.



(a) $\delta = \sigma_\epsilon$



(b) Envelope

Figure 3.14: **[Top]** Histograms and shock relaxation function for a one-standard-deviation shock in interest rate uncertainty. Original (blue) versus shocked (orange) distribution of the sample realization of ϵ_{1t}. The dashed vertical line is the mean of the original distribution, while the solid vertical line is the mean after the shock. **[Bottom]** Envelope (min-max) of shocked paths for one-standard-deviation impulse response.

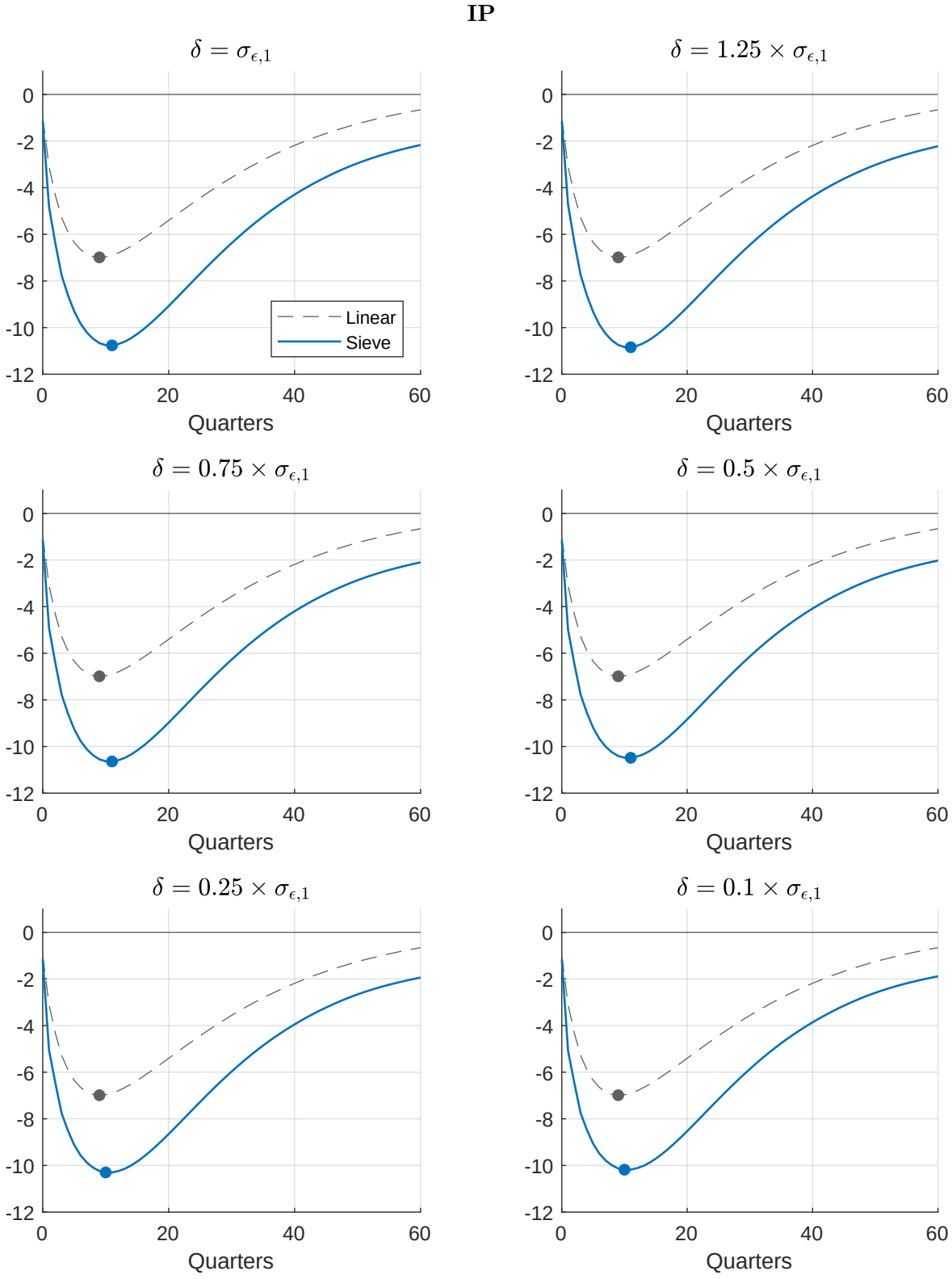


Figure 3.15: Relative changes in the industrial production impulse responses function when the size of the shock is reduced from that used in Figure 3.7. The standard deviation of $\equiv \epsilon_{1t}$ is $\sigma_{\epsilon,1} \approx 0.0389$. Linear IRFs are re-scaled such that for all values of δ the linear response at $h = 0$ is one in absolute value. Nonlinear IRFs are re-scaled by δ times the linear response scaling factor.

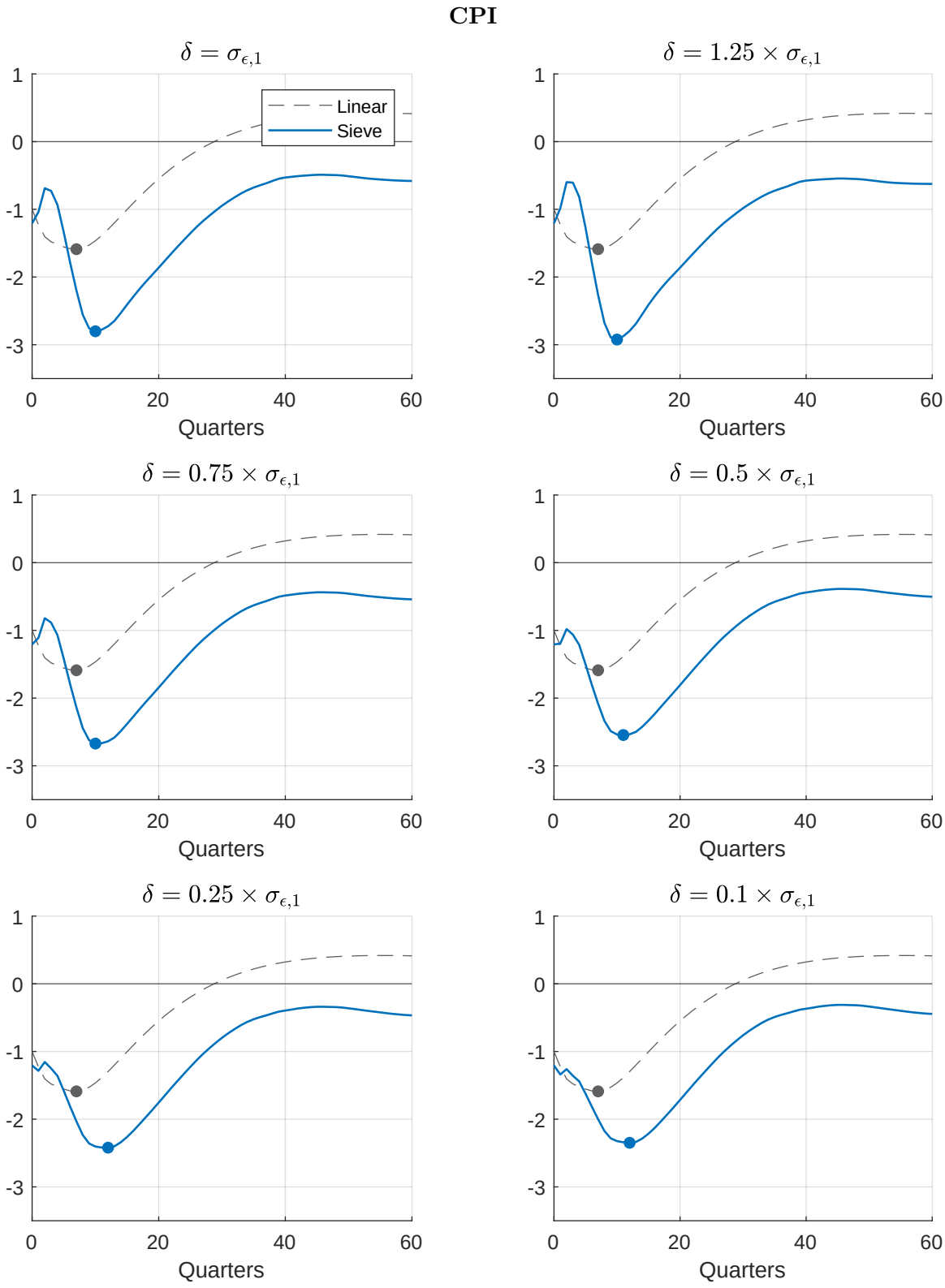


Figure 3.16: Relative changes in the CPI impulse responses function when the size of the shock is reduced from that used in Figure 3.7. The standard deviation of $\equiv \epsilon_{1t}$ is $\sigma_{\epsilon,1} \approx 0.0389$. Linear IRFs are re-scaled such that for all values of δ the linear response at $h = 0$ is one in absolute value. Nonlinear IRFs are re-scaled by δ times the linear response scaling factor.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, 33(1):178–196.

Andreou, E., Ghysels, E., and Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, 31(2):240–251.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, 72:269–342.

Aparicio, D. and de Prado, M. L. (2018). How hard is it to pick the right model? MCS and backtest overfitting. *Algorithmic Finance*, 7(1-2):53–61.

Arcomano, T., Szunyogh, I., Wikner, A., Pathak, J., Hunt, B. R., and Ott, E. (2022). A hybrid approach to atmospheric modeling that combines machine learning with a physics-based numerical model. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002712.

Armesto, M. T., Engemann, K. M., and Owyang, M. T. (2010). Forecasting with Mixed Frequencies. *Federal Reserve Bank of St. Louis Review*, 92(6):521–536.

Arora, S., Little, M. A., and McSharry, P. E. (2013). Nonlinear and nonparametric modeling approaches for probabilistic forecasting of the US gross national product. *Studies in Nonlinear Dynamics and Econometrics*, 17(4):395–420.

Aruoba, S. B., Diebold, F. X., and Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, 27(4):417–427.

Auerbach, A. J. and Gorodnichenko, Y. (2012). Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy*, 4(2):1–27.

Babii, A., Ghysels, E., and Striaukas, J. (2021). Machine Learning Time Series Regressions With an Application to Nowcasting. *Journal of Business & Economic Statistics*, 40(3):1–23.

Babii, A., Ghysels, E., and Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 40(3):1094–1106.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.

Bai, J., Ghysels, E., and Wright, J. H. (2013). State space models and MIDAS regressions. *Econometric Reviews*, 32(7):779–813.

Bai, J. and Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, 25(1):52–60.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Bai Zhang, Miller, D. J., and Yue Wang (2012). Nonlinear system modeling with random matrices: echo state networks revisited. *IEEE Transactions on Neural Networks and Learning Systems*, 23(1):175–182.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Ballarin, G. (2023). Ridge regularized estimation of VAR models for inference. Preprint.

Ballarin, G., Grigoryeva, L., and Ortega, J.-P. (2023). Memory of recurrent networks: Do we compute it right? Preprint.

Bańbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-Casting and the Real-Time Data Flow. In *Handbook of Economic Forecasting*, pages 195–237. Elsevier.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Bańbura, M. and Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, 29(1):133–160.

Bańbura, M. and Rünstler, G. (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting*, 27(2):333–346.

Barnichon, R. and Brownlees, C. (2019). Impulse Response Estimation by Smooth Local Projections. *The Review of Economics and Statistics*, 101(3):522–530.

Bauwens, L., Chevillon, G., and Laurent, S. (2023). We modeled long memory with just one lag! *Journal of Econometrics*, 236(1):105467.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015a). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015b). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018a). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018b). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Bhatia, R. (1997). *Matrix Analysis*. Springer, New York, NY, New York, NY, USA.

Blasques, F., Koopman, S. J., Mallee, M., and Zhang, Z. (2016). Weighted maximum likelihood for dynamic factor analysis and forecasting with mixed frequency data. *Journal of Econometrics*, 193(2):405–417.

Boivin, J. and Ng, S. (2005). Understanding and comparing factor-based forecasts. Technical report, National Bureau of Economic Research.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Borio, C. (2011). Rediscovering the macroeconomic roots of financial stability policy: Journey, challenges, and a way forward. *Annual Review of Financial Economics*, 3(1):87–117.

Borio, C. (2013). The Great Financial Crisis: Setting priorities for new statistics. *Journal of Banking Regulation*, 14(3-4):306–317.

Borio, C. and Lowe, P. W. (2002). Asset prices, financial and monetary stability: Exploring the Nexus. *SSRN Electronic Journal*.

Boubacar Mainassara, Y. and Francq, C. (2011). Estimating structural VARMA models with uncorrelated but non-independent error terms. *Journal of Multivariate Analysis*, 102(3):496–505.

Boyd, S. and Chua, L. (1985). Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161.

Bräuning, F. and Koopman, S. J. (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *International Journal of Forecasting*, 30(3):572–584.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer Science + Business Media.

Brüggemann, R., Jentsch, C., and Trenkler, C. (2016). Inference in VARs with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, 191(1):69–85.

Buehner, M. and Young, P. (2006). A tighter bound for the echo state property. *IEEE Transactions on Neural Networks*, 17(3):820–824.

Buell, B., Cherif, R., Chen, C., Hyeon, Tang, J., and Wendt, N. (2021). Impact of COVID-19: Nowcasting and big data to track economic activity in Sub-Saharan Africa. *IMF Working Paper*, 124:1–61.

Burman, P., Chow, E., and Nolan, D. (1994). A Cross-Validatory Method for Dependent Data. *Biometrika*, 81(2):351–358.

Caggiano, G., Castelnuovo, E., Colombo, V., and Nodari, G. (2015). Estimating Fiscal Multipliers: News From A Non-linear World. *The Economic Journal*, 125(584):746–776.

Caggiano, G., Castelnuovo, E., and Figueres, J. M. (2017). Economic policy uncertainty and unemployment in the United States: A nonlinear approach. *Economics Letters*, 151:31–34.

Caggiano, G., Castelnuovo, E., and Pellegrino, G. (2021). Uncertainty shocks and the great recession: Nonlinearities matter. *Economics Letters*, 198:109669.

Camacho, M. and Pérez-Quirós, G. (2010). Introducing the euro-sting: Short-term indicator of euro area growth. *J. Appl. Econ.*, 25:663–694.

Cao, W., Wang, X., Ming, Z., and Gao, J. (2018). A review on neural networks with random weights. *Neurocomputing*, 275:278–287.

Carriero, A., Galvão, A. B., and Kapetanios, G. (2019). A comprehensive evaluation of macroeconomic forecasting methods. *International Journal of Forecasting*, 35(4):1226–1239.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020). Large sample properties of partitioning-based series estimators. *The Annals of Statistics*, 48(3):1718–1741.

Cattaneo, M. D., Masini, R. P., and Underwood, W. G. (2022). Yurinskii's Coupling for Martingales. *Working Paper*.

Cavaliere, G., Gonçalves, S., and Nielsen, M. Ø. (2022). Bootstrap inference in the presence of bias. *Working Paper*.

Chauvet, M., Senyuz, Z., and Yoldas, E. (2015). What does financial volatility tell us about macroeconomic fluctuations? *Journal of Economic Dynamics and Control*, 52:340–360.

Chen, X. (2007). Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6, pages 5549–5632. Elsevier.

Chen, X. (2013). Penalized Sieve Estimation and Inference of Seminonparametric Dynamic Models: A Selective Review. In Acemoglu, D., Arellano, M., and Dekel, E., editors, *Advances in Economics and Econometrics*, pages 485–544. Cambridge University Press, 1 edition.

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.

Chen, X. and Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84.

Chen, X. and Ghysels, E. (2010). News - good or bad - and its impact on volatility predictions over multiple horizons. *Review of Financial Studies*, 24(1):46–81.

Chen, X., Shao, Q.-M., Wu, W. B., and Xu, L. (2016). Self-normalized Cramér-type moderate deviations under dependence. *The Annals of Statistics*, 44(4):1593–1617.

Chen, X. and Shen, X. (1998). Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica*, 66(2):289.

Chetverikov, D., Santos, A., and Shaikh, A. M. (2018). The Econometrics of Shape Restrictions. *Annual Review of Economics*, 10(1):31–63.

Clements, M. and Galvão, A. (2008). Macroeconomic forecasting with mixed-frequency data: forecasting output growth in the United States. *Journal of Business & Economic Statistics*, 26:546–554.

Clements, M. P. and Galvão, A. (2009). Forecasting US output growth using leading indicators: an appraisal using MIDAS models. *Journal of Applied Econometrics*, 7(7):1187–1206.

Crutchfield, J. P., Ditto, W. L., and Sinha, S. (2010). Introduction to focus issue: Intrinsic and designed computation: Information processing in dynamical systems - beyond the digital hegemony. *Chaos*, 20(3):037101.

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

De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328.

Debortoli, D., Forni, M., Gambetti, L., and Sala, L. (2020). Asymmetric Effects of Monetary Policy Easing and Tightening. *Working Paper*.

Delle Monache, D. and Petrella, I. (2019). Efficient matrix approach for classical inference in state space models. *Economics Letters*, 181:22–27.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1):1–38.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., and Saurous, R. A. (2017). Tensorflow distributions. *arXiv:1711.10604*.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Annals of Statistics*, 46(1):247–279.

Douc, R., Moulines, E., Olsson, J., and Van Handel, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 39(1):474–513.

Doucet, A., de Freitas, N., and Gordon, N. J., editors (2001). *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer.

Doya, K. (1992). Bifurcations in the learning of recurrent neural networks. In *Proceedings of IEEE International Symposium on Circuits and Systems*, volume 6, pages 2777–2780.

Doz, C., Giannone, D., and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164(1):188–205.

Fan, J. and Yao, Q. (2003). *Nonlinear time series: nonparametric and parametric methods*, volume 20. Springer.

Farkas, I., Bosak, R., and Gergel, P. (2016). Computational analysis of memory capacity in echo state networks. *Neural Networks*, 83:109–120.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213.

Feng, B. Q. (2003). Equivalence constants for certain matrix norms. *Linear Algebra and its Applications*, 374:247–253.

Ferrara, L., Marsilli, C., and Ortega, J.-P. (2014). Forecasting growth during the Great Recession: is financial volatility the missing ingredient? *Economic Modelling*, 36:44–50.

Forni, M., Gambetti, L., Maffei-Faccioli, N., and Sala, L. (2023a). Nonlinear transmission of financial shocks: Some new evidence. *Journal of Money, Credit and Banking*.

Forni, M., Gambetti, L., and Sala, L. (2023b). Asymmetric effects of news through uncertainty. *Macroeconomic Dynamics*, pages 1–25.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100(471):830–840.

Foroni, C. and Marcellino, M. (2011). A comparison of mixed approaches for modelling euro area macroeconomic variables. Technical report, EUI.

Frale, C., Marcellino, M., Mazzi, G. L., and Proietti, T. (2011). EUROMIND: a monthly indicator of the euro area economic conditions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174:439–470.

Francis, N., Ghysels, E., and Owyang, M. T. (2011). The low-frequency impact of daily monetary policy shocks. Technical report, Federal Reserve Bank of St. Louis.

Freyberger, J. and Reeves, B. (2018). Inference under Shape Restrictions. *Working Paper*.

Fu, W. and Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Fuleky, P., editor (2020a). *Macroeconomic Forecasting in the Era of Big Data*. Springer International Publishing.

Fuleky, P., editor (2020b). *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*, volume 52 of *Advanced Studies in Theoretical and Applied Econometrics*. Springer International Publishing, Cham.

Gagliardini, P., Ghysels, E., and Rubin, M. (2017). Indirect inference estimation of mixed frequency stochastic volatility state space models using MIDAS regressions and ARCH models. *Journal of Financial Econometrics*, 15(4):509–560.

Galvão, A. B. (2013). Changes in predictive ability with mixed frequency data. *International Journal of Forecasting*, 29(3):395–410.

Galvão, A. B. and Marcellino, M. (2010). Endogenous monetary policy regimes and the great moderation. Technical report, EUI.

Gambetti, L., Maffei-Faccioli, N., and Zoi, S. (2022). Bad News, Good News: Coverage and Response Asymmetries. *Working Paper*.

Gao, J. (2007). *Nonlinear Time Series: Semiparametric and Nonparametric Methods*. Chapman and Hall/CRC.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent variables in socio-economic models*.

Ghosh, S., Khare, K., and Michailidis, G. (2019). High-Dimensional Posterior Consistency in Bayesian Vector Autoregressive Models. *Journal of the American Statistical Association*, 114(526):735–748.

Ghysels, E. (2016). Macroeconomics and the reality of mixed frequency data. *Journal of Econometrics*, 193(2):294–314.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models. Technical Report 919, UCLA: Finance.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26(1):53–90.

Ghysels, E. and Wright, J. H. (2009). Forecasting professional forecasters. *Journal of Business & Economic Statistics*, 27(4):504–516.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior Selection for Vector Autoregressions. *The Review of Economics and Statistics*, 97(2):436–451.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Godsill, S. J., Doucet, A., and West, M. (2004). Monte Carl smoothing for nonlinear time series. *Journal of the American Statistical Association*, 99(465):156–168.

Gonçalves, S., Herrera, A. M., Kilian, L., and Pesavento, E. (2021). Impulse response analysis for structural dynamic models with nonlinear regressors. *Journal of Econometrics*, 225(1):107–130.

Gonon, L., Grigoryeva, L., and Ortega, J.-P. (2020a). Memory and forecasting capacities of nonlinear recurrent networks. *Physica D*, 414(132721):1–13.

Gonon, L., Grigoryeva, L., and Ortega, J.-P. (2020b). Risk bounds for reservoir computing. *Journal of Machine Learning Research*, 21(240):1–61.

Gonon, L., Grigoryeva, L., and Ortega, J.-P. (2023a). Approximation error estimates for random neural networks and reservoir systems. *The Annals of Applied Probability*, 33(1):28–69.

Gonon, L., Grigoryeva, L., and Ortega, J. P. (2023b). Infinite-dimensional reservoir computing. *Arxiv preprint*.

Gonon, L. and Ortega, J.-P. (2020). Reservoir computing universality with stochastic inputs. *IEEE Transactions on Neural Networks and Learning Systems*, 31(1):100–112.

Gonon, L. and Ortega, J.-P. (2021). Fading memory echo state networks are universal. *Neural Networks*, 138:10–13.

Goudarzi, A., Marzen, S., Banda, P., Feldman, G., Lakin, M. R., Teuscher, C., and Stefanovic, D. (2016). Memory and information processing in recurrent neural networks. Technical report, Portland State University.

Goulet Coulombe, P. (2023). Time-Varying Parameters as Ridge Regressions. *Working Paper*.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *J. Appl. Econ.*, 37(5):920–964.

Gourieroux, C. and Jasiak, J. (2005). Nonlinear Innovations and Impulse Responses with Application to VaR Sensitivity. *Annales d'Économie et de Statistique*, pages 1–31.

Gourieroux, C. and Lee, Q. (2023). Nonlinear impulse response functions and local projections. *Working Paper*.

Gramlich, D., Miller, G. L., Oet, M. V., and Ong, S. J. (2010). Early warning systems for systemic banking risk: Critical review and modeling implications. *Banks and Bank Systems*, 5(2):199–211.

Grigoryeva, L., Hart, A. G., and Ortega, J.-P. (2021). Learning strange attractors with reservoir systems. *arXiv preprint arXiv:2108.05024*.

Grigoryeva, L., Henriques, J., Larger, L., and Ortega, J.-P. (2015). Optimal nonlinear information processing capacity in delay-based reservoir computers. *Scientific Reports*, 5(12858):1–11.

Grigoryeva, L., Henriques, J., Larger, L., and Ortega, J.-P. (2016). Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals. *Neural Computation*, 28:1411–1451.

Grigoryeva, L. and Ortega, J.-P. (2018a). Echo state networks are universal. *Neural Networks*, 108:495–508.

Grigoryeva, L. and Ortega, J.-P. (2018b). Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems. *Journal of Machine Learning Research*, 19(24):1–40.

Grigoryeva, L. and Ortega, J.-P. (2019). Differentiable reservoir computing. *Journal of Machine Learning Research*, 20(179):1–62.

Grigoryeva, L. and Ortega, J.-P. (2021). Dimension reduction in recurrent networks by canonicalization. *Journal of Geometric Mechanics*, 13(4):647–677.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617.

Hamilton, J. D. (1994a). State-space models. *Handbook of Econometrics*, 4:3039–3080.

Hamilton, J. D. (1994b). *Time Series Analysis*. Princeton University Press.

Hansen, B. E. (2016a). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132.

Hansen, B. E. (2016b). Stein Combination Shrinkage for Vector Autoregressions. *Working Paper*.

Hansen, P. R., Huang, Z., and Shek, H. H. (2011). Realized GARCH: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, 27(6):877–906.

Härdle, W., Lütkepohl, H., and Chen, R. (1997). A Review of Nonparametric Time Series Analysis. *International Statistical Review*, 65(1):49–72.

Hart, A. G., Hook, J. L., and Dawes, J. H. P. (2021). Echo State Networks trained by Tikhonov least squares are L2(μ) approximators of ergodic dynamical systems. *Physica D: Nonlinear Phenomena*, 421:132882.

Harvey, A. C., Koopman, S. J., and Penzer, J. (1998). Messy time series: A unified approach. *Advances in Econometrics*, 13:103–144.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291.

Hastie, T. (2020). Ridge Regularization: An Essential Concept in Data Science. *Technometrics*, 62(4):426–433.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022a). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022b). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition.

Hatzius, J., Hooper, P., Mishkin, F., Schoenholtz, K., and Watson, M. (2010). Financial conditions indexes: A fresh look after the financial crisis. Technical report, National Bureau of Economic Research.

Hausman, J. A. (1983). Chapter 7: Specification and estimation of simultaneous equation models. In *Handbook of Econometrics*, volume 1, pages 391–448. Elsevier.

Heaps, S. E. (2023). Enforcing Stationarity through the Prior in Vector Autoregressions. *Journal of Computational and Graphical Statistics*, 32(1):74–83.

Hihi, S. and Bengio, Y. (1995). Hierarchical recurrent neural networks for long-term dependencies. *Advances in neural information processing systems*, 8.

Hindrayanto, I., Koopman, S. J., and de Winter, J. (2016). Forecasting and nowcasting economic growth in the euro area using factor models. *International Journal of Forecasting*, 32(4):1284–1305.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82.

Hong, H. and Yogo, M. (2012). What does futures market interest tell us about the macroeconomy and asset prices? *Journal of Financial Economics*, 105(3):473–490.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, second edition.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer, New York, NY, USA.

Horowitz, J. L. and Lee, S. (2017). Nonparametric estimation and inference under shape restrictions. *Journal of Econometrics*, 201(1):108–126.

Huang, Y., Chen, X., and Wu, W. B. (2014). Recursive Nonparametric Estimation for Time Series. *IEEE Transactions on Information Theory*, 60(2):1301–1312.

Huber, F. and Koop, G. (2023). Subspace shrinkage in conjugate Bayesian vector autoregressions. *Journal of Applied Econometrics*, 38(4):556–576.

Huber, F., Koop, G., Onorante, L., Pfarrhofer, M., and Schreiner, J. (2021). Nowcasting in a pandemic using non-parametric mixed frequency VARs. *ECB Working Paper Series*, 2510:1–40.

Ingenito, R. and Trehan, B. (1996). Using monthly data to predict quarterly output. *Econometric Reviews*, pages 3–11.

Inoue, A., Jin, L., and Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics*, 196(1):55–67.

Inoue, A. and Kilian, L. (2008). How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation. *Journal of the American Statistical Association*, 103(482):511–522.

Ishwaran, H. and Rao, J. (2014). Geometry and Properties of Generalized Ridge Regression in High Dimensions. In Ahmed, S., editor, *Contemporary Mathematics*, volume 622, pages 81–93. American Mathematical Society, Providence, Rhode Island.

Istrefi, K. and Mouabbi, S. (2018). Subjective interest rate uncertainty and the macroeconomy: A cross-country analysis. *Journal of International Money and Finance*, 88:296–313.

Jaeger, H. (2010). The 'echo state' approach to analysing and training recurrent neural networks with an erratum note. Technical report, German National Research Center for Information Technology.

Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80.

Jardet, C. and Meunier, B. (2022). Nowcasting world GDP growth with high-frequency data. *Journal of Forecasting*, 41(6):1181–1200.

Jordà, Ò. (2005). Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review*, 95(1):161–182.

Jungbacker, B. and Koopman, S. J. (2015). Likelihood-based dynamic factor analysis for measurement and forecasting. Technical report, Tinbergen Institute Discussion Paper.

Kadiyala, K. R. and Karlsson, S. (1997). Numerical methods for estimation and inference in bayesian var-models. *Journal of Applied Econometrics*, 12(2):99–132.

Kanazawa, N. (2020). Radial basis functions neural networks for nonlinear time series analysis and time-varying effects of supply shocks. *Journal of Macroeconomics*, 64:103210.

Kang, B. (2021). Inference In Nonparametric Series Estimation with Specification Searches for the Number of Series Terms. *Econometric Theory*, 37(2):311–345.

Kang, J. and Kwon, K. Y. (2020). Can commodity futures risk factors predict economic growth? *Journal of Futures Markets*, 40(12):1825–1860.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., Chopin, N., and Others (2015). On particle methods for parameter estimation in state-space models. *Statistical science*, 30(3):328–351.

Kilian, L. and Kim, Y. J. (2011). How Reliable Are Local Projection Estimators of Impulse Responses? *Review of Economics and Statistics*, 93(4):1460–1466.

Kilian, L. and Lütkepohl, H. (2017a). *Structural vector autoregressive analysis*. Cambridge University Press.

Kilian, L. and Lütkepohl, H. (2017b). *Structural Vector Autoregressive Analysis*. Themes in Modern Econometrics. Cambridge University Press, Cambridge.

Kilian, L. and Vega, C. (2011). Do energy prices respond to us macroeconomic news? a test of the hypothesis of predetermined energy prices. *Review of Economics and Statistics*, 93(2):660–671.

Kilian, L. and Vigfusson, R. J. (2011). Are the responses of the us economy asymmetric in energy price increases and decreases? *Quantitative Economics*, 2(3):419–453.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

Kock, A. B., Medeiros, M., and Vasconcelos, G. (2020). Penalized Time Series Regression. In Fuleky, P., editor, *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*, Advanced Studies in Theoretical and Applied Econometrics, pages 193–228. Springer International Publishing, Cham.

Koop, G., Pesaran, M. H., and Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics*, 74(1):119–147.

Kostrov, A. (2021). *Essays on the use of MIDAS regressions in banking and finance*. PhD thesis, Universität St. Gallen.

Lanne, M. and Nyberg, H. (2023). Nonparametric Impulse Response Analysis in Changing Macroeconomic Conditions. *Working Paper*.

Legenstein, R. and Maass, W. (2007). What makes a dynamical system computationally powerful? In Haykin, S., editor, *New directions in statistical signal processing: from systems to brain*. MIT Press, Cambridge, MA.

LeGland, F. and Mevel, L. (1997). Recursive estimation in hidden Markov models. In *Proceedings of the 36th IEEE Conference on Decision and Control*, volume 4, pages 3468–3473.

Leippold, M. and Yang, H. (2019). Particle filtering, learning, and smoothing for mixed-frequency state-space models. *Econometrics and Statistics*, 12:25–41.

Li, D., Plagborg-Møller, M., and Wolf, C. K. (2023). Local Projections vs. VARs: Lessons From Thousands of DGPs. *Working Paper*.

Li, J. and Liao, Z. (2020). Uniform nonparametric inference for time series. *Journal of Econometrics*, page 14.

Li, Q. and Racine, J. S. (2009). *Nonparametric econometric methods*. Emerald Group Publishing.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions five years of experience. Working Papers 274, Federal Reserve Bank of Minneapolis.

Liu, S. and Dobriban, E. (2020). Ridge Regression: Structure, Cross-Validation, and Sketching. In *International Conference on Learning Representations*.

Lukoševičius, M. (2012). A Practical Guide to Applying Echo State Networks. In Montavon, G., Orr, G. B., and Müller, K.-R., editors, *Neural Networks: Tricks of the Trade: Second Edition*, Lecture Notes in Computer Science, pages 659–686. Springer, Berlin, Heidelberg.

Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Lütkepohl, H. (1990). Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models. *The Review of Economics and Statistics*, 72(1):116–125.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. New York : Springer, Berlin.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation*, 14:2531–2560.

Maillard, O.-A. and Munos, R. (2012). Linear regression with random projections. *Journal of Machine Learning Research*, 13(89):2735–2772.

Manjunath, G. and Jaeger, H. (2013). Echo state property linked to an input: exploring a fundamental characteristic of recurrent neural networks. *Neural Computation*, 25(3):671–696.

Manjunath, G. and Ortega, J.-P. (2023). Transport in reservoir computing. *Physica D: Nonlinear Phenomena*, 449:133744.

Marcellino, M. and Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics*, 72(4):518–550.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2):499–526.

Mariano, R. S. and Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of applied Econometrics*, 18(4):427–443.

Marsilli, C. (2014). *Mixed-Frequency Modeling and Economic Forecasting*. PhD thesis, Université de Franche-Comté.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

McCracken, M. W. and Ng, S. (2020). FRED-QD: A quarterly database for macroeconomic research. Technical report, Federal Reserve Bank of St. Louis.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., and Zilberman, E. (2021). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics*, 39(1):98–119.

Mikusheva, A. (2007). Uniform Inference in Autoregressive Models. *Econometrica*, 75(5):1411–1452.

Mikusheva, A. (2012). One-Dimensional Inference in Autoregressive Models With the Potential Presence of a Unit Root. *Econometrica*, 80(1):173–212.

Monteforte, L. and Moretti, G. (2012). Real-time forecasts of inflation: The role of financial variables. *Journal of Forecasting*, 32(1):51–61.

Morley, J. (2015). Macro-finance linkages. *Journal of Economic Surveys*, 30(4):698–711.

Movahedifar, M. and Dickhaus, T. (2023). On the closed-loop Volterra method for analyzing time series. *Working Paper*.

Nowzohour, L. and Stracca, L. (2020). More than a feeling: Confidence, uncertainty, and macroeconomic fluctuations. *Journal of Economic Surveys*, 34(4):691–726.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Paranhos, L. (2021). Predicting inflation with neural networks.

Park, J. Y. and Phillips, P. C. (1988). Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory*, 4(3):468–497.

Park, J. Y. and Phillips, P. C. (1989). Statistical inference in regressions with integrated processes: Part 2. *Econometric Theory*, pages 95–131.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, volume 28, pages 1310–1318. PMLR.

Pathak, J., Hunt, B., Girvan, M., Lu, Z., and Ott, E. (2018). Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical Review Letters*, 120(2):24102.

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., and Ott, E. (2017). Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. *Chaos*, 27(12).

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. (2021). Uniform Consistency of Cross-Validation Estimators for High-Dimensional Ridge Regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 3178–3186. PMLR.

Pellegrino, G. (2021). Uncertainty and monetary policy in the US: A journey into nonlinear territory. *Economic Inquiry*, 59(3):1106–1128.

Pesavento, E. and Rossi, B. (2006). Small-sample confidence intervals for multivariate impulse response functions at long horizons. *Journal of Applied Econometrics*, 21(8):1135–1155.

Pettenuzzo, D., Timmermann, A., and Valkanov, R. (2016). A MIDAS approach to modeling first and second moment dynamics. *Journal of Econometrics*, 193(2):315–334.

Phillips, P. C. B. (1988). Regression Theory for Near-Integrated Time Series. *Econometrica*, 56(5):1021–1043.

Plagborg-Møller, M. (2016). *Essays in Macroeconometrics*. PhD thesis, Harvard University.

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

Potter, S. M. (2000). Nonlinear impulse response functions. *Journal of Economic Dynamics and Control*, 24(10):1425–1446.

Pratt, J. W. (1961). Length of Confidence Intervals. *Journal of the American Statistical Association*, 56(295):549–567.

Proietti, T. and Moauro, F. (2006). Dynamic factor analysis with non-linear temporal aggregation constraints. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(2):281–300.

Qin, D., van Huellen, S., Wang, Q. C., and Moraitis, T. (2022). Algorithmic modelling of financial conditions for macro predictive purposes: Pilot application to USA data. *Econometrics*, 10(2):22.

Quaedvlieg, R. (2021). Multi-horizon forecast comparison. *Journal of Business & Economic Statistics*, 39(1):40–53.

Ramey, V. A. and Zubairy, S. (2018). Government spending multipliers in good times and in bad: evidence from us historical data. *Journal of political economy*, 126(2):850–901.

Rio, E. (2017). *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80 of *Probability Theory and Stochastic Modelling*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rodan, A. and Tino, P. (2011). Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, 22(1):131–44.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American economic review*, 94(4):1055–1084.

Salehinejad, H., Baarbe, J., Sankar, S., Barfett, J., Colak, E., and Valaee, S. (2017). Recent advances in recurrent neural networks.

Sargent, T. J., Sims, C. A., and Others (1977). Business cycle modeling without pretending to have too much a priori economic theory. *New methods in business cycle research*, 1:145–168.

Schorfheide, F., Song, D., and Yaron, A. (2018). Identifying long-run risks: A Bayesian mixed-frequency approach. *Econometrica*, 86(2):617–654.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1):1–48.

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica: Journal of the Econometric Society*, pages 113–144.

Sirotko-Sibirskaya, N., Franz, M. O., and Dickhaus, T. (2020). Volterra bootstrap: Resampling higher-order statistics for strictly stationary univariate time series. *Working Paper*.

Smeekes, S. and Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting*, 34(3):408–430.

Stapleford, T. A. (2021). Revisiting the Past?: Big Data, Interwar Statistical Economics, and the Long History of Statistical Inference in the United States. *History of Political Economy*, 53(S1):175–203.

Stephenson, W., Frangella, Z., Udell, M., and Broderick, T. (2021). Can we globally optimize cross-validation loss? Quasiconvexity in ridge regression. In *Advances in Neural Information Processing Systems*, volume 34, pages 24352–24364. Curran Associates, Inc.

Stock, J. H. and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1):11–30.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2006). Forecasting with Many Predictors. In Elliot, G., Granger, C. W., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1. Elsevier edition.

Stock, J. H. and Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics*, volume 2, pages 415–525. Elsevier.

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., Numata, H., Nakano, D., and Hirose, A. (2019). Recent advances in physical reservoir computing: A review. *Neural Networks*, 115:100–123.

Tenreyro, S. and Thwaites, G. (2016). Pushing on a string: US monetary policy is less powerful in recessions. *American Economic Journal: Macroeconomics*, 8(4):43–74.

Teräsvirta, T., Tjøstheim, D., and Granger, C. W. J. (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press.

Tong, H. (1990). *Non-linear Time Series: a Dynamical System Approach*. Oxford University Press.

Tropp, J. A. (2012). User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Tsay, R. S. and Chen, R. (2018). *Nonlinear Time Series Analysis*, volume 891. John Wiley & Sons.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York ; London.

van der Maaten, L. (2009). Learning a parametric embedding by preserving local structure. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 384–391. PMLR.

van Huellen, S., Qin, D., Lu, S., Wang, H., Wang, Q. C., and Moraitis, T. (2020). Modelling opportunity cost effects in money demand due to openness. *International Journal of Finance & Economics*, 27(1):697–744.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*, pages 1–11.

Wainrib, G. and Galtier, M. N. (2016). A local echo state property through the largest Lyapunov exponent. *Neural Networks*, 76:39–45.

Watson, M. W. and Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23(3):385–400.

Whitney, H. (1972). *Complex Analytic Varieties*, volume 131. Addison-Wesley Reading.

Wikner, A., Pathak, J., Hunt, B. R., Szunyogh, I., Girvan, M., and Ott, E. (2021). Using data assimilation to train a hybrid forecast system that combines machine-learning and knowledge-based components. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(5):53114.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and its Interface*, 4(2):207–226.

Wu, W. B., Huang, Y., and Huang, Y. (2010). Kernel estimation for time series: An asymptotic theory. *Stochastic Processes and their Applications*, 120(12):2412–2431.

Yildiz, I. B., Jaeger, H., and Kiebel, S. J. (2012). Re-visiting the echo state property. *Neural Networks*, 35:1–9.

Author's Declaration

Eidesstattliche Versicherung gemäß §8 Absatz 2 Buchstabe a) der Promotionsordnung der Universität Mannheim zur Erlangung des Doktorgrades der Volkswirtschaftslehre (Dr. rer. pol.)

1. Bei der eingereichten Dissertation mit dem Titel „Essays in Time Series Econometrics and Machine Learning“ handelt es sich um mein eigenständig erstelltes Werk, das den Regeln guter wissenschaftlicher Praxis entspricht.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtliche und nicht wörtliche Zitate aus anderen Werken als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslandsals Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärung bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Die eingereichten Dissertationsexemplare sowie der Datenträger gehen in das Eigentum der Universität über.

Mannheim, 2024

Giovanni Ballarin

Curriculum Vitae

2018 – 2024	University of Mannheim Ph.D. in Economics
2017 – 2018	University of Konstanz M.Sc. in Economics
2016 – 2018	University of Tor Vergata M.Sc. in Economics
2013 – 2016	University of Milano–Bicocca B.Sc. in Mathematics