CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

DISSERTATION

Essays on Data-driven, Non-parametric Modelling of Time-Series

Author: Mgr. Luboš Hanus

Supervisor: Mgr. Lukáš Vácha, Ph.D.

Year of defense: 2024

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, January 17, 2024

Mgr. Luboš Hanus

Abstract

This thesis consists of four contributions to the literature on data-driven and non-parametric modelling of time series. In the first paper, we study the synchronisation of business cycles and propose a multivariate co-movement measure based on time-frequency cohesion. We suggest that economic integration may lead to increased co-movement of business cycles, which may reflect the benefits of convergence and coordination of economic policies. The second paper presents a new methodology for identifying persistence in macroeconomic variables. Using time-varying frequency response functions, we identify heterogeneous persistence effects in US macroeconomic variables. The third and fourth papers propose data-driven techniques for probabilistic forecasting of time series using deep learning. We introduce a multi-output neural network that selects the most appropriate distribution for the data. The distributional neural network is valuable for modelling data with non-linear, non-Gaussian and asymmetric structures. The third paper demonstrates the usefulness of the method by estimating information-rich macroeconomic fan charts and distributional forecasts of asset returns. In the last paper, we present the distributional neural network to obtain the probability distribution of electricity price forecasts. We forecast hourly day-ahead data for German electricity and make comparisons with state-of-the-art models in the electricity price forecasting literature.

JEL Classification	C11, E47, E32, C53
Keywords	Non-parametric, Data-driven, Time-series, Time-varying, Neural networks, Frequency domain, Business cycles, Forecasting, Distribution
Title	Essays on Data-driven, Non-parametric Modelling of Time-Series

Author's e-mail	`lubos.hanus@gmail.com`
Supervisor's e-mail	`vachal@utia.cas.cz`

Abstrakt

Tato dizertační práce se skládá ze čtyř článků přispívající k literatuře o datově řízeném a neparametrickém modelování časových řad. V prvním příspěvku studujeme synchronizaci hospodářských cyklů a navrhujeme vícerozměrnou míru sladěnosti založenou na časové frekvenční kohezi. Naznačujeme, že ekonomická integrace může vést k vyšší sladěnosti hospodářských cyklů, což může odrážet výhody konvergence a koordinace hospodářských politik. Druhý článek představuje novou metodiku pro identifikaci perzistence makroekonomických proměnných. Pomocí časově proměnných funkcí frekvenční odezvy identifikujeme heterogenní efekty perzistence v makroekonomických proměnných USA. Třetí a čtvrtý článek navrhují metody založené na datech pro předpovídání distribucí časových řad s využitím strojového učení. Zavádíme vícevýstupovou neuronovou síť, která pro data vybírá nejvhodnější rozdělení. Distribuční neuronová síť je přínosná pro modelování dat s nelineární, negaussovskou a asymetrickou strukturou. Třetí článek demonstruje užitečnost této metody k odhadu informačně bohatých makroekonomických vějířových grafů a pravděpodobnostních předpovědí výnosů akcií. V posledním článku představujeme distribuční neuronovou síť k získání pravděpodobnostního rozdělení předpovědí cen elektřiny. Předpovídáme hodinové ceny pro období příštího dne pro německou elektřinu a provádíme srovnání se současnými modely k předpovídání cen elektřiny.

Klasifikace JEL	C11, E47, E32, C53
Klíčová slova	Neparametrické, daty řízené, časové řady, proměnné v čase, neuronové sítě, frekvenční doména, hospodářské cykly, předpovídání, distribuce
Název práce	Eseje o neparametrickém a datech řízeném modelování časových řad

E-mail autora	`lubos.hanus@gmail.com`
E-mail vedoucího práce	`vachal@utia.cas.cz`

Acknowledgments

I would like to express my gratitude to Lukáš Vácha and Jozef Baruník for their continuous support and positive influence during my graduate studies. Their guidance and friendship have made this journey both enjoyable and unexpected in many aspects.

I am also grateful to my family for their unconditional support. Further many thanks go to all who have helped and influenced me in various ways, friends and colleagues particularly.

Throughout my studies, I participated in several conferences and seminars, where I presented my research and developed valuable connections. I wish to express my gratitude to the individuals who attended these events and provided valuable feedback that helped enhance the quality of my work.

This dissertation was financially backed by the Czech Science Foundation, under the EXPRO GX19-28231X and GA16-14151S projects. Furthermore, it was partially funded by projects from the Grant Agency of Charles University (GAUK), specifically project numbers 366015 and 1390218.

All errors are solely the author's responsibility.

Typeset in LaTeXusing the IES Thesis Template with additional changes.

Bibliographic Record

Hanus, Luboš: *Essays on Data-driven, Non-parametric Modelling of Time-Series*. DISSERTATION. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2024, pages 168. Advisor: Mgr. Lukáš Vácha, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BVAR	Bayesian Vector Autoregression
BMA	Bayesian Model Averaging
CDF	Cumulative Distribution Function
CPI	Consumer Price Index
DistrNN	Distributional Neural Network
GDP	Gross Domestic Product
EU	European Union
EMU	Economic and Monetary Union
EPF	Electricity Price Forecasting
OCA	Optimum Currency Area
QRA	Quantile Regression Averaging
QRM	Quantile Regression committee Machine
VAR	Vector Auto-Regression
V4	Visegrad Four

Chapter 1

Introduction

Understanding the structure of time-series and their driving forces is an important and fascinating task, especially for an economist. The dissertation consists of four papers that are bound by non-parametric methodologies that aim to understand better and model time-series data. The thesis can be methodologically split into two parts: the first two chapters employ frequency domain non-parametric and data-driven measures, while the latter two focus on learning and predicting probabilistic distributions using up-to-date data-driven approaches. Hence, this work is about data-driven non-parametric modelling of time-series.

The initial two papers of the dissertation make contribution to macroeconomic literature by examining the cyclical and persistence properties of time-series. In this respect, the thesis contributes to the time-series literature by utilizing frequency domain approaches to observe patterns and dynamics from perspectives other than time-domain. The first paper provides a detailed analysis of the synchronisation of business cycles between the Visegrad Four (V4) countries and the European Union (EU). The second text presents an approach to measure the persistence identification of macroeconomic variables. These measures, also varying over time, provide researchers with localized information about interactions between macroeconomic variables such as business cycles between countries or within a system of Gross Domestic Product (GDP) growth, inflation and unemployment, for instance. Additionally, with the shift to data-driven methods, the third paper proposes a novel approach to probabilistic forecasting of macroeconomic and financial time-series using deep learning. The fourth paper modifies the approach used in the third chapter and uses this machine learning for

probabilistic forecasting of hourly day-ahead electricity prices.

In the forthcoming text and all chapters, I use the first-person plural pronoun "we", as it is customary in academic writing. It is also due to the fact that all chapters are the result of collaboration. The first two chapters were co-written with my supervisor, Lukáš Vácha, while I collaborated with and received guidance from Jozef Baruník for the subsequent two. My contribution to all chapters is substantial throughout all stages of the research. In the remainder of the introduction, we present the contribution of all chapters in detail.

In Chapter 2 - *Growth cycle synchronisation of the Visegrad Four and the European Union*[1] - we have shown that the increasing comovement of the Visegrad Four countries with the European Union may reflect the benefits of economic integration and convergence. We provide an analysis of the synchronisation of business cycles between the Visegrad Four countries and the European Union, where we employ a wavelet cohesion measure to depict the comovement of GDP growth rates.

This chapter focuses on the membership of the Visegrad Four in the European Union since 2004 and analyses their integration into the European Economic and Monetary Union (EMU) through the business cycle synchronisation. Business cycle synchronisation is one of EMU's essential elements motivated by the Optimum Currency Area (OCA) theory (Mundell, 1961). Increased business cycle synchronisation may decrease the costs of joining the EMU. Alternatively, countries can derive benefits from joining the OCA, where synchronisation increases ex-post. Either way, assessing the degree of synchronisation come in handy with one of the most challenging tasks in economics – to identify, understand, and disentangle the factors and mechanisms that impact the dynamics of macroeconomic variables.

Adopting wavelet methodology, we have overcome the problems of traditional measures, such that it operates in time or frequency domain only and does not require the time-series stationarity. Popular tools assessing the degree of synchronisation, such as the Pearson correlation coefficient or the bandpass filter of Christiano and Fitzgerald (2003), use only a specific part of the time-frequency domain. In contrast to these tools, we benefit from the use of wavelet methods to obtain complete localized information in both domains. Wavelets are not new to the economic literature and macroeconomics. For example, Crowley et al. (2006) already used wavelets to study

[1]This text is published in Empirical Economics (Hanus & Vácha, 2020).

growth cycles of the euro area, or Aguiar-Conraria et al. (2018) employed the methods to analyse the Taylor rule behaviour. To overcome the limitations and to combine both time and frequency domains, we propose the wavelet cohesion measure as a means of assessing multivariate comovement employing time-varying weights. Taking Croux et al. (2001) and Rua and Lopes (2015) we define this measure of cohesion as a weighted average of pairwise comovement, where the weights are attached to each pair of time-series. This allows us to localise and observe the dynamics of the relationship among the countries over time, which is particularly important given the changing economic conditions and policies in the region. Additionally, we use the wavelet coherence (Aguiar-Conraria & Soares, 2011) to assess paired synchronisations.

In this study, we employ quarterly GDP data from 1995 to 2017 as a proxy for aggregated economic activity. Both GDP growth and nominal values are used in the cohesion measurement, with the former representing the strength of economies and the latter indicating their impact on the cycle via a weighting scheme. To proxy EU economic activity, we utilise EU-28 GDP data. Furthermore, we present a broader picture of the synchronisation of the business cycles of states that are already members of the EMU, and we also split them into groups in order to disentangle the dynamics. Alongside the Visegrad Four, we utilise both the EU core group and the EU periphery group of five countries.

The findings confirmed previously established interesting patterns. Slovakia demonstrated poor synchronisation with the EU prior to its EU accession. However, the comovement strengthened after 2005, supporting the theory of the endogeneity of the OCA and the adoption of the euro. Additionally, we demonstrate that the strongest synchronisation between the EU and both the Czech Republic and Hungary started in 2001. This might imply the readiness of these countries to adopt the euro, taking into account one of the characteristics - the coherence of growth cycles. Regarding the Visegrad Group's position within the EU, we have identified strong pro-cyclical behaviour in cycles that are longer than two years. Regarding EU core countries, our research reveals a weak synchronisation of short-term dynamics. Policy-wise, it is reported that the adoption of the euro has a considerable effect on the country's comovement with the rest of the euro area, leading to an increased comovement of business cycles. Such an outcome may imply the advantages of a currency union, but it may also create difficulties

for countries with distinct economic structures and policy preferences. The comparable growth cycle reactions to external shocks of both the V4 and the EU could be a pertinent aspect to contemplate for potential inclusion in the monetary union.

In Chapter 3 - *Identification persistence in macroeconomic responses* - we study the evolving macroeconomic dynamics approached to infer the persistence of variables. The study builds on traditional linear (vector autoregressive) models with impulse response functions. However, impulse responses are a prominent tool in the analysis of monetary policy; we focus on a non-linear and data-driven identification of persistence for macroeconomic variables.

This paper contributes to the literature of macroeconomic modelling and time-series filtering. We propose frequency-specific methodology to provide an additional way of looking at the monetary policy transmission dynamics. To assess the persistence of monetary transmission using frequency response functions. The TVP-VAR model with time-varying coefficients and covariance structure allows us to quantify dependence measures in both the time and frequency domains. Compared to the rolling window Vector Auto-Regression (VAR), which may suffer from outlier observations with a small window or lack of variability with a wide window, we opt for the TVP-VAR for macroeconomic modelling.

We study a traditional monetary system of three variables: GDP growth, inflation and interest rate of the US data. The findings show that the monetary policy transmission effects on macroeconomic variables vary in time, frequency and intensity. We find a substantial variation in output and inflation in response to shocks in time, which aligns with the literature. The frequency transmission of monetary policy pronounces the largest positive impacts in output at economic cycles longer than eight years. We show that even by controlling for the negative impulse response of inflation to monetary policy, the response is both negative and positive. Thus, the impacts are frequency-dependent, and the direction changes. Additionally, we quantify cyclical behaviour using a time-frequency correlation that captures the dependence between output, inflation, and interest rates, which is rather weak at short cycles and much stronger during business cycles.

Moreover, using the time-varying approach, we observe a negative propagation of shocks to inflation, meaning that the price puzzle phenomenon is not observed for business cycles after 1985. However, the average frequency transmission varies, and in every decade, we see that prices rise in response

to monetary policy, except in the years between 2000 and 2010. This leads us to a new result that the price puzzle phenomenon may have a frequency-dependent effect and be propagated at different cycles, which is not observed using only impulse response functions, even with sign restrictions imposed.

The following two chapters demonstrate a shift from traditional non-parametric methods to modern data-driven machine learning, in particular neural networks. The approximation theorem (Hornik et al., 1989) and neural networks have been acknowledged for many years. Recently, economists have sought to move beyond relying exclusively on models and towards utilising machine learning techniques (Athey & Imbens, 2019). Additionally, we are turning our focus from point forecasts to probabilistic forecasting using extensive data sets. In light of these data-rich environments, the world of data science and econometrics has considered and adopted more data-driven approaches, given the availability of computational power to work with such data.

In Chapter 4 - *Taming data-driven probability distributions* - we present a comprehensive and detailed approach to probabilistic forecasting of macroeconomic and financial time-series using deep learning. We argue that our approach is useful for decision-making that depends on the uncertainty of a large number of economic outcomes, and that it has several advantages over traditional methods. We contend that deep learning, specifically recurrent neural networks, provides a valuable method for predicting distributions without requiring model specification; it simply learns the distributions from the data in the spirit of a non-parametric approach. Economic time-series data show many characteristics such as heavy tails, asymmetries, irregularities, spikes or regime shifts that cannot be fully characterized by methods assuming Gaussian distribution estimating its mean and variance. We oppose this approach with our contribution to the literature and propose how to use deep learning techniques as a useful tool for approximation and prediction of conditional distributions in a data-rich environment.

In terms of distributions, many studies focus on the prediction of conditional return distribution to characterize the cumulative conditional distribution by a collection of conditional quantiles (Engle & Manganelli, 2004; Žikeš & Baruník, 2016). On the contrary, Leorato and Peracchi (2015) argue that collection of conditional probabilities describing the cumulative distribution function using a set of separate logistic regressions (Foresi & Peracchi, 1995)

provides a better approach. In this manner, Anatolyev and Baruník (2019) proposed an ordered logistic parametrization to forecast excess asset returns distribution.

We draw inspiration from Anatolyev and Baruník (2019) parametric binary choice model and generalise it into a novel multiple output neural network, i.e. a Distributional Neural Network (DistrNN) capable of learning complete information about the probability of future outcomes given past information. Like logistic regression, the distributional neural network uses binary cross-entropy as its loss function. We offer an additional and significant contribution to this innovative approach by adding a penalty function for deviating from monotonic behaviour to the objective function. This penalty adjusts loss towards the monotonicity of the distributional predictions. The issue of monotonicity, or quantile crossing, is a known problem in quantile regression.

Identifying the optimal model is a complex undertaking, and each problem necessitates a specific set of hyper-parameters to mitigate risks. As is customary with neural networks, the selection of hyper-parameters, in combination with regularisation methods, plays a crucial role in reducing the risk of estimation. We use hyper-optimisation algorithms to select the best set of hyper-parameters, which are also specific to a particular experiment. We illustrate the usefulness of the approach on two different datasets.

First, distributional neural networks are employed to construct data-driven macroeconomic fan charts based on the information contained in a large number of variables. A high-dimensional data set of quarterly data from the Federal Reserve Bank of St. Louis is used (McCracken & Ng, 2020). We have produced data-rich fan charts representing real Gross Domestic Product growth, inflation, and the unemployment rate. These fan charts are the first of their kind to capture information from 216 relevant variables. They are significant for policy-makers because they show the underlying structure in the data, independent of any model choice. We compare the distributional fan charts to the Bayesian vector autoregression, which is the state-of-the-art macroeconomic model. We assess models by using out-of-sample quarterly prediction intervals for horizons of 1 to 6 quarters. The results indicate that neural networks provide forecasts with lower quantile losses for most considered probability levels and horizons.

In the second empirical application, we forecast conditional distributions of asset returns using the distributional neural network. We consider the

most liquid 29 US stocks in the S&P500. Such data are known to have heavy tails and a low signal-to-noise ratio. We take the ordered logistic regression model of Anatolyev and Baruník (2019) as a benchmark. Initially, we utilise the original predictors as input data and utilise a distributional neural network as a direct non-parametric and potentially non-linear alternative. To further benchmark setups, we generate an additional five realised measures (volatility, skewness, kurtosis, and positive and negative semi-variances) using high-frequency one-minute intra-day data. While these predictors offer valuable information for distributional forecasts, their inclusion in the original (benchmark) ordered logit model would over-parametrize the model, making it infeasible. As such, our approach offers a flexible and comprehensive method of predicting distributions in data-rich environments and exploring potential non-linearities within the data. We utilize a rolling scheme that results in 327 day-ahead forecasts. By employing statistical measures to assess forecasts, we document that the neural network is capable of predicting the conditional distribution of asset returns well, and by providing informative variables, it provides improved predictions.

Building on the previous chapter, in Chapter 5 - *Learning probability distributions of day-ahead electricity prices* - we reframe the novel machine learning approach to a problem of probabilistic hourly Electricity Price Forecasting (EPF). As electricity is a vital commodity and future price uncertainty is crucial for economic agents, it has been extensively studied. Moving away from point estimates of electricity prices, probabilistic forecasting has become essential for those needing to assess uncertainty and improve optimal strategies (Bunn et al., 2016). Our paper contributes to the emerging literature on machine learning methods in the context of hourly electricity price forecasting. To name a few Lago et al. (2021), Lehna et al. (2022), Marcjasz et al. (2023), Mashlakov et al. (2021), Nowotarski and Weron (2018), and Zhang et al. (2022), however, all these methods rely on models and assumptions about the data or its distribution. Typically, the literature suggests learning only some properties of the distribution, such as moments. In contrast, our approach selects the best distribution from all possible empirical distributions learned from the data.

We construct a distributional neural network to forecast German hourly day-ahead electricity prices using the 221 characteristics, including lagged prices, total load, external variables such as EU allowance prices, and fuel prices, in particular coal, gas and oil. The proposed network aims to predict

the output variable of the target price, and its multiple outputs enable us to approximate the conditional distribution function as a set of joint probabilities.

We employ a daily forward rolling scheme utilizing data to forecast 554 days of the out-of-sample period between 2019-06-07 and 2020-12-31. Our analysis involves a four-and-a-half-year training and validation period preceding the out-of-sample. To train and analyse neural networks, we used all essential components, such as optimisation algorithms and techniques, to prevent over-fitting. We performed the hyper-optimization search for hyperparameters. Further, to decrease the variance of our predictions, we utilise ensembles.

To conform with the literature, we utilize state-of-the-art benchmarks to compare our approach. We make use of the naive model and two quantile regression-based linear models with autoregressive and exogenous variables. The two parametric methods of distributional forecasting are widely used in the literature. Both quantile regression models require point forecasts of the price to provide probabilistic forecasts; one is the Quantile Regression Averaging (QRA) (Nowotarski & Weron, 2015) and the other is the Quantile Regression committee Machine (QRM) (Marcjasz et al., 2020).

We employ both the continuous rank probability score and the Diebold-Mariano test to present the models' performance. Our comparison of the distributional neural network and state-of-the-art frameworks showed that, for multivariate loss, we accept the alternative hypothesis that the distributional neural network has significantly superior probabilistic prediction accuracy. This is primarily attributed to the fact that it does not rely on restrictive model assumptions and allows for non-Gaussian, heavy-tailed data and their non-linear interactions. Moreover, we offer an efficient computational package that researchers can utilize.

References

Aguiar-Conraria, L., Martins, M. M., & Soares, M. J. (2018). Estimating the taylor rule in the time-frequency domain. *Journal of Macroeconomics*, *57*, 122–137.

Aguiar-Conraria, L., & Soares, M. J. (2011). *The continuous wavelet transform: A primer* (tech. rep. No. 16/2011). NIPE-Universidade do Minho.

Anatolyev, S., & Baruník, J. (2019). Forecasting dynamic return distributions based on ordered binary choice. *International Journal of Forecasting*, *35*(3), 823–835.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Bunn, D., Andresen, A., Chen, D., & Westgaard, S. (2016). Analysis and forecasting of electricty price risks with quantile factor models. *The Energy Journal*, *37*(1).

Croux, C., Forni, M., & Reichlin, L. (2001). A measure of comovement for economic variables: Theory and empirics. *The Review of Economics and Statistics*, *83*(2), 232–241.

Crowley, P. M., Maraun, D., & Mayes, D. (2006). How hard is the euro area core?: An evaluation of growth cycles using wavelet analysis. *Bank of Finland Research Discussion Papers*, (No. 18/2006).

Engle, R. F., & Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367–381.

Foresi, S., & Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, *90*(430), 451–466.

Hanus, L., & Vácha, L. (2020). Growth cycle synchronization of the Visegrad Four and the European Union. *Empirical Economics*, *58*(4), 1779–1795.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Christiano, L. J., & Fitzgerald, T. J. (2003). The band pass filter. *International Economic Review*, *44*(2), 435–465.

Lago, J., Marcjasz, G., De Schutter, B., & Weron, R. (2021). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, *293*, 116983.

Lehna, M., Scheller, F., & Herwartz, H. (2022). Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account. *Energy Economics*, *106*, 105742.

Leorato, S., & Peracchi, F. (2015). Comparing distribution and quantile regression. *EIEF Working Papers Series*, (No. 1511).

Marcjasz, G., Narajewski, M., Weron, R., & Ziel, F. (2023). Distributional neural networks for electricity price forecasting. *Energy Economics*, *125*, 106843.

Marcjasz, G., Uniejewski, B., & Weron, R. (2020). Probabilistic electricity price forecasting with narx networks: Combine point or probabilistic forecasts? *International Journal of Forecasting*, *36*(2), 466–479.

Mashlakov, A., Kuronen, T., Lensu, L., Kaarna, A., & Honkapuro, S. (2021). Assessing the performance of deep learning models for multivariate probabilistic energy forecasting. *Applied Energy*, *285*, 116405.

McCracken, M., & Ng, S. (2020). *FRED-QD: A quarterly database for macroeconomic research* (tech. rep.). National Bureau of Economic Research NBER Working Paper No. 26872.

Mundell, R. A. (1961). A theory of optimum currency areas. *The American Economic Review*, *51*(4), 657–665.

Nowotarski, J., & Weron, R. (2015). Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, *30*(3), 791–803.

Nowotarski, J., & Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, *81*, 1548–1568.

Rua, A., & Lopes, A. S. (2015). Cohesion within the euro area and the US: A wavelet-based view.

Zhang, F., Fleyeh, H., & Bales, C. (2022). A hybrid model based on bidirectional long short-term memory neural network and catboost for short-term electricity spot price forecasting. *Journal of the Operational Research Society*, *73*(2), 301–325.

Žikeš, F., & Baruník, J. (2016). Semi-parametric conditional quantile models for financial returns and realized volatility. *Journal of Financial Econometrics*, *14*(1), 185–226.

Chapter 2

Growth cycle synchronization of the Visegrad Four and the European Union

Abstract

In this paper, we map the growth cycle synchronization across the European Union, specifically focusing on the position of the Visegrad Four countries. We study the synchronization using frequency and time-frequency domain. To accommodate for dynamic relationships among the countries we propose a wavelet cohesion measure with time-varying weights. Analyzing quarterly data from 1995 to 2017, we show an increasing comovement of the Visegrad Four countries with the European Union after the countries have accessed the European Union. We show that participation in a currency union increases the comovement of the country adopting the Euro. Furthermore, we find a high degree of synchronization at business cycle frequencies of the Visegrad Four and countries of the European monetary union.

This chapter was co-authored with Lukáš Vácha and published in the *Empirical Economics*. We gratefully acknowledge the financial support from the Grant Agency of Charles University (GAUK), No. 366015. This article is part of a research initiative launched by the Leibniz Community. Support from the Czech Science Foundation under the GA16-14151S project is gratefully acknowledged. We thank the referees and seminar participants at FMND2015 Paris and at SEAM Košice 2015 for helpful comments.

2.1 Introduction

It has been more than two decades since the break-up of the Eastern Bloc;[1] following its disintegration, those countries began their independent economic and political journeys. While undertaking their economic transformations during this time, the Czech Republic, Hungary, Poland, and Slovakia began discussing mutual cooperation – Visegrad Four (V4), established in 1991. Despite their initially different levels of economic maturity and development, historical and regional proximity connected them to achieve several goals to return from the East back to Europe. We take a closer look at their aim for faster convergence and integration into the European Union.[2]

In 2004, the Visegrad Four countries became members of the EU, which obliges them to adopt the Euro currency as part of the integration process. One of the concerns of successful integration into the European Economic and Monetary Union (EMU) is business cycle synchronization, which is motivated by the theory of Optimum Currency Area (OCA) (Mundell, 1961). The common currency can be beneficial for both new and former countries in terms of trade transaction costs. A country joining the OCA needs to be well integrated at the level of macroeconomic variables to balance the costs and benefits of future unified monetary policies (De Haan et al., 2008). If the country is not integrated enough at the European level, then the policies of the European Central Bank that apply to all member states may be countercyclical for countries with low business cycle synchronization (Kolasa, 2013). On the one hand, these policies may create difficulties for those countries. On the other hand, countries with low levels of synchronization may benefit from being members of the OCA ex-post, because the business cycle synchronization may appear as an endogenous criterion. This endogeneity of OCA means that forming a monetary union may make its members more synchronized (Frankel & Rose, 1998).[3]

Assessing the degree of synchronization comes in hand with one of the most challenging tasks in economics – to identify, understand, and disentangle the factors and mechanisms that impact the dynamics of macroeconomic

[1] The Eastern Bloc was generally formed of the countries of the Warsaw Pact (as Central and Eastern European countries) and the Soviet Union.

[2] The Visegrad Four countries also joined the North Atlantic Treaty Organization in 1999 and applied for membership in the European Union in 1995-1996.

[3] The literature focusing on the evolution and determinants of business cycle synchronization between Central and Eastern European (CEE) countries and the EU is extensive, see, e.g., Artis et al. (2004), Backus et al. (1992), and Darvas and Szapáry (2008).

variables. Many quantitative econometric techniques have been developed to study the regular fluctuations of macroeconomic indicators and business cycles, e.g., Baxter and King (1999), Harding and Pagan (2002), and Hodrick and Prescott (1997). Our work investigates the decomposition of business cycle into growth cycles over different time horizons. In order to disentangle the desired information, we apply wavelet methodology working in a time-frequency space. The analysis considers the case of the V4, both regarding the internal relationships among its constituent countries and regarding the relationships established within the framework of the European Union (EU).

We find different levels of comovement between V4 countries and the EU during the 1995-2017 period. The V4 countries show strong comovement concerning cycles longer than 3 years. The pairwise synchronization of V4 countries with the EU appears to be significant even at longer cycles from 2004 onward. Studying common economic cycles shows that the V4 countries are well-synchronized for growth cycle with a periodicity of 2-8 years. Similarly, we observe higher synchronization for seven European core countries for business cycles of 3-8 years, and the relationship becomes even stronger after 2004. On the contrary, all countries together show no considerable relationship for cycles shorter than 1 year, which may reflect some short-term policy heterogeneity.

The contribution of this paper is twofold. We contribute to the literature with an empirical analysis of 23 years period studying the Visegrad Four within the framework of the European Union. Secondly, we propose the novel measure of cohesion with time-varying weights which better explains the relationship among countries.

The remainder of the paper is structured as follows. The following section reviews the relevant literature. Section 2.3 describes the methodology and introduces the cohesion measure with time-varying weights. Section 2.4 provides the data description. In section 2.5, we provide the results. Finally, Section 2.6 concludes.

2.2 Literature review

Regarding the EU integration – and particularly the economic integration of the Central and Eastern European (CEE) countries – the literature has grown rapidly. Fidrmuc and Korhonen (2006) conduct a meta-analysis of

35 studies involving the synchronization of the EU and CEE countries and find a significant synchronization between new member states and the EU. However, only Hungary and Poland among the V4 countries reached a high level of synchronization. Artis et al. (2004) and Darvas and Szapáry (2008) obtained similar results studying correlations between the business cycles of the EU and Hungary and Poland. Results of Jagrič (2002) also implies that the economic comovement of Hungary and Poland is high. Analogously, Bruzda (2011) shows that Poland's economic synchronization with the EU rises when intra-EU synchronization is stable. Recently, Aguiar-Conraria and Soares (2011a) study the business cycle synchronizations, using the industrial production index of Euro-12 countries,[4] taking into account the distances between regions. They show that countries that are closer to one another show higher synchronization. Moreover, Hungary and the Czech Republic exhibit a high level of business cycles comovement with the EU after 2005. However, Slovakia, a member of the Euro area, has only minimal synchronization with the EU. Jiménez-Rodríguez et al. (2013) also find high correlations of CEE countries (except for the Czech Republic) with the EU business cycle, they show that together these countries exhibit a lower level of concordance when a factor model is employed. Crespo-Cuaresma and Fernández-Amador (2013) look at second moments of business cycles in the EU and they report a significant convergence since the 90s. They show there is no decrease in the optimality of the currency area after the EU enlargements. Further, Bekiros et al. (2015) study business cycle of industrial production indexes at two sample periods; first only up to the crises (2007M12) and second including the crises. They use Germany as a proxy to the EU and show increased coherence during the crisis.

One of the most popular tools to assess the degree of synchronization is the Pearson correlation coefficient, which solely measures the degree of comovement in a time domain. However, market-based economies are structured over different time horizons. For this reason, the interest in frequency domain techniques has grown. Christiano and Fitzgerald (2003) proposed a model based on a bandpass filter. Further, Croux et al. (2001) introduced a measure of comovement, the dynamic correlation, based on a spectral analysis. This measure estimates correlation on a filtered time-series. Nev-

[4]These group consists of Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, and Spain. We analyze this group and the V4 countries.

ertheless, both the time (static) Pearson correlation and the spectral domain dynamic correlation have several caveats. The first method loses time information, and the latter omits the comovement dependence in time. Wavelet analysis overcomes these limitations since it combines both time and frequency domain. It also relaxes the assumption of covariance-stationarity; hence the analyzed time-series can be locally stationary (Nason et al., 2000; Raihan et al., 2005).[5]

The literature acquaints with many studies that use wavelets. To name few compelling in economics, we refer to Crowley et al. (2006) who studied growth cycles of euro area core, Aguiar-Conraria et al. (2008) who analyzed the evolution of monetary policy in the US, Vacha and Barunik (2012) studying energy markets relationships, and Yogo (2008) who apply wavelet analysis to determine peaks and valleys of business cycles that correspond to the definition of the National Bureau of Economic Research. Further, Crowley and Hallett (2015) have used wavelet techniques to disentangle the relationship of the real Gross Domestic Product (GDP) growth at different frequencies. Recently, Aguiar-Conraria et al. (2018) analyzed the cyclical behavior of the Taylor rule using the wavelet framework.

2.3 Methodology

To capture the comovement of the growth cycles we use wavelet measures. The wavelet transform has been developed to find a better balance between time and frequency dimensions and it also overcomes problems of stationarity.[6] The wavelet functions (filters) used for the decomposition are narrow or wide when we analyze high or low frequencies, respectively (Daubechies, 1992). Thus, wavelet analysis is suitable for different types of stochastic processes using optimal time-frequency resolution (Cazelles et al., 2008; Crowley, 2007).[7]

In the first part of our empirical study, we use the wavelet coherence (Grinsted et al., 2004; Torrence & Compo, 1998) to quantify the pairwise co-

[5]Characteristics of locally stationary time-series are close to the stationary ones at each point of time or shorter periods.

[6]This also overcomes the problem of short-time Fourier transform, or windowed Fourier transform (Gabor, 1946).

[7]It is possible to use methods of evolutionary spectra of non-stationary time-series developed by Priestley (1965). However, to study time-varying dynamics we need to give up some frequency resolution, which is not the case when using wavelet techniques.

movement of countries. For more details about the wavelet coherence see Appendix 2.A.1. The rest of the study is focused on relationships among multiple countries. We begin with the frequency domain cohesion of Croux et al. (2001). Further, extending the work of Rua (2010) and Rua and Silva Lopes (2012) we propose wavelet cohesion estimator with time-varying weights.

2.3.1 Measurement of common cycles

Many time or frequency domain comovement measures rely on the bivariate correlation. Based on the bivariate correlation, Croux et al. (2001) proposed a powerful tool for studying the relationship of multiple time-series over the frequencies, coined cohesion. For the multiple time-series $x_t = (x_{1t}, \ldots, x_{nt})$, $n \geq 2$, the cohesion in the frequency domain is defined as:

$$coh(\lambda) = \frac{\sum_{i \neq j} w_i w_j \rho_{x_i x_j}(\lambda)}{\sum_{i \neq j} w_i w_j}, \quad coh(\lambda) \in [-1, 1], \tag{2.1}$$

where λ is the frequency, $-\pi \leq \lambda \leq \pi$, w_i denotes a weight associated with time-series x_{it}.

Following Croux et al. (2001), Rua and Silva Lopes (2012) extend the frequency measure to the time-frequency domain. Using wavelets they analogically define the wavelet cohesion as:

$$coh(\tau, s) = \frac{\sum_{i \neq j} \bar{\omega}_{ij} \rho_{x_i x_j}(\tau, s)}{\sum_{i \neq j} \bar{\omega}_{ij}}, \quad coh(\tau, s) \in [-1, 1], \tag{2.2}$$

where $\rho_{x_i x_j}(\tau, s)$ is a real wavelet-based measure of comovement (Rua, 2010), interpretable as a contemporaneous correlation coefficient around each point in the time-frequnecy plane, defined as:

$$\rho_{x_i x_j}(\tau, s) = \frac{\Re(W_{x_i x_j}(\tau, s))}{\sqrt{|W_{x_i}(\tau, s)|^2 |W_{x_j}(\tau, s)|^2}}, \quad \rho_{x_i, x_j}(\tau, s) \in [-1, 1], \tag{2.3}$$

where $\Re(W_{x_i x_j}(\tau, s))$ is the real part of the wavelet cross-spectrum, known as co-spectrum, of two time-series. The co-spectrum is normalized by the squared roots of two wavelet power spectra. The measure captures both positive and negative comovements of time-series, which is inherited by the

cohesion measure.[8] The wavelet cohesion, Eq. 2.2, is a weighted average of pairwise comovement, where the weights, $\bar{\omega}_{ij}$, are attached to the pair of series, (i, j), i.e., a share of each pair among all time-series. As a measure of comovement of multiple time-series, the cohesion uncovers their common cyclical behavior.

The wavelet cohesion (Eq. 2.2) employs fixed weights that represent constant shares of each pair. However, we see that weights (e.g. GDP, population size) often change over time. This would reflect, for instance, that the developing or emerging countries have different speed of development. In our case, emerging countries show higher growth as they converge to the developed countries; hence, the weight of each pair significantly changes over time.

To reveal the dynamics and development of economies, we propose a new approach to map a multivariate relationship using the time-varying weights in the wavelet cohesion measure:

$$coh^{TV}(\tau, s) = \frac{\sum_{i \neq j} \omega_{ij}(\tau) \rho_{x_i x_j}(\tau, s)}{\sum_{i \neq j} \omega_{ij}(\tau)}, \quad coh^{TV}(\tau, s) \in [-1, 1], \tag{2.4}$$

where $\omega_{ij}(\tau)$ is the weight attached to the pair of time-series (i, j) at given time τ. Similarly, the weights are $\omega_{ij}(\tau) = w_i(\tau) w_j(\tau)$ as in frequency cohesion, Eq.2.1. The wavelet cohesion with time-varying weights allows using different types of weights. For example, using GDP as a weight representing the size of an economy, a country with smaller or larger GDP can have smaller or larger effects on the comovement than other countries. Additionally, as for the wavelet coherence, we test the statistical significance of wavelet cohesion estimates using Monte Carlo simulation methods (Aguiar-Conraria & Soares, 2014).[9]

Additionally, we support our wavelet cohesion using frequency domain cohesion of Croux et al. (2001), which we use for two time-invariant periods and it ideally complements the wavelet results.[10]

[8] In the appendix we demonstrate the wavelet-based measure (Eq. 2.3) in two particular cases, Fig. 2.9.

[9] Another possibility for testing the significance is area-wise test approach of Maraun et al. (2007).

[10] To obtain the confidence intervals of frequency cohesion; we follow the procedures of Franke and Hardle (1992) and Berkowitz and Diebold (1998), where instead of bootstrapping the cohesion measure we bootstrap each (cross-)spectrum. Schüler et al. (2017) used this approach in their power cohesion measure while studying financial cycles for G-7 countries.

2.4 Data

To study the synchronization, we use the GDP from the database of OECD (2018).[11] We consider the GDP data as a measure of aggregated economic activity. We use both percentage changes from the previous period and the nominal value in EUR. We employ the nominal GDP in cohesion estimation as time-varying weights to measure the power of economies and their impact on the cycles. The dataset includes quarterly data, covering the period from 1995Q1 to 2017Q4.

The Visegrad region consists of the Czech Republic, Hungary, Poland and Slovakia, where only Slovakia is an EMU member. To study the comovement of the Visegrad countries with the EU, we use the GDP data of EU-28.[12] Furthermore, we measure the common cycles of the V4 countries with the EU core group: Austria, Belgium, Germany, Finland, France, Luxembourg, and the Netherlands. In the literature, there is evidence of business cycles synchronization of EMU-12 during the 1990s, e.g., Crespo-Cuaresma and Fernández-Amador (2013). However, several countries form the EU core; it is always France and Germany, and additional countries on which studies are not consistent when specifying the EU core and periphery. We separate Greece, Ireland, Italy, Portugal, and Spain as five peripheral countries (Aguiar-Conraria & Soares, 2011b; Ferreira-Lopes & Pina, 2011; Grigoraş & Stanciu, 2016). The EU core group is supposed to be the target for the V4 countries both economically and politically.

2.5 Results

2.5.1 Synchronization of the Visegrad Four countries

We begin our analysis of growth cycle synchronization within the group of V4 countries. The extent of synchronization and comovement of Gross domestic products is measured by the wavelet coherence. The wavelet coherence, depicted in Fig. 2.1, show regions of comovement localized in time-frequency; on x and y-axes we have time and corresponding cyclical component, respec-

[11] Data was obtained via OECD Database, May, 2018.

[12] The V4 countries are included in the EU-28; however, the contribution is minimal to change the EU GDP growth. For robustness check, we analyzed the comovement of the V4 and the EA-19 GDPs, and these results are almost identical to those we report.

tively. The yellow color represents the strongest coherence, while the blue color indicates no coherence.

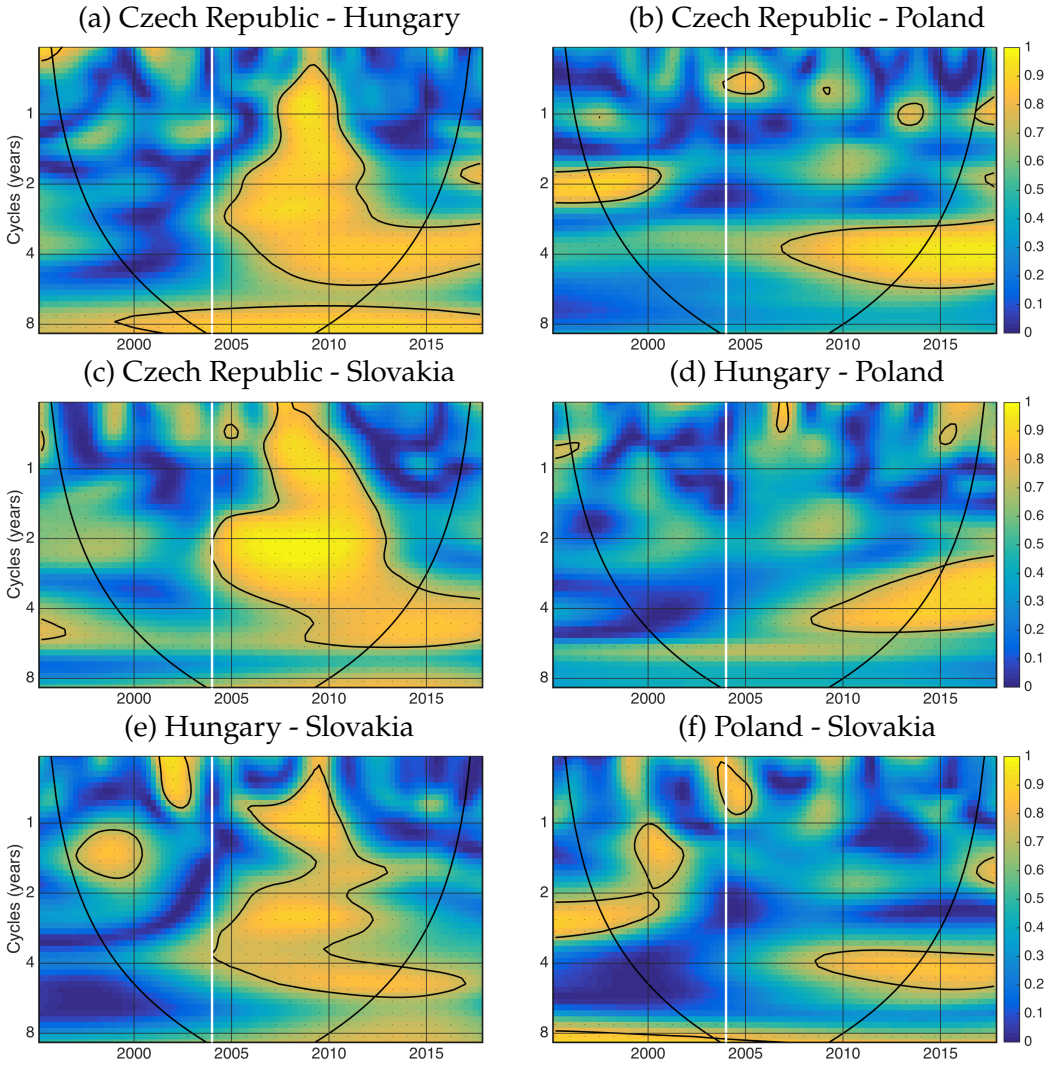

Figure 2.1: Wavelet coherence within the Visegrad Four countries. The solid black line contours the significance level of 5% against the red noise. The area below the black curve is the cone of influence. The vertical solid white line indicates 2004 – the year of the enlargement of the EU.

The beginning of the transition period in the 90s presents weak synchronization at all cycles. This reflects the situation of the V4 countries which started their transition to market-based economies after the break-up of the Eastern Bloc. The low comovement, except several regions around 1999 and 2000, may be caused by Slovakia's cold-shouldered participation in the political discussions during 1993-1997, which translated into the economic performance with a delay. Another possible reason is that even after a few

years of formal and intensive cooperation the monetary and fiscal policies started to diverge. Many countries went through a financial crisis in the years around 1997. For instance, in the late 1990s, the Czech Republic had been through difficult years of stabilization (Antal et al., 2008). This divergence in economic environments might cause significant asynchrony in growth cycle behavior over both shorter and longer cycle periods. Related to policies of sovereign states, the low synchronization may also come from the low level of convergence of other macroeconomic variables (Kutan & Yigit, 2004).

Nevertheless, this characteristic feature of a weak relationship for all pairs within the V4 countries lasts until 2004, with the exception of several regions of short periods of strong coherence.[13] We observe a high degree of synchronization of 1-5 year cycles beginning around 2004 for Hungary with Slovakia, the Czech Republic with both Hungary and Slovakia. These are the strongest coherences among the V4 countries. Furthermore, all V4 countries pairs comove at cyclical component around 3-5 years beginning approximately in 2009. Interestingly, the overall relationship between Poland's and other V4 countries is notably weak, see the right column of Fig 2.1. Additionally, the coherence of the growth cycles up to 1 year are low during most the sample period.

2.5.2 Synchronization of V4 and the EU

In this section, we analyze the comovement of the Visegrad Four countries and the European Union. We take the GDP growth of all 28 countries of the European Union. The reason is straightforward since once states are members of the EU they should support the economic aims of the EU and coordinate policies they make towards these aims.[14]

We observe a strong comovement of the Czech Republic and Hungary with the EU starting around 2003 at 2-6 year cycles. In contrast, Slovakia and Poland are less synchronized with the EU. These findings are in line with the results of Aguiar-Conraria and Soares (2011a). A significant synchronization of Slovakia with the EU starts right after its accession to the EU. Synchronization increases gradually from 2004 and spreads from 2-4 years to 1-6

[13] Short periods of comovement appear around and prior to 2000 at 1-2 years, and 2 years cycles, respectively between Hungary and Slovakia, and Poland with the Czech Republic and Slovakia.

[14] We have additionally checked the comovement of the V4 countries and the Euro area of 19 countries (EA-19) as a proxy of the EU. The results are almost indistinguishable.

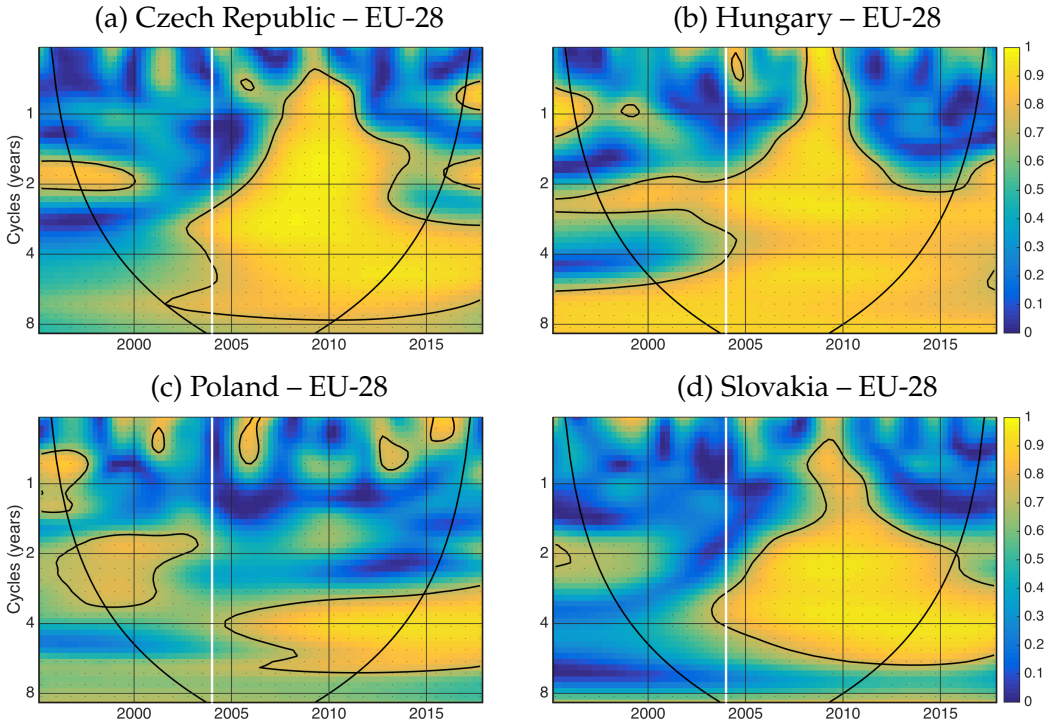

Figure 2.2: Wavelet coherences of the Visegrad Four and EU-28. The solid black line contours the significance level of 5% against the red noise. The area below the black curve is the cone of influence. The vertical solid white line indicates 2004 – the year of the enlargement of the EU.

years cycles around 2008 that is precisely at the time when Slovakia adopted the Euro, on January 1, 2009. Eventually, Slovakia may be considered as an example where the degree of synchronization increases after accession to the EU and EMU, which is consistent with the theory of endogeneity for optimum currency areas. Moreover, in comparison to the other 3 countries, Slovakia has not experienced any significant synchronization before 2004. On the other hand, the high synchronization around 2008 may also be a reaction to the global financial crisis hitting all countries. Nevertheless, this may be in line with the OCA theory when the crisis spills over all highly synchronized European countries even to those of V4.

Preparation of the Visegrad countries for the EU accession, which began shortly before 2000, is also one of the reasons for increased synchronization with EU. This is in line with findings of Kolasa (2013) who reports a substantial convergence with the Eastern enlargement of the EU. The high degree of synchronization of Hungary also supports the results of Fidrmuc and Korhonen (2006). For Poland, we do not see a strong comovement, we observe

higher coherence only for a small region in comparison with other countries in the group, and it is around 4-6 year cycles beginning 2006. This low level of synchronization of most of Poland's and EU's growth cycles may be due to the different orientation structures of Poland's economy.[15]

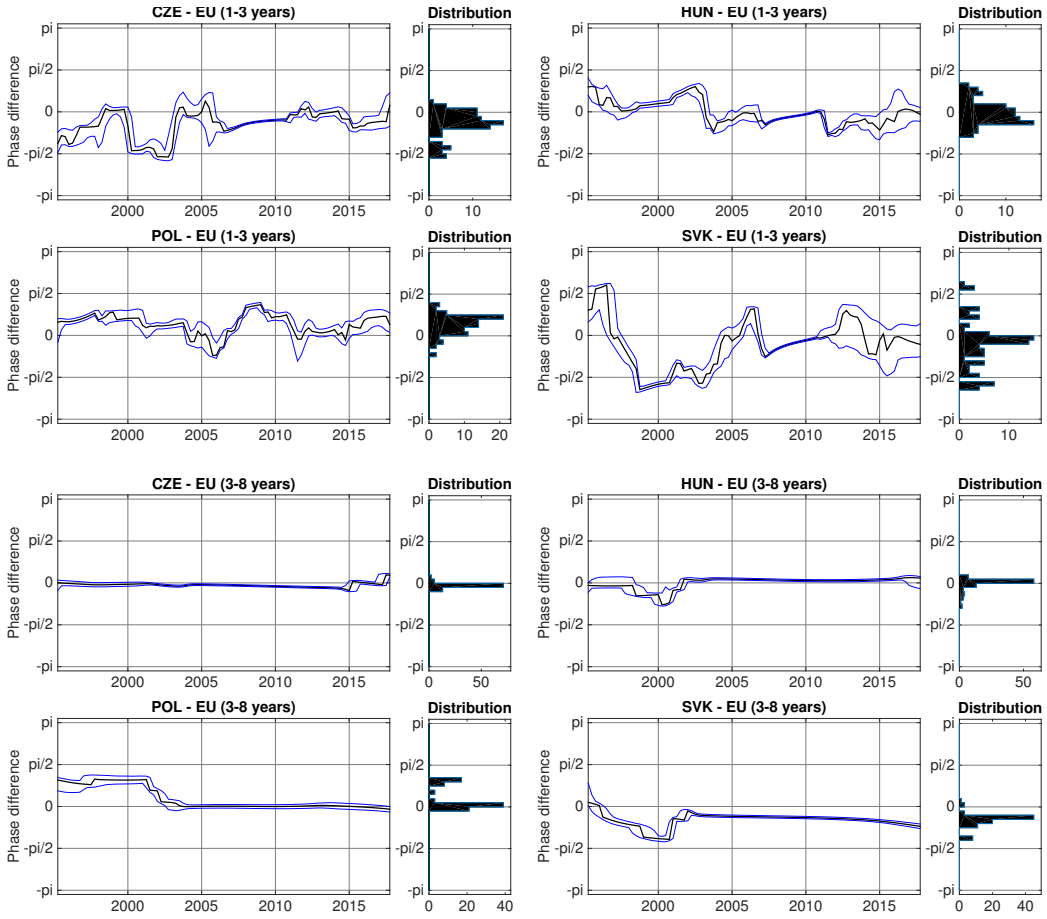

Figure 2.3: Phase differences of short- and business cycle frequencies, 1-3 and 3-8 years respectively. The solid black line is the true phase difference of two time-series. The blue solid line is the 95% bootstrapped confidence interval. For each phase difference, its distribution is provided.

Additionally, we provide the analysis of phase differences between each of V4 countries and the EU's GDP. The phase difference presents the information about the position of cycles of two economies, i.e., whether one leads the other. In the Fig. 2.3, the phase differences show lead or lag position of the 1-3 and 3-8 years cycles between V4 countries and the EU.[16] There are many periods of time where phase differences are not significantly different

[15]Poland's economy share of agriculture in GDP is one of the higher.

[16]Two countries are in-phase if the phase difference belongs to $[-\pi/2, \pi/2]$; otherwise,

from zero, indicating there is no country in the lead position, which also means the countries are in-phase.[17] Nevertheless, we observe periods with significant phase differences such as between 2006 and 2010. Which for all V4 countries means following EU's growth cycles at 1-3 years frequency before and during the recession, except for Poland. When the phase difference belongs to the $[0, \pi/2]$ interval, the 1-3 years growth cycles of Hungary and Poland lead the EU cycles at these frequencies. This is most of the time for Poland and between 1998 and 2004 for Hungary. This lead/lag situation of the V4 region is puzzling, as the countries have been tightly connected to the EU economy and it happens only for Poland and Hungary at the beginning of the sample. One possible explanation for this counter-intuitive finding is that recessions or rebounds of the economies occur sooner compared to EU. For example, during the recession, the lead of V4 countries could be the negative growth lead since these countries are often at the beginning of the chain of outsourced production. In the time of crises, the cuts may start by subcontracted production. This may be the case with the debt crisis in 2013 where we see more volatile phase differences at 1-3 years cycles for all V4 countries. The Czech Republic was in-phase before 2013, then it follows in 2015, similarly for Hungary. On the other hand, in case of Slovakia, we have the significant leading position of the EU during 2004 and 2007 to 2010 for 1-3 years cycles and during whole sample period at 3-8 years cycles, which supports the concept of the endogeneity of OCA.

Observing the phase differences, the phase differences look more stable at business cycles horizons. This observation is due to high and significant coherences at these cycles.[18] Fig. 2.3 shows that Slovakia follows the EU at cycles of 3-8 years period over the whole sample. Poland's 3-8 years growth cycles are in-phase and lead those of the EU from 1995 until 2003 then the phase is not significant from zero. Cycles of Hungary were lagging around 2000, but after 2001 Hungary is in-phase and leads the EU cycles. The business cycle growth component of the Czech Republic is lagging the EU after 2001 when it is significant, and their coherence is strong. At 3-8 years business cycles there are no directional changes of growth phase we may surely link to crisis periods.

they are in the anti-phase. Moreover, the first country leads the second, x_j, if the phase is in $[0, \pi/2]$ and $[-\pi, -\pi/2]$; when in $[-\pi/2, 0]$ and $[\pi/2, \pi]$, the second country is leading.

[17] Furthermore, the phase is more volatile when the coherence is low.

[18] We should also carefully interpret the phase difference at 3-8 years cycles because of the cone of influence, which affects influences results at 8 years from both sides of the sample.

2.5.3 Common economic cycle within Europe

In this part, we investigate the multivariate relationship of countries in the EU. As some of the countries experienced the very dynamic development of GDP, it is natural to construct a measure that takes these changes into account. We propose the wavelet cohesion with time-varying weights that precisely quantifies commonalities among cycles. In contrast to the coherence, the cohesion may be negative; it can capture a counter-cyclical comovement of time-series. For the weights, we use nominal GDP in EUR, which relate to the size of countries' economies and their wealth, respectively. For the measuring the synchronization, we continue using GDP growth data, and we analyze a period of 23 years spanning from 1995Q1 to 2017Q4.

Does the size of economies affect economic cycle cohesion?

Let us now take a closer look at the GDP development of the V4 economies with respect to the European Union (EU-28), the Euro area (EU-12) and their close EU partner – Germany. We highlight Germany as well since its economic relationship with the V4 is strong. Germany's international trade with the V4 countries is larger than the trade of Germany and China. Since the transition period of the V4 countries, we have observed notable differences in GDP growth. Czech Republic's, Slovakia's, and Poland's GDP have grown to more than 400% of their 1995 level, the GDP of Hungary is approximately at its 313%, whereas Germany's and EU-28 GDP increases are only to their 168% and 209%, respectively, see Table 2.1.

Table 2.1: Change of Gross Domestic Products

1995 vs. 2017	EU-28	EU-12	Germany	Czech Rep.	Hungary	Poland	Slovakia
Δ GDP EUR (in %)	209	194	168	421	313	428	544

Note: The values show how the GDPs of given countries have grown between 1995 and 2017, 1995 = 100%.

Hence, considering the GDP differences, the adoption of weights is beneficial for our analysis. From the OCA's point of view, we are interested in the synchronization of economic cycles and the nominal GDP tells by how much power can each of the economies affect the whole system. The nominal GDP takes into account a gravitational attraction of the countries and their contribution to the common cycle.

Common cycles of the Visegrad Four and the EU

Despite many contributions, there is no general consensus to the question of EMU synchronization and the Euro adoption (Aguiar-Conraria & Soares, 2011b; Crespo-Cuaresma & Fernández-Amador, 2013; De Haan et al., 2008). Here, we proceed to study the strength of relationship of the V4 within the EU looking at growth cycle similarities. We focus on the common cycle of seven already integrated (EMU) countries, the EU core, and the V4 countries which are in the process of integration. Moreover, we also look at common cycles of both groups individually.

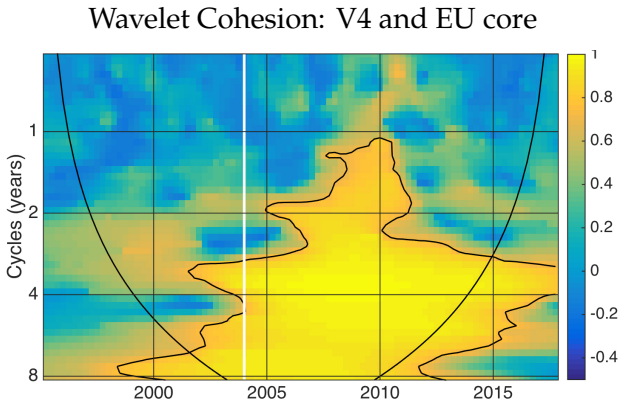

Figure 2.4: Wavelet cohesion of the Visegrad Four and the EU core. The solid black line contours the significant cohesion (95%). The area below the black curve is the cone of influence. The vertical solid white line indicates 2004 – the year of the enlargement of the EU.

Using the wavelet cohesion with time-varying weights, we observe that V4 and EU core are significantly cohesive at cycles corresponding to cycles longer than 2 years from 2005. Surprisingly and contrary to the cohesive business cycles, we see very small common movements at the short-term cycles, 0.5-1 year over the whole sample. The strongest relationship appears during the period after 2002 at cycles of 3 to 8 years, see Fig. 2.4.[19] Knowing the case of Slovakia, which experienced a gradual increase of comovement with EU after the Euro adoption, the high cohesion of V4 and EU core may signal a potential benefit from joining the EMU for the Czech Republic, Poland, and Hungary.

[19]The figures of wavelet cohesion, heatmaps, display the results the same way as those of the wavelet coherence, except that the scale of the cohesion may be negative. Hence, the blue color depicts the negative relationship between economies, which may also be strong.

(a) Positive Cohesion of V4 and EU core (b) Negative Cohesion of V4 and EU core

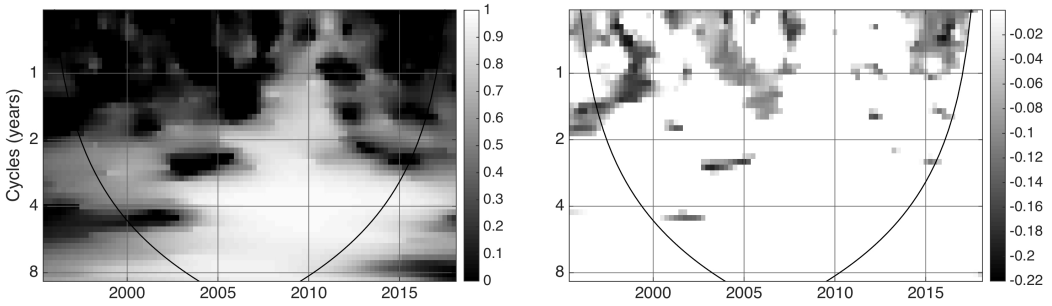

Figure 2.5: Positive (a) and negative (b) cohesion of the Visegrad Four and EU core using time-varying weights in terms of Nominal GDP. The area below the black curve is the cone of influence.

To assess the common cycles of the V4 and EU core in a detailed perspective, we extract the cohesion (Fig. 2.4) into two parts: positive and negative. It indicates that the positive (pro-cyclical) cohesion dominates in this relationship and it is mostly for cycles longer than 2 years, Fig. 2.5-(a). On the contrary, when countries co-move counter-cyclically it is at cycles shorter than 2 years, Fig. 2.5-(b). Hence, the countries have common positive movements at business cycle frequencies, and in short periods they may go in the opposite direction.

(a) Visegrad Four (b) EU core

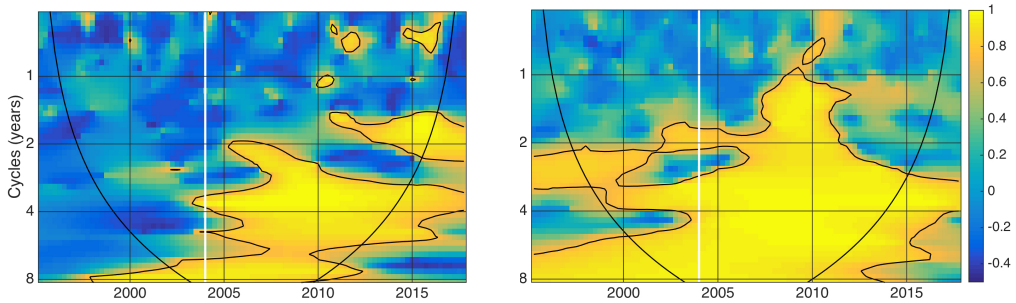

Figure 2.6: Wavelet cohesion of Visegrad Four (a) and EU core (b) countries. The nominal GDP data used as time-varying weights. The solid black line contours the significance level of 5% against the red noise. The area below the black curve is the cone of influence. The vertical solid white line indicates 2004 – the year of the enlargement of the EU.

Further, we support our findings with an analysis of common cycles of both subgroups separately. The degree of synchronization of the V4 is high and pro-cyclical at business cycle frequencies (3-8 year), especially from 2005

to 2015 for 3-6 year cycles, Fig. 2.6-(a). An area of high cohesion also appears at cycles around 2 years and begins in 2011. The short-term outlook provides some insights that Visegrad countries react weakly and counter-cyclically at cycles up to 2 years. Also an overall weak and negative synchronization covers the first part of the sample from 1995 to 2004.

Although we observe similar patterns of comovement over the longer horizons, the overall synchronization of the EU core is much stronger than the one of the V4 countries, Fig. 2.6-(b). This result is in line with Rua and Silva Lopes (2012), who find a high cohesion of the business cycle dynamics of the EU countries. In the second half of the sample, for the EU core, the relationship slightly increases at the shorter cycles.[20]

(a) Frequency Cohesion: 1995–2004 (b) Frequency Cohesion: 2005–2017

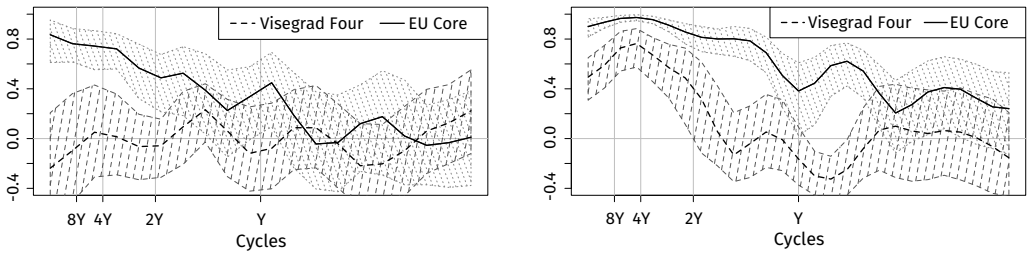

Figure 2.7: Frequency cohesion of the Visegrad Four and EU core countries during two periods: 1995-2004 (a) and 2005-2017 (b). On the x-axis we depict cycle length and label four cyclical components of 1 year, 2 years, 4 years, and 8 years using the gray vertical lines. The longest observed cycles are 10 and 13 years, for Figs. (a) and (b), respectively. Dashed areas are 90% bootstrapped confidence intervals.

Moreover, we complement the analysis with the frequency cohesion. Although it is time-invariant, it helps us depict the very low synchronization of all cyclical components of the V4 prior its accession to the EU. We find almost none existent common cycle among the V4 countries between 1994 and 2004, see Fig. 2.7-(a). On the contrary, the EU core countries significantly co-move at cycles longer than 1.5 years during that period. Looking at the period of 2005-2017, Fig. 2.7-(b), we see a different situation. The cohesion of V4 countries is greater for cycles longer than 2 years. The synchronization of the EU core during 2005-2016 is strong at almost all growth cycles. These

[20] In the appendix Fig. 2.8, we provide complementary results showing cohesion of the EU-12 countries (a) and peripheral countries (b), where both show much lower synchronization than the EU core in the Fig. 2.6.

findings are in line with our previous results from sections 2.5.1 and 2.5.2 of the pairwise coherence. The coherence of the V4 countries is low during 1995-2004 in both situations.

2.6 Concluding remarks and Policy implications

Growth cycle synchronization is a central question of economic integration, and thus it needs a rigorous examination. Adopting wavelet methodology, we have overcome the problems of traditional measures, such as operation in time or frequency domain only and of the necessity of time-series stationarity. In this paper, we have proposed a measure of multivariate comovement using the wavelet cohesion with time-varying weights that allows for precise localization of cyclical comovement.

We have investigated the impact of V4 cooperation, which has one of its main aims to converge faster towards the EU. We have found very low levels of synchronization for the first years of their cooperation, which might be linked to the economic turbulence of the late 90s. The coherence is low for each country paired with Poland, except cycles at 3-5 years period beginning around 2008. Nevertheless, the comovement and cohesion of the V4 economies are strong, particularly after 2005.

Further, we have studied the growth cycle synchronization of the V4 with the EU. The results confirmed some already known interesting patterns. Slovakia's synchronization with the EU was poor before its accession to the EU. However, the relationship gets stronger after 2005, which supports the theory of the endogeneity of the OCA and the adoption of Euro. We have revealed that the highest coherence is between EU and both the Czech Republic and Hungary beginning in 2001. This might imply readiness of these countries for the Euro adoption considering one of the features – the coherence of growth cycles. By contrast, the degree of synchronization of the business cycles of Poland and EU is the lowest among V4.

Employing wavelet cohesion with time-varying weights, we have uncovered relationships in both time and frequency domains for multiple time-series. Regarding the V4 its position within the EU, we have shown strong pro-cyclical behavior at cycles longer than 2 years. Concerning the EU core countries, we show that there is a weak synchronization of short-term dynamics. Conversely, we have demonstrated that the EU is highly cohesive

at longer economic cycles, such as business cycle frequencies at 3-8 years. Finally, we have found high comovement of the business cycle frequencies of the V4 and the EU core countries for the sample period when the V4 countries have been part of the EU. The similar growth cycles reactions to exogenous shocks of the V4 and the EU may be a relevant feature for further consideration of joining the monetary union. Higher the cohesion of the EU and its members, more efficient and coherent all policies might be.

References

Aguiar-Conraria, L., Martins, M. M., & Soares, M. J. (2018). Estimating the taylor rule in the time-frequency domain. *Journal of Macroeconomics*, *57*, 122–137.

Aguiar-Conraria, L., Azevedo, N., & Soares, M. J. (2008). Using wavelets to decompose the time–frequency effects of monetary policy. *Physica A: Statistical Mechanics and its Applications*, *387*(12), 2863–2878.

Aguiar-Conraria, L., & Soares, M. J. (2011a). Business cycle synchronization and the Euro: A wavelet analysis. *Journal of Macroeconomics*, *33*(3), 477–489.

Aguiar-Conraria, L., & Soares, M. J. (2011b). *The continuous wavelet transform: A primer* (tech. rep. No. 16/2011). NIPE-Universidade do Minho.

Aguiar-Conraria, L., & Soares, M. J. (2014). The continuous wavelet transform: Moving beyond uni- and bivariate analysis. *Journal of Economic Surveys*, *28*(2), 344–375.

Antal, J., Hlaváček, M., & Holub, T. (2008). Inflation target fulfillment in the Czech Republic in 1998–2007: Some stylized facts. *Czech Journal of Economics and Finance (Finance a uver)*, *58*(09-10), 406–424.

Artis, M. J., Marcellino, M., & Proietti, T. (2004). *Characterizing the business cycle for accession countries* (tech. rep. No. 4457). C.E.P.R. Discussion Papers.

Backus, D. K., Kehoe, P. J., & Kydland, F. E. (1992). International real business cycles. *Journal of Political Economy*, *100*(4), 745–775.

Baxter, M., & King, R. G. (1999). Measuring business cycles: Approximate band-pass filters for economic time series. *The Review of Economics and Statistics*, *81*(4), 575–593.

Bekiros, S., Nguyen, D. K., Uddin, G. S., & Sjö, B. (2015). Business cycle (de) synchronization in the aftermath of the global financial crisis: Implications for the euro area. *Studies in Nonlinear Dynamics & Econometrics*, *19*(5), 609–624.

Berkowitz, J., & Diebold, F. X. (1998). Bootstrapping multivariate spectra. *The Review of Economics and Statistics*, *80*(4), 664–666.

Bruzda, J. (2011). Business Cycle Synchronization According to Wavelets-the Case of Poland and the Euro Zone Member Countries. *Bank i Kredyt*, (3), 5–31.

Cazelles, B., Chavez, M., Berteaux, D., Ménard, F., Vik, J. O., Jenouvrier, S., & Stenseth, N. C. (2008). Wavelet analysis of ecological time series. *Oecologia*, *156*(2), 287–304.

Crespo-Cuaresma, J., & Fernández-Amador, O. (2013). Business cycle convergence in EMU: A first look at the second moment. *Journal of Macroeconomics*, *37*, 265–284.

Croux, C., Forni, M., & Reichlin, L. (2001). A measure of comovement for economic variables: Theory and empirics. *The Review of Economics and Statistics*, *83*(2), 232–241.

Crowley, P. M. (2007). A guide to wavelets for economists. *Journal of Economic Surveys*, *21*(2), 207–267.

Crowley, P. M., & Hallett, A. H. (2015). Great moderation or "Will o' the Wisp"? A time–frequency decomposition of GDP for the US and UK. *Journal of Macroeconomics*, *44*, 82–97.

Crowley, P. M., Maraun, D., & Mayes, D. (2006). How hard is the euro area core?: An evaluation of growth cycles using wavelet analysis. *Bank of Finland Research Discussion Papers*, (No. 18/2006).

Darvas, Z., & Szapáry, G. (2008). Business cycle synchronization in the enlarged EU. *Open Economies Review*, *19*(1), 1–19.

Daubechies, I. (1992). *Ten lectures on wavelets* (Vol. 61). SIAM.

De Haan, J., Inklaar, R., & Jong-A-Pin, R. (2008). Will business cycles in the euro area converge? A critical survey of empirical research. *Journal of Economic Surveys*, *22*(2), 234–273.

Ferreira-Lopes, A., & Pina, Á. M. (2011). Business cycles, core, and periphery in monetary unions: Comparing europe and north america. *Open Economies Review*, *22*(4), 565–592.

Fidrmuc, J., & Korhonen, I. (2006). Meta-analysis of the business cycle correlation between the euro area and the CEECs. *Journal of Comparative Economics*, *34*(3), 518–537.

Franke, J., & Hardle, W. (1992). On bootstrapping kernel spectral estimates. *The Annals of Statistics*, 121–145.

Frankel, J. A., & Rose, A. K. (1998). The Endogeneity of the Optimum Currency Area Criteria. *Economic Journal*, *108*(449), 1009–25.

Gabor, D. (1946). Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, *93*(26), 429–441.

Ge, Z. (2008). Significance tests for the wavelet cross spectrum and wavelet linear coherence. *Annales Geophysicae*, *26*(12), 3819–3829.

Grigoraş, V., & Stanciu, I. E. (2016). New evidence on the (de) synchronisation of business cycles: Reshaping the European business cycle. *International Economics*, *147*, 27–52.

Grinsted, A., Moore, J. C., & Jevrejeva, S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics*, *11*(5/6), 561–566.

Harding, D., & Pagan, A. (2002). Dissecting the cycle: A methodological investigation. *Journal of Monetary Economics*, *49*(2), 365–381.

Hodrick, R. J., & Prescott, E. C. (1997). Postwar US business cycles: An empirical investigation. *Journal of Money, Credit, and Banking*, 1–16.

Christiano, L. J., & Fitzgerald, T. J. (2003). The band pass filter. *International Economic Review*, *44*(2), 435–465.

Jagrič, T. (2002). Measuring business cycles–a dynamic perspective. *Banka Slovenije, Prikazi in analize X/1, Ljubljana*.

Jiménez-Rodríguez, R., Morales-Zumaquero, A., & Égert, B. (2013). Business cycle synchronization between Euro Area and Central and Eastern European countries. *Review of Development Economics*, *17*(2), 379–395.

Kolasa, M. (2013). Business cycles in EU new member states: How and why are they different? *Journal of Macroeconomics*, *38, Part B*, 487–496.

Kutan, A. M., & Yigit, T. M. (2004). Nominal and real stochastic convergence of transition economies. *Journal of Comparative Economics*, *32*(1), 23–36.

Maraun, D., Kurths, J., & Holschneider, M. (2007). Nonstationary gaussian processes in wavelet domain: Synthesis, estimation, and significance testing. *Physical Review E*, *75*(1), 016707.

Mundell, R. A. (1961). A theory of optimum currency areas. *The American Economic Review*, *51*(4), 657–665.

Nason, G. P., Von Sachs, R., & Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(2), 271–292.

OECD. (2018). Quarterly GDP (indicator). *April 2018*.

Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, *27*(2), 204–237.

Raihan, S. M., Wen, Y., & Zeng, B. (2005). Wavelet: A new tool for business cycle analysis. *Federal Reserve Bank of St. Louis Working Paper Series*, (2005-050).

Rua, A. (2010). Measuring comovement in the time-frequency space. *Journal of Macroeconomics*, *32*(2), 685–691.

Rua, A., & Silva Lopes, A. (2012). Cohesion within the euro area and the US: A wavelet-based view. *Banco de Portugal, Working Paper*, (4).

Schüler, Y. S., Hiebert, P., & Peltonen, T. A. (2017). Coherent financial cycles for g-7 countries: Why extending credit can be an asset [Working Paper Series]. *ESRB Working Papers*.

Torrence, C., & Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, *79*(1), 61–78.

Vacha, L., & Barunik, J. (2012). Co-movement of energy commodities revisited: Evidence from wavelet coherence analysis. *Energy Economics*, *34*(1), 241–247.

Yogo, M. (2008). Measuring business cycles: A wavelet analysis of economic time series. *Economics Letters*, *100*(2), 208–212.

Appendix

2.A Supplementary text and figures

2.A.1 Wavelet coherence and phase difference

For the bivariate analysis in Section 2.5.2, we employ the wavelet coherence. Following Torrence and Compo (1998) and Grinsted et al. (2004), let us state the wavelet coherence definition, which can be viewed as a local linear correlation of two time-series, x_i and x_j as:

$$R^2(\tau,s) = \frac{|S(s^{-1}W_{x_ix_j}(\tau,s))|^2}{S(s^{-1}|W_{x_i}(\tau,s)|^2)\cdot S(s^{-1}|W_{x_j}(\tau,s)|^2)}, \quad R^2 \in [0,1], \tag{2.5}$$

where $W_{cdot}(\tau,s)$ is a particular wavelet (co-)spectrum, S is a smoothing function as $S(W) = S_{scale}(S_{time}(W_n(s)))$ (Grinsted et al., 2004; Torrence & Compo, 1998). We refer a reader interested in the properties of the wavelet transform and wavelet spectrum to consult Grinsted et al. (2004) and Torrence and Compo (1998). Since a wavelet power spectrum containing non-reliable estimates at the beginning and end of the data. The data at the edges are treated by padding both ends with a sufficient number of zeros. The area affected by zero-padding is called the cone of influence (COI). Further, we use the Morlet wavelet, and thus, the COI is e^{-2}-folding.

To obtain coherences and phases, we adjust and use the package developed by Grinsted et al. (2004). Since the wavelet coherence has only positive values, we use the phase difference to describe the sign and position of two countries' relationship.

$$\phi_{x_i,x_j} = \tan^{-1}\left(\frac{\Im\{W_{x_ix_j}(\tau,s)\}}{\Re\{W_{x_ix_j}(\tau,s)\}}\right), \quad \phi_{x_i,x_j} \in [-\pi,\pi], \tag{2.6}$$

where $\Im\{W_{x_ix_j}(\tau,s)\}$ and $\Re\{W_{x_ix_j}(\tau,s)\}$ are the imaginary and real parts of a cross wavelet transform, respectively. The two time-series are positively correlated if $\phi_{x_i,x_j} \in [-\pi/2,\pi/2]$, otherwise the correlation is negative. Moreover, the first variable, x_i, leads the second, x_j, if the phase is in $[0,\pi/2]$ and $[-\pi,-\pi/2]$; when in $[-\pi/2,0]$ and $[\pi/2,\pi]$, the second variable is leading.

The significance of both measures is obtained using Monte Carlo methods. We follow Torrence and Compo (1998) to assess the statistical significance,

which is depicted in figures as a black contour and the 5% significance level. Aguiar-Conraria and Soares (2014) indicate there are no good statistical tests for the phase difference. The significance of the phase should be connected with the significance of the power spectrum or coherence (Ge, 2008). That said, we obtain the confidence intervals using bootstrap techniques. We add 5% noise to each analyzed series and do a Monte Carlo study, which provides us the 95% confidence interval of the phase difference.

2.A.2 Additional results of the EU cohesion

In this section, we present complementary results to our analysis in Section 2.5.3 regarding the common cycles within the European Union. We showed a very strong cohesion of the EU core and here we show the appearance of the cohesion of the EU core and the group of peripheral countries (PT, IR, IT, GR, ES), Fig 2.8. Countries such as Portugal, Ireland, Italy, Greece, and Spain show low degree of synchronization (Fig.2.8-(b)), except 4 years cycle from 1995 until 2010. When we join the EU core group with the peripheral countries into the former EU-12, we observe this common 4 years cycle still strong for whole sample. Further, the EU-12 shows synchronization around the crisis period, 2005 – 2010, for cycles from 1 to 8 years. Nevertheless, enlarging the EU core group by the five peripheral countries gives us weaker group towards which the V4 countries aim to integrate and be compared.

(a) Cohesion of the EU-12 (b) Cohesion of the peripheral group

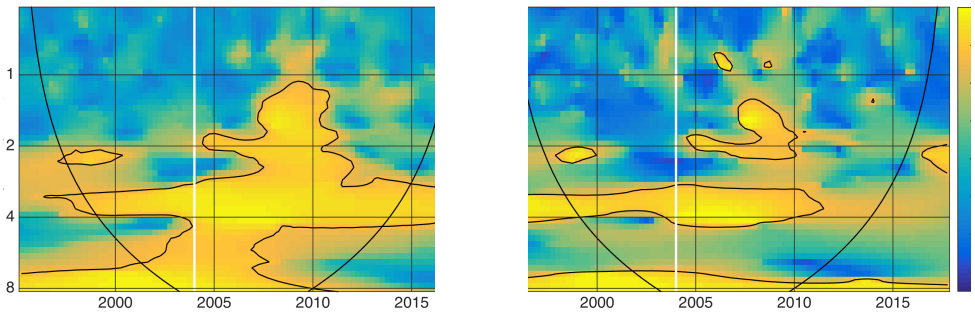

Figure 2.8: Wavelet cohesion of the EU-12 (a) and peripheral group of countries (b). The nominal GDP data used as time-varying weights. The solid black line contours the significance level of 5% against the red noise. The area below the black curve is the cone of influence. The vertical solid white line indicates 2004 – the year of the enlargement of the EU.

2.A.3 Real wavelet-based measure of comovement demonstrations

We show the usefulness of the wavelet-based measure (Eq. 2.3) when studying a comovement of a white noise (WN), $u_t \sim N(0,1)$ and its lagged values, u_{t-1}, u_{t-4}, and u_{t-8}. For the first 200 observations in Fig. 2.9-(a), we see the

(a) comovement of WN and its lags (b) comovement of WN and its random walk

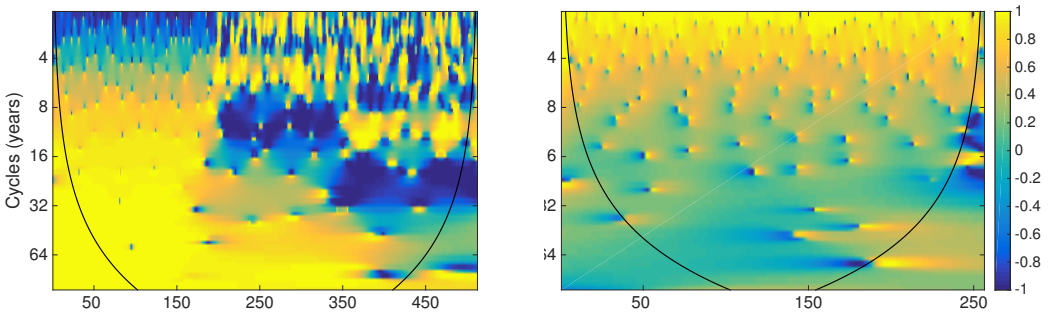

Figure 2.9: Real wavelet-based measure of comovement for two series: (a) $a_t = u_t$ for $t = [0, 511]$; and $b_t = u_{t-1}$ for $t = [0, 200]$, $b_t = u_{t-4}$ for $t = [201, 350]$, and $b_t = u_{t-8}$ for $t = [351, 511]$; (b) $a_t = u_t$ and $b_t = b_{t-1} + u_t$.

negative correlation equal to minus one in the shortest cycles, which changes to a positive correlation equal to one in the long-term. If we averaged this part of Fig. 2.9-(a) over time, the obtained result would be same as the dynamic correlation of Croux et al. (2001), which follows a curve that maps $(1, -1)$ on $(0, \pi)$ frequencies. In the second part of Fig. 2.9-(a), the series have more lags, u_{t-4} and u_{t-8}, whose relationship is negative with the original u_t at longer horizons, which demonstrates the possibility of the well-localized information in the time-frequency plane. In Fig. 2.9-(b), we plot the dynamic correlation using wavelet-based measure of the white noise, u_t, with its cumulative sum. Contrary to Fig. 2.9-(a) for lagged noises, these two series are positively correlated in the short-term cycles and not-correlated in the long-term.

Chapter 3

Identification Persistence in Macroeconomic Responses

Abstract

In our study, we consider a new approach to quantify the effects of economic shocks on monetary transmission. We analyse the widely known phenomenon of price puzzle in a time-varying environment using the frequency decomposition. We use the frequency response function to measure the power of shocks transferred to different economic cycles. Considering both time and frequency domains, we quantify the dynamics of shocks implied by monetary policy within an economic system. While studying the monetary policy transmission of the U.S., the empirical evidence shows that low-frequency cycles of output are prevalent and have positive transfers. Examination of the inflation reveals that the frequency responses vary significantly in time and alter the direction of transmission for all cyclical lengths.

This chapter was co-authored with Lukáš Vácha. Previous version is part of the IES Working Papers (2018). The current version is a work in progress in its pre-final stage for the pre-defense. We thank to seminar participant at ISCEF2016 Paris, France, FMND2017, Paris, France, CFE 2017, Sevilla, Spain. Luboš Hanus gratefully acknowledges the grant support provided by Grant Agency of Charles University (GAUK) no. 1390218.

3.1 Introduction

Considerable research has been conducted on the transmission of monetary policy, with the aim to explain the interaction between economic variables. A particular increase in interest in monetary policy transmission is linked to the study by Sims (1980), by which the vector autoregression (VAR) models became fundamental to contemporary macroeconomic studies (Castelnuovo & Surico, 2010; Sims, 1992). The VARs, together with their impulse response functions (IRFs), are prominent tools in the analysis of macroeconomics. Impulse response functions have proliferated the phenomenon known as the "price puzzle": an increase in prices follows a monetary tightening, which has been found in many studies that use a VAR to monetary policy (Rusnák et al., 2013). Naturally these numerous attempts to properly quantify the monetary policy responses have varied dramatically in terms of estimation modifications, identification schemes, and data used and led to multiple results and implications. Therefore, the final explanation is still uncertain and there is a room to approach this story from a different perspective.

Our contribution is to look at the monetary policy transmission across economic cycles – from the perspective of persistence. Economists using the traditional impulse response analysis are unsuccessful in quantifying the transmission mechanism across horizons. Studying economic cycles is particularly important for nations that use a short-term interest rate to control inflation over the medium term, such as the US and those implementing inflation-targeting policies. While the impulse response functions handle the transmission from a time-domain perspective, we decompose responses of structural monetary policy and measure the power of shocks transferred to different economic cycles using frequency domain (Dew-Becker & Giglio, 2016). We can represent an economic model as a linear combination of an infinite number of shocks, which is the way we can identify the persistence of a macroeconomic variable responses. We decompose these shocks using the Fourier transformation, which identifies the persistence and economic cycles, instead of cumulative responses of the well-established IRFs.

In order to quantify the frequency dependency of economic dynamics, we need to estimate the structure of a model. Further, to account for a non-negligible time-variation of macroeconomic variables, the literature recognises the importance of identifying the evolving structure of economies (Cogley & Sargent, 2005; Koop et al., 2009; Primiceri, 2005). Authors

have estimated structural models using rolling windows. However, the use of small subsamples in rolling windows may result in outlier observations. Conversely, if the window is too wide, it may not capture enough variability (Geraci & Gnabo, 2018). Therefore, we opt for a time-varying parameter vector autoregressive model (TVP-VAR) to capture the evolution in macroeconomic dynamics. Furthermore, we employ a structural vector autoregressive model in which structural responses are identified using sign restrictions.

Our findings demonstrate that the effects of monetary policy transmission on macroeconomic variables vary in time, frequency, and intensity. We use frequency response functions to describe the impact of monetary policy on inflation and output. It is also shown that, even when controlling for the negative impulse response of inflation to monetary policy, the response is both negative and positive. Furthermore, cyclical behaviour is quantified using time-frequency correlation, which captures the dependence between output, inflation, and interest rates. This correlation is relatively weak in short cycles and stronger in business cycles. Localised representations are employed to allow for time and frequency variations.

The literature on the vector autoregressive model and monetary policy is extensive, with multivariate modelling dating back several decades (Sims, 1980, 1992). However, the proper identification of shocks in models using zero restrictions Christiano et al. (1999) or sign-restriction identification schemes Rubio-Ramirez et al. (2010) and Uhlig (2005) or a combination remains an ongoing challenge. We focus our monetary policy analysis, particularly on the responses of endogenous variables as the output, prices, and the interest rate. This analysis focuses on monetary policy, specifically on the responses of endogenous variables such as output, prices, and interest rates. The increase in prices after monetary tightening, known as the "price puzzle", has been a long-term research subject since Sims (1992). However, Rusnák et al. (2013) compared over 70 studies and found that the price puzzle is more likely to result from model misspecification than from price behaviour in the economy. There are various methods for addressing this issue. For instance, in a recent study, Koop et al. (2009) compare time-variant and time-invariant settings of the Primiceri (2005), and they found a strong preference for a model that allows for both time-varying coefficient and covariance matrix. Alternatively, Baumeister and Hamilton (2015) recommend that the researcher should acknowledge their prior beliefs about structural conclusions. The results suggest a Bayesian vector autoregression. There-

fore, we will incorporate both modelling features and use a conventional TVP-VAR setup with stochastic volatility and sign restrictions.

The time-series filtering is closely related to the frequency domain. This text aims to work with frequency response functions as counterparts to IRFs. To begin, we refer to the conceptual works of Baxter and King (1999), Christiano and Fitzgerald (2003), and Murray (2003), where readers can find spectral domain[1] representations of economic variables, along with filters used to define cyclical components. The tool we wish to highlight is the frequency response function,[2] which provides a way to filter data or quantify how the data has been filtered (Dew-Becker & Giglio, 2016). We find potential in this interpretation and employ frequency responses to attempt to quantify information in macroeconomic dynamics. The dynamics can be studied through exogenous shocks to the economy and depicted in both time and frequency domains. Additionally, we use spectral measures such as dynamic correlation to examine cyclical dependencies between monetary variables.

Additionally, Gehrke and Yao (2017) examine the role of supply shocks in real exchange rates by making another connection between frequency domain and vector autoregression modelling. To identify structural shocks, they use a structural VAR model with sign restriction. Their spectral variance decomposition (SVD) shows the importance of productivity shock in the persistence of the real exchange rate. Ellington (2018) uses TVP-VAR and time-frequency coherence to study the dynamics between monetary variables with Divisia money. Geraci and Gnabo (2018) employ Bayesian TVP-VAR and define time-varying spillovers, which is a new concept in financial literature. Geraci's (2016) time-varying work also relates to the rolling window frequency connectedness measure of Baruník and Křehlík (2018). Lovcha and Perez-Laborda (2018) employ a spectral matrix of the reduced-form model in the frequency domain Whittle log-likelihood function to estimate a fractionally integrated VAR for monetary policy.

The structure of the text is as follows. Section 3.2 begins with relevant methodological approaches. We introduce today's traditional time-varying vector autoregression framework. We present its localised moving average representation. Finally, we propose related frequency dependent measures.

[1] The names "spectral" and "frequency" domain are interchangeable.

[2] In this text, we continue to work the term "impulse transfer function", which is equivalent to the "frequency response function".

In section 3.3.1 we present the data and the estimation procedure. Section 3.4 reports the results and the last section concludes.

3.2 Methodology

The study of the dynamics of an economic system and cyclical responses to monetary policy requires an estimation framework. As the literature has shown that economic systems evolve in time, e.g. (Canova & Gambetti, 2009), we build on the time-varying parameters vector autoregressive model of Cogley and Sargent (2005) and Primiceri (2005), which allows for the capture of gradual evolution parameters. Furthermore, we use the specification of Benati and Mumtaz (2007), who, among other things, introduced the model using the identification scheme of sign restrictions of Rubio-Ramirez et al. (2010).

First, we define the model with time-varying parameters and stochastic volatility, for which we introduce the central concept of localised frequency response functions. The concept relates to the traditional measure of impulse responses, however, it is decomposed into frequencies corresponding to economic cycles that relate to the monetary policy. Second, we present spectral measures that capture the comovement between variables from the reduced form of the model. The time-frequency representation of the correlation measure is informative because it provides detailed information about the different length of business cycles. Finally, we describe the data and estimation strategy.

3.2.1 TVP-VAR framework

Since our aim is to capture time-frequency dynamics, we consider a time-varying parameters VAR with stochastic volatility. We work with TVP-VAR of n variables and k lags:

$$y_t = B_{0,t} + B_{1,t}y_{t-1} + \cdots + B_{k,t}y_{t-k} + u_t, \text{ for } t = 1, ..., T, \tag{3.1}$$

$$= X_t'\theta_t + u_t, \tag{3.2}$$

where y_t is an $n \times 1$ vector of endogenous variables, $B_{0,t}$ are $n \times 1$ time varying intercepts, $B_{i,t}$ are $n \times n$ matrices of time-varying coefficients for $i = 1, ..., k$ lags, and $u_t \sim N(0, \Omega_t)$ are unobservable shocks with time-

varying covariance matrix Ω_t. The specification of the model follows Cogley and Sargent (2005) and Primiceri (2005) and Benati and Mumtaz (2007). It is assumed that parameters θ_t evolve as a drift-less random walk constrained by $p(\theta_t|\theta_{t-1}, Q) = I(\theta_t)f(\theta_t|\theta_{t-1}, Q)$, where $I(\theta_t)$ is an indicator function that rejects unstable draws and $f(\theta_t|\theta_{t-1}, Q) \sim N(\theta_t, Q)$ (Cogley & Sargent, 2005). Hence, parameters θ_t follow

$$\theta_t = \theta_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, Q). \tag{3.3}$$

The covariance matrix

$$Var(u_t) = \Omega_t = A_t^{-1} H_t (A_t^{-1})' \tag{3.4}$$

can be factorized such that $u_t = A_t^{-1} H_t^{\frac{1}{2}} \zeta_t$, with $\zeta_t \sim N(0, I)$. Further, the matrix A_t is a lower triangular matrix with elements $\alpha_t = [\alpha_{2,1}, \alpha_{3,1}, \alpha_{3,2}]$ below diagonal that depict contemporaneous relations and evolve as a random walk

$$\alpha_t = \alpha_{t-1} + \xi_t, \quad \xi_t \sim N(0, S). \tag{3.5}$$

The matrix H_t is diagonal such that

$$H_t = \begin{pmatrix} h_{1t} & 0 & 0 \\ 0 & h_{2t} & 0 \\ 0 & 0 & h_{3t} \end{pmatrix}, \tag{3.6}$$

$$\log(h_{i,t}) = \log(h_{i,t-1}) + \eta_{i,t}, \quad \eta_t \sim N(0, W). \tag{3.7}$$

and its elements on the diagonal evolve as a geometric random walk. All innovations of in the model form a matrix S, which is a block diagonal. And all the innovations that are serially uncorrelated follow a joint normal distribution with the matrix S with the elements $\text{Var}([u_t,\ \varepsilon_t,\ \xi_t,\ \eta_t]^T)$. Since, we follow the specification of Benati and Mumtaz (2007) we refer a reader to consult Benati and Mumtaz (2007) and Del Negro and Primiceri (2015) for further specification of priors and calibration of the models.

3.2.2 Local moving average representation

Considering the time-varying vector autoregresive process stated in (3.2), we might re-write it into a local MA representation (Canova & Gambetti, 2009;

Dahlhaus, 1997) such that $y_{t|T}$ for $t = 1, \ldots, T$ is

$$y_{t|T} = \sum_{h=-\infty}^{\infty} \Psi_{t|T}(h)\varepsilon_{t-h}, \tag{3.8}$$

where $\Psi_{t|T}(h)$ is a $n \times n \times h$ matrix of parameters localized at $t|T$, satisfying condition $\sup_{t|T} |\Psi_{t|T}(h)| \leq \frac{K}{l(h)}$, such functions $\Psi(h) : (0, 1] \to \mathbb{R}$ exist, and ε_t are independent and identically distributed with $\mathbb{E}\varepsilon_t = 0$ and $\mathbb{E}\varepsilon_t^2 = 0$. Having this time-varying MA representation, we can find its time-varying spectral representation (Dahlhaus, 1997)

$$y_{t|T} = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \exp(\mathrm{i}\omega t) G_{t|T}(\omega) \mathrm{d}\xi(\omega) \tag{3.9}$$

with

$$G_{t|T}(\omega) := \sum_{h=-\infty}^{\infty} \Psi_{t|T}(h) \exp(-\mathrm{i}\omega h) \tag{3.10}$$

where $\xi(\omega)$ is a process with mean 0 and orthonormal increments (Dahlhaus, Polonik, et al., 2009).

For our approch, Eq. (3.10) is the key. We can define time-varying spectral density

$$f_{t|T}(\omega) := \frac{1}{2\pi} |G_{t|T}(\omega)|, \tag{3.11}$$

and more importantly, the functions $G_{t|T}(\omega)$ are frequency representations of the $MA(\infty)$ coefficients, in other words, the (local) time-varying frequency response functions.

3.2.3 Frequency response function

We look at an economic system as a filter formed by economic variables, which carries on the transmission of new information from one variable to the other. To quantify the effects transmitted through such a system, we propose to use the apparatus known in the filtering domain as the impulse transfer function (ITF) or frequency response function (FRF).[3]

First, let us define the frequency response function of the system in Eq. (3.2) and (3.8). The moving average representation is important since the coefficient matrices, $\Psi_{t|T}(h)$, capture the dynamics of the system. At one locat

[3]In this section, we define the impulse transfer function (ITF) because of the filtering literature. And further, we equivalently re-label it as frequency response function (FRF), which is going to be used throughout the text.

time point, $t|T$, the frequency response function is a spectral representation of the $\Psi_{h\to\infty}$ coefficients

$$\Psi(e^{-i\omega}) = \sum_h \Psi_h e^{-i\omega h}, \tag{3.12}$$

where h corresponds to infinite lags of the $MA(\infty)$ representation. This is equivalent to Eq. (3.10).[4]

Traditionally, the VAR provides to researchers impulse response functions to study the propagation of shock for a given time horizon. The frequency response function[5] is a spectral counterpart of the impulse response function and provides a measure of persistence. Hence, in the structural analysis, we define the time-varying frequency response function of i^{th} variable to a shock to the j^{th} variable as

$$G_{t|T,ij}(\omega) = \sum_h \Psi_{t|T,ij}(h) e^{-i\omega h},$$

Since, we work with real economic variables, we have $\Psi_{ij} = 0$ for the horizon $h < 0$, then the frequency response functions, $FRF_{t|T.ij}(\omega)$, can be written as the real part of the Fourier transform of the MA coefficients (Dew-Becker & Giglio, 2016),

$$FRF_{t|T,ij}(\omega) = Re\left(\sum_h \Psi_{t|T,ij}(h) e^{-i\omega h}\right) = \sum_{h=0}^{\infty} \cos(\omega h)\Psi_{t|T,ij}(h) \tag{3.13}$$

The frequency response measures how the filter (coefficients) processes the information through itself at each frequency. In the time-series analysis, frequencies are mostly viewed as economic cycles of different lengths. This allows us to obtain economic cycles of interest corresponding to particular frequencies. Cyclical averages are used over a subinterval of frequencies, $[\omega_1, \omega_2]$, which correspond to cycles from 2 to 8 years, for example. The frequencies, ω, are defined on the interval of $[0, \pi]$.

Spectral representation of a time-series process known as the power spectrum is a common measure to understand distribution of variance across frequencies, defined in Eq. (3.11). Above all, for a general reader, the transfer

[4]We approximate the coefficient matrices for sufficient number of lags H allowing the calculation of the frequency response functions.

[5]Filtering techniques use the power transfer function, or equivalently the gain function, $H(\omega) = |\Psi(e^{-i\omega})|^2$.

function $\Psi(e^{-i\omega h})$ forms the power spectrum of a VAR model (Stiassny, 1996), such that

$$f(\omega) = \Psi(e^{-i\omega h})\Omega\Psi'(e^{+i\omega h}), \tag{3.14}$$

where Ω is the covariance matrix. We present similar representation of the power spectrum of reduced form TVP-VAR in Section 3.2.5.

3.2.4 Examples of frequency responses

We demonstrate the relationship between the impulse response function and the frequency response function in Fig 3.1. We use two shocks with transitory and two with persistent effects.

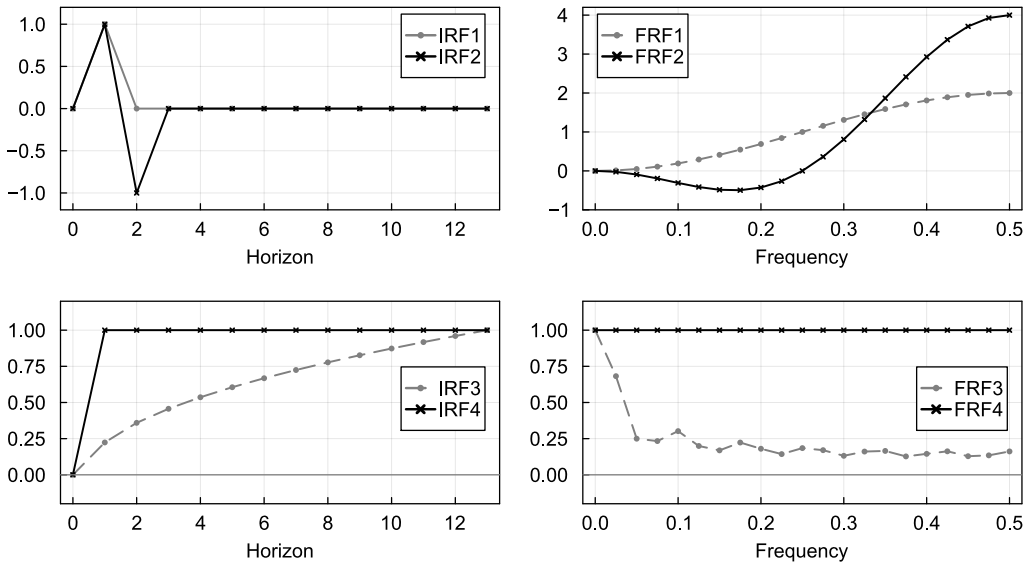

Figure 3.1: Illustration of impulse and frequency responses
We show the impulse response functions on the left and corresponding frequency response functions on the right. The top row shows transitory impulse responses with a stronger frequency response at higher frequencies. The bottom row shows persistent responses with a strong frequency response at all frequencies (FRF4) and close to lower frequencies (FRF3), indicating long cycles.

In Figure 3.1 (left), the impulse response 1 (IRF1) reacts to a shock of size 1 at time $h = 0$ and size of -1 at time $h = 1$, which translates the shock into IRF1 values of 1 and 0 a times $h = 1$ and $h = 2$. A reader might recognise the representation of the first difference filter of the time domain. In the frequency domain, it filters out the information at the lowest frequency

components or the longest cycles. IRF2 reacts to a different shock that lasts longer, specifically three periods. In the frequency domain, this shock has a quadrupling effect on the highest frequencies (at frequency 0.5), reduces the low frequencies, and applies negative weight to frequencies approximately between 0.1 and 0.2.

In contrast, impulse responses 3 and 4 have a more persistent effect and do not return to zero. IRF3 experiences the shock fully upon impact, and its frequency response is the same at all frequencies (all cycle lengths). IRF4 reacts gradually to the impact before the shock remains persistent. Therefore, its reaction is lower at higher frequencies (shorter cycle lengths) and largest at the lowest frequency (longest cycle).

The benefit of frequency responses, in comparison to IRFs, is that using IRFs we only observe reactions to shocks and their evolution in time for a given horizon $1, \ldots, H$. However, we do not distinguish the real not-cumulative transfer for given economic cycles. The frequency responses provide insights upon frequency-specific decomposed information of the effects of shocks on the system transmission.

3.2.5 Spectral dependent measures

For further analysis of frequency dependent dynamics, we describe the time-varying (temporary) spectrum following Cogley and Sargent (2005) and Primiceri (2005) of the reduced form model such as

$$f_{t|T}^{(ii)}(\omega) = s_i(I - B_{t|T}e^{-i\omega})^{-1}\frac{\Omega_{t|T}}{2\pi}((I - B_{t|T}e^{+i\omega})^{-1})'s_i' \tag{3.15}$$

where s_i is selecting vector of variable i, $B_{t|T}e^{-i\omega} = B_{t|T,p=1}e^{-i\omega 1} + \cdots + B_{t|T,p=k}e^{-i\omega k}$, $B_{t|T}e^{+i\omega}$ respectively, and $\Omega_{t|T}$ is a time-varying reduced-form VAR covariance matrices. Studies in the literature use the temporal spectral density to access the persistence of variables, which is studied around frequency zero, for example, in Cogley and Sargent (2005).

We study connections between variables at different frequencies that are known in the economic literature. Traditionally, the link is assessed via correlations coefficient on filtered series to see the dependence at different horizons, for example. Croux et al. (2001) provide frequency-dependent measures such as a dynamic correlation and coherence. For instance, to study the comovement between output and inflation using the coherence

measure, which depicts strength of the relationship. The time-varying coherence based on the temporal spectra is defined (Ellington, 2018; Mumtaz & Sunder-Plassmann, 2013) as

$$\hat{h}_{t|T}^{(ij)}(\omega) = \sqrt{\frac{Re\{\hat{f}_{t|T}^{(ij)}(\omega)\}^2 + Im\{\hat{f}_{t|T}^{(ij)}(\omega)\}^2}{\hat{f}_{t|T}^{(ii)}(\omega)\hat{f}_{t|T}^{(jj)}(\omega)}}, \tag{3.16}$$

where $Re\{\}$ is the real part of the spectrum (co-spectrum), $Im\{\}$ is the imaginary part of the spectrum (quadrature spectrum), and hence the dynamic coherence $\hat{h}_{t|T}^{(ij)}(\omega) \in [0,1]$. Similar type of coherence measure is nowadays often in time-frequency (wavelet) analysis to study comovement between variables.[6] However, the coherence measure depicts only the strength not the direction of the relationship, thus we present the time-frequency correlation. Accordingly to the time-invariant dynamic correlation of Croux et al. (2001), we write

$$\rho_{t|T}^{(ij)}(\omega) = \frac{Re\{\hat{f}_{t|T}^{(ij)}(\omega)\}}{\sqrt{\hat{f}_{t|T}^{(ii)}(\omega)\hat{f}_{t|T}^{(jj)}(\omega)}}, \tag{3.17}$$

which range is as for the traditional correlation measure, such that $-1 \leq \rho_{ij,t|T}^{(ii)}(\omega) \leq 1$ for all ω. The local correlation provides unconditional prospects to see relationship between the output, inflation, and interest rate at given time, t, and frequency horizon, ω. Further, we use the time-frequency correlation averaged at given frequency band. Such a correlation averages values of $\rho_{t|T}^{(ij)}(\omega_{band})$ for given $\omega_{band} \in [\omega_1, \omega_2]$, where ω_1, ω_2 correspond to desired frequencies.

3.3 Data and estimation

3.3.1 Data

We use the data standard for the literature (Canova & Gambetti, 2009; Cogley & Sargent, 2005; Koop et al., 2009; Primiceri, 2005). For the analysis, we employ the output growth of real GDP, GDP deflator inflation, and the effective Federal Funds rate, Figure 3.2. The data are at the quarterly frequency and

[6]For example, Aguiar-Conraria et al. (2008) and Ramsey and Lampart (1998)

spans from 1955:Q3 to 2018:Q2.[7] Altough, th data set is small, it provides a sufficient information to the monetary policy analysis and the frequency measures we focus on.

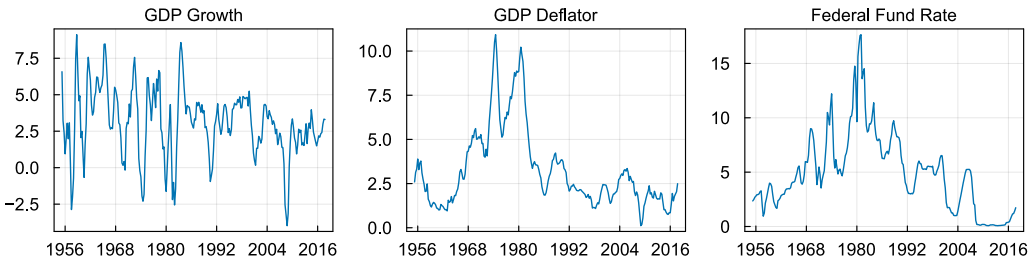

Figure 3.2: US macroeconomic data spanning 1955:Q3–2018:Q2.

3.3.2 Estimation

We estimate the TVP-VAR model in line with Benati and Mumtaz (2007)[8], variables include the output growth, inflation, and the interest rate, in this order, which matters (Del Negro & Primiceri, 2015). The number of lags is set to 2, consistently with the literature due to the parsimonious and computational reasons. Moreover, the two lags allow capturing the essential economic dynamics of the multivariate system, of which we estimate frequency-specific features.

The model is estimated using Bayesian techniques that have shown to be useful in vector autoregressions, especially when the number of parameters is large. We take the first ten years of the data as a training sample to obtain OLS estimates for calibration of the TVP-VAR model. Subtracting those ten years, we analyse the rest of the sample spanning from 1964:Q2 to 2018:Q2. We use the MCMC algorithms of Cogley and Sargent (2005) to draw parameters sequentially from different distributions conditional on remaining parameters of the model (Koop et al., 2009). We simulate posterior distributions of parameters using 50000 iterations.

To identify the structural monetary policy shocks, we stay in line with the literature and theory (Rubio-Ramirez et al., 2010) and impose sign restrictions on every quarter in the time-varying scheme, meaning the restriction is put on the first horizon, in other words on impact. We impose contemporaneous

[7]The data used are downloaded from the FRED St. Louis website.

[8]We use and modify Matlab and Julia codes of Harood Mumtaz, available at `https://sites.google.com/site/hmumtaz77/`.

restrictions in a way that positive monetary policy shocks have non-negative effects on interest rates and non-positive effects on inflation and output. The restrictions of structural shocks are outlined in Table 3.1.

Table 3.1: Sign restrictions

	GDP, y_t	Inflation, π_t	Interest rate, i_t
Monetary policy shock	$\leq$	$\leq$	$\geq$

Note: The restrictions are imposed contemporaneously on the endogenous variables.

3.4 Results

In this section, we present the monetary policy analysis in time and frequency domains. All results are based on medians of posterior estimates computed via the time-varying parameters VAR model. Our primary focus is on the decomposition of dynamic responses to different frequency horizons – economic cycles. Traditionally impulse responses often do not look further than a given time horizon, at which researchers aim their stury, although, they observe limited information. In our case, we ideally decompose an infinite horizon coefficient ($H \to \infty$) to capture the longest relationships between variables. We follow this intuition when picturing the extent of local correlations between macroeconomic variables in time and frequency meaning that the longest possible cycle that can be measures equals the half of our sample length.[9]

3.4.1 Monetary policy in time and frequency

In this section, we present propagation of identified monetary policy shocks in the form of impulse response functions in Figure 3.3. The IRFs are computed as medians at each point of time based on posterior distributions. Shocks might be considered as a policy action, which is in the case of US data, controlled by the Federal Reserve Board.

[9]We work with frequencies ω on the interval of $[0, \pi]$. Accordingly, we obtain economic cycles of interest corresponding to particular frequencies for quarterly data such that business cycles (2-8 years): $\omega \in [\frac{\pi}{16}, \frac{\pi}{4}]$, long cycles ($>$ 8 years): $\omega < \frac{\pi}{16}$, and short cycles ($<$ 2 years): $\omega > \frac{\pi}{4}$.

Figure 3.3 depicts time-varying impulse respons funcstion to a unitary increase in interest rates. We plot 40 horizons to provide results similar to the literature. The results concord with the assumed contemporaneous restrictions we imposed on the model such that the impulse response functions have the expected signs with. In Figures 3.4 and 3.5, we depict the frequency response functions of the monetary policy transmission related to the impulse response functions in Figure 3.3. The impulse responses of interest rates are positive and stable over time with a slow decay. The propagation of the monetary policy shock is also positive at most economic cycles, except the long term where we do not observe it being significantly different from zero, see Figure 3.6.

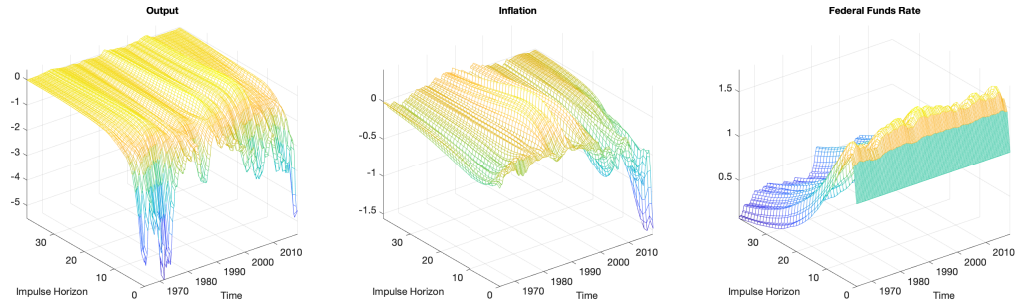

Figure 3.3: Time-varying median impulse responses of *GDP growth*, *inflation* and *interest rate* to a unitary *monetary policy* shocks.

Impulse responses of the output vary during the sample and react extensively to the policy shocks, especially during the 1960s and 1970s. After this period, the relationship stabilises until the 2000s. Similarly, Castelnuovo and Surico (2010) find more significant falls of output gap before 1979 than in the sub-sample afterwards. During the Great Moderation, we observe the lowest effects of shocks on output, which appear to recover quickly and increase the output after several quarters. These findings are in line with Belongia and Ireland (2016) who show practically no differences in output responses between the years 2000 and 2007. While the transmission depicted in the impulse response functions returns the economy to the origin,[10] observing the frequency response functions we decompose the frequency-horizons, which determine the adjustment. The monetary policy has the strongest impact on the output in the period before 1980. The output frequency response of the policy was the lowest during late 1980s and 1990s.

[10] Assuming estimation of a stable model.

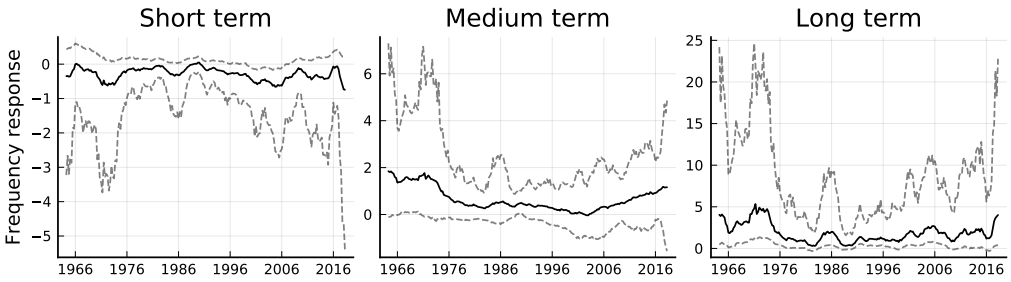

Figure 3.4: Time-varying frequency response functions of *output* in response to *monetary policy*.
Frequency responses are shown at different cyclical intervals, long-, medium-, short-terms corresponding to averages over intervals of longer than 8 years, between 2 to 8 years, and shorter than 2 years. Results are posterior medians with 1σ confidence intervals.

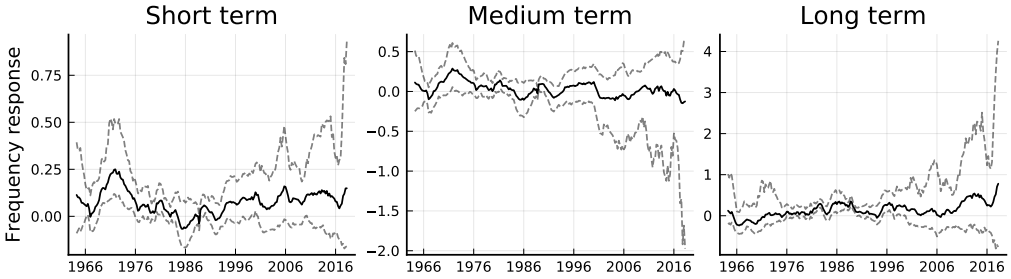

Figure 3.5: Time-varying frequency response functions of *inflation* in response to *monetary policy*.
Frequency responses are shown at different cyclical intervals, long-, medium-, short-terms corresponding to averages over intervals of longer than 8 years, between 2 to 8 years, and shorter than 2 years. Results are posterior medians with 1σ confidence intervals.

While examining frequency responses of the inflation to monetary policy shocks, there are evident patterns of switching the direction in time and at different horizons. During the first years, the inflations responses negatively at economic cycles longer than eight years. Before 1980 the situation changes and the responses are all significantly positive, which corresponds to the debated empirical evidence of price puzzle prior to 1979. In 1980s we observer negative responses (or not significant from zero) of inflation at short and medium term cycles. This frequency observations may explain the post-Volcker era without the price puzzle since researchers draw such a conclusion from IRFs based on the first several quarters, which are mostly not induced by long cycles. Combining the results from IRFs and frequency responses during late 1980, we see inflation positive response at long cycles together with its gradual increase after 15 quarters in IRFs, Figure 3.3. Similar explanation follows for first quarters and short cycles. Nevertheless, this

comparison cannot be used in general. For example in years after 2010, the inflation responses positively at short cycles, even there is a large fall for the first horizons and no puzzling result in the time domain, but all other cycles, it is. The alternating behaviour of inflation frequency responses creates evidence that at different point of time even thought the impulse response functions fulfil the sign restriction of no price puzzle and yet in the frequency domain, we observe with both positive and negative transfers.

Hence, given the frequency-specific transfers, these findings provide new insights to the understanding of the impact of structural shocks for policy-making. Nonetheless, we observe that during turbulent times around 2008 and after the estimation and stability of the model is crucial and confidence intervals of simulated coefficients wider. This should be dealt with caution.

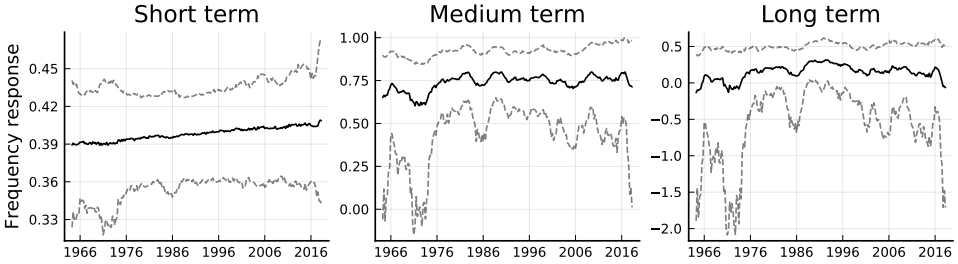

Figure 3.6: Time-varying frequency response functions of *interest rate* in response to *monetary policy*.

Frequency responses are shown at different cyclical intervals, long-, medium-, short-terms corresponding to averages over intervals of longer than 8 years, between 2 to 8 years, and shorter than 2 years.

3.4.2 Time-varying frequency dependence

We complement the previous structural analysis with the evidence about local correlations of output and inflation with interest rates. We draw the correlations from the spectral matrix (Eq. 3.15) obtained from posterior time-varying coefficients and covariance matrix. Figures 3.7 and 3.8 depict time-frequency correlations correlations at specific frequency horizons – economic cycles. The correlations depict the link and direction between the variables, which are not causal as the frequency response functions. We observe pro-cyclical to counter-cyclical behaviour with changing magnitudes in both time and frequency.

Correlations between inflation and interest rates show similar patterns of the strong relationship at the beginning of the sample, in the mid-1980s, and

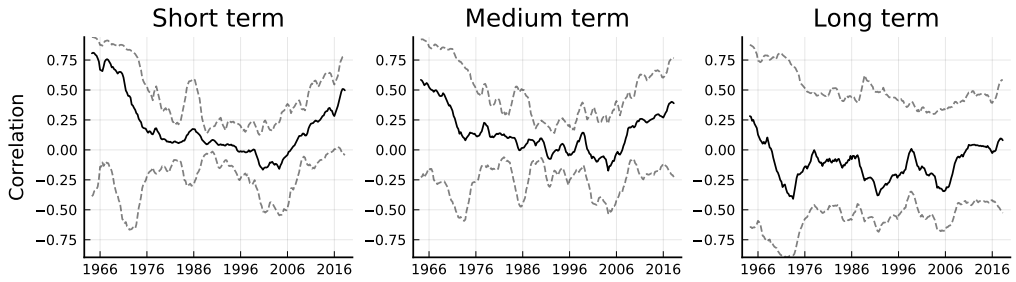

Figure 3.7: Time-varying correlations of *inflation* and *interest rate*.
Figure depicts average information over particular frequency bands – different cyclical lengths: long cycles – greater than 8Y, medium cycles – from 2Y to 8Y, and short cycles – lower than 2Y. Results are posterior medians (unconditional) with 1σ confidence interval.

after 2006 at economic cycles up to 8 years. During the Great Moderation, particularly in the 1990s, we see that correlations are more stable, it is also when the frequency responses of interest rates (Figure 3.4) are weaker and stable. One might link the stability of correlations during 1980 to the Volcker's disinflation because the correlation stabilized at weak level. However, after 2006 the linke between inflation and interest rates has increased significantly. Moreover, the three correlations share the same pattern at business cycle frequencies, which is strongest in the 1970s, lowest in 1990s, and from 2000 rises again.

Empirical literature often presents time-varying correlations between monetary variables calculated directly from the posterior distribution of the time-varying covariance matrix. For example, the correlations between inflation and interest rates are similar in shape to Cogley and Sargent (2005), but in our case, on average, the inflation is less correlated with the interest rate. Unfortunately, their sample ended in 2000. Thus, we do not compare the phase of strengthening.

Interestingly, related to the price puzzle phenomenon, we see the relatively clear pattern in time-frequency dependence between the inflation and the interest rates, Figure 3.7. The horizon specific outlook might help in policy-making that at short and medium (business) cycles inflation co-moves with interest rates. In contrary to Benati and Mumtaz (2007), we do not observe a sign change in the correlation between output and inflation in the mid-1970s. The correlation betwen GDP growth and inflation (Figure 3.8) is negative for cycles shorther than 2 years and positive and strong for economic cycles between 2 to 8 years; and it is negative for whole sample, however the

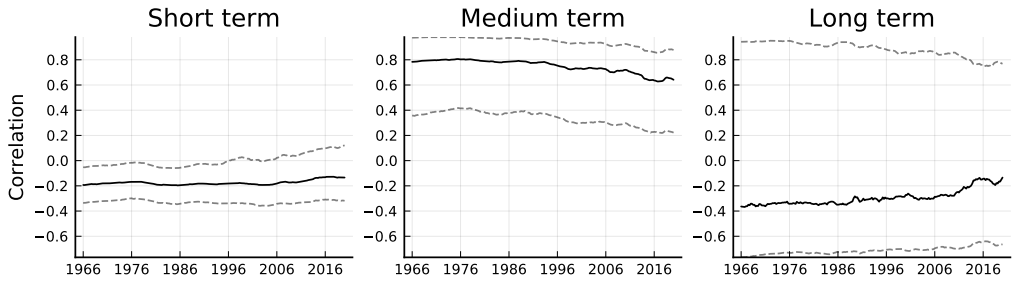

Figure 3.8: Time-varying correlations of *GDP growth* and *inflation*.
Figure depicts average information over particular frequency bands – different cyclical lengths: long cycles – greater than 8Y, medium cycles – from 2Y to 8Y, and short cycles – lower than 2Y. Results are posterior medians (unconditional) with 1σ confidence interval.

confidence interval covers complete range.[11]

We observe the time-varying frequency responses of inflation and interest rates that change vividly (Figure 3.5), and when looking at time-varying correlations, we do not observe changing behaviour that dramatic. The reason is that the two measures use different underlying information. The covariance structure measures correlations, and the frequency response functions capture the filtering process in coefficients of the model.

3.5 Conclusion

In this paper, we have used the frequency-specific methodology to provide an additional way of looking at the monetary policy transmission dynamics. We employ the frequency response functions to decompose the transfers of monetary policy in the framework that allows for time-varying coefficients and covariance structure. The time-varying model empowers us to quantify dependence measures in both time and frequency domains. To assess the frequency domain monetary transmission, we estimate a traditional TVP-VAR identified with sign restrictions. The propagations are studied on the US data.

Firstly, we find a substantial variation in output and inflation in response to shocks in time, which is in line with the literature. The frequency transmission of monetary policy pronounces the most considerable positive impacts in output at economic cycles longer than eight years. The frequency response of output is overall positive, only for cycles shorter than two years is nega-

[11] This might be due to an uncertainty at lowest frequencies, $< \frac{\pi}{32}$, for which we have less observations than for other cycles.

tive over time. The shocks affect the interest rates most at business cycles of length from 2 to 8 years, with positive impact at all frequencies.

Moreover, using the time-varying approach, we observe a negative propagation of shocks to inflation, meaning that the price puzzle phenomenon is not observed for business cycles after 1985. However, the average frequency transmission varies, and in every decade we see that prices rise in response to monetary policy, except the years between 2000 and 2010 for cycles of 2-8 years length. This leads us to a new result that the price puzzle phenomenon may have frequency-dependent effect and be propagated at different cycles, which is not observed using only impulse response functions, even with sign restrictions imposed.

Lastly, we estimated reduced-form time-frequency correlation at each point of time that quantifies local dependence structure between variables. We find a characteristic pattern of strong dependence at the beginning and the end of the studied period and smaller dependence from 1980 to 2000 for inflation and interest rates. The relationship between inflation and interest rate is strong, they move pro-cyclically at short and medium term cycles.

References

Aguiar-Conraria, L., Azevedo, N., & Soares, M. J. (2008). Using wavelets to decompose the time–frequency effects of monetary policy. *Physica A: Statistical mechanics and its Applications*, *387*(12), 2863–2878.

Baruník, J., & Křehlík, T. (2018). Measuring the Frequency Dynamics of Financial Connectedness and Systemic Risk. *Journal of Financial Econometrics*, *16*(2), 271–296.

Baumeister, C., & Hamilton, J. D. (2015). Sign restrictions, structural vector autoregressions, and useful prior information. *Econometrica*, *83*(5), 1963–1999.

Baxter, M., & King, R. G. (1999). Measuring business cycles: Approximate band-pass filters for economic time series. *The Review of Economics and Statistics*, *81*(4), 575–593.

Belongia, M. T., & Ireland, P. N. (2016). The evolution of US monetary policy: 2000–2007. *Journal of Economic Dynamics and Control*, *73*, 78–93.

Benati, L., & Mumtaz, H. (2007). US evolving macroeconomic dynamics: A structural investigation.

Canova, F., & Gambetti, L. (2009). Structural changes in the US economy: Is there a role for monetary policy? *Journal of Economic Dynamics and Control*, *33*(2), 477–490.

Castelnuovo, E., & Surico, P. (2010). Monetary policy, inflation expectations and the price puzzle. *The Economic Journal*, *120*(549), 1262–1283.

Cogley, T., & Sargent, T. J. (2005). Drift and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S. *Review of Economic Dynamics*, *8*(2), 262–302.

Croux, C., Forni, M., & Reichlin, L. (2001). A measure of comovement for economic variables: Theory and empirics. *The Review of Economics and Statistics*, *83*(2), 232–241.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, *25*(1), 1–37.

Dahlhaus, R., Polonik, W., et al. (2009). Empirical spectral processes for locally stationary time series. *Bernoulli*, *15*(1), 1–39.

Del Negro, M., & Primiceri, G. E. (2015). Time varying structural vector autoregressions and monetary policy: A corrigendum. *The Review of Economic Studies*, *82*(4), 1342–1345.

Dew-Becker, I., & Giglio, S. (2016). Asset pricing in the frequency domain: Theory and empirics. *The Review of Financial Studies*, *29*(8), 2029–2068.

Ellington, M. (2018). The case for Divisia monetary statistics: A Bayesian time-varying approach. *Journal of Economic Dynamics and Control*, *96*, 26–41.

Gehrke, B., & Yao, F. (2017). Are supply shocks important for real exchange rates? a fresh view from the frequency-domain. *Journal of International Money and Finance*, *79*, 99–114.

Geraci, M. V., & Gnabo, J.-Y. (2018). Measuring interconnectedness between financial institutions with bayesian time-varying vector autoregressions. *Journal of Financial and Quantitative Analysis*, *53*(3), 1371–1390.

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? *Handbook of macroeconomics*, *1*, 65–148.

Christiano, L. J., & Fitzgerald, T. J. (2003). The band pass filter. *International Economic Review*, *44*(2), 435–465.

Koop, G., Leon-Gonzalez, R., & Strachan, R. W. (2009). On the evolution of the monetary policy transmission mechanism. *Journal of Economic Dynamics and Control*, *33*(4), 997–1017.

Lovcha, Y., & Perez-Laborda, A. (2018). Monetary policy shocks, inflation persistence, and long memory. *Journal of Macroeconomics*, *55*, 117–127.

Mumtaz, H., & Sunder-Plassmann, L. (2013). Time-varying dynamics of the real exchange rate: An empirical analysis. *Journal of Applied Econometrics*, *28*(3), 498–525.

Murray, C. J. (2003). Cyclical properties of baxter-king filtered time series. *The Review of Economics and Statistics*, *85*(2), 472–476.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, *72*(3), 821–852.

Ramsey, J. B., & Lampart, C. (1998). Decomposition of economic relationships by timescale using wavelets. *Macroeconomic Dynamics*, *2*(1), 49–71.

Rubio-Ramirez, J. F., Waggoner, D. F., & Zha, T. (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies*, *77*(2), 665–696.

Rusnák, M., Havranek, T., & Horváth, R. (2013). How to solve the price puzzle? A meta-analysis. *Journal of Money, Credit and Banking*, *45*(1), 37–70.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, *48*(1), 1–48.

Sims, C. A. (1992). Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review*, *36*(5), 975–1000.

Stiassny, A. (1996). A spectral decomposition for structural var models. *Empirical Economics*, *21*(4), 535–555.

Uhlig, H. (2005). What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics*, *52*(2), 381–419.

Chapter 4

Taming data-driven probability distributions

Abstract

We propose a deep learning approach to probabilistic forecasting of macroeconomic and financial time-series. Allowing to learn complex patterns from a data rich environment, our approach is useful for a decision making that depends on uncertainty of large number of economic outcomes. Specifically, it is informative to agents facing asymmetric dependence of their loss on outcomes from possibly non-Gaussian and non-linear variables. We show the usefulness of the proposed approach on the two distinct datasets where a machine learns the pattern from data. First, we construct macroeconomic fan charts that reflect information from high-dimensional data set. Second, we illustrate gains in prediction of stock return distributions which are heavy tailed, asymmetric and suffer from low signal-to-noise ratio.

This chapter was co-authored with Jozef Baruník and its part has been published in *Finance Research Letters*, Baruník and Hanus (2024). We are grateful to Wolfgang Hardle, Lukáš Vácha, Martin Hronec, František Čech, and the participants at various conferences and research seminars for many useful comments, suggestions, and discussions. We gratefully acknowledge the support from the Czech Science Foundation under the EXPRO GX19-28231X project. We provide the computational package `DistrNN.jl` in JULIA available at https://github.com/barunik/DistrNN.jl that allows one to obtain our measures on data the researcher desires.

4.1 Introduction

Despite advances in data availability, theory, and computational power, economics have not enjoyed dramatic improvements in forecast accuracy of economic variables over past decades (Stock & Watson, 2017). A fundamental problem underlying the lack of success is that economic variables are difficult to forecast in nature. Economic forecasters whose decisions depend on such uncertainty need to focus on communicating full predictive distribution of the variable surrounding point estimates.[1] At the same time, economists being keen to use a large number of series to understand fluctuations in economic data, collect data unimaginable decades ago.[2] Enlocking the information hidden in big data is becoming a key theme in economics (Diebold, 2021). Economists have turned to machine learning to explore the rich information content of new datasets in response to heavy criticism of arbitrarily chosen restrictions on reduced and structured models in past decades and the challenge posed by the proliferation of parameters. (Mullainathan & Spiess, 2017).

In this paper, we explore the use of machine learning for information-rich uncertainty forecasts. We develop a distributional machine learning method based on deep learning and recurrent network techniques to provide probabilistic forecasts that reflect time-series dynamics of possibly large amounts of available information. Such data-driven probabilistic forecasts are not possible with classical methods without a set of restricting assumptions.

Our main contribution to the literature is that we propose how to use deep learning techniques as a useful tool for the approximation and prediction of conditional distributions in a data-rich environment. Our distributional neural network takes advantage of deep learning (especially recurrent neural networks), it is capable of predicting an entire distribution of a time-series and allows the use of large amounts of variables. We frame our approach as a multi-output neural network, which returns approximate probability functions of the distribution. Our approach's novelty lies in learning the entire conditional distribution using (deep) recurrent networks from big data. The proposed network is also capable of capturing time-variation of distributions

[1] The Bank of England was an early leader in recognizing this need and started to communicate the uncertainty as fan charts to the public.

[2] Already a century ago, in the 1920s, Harvard Economic Service provided economic indexes and forecasts based on available data (Friedman, 2009). Later, 1277 time-series were used to study business cycles by Lerner (1947).

when, for example, dealing with highly dynamic data and recovering longer and more complex time-dependence structures present in data. Our framework generalizes binary choice models (Anatolyev & Baruník, 2019; Foresi & Peracchi, 1995) and, together with the state-of-the-art machine learning tools, form a new toolkit for economists interested in describing future uncertainty of economic variables. An important contribution is also our novel approach to the construction of an objective function that fulfils the monotonicity of distributional forecasts by introducing penalty function for divergence from monotone behavior.

Two distinct and important economic datasets illustrate how the machine learning approach to probabilistic forecasting may help a decision-maker facing uncertainty. First, we use deep learning to construct data-driven macroeconomic fan charts reflecting information contained in a large number of variables. Such data-rich fan charts are the first of their kind to reflect high-dimensional information on 216 relevant variables and are of great importance for policymakers as they reflect the structures in data and are not influenced by the choice of the model. A forecasting model is, in contrast, learned from data. Such data-rich fan charts, moreover, can not be obtained with traditional methods. Second, we study the set of most liquid U.S. stock returns that display asymmetric, heavy-tailed, dynamically evolving distributions that are hard to predict due to a very low signal-to-noise ratio.

Understanding uncertainty in decision-making is crucial for financial operations, central and retail banking, as well as researchers and practitioners seeking to minimize risk, develop appropriate plans, and assist in designing and implementing economic policies. However, even after decades of research, a conditional mean forecast often serves economists as a convenient tool for measuring the central tendency of a target variable or simply as the best guess about the future outcomes of a variable. Variance forecast accompanying it often serves as the best expectation about uncertainty and future risk. Nevertheless, such predictions are not fully informative in case a decision maker is facing asymmetric dependence of her loss on outcomes from possibly non-Gaussian variables. An uncertainty is a key ingredient in economic decision-making, a shift to probabilistic forecasting also shifts our hopes towards obtaining better expectations about entire distributions of economic variables. A non-trivial question is how we make such forecasts, especially utilizing available data.

Traditionally, distribution forecasts are made using time-series models

and surveys or are collected in real-time.[3] With rapid improvements in accessibility and availability of large datasets, we believe one can improve the description of uncertainty substantially utilizing methods that focus on learning patterns from data. In line with recent endeavours of economists to move away from exclusive dependence on models towards machine learning approaches (Athey & Imbens, 2019) when it makes sense to utilize data and improve our understanding of the problem (Mullainathan & Spiess, 2017), we propose to use deep machine learning to learn the complex patterns in data and return the user a prediction of an entire distribution.

A key idea of (machine) learning that can be thought of as inferring plausible models to explain observed data recently attracted number of researchers who document how learning patterns from data can be useful.[4] Surge in the literature and increasing number of applications in economics focus mostly on cross-sectional data and ultimately on point forecasts. While machines can use such models to make predictions about future data, uncertainty plays a fundamental part. At the same time, data, being a key ingredient of all machine-learning systems, are useless on their own until one extracts knowledge or inferences from them. Shifting focus from point forecasts towards probabilistic forecasting using big data is an essential next step for economists wishing to explore what computer science has to offer.

We contribute to this debate by exploring machine learning in a time-series context and developing a machine learning strategy to forecast the full distributions in possibly large dimensional settings. We argue that deep learning, in particular recurrent neural network offers a useful tool for distribution prediction without the need for model specification, simply learning the distributions from data. While the ability to outperform alternative methods on specific data sets in terms of out-of-sample predictive power is valuable in practice, such performance is rarely explicitly acknowledged as a goal to be addressed in econometrics. As Mullainathan and Spiess (2017) highlights, some substantive problems are naturally cast as prediction problems, and assessing their goodness of fit on a test set may be sufficient for the purpose of the analysis. We believe that the distribution prediction task in

[3]Methods for constructing distribution forecasts are reviewed in a special issue on "Density Forecasting in Economics and Finance" (Timmermann, 2000) and "Probability Forecasting" (Gneiting, 2008) for collection of papers.

[4]Bianchi et al. (2021), Feng et al. (2018), Goulet Coulombe et al. (2022), Gu et al. (2020), Heaton et al. (2017), Israel et al. (2020), Iworiso and Vrontos (2020), Mullainathan and Spiess (2017), Sadhwani et al. (2020), and Tobek and Hronec (2020)

a data-rich environment is one of such an important problems in economics where machine learning could be helpful for a researcher, policy maker or practitioner.

What are the challenges specific to probabilistic forecasting of economic variables? Time-series such as stock returns, electricity prices, traffic data, or macroeconomic series display distributions that can not be captured by convenient Gaussian distribution and hence are not fully characterized by means and variances. These distributions show heavy tails, are asymmetric and often violate stationarity. Further, the data contain irregularities, difficult to predict spikes, and regime shifts. Hence, complete information about the probability of future outcomes, given the past information that can be mapped into different representations to construct prediction intervals or probability distribution functions reflecting the data, is needed. Such a fully approximated distribution function provides comprehensive information about the uncertainty of future observations.

Most studies that focus on the prediction of conditional return distributions characterize the cumulative conditional distribution by a collection of conditional quantiles (Engle & Manganelli, 2004; Žikeš & Baruník, 2016). In contrast, Leorato and Peracchi (2015) argue that collection of conditional probabilities that describe the cumulative distribution function using a set of separate logistic regressions (Foresi & Peracchi, 1995) provide a better approach. Following decades resulted in few contributions exploring distributional regressions (Fortin et al., 2011; Chernozhukov et al., 2013; Rothe, 2012) including attempts to overcome the problem of monotonicity of the forecasts (Anatolyev & Baruník, 2019) using ordered logistic parametrization. Another important strand of literature focuses on Bayesian forecasting, where uncertainty is characterized by probabilities automatically (Geweke & Whiteman, 2006; Lahiri, Martin, et al., 2010). Further, the literature offers model averaging to deal with forecasts' uncertainty. The model uncertainty can be lowered with averaging of multiple models using the Bayesian Model Averaging (BMA) method (Liu & Maheu, 2009; Wright, 2008), for instance.[5]

At the same time, literature in computer science attempts to use machine learning in the prediction of distributions. These attempts are similar to traditional methods and mostly rely on an approximation of some pre-specified distribution, such as the first two moments. Duan et al. (2019) applies the natural gradient boosting algorithm to estimate parameters for the condi-

[5]For a good overview of this literature, see Steel (2020).

tional probability distribution while assuming homoskedasticity. Salinas et al. (2020) build an autoregressive recurrent neural network, which learns mean and standard deviation for Gaussian and mean and shape parameters for Negative binomial. Lim and Gorse (2020) classifies price movements for high-frequency trading via deep probabilistic modelling when optimizing parameters of different families of distribution. Although similar to Salinas et al. (2020), Y. Chen et al. (2020) proposes to use a deep temporal convolutional neural network to estimate parameters of Gaussian distribution to model probabilistic forecast, and they further propose to use the same architecture for non-parametric estimation of quantile regression. An alternative approach is distribution-free and can produce more robust results. Wen et al. (2017) performs multi-horizon predictions to study forecasts distributions via direct quantiles using recurrent neural network. Quantile function represented by spline combined with recurrent neural network proposed by Gasthaus et al. (2019) is a distribution-free approach with objective function based on CRPS score (Gneiting & Raftery, 2007) constructed with respect to monotonicity of quantile function. Hu et al. (2019) build deep neural networks to obtain distribution-free probability distribution where one of the steps in the procedure is to obtain cumulative distribution estimates.[6] To the best of our knowledge, the literature has not yet moved to fully non-parametric approaches approximating the data structures in the context of distributional forecasting in economics and finance.

Why should we believe that machine learning can improve probability forecasts? Classical time-series econometrics (Box et al., 2015; Hyndman et al., 2008) mainly focuses on predetermined autocorrelation or seasonality structures in data that are parametrized. Having a large amount of time-series available to researchers, these methods quickly become infeasible and unable to explore more complex data structures. Keeping in mind a famous wisdom that *"all models are wrong..., but some of them are useful."* (Box, Draper, et al., 1987), modern machine learning methods are able to overcome these problems easily. Being a powerful tool for approximation of complex and unknown data structures (Kuan & White, 1994), these methods can be useful in a number of application problems where data contain rich information structure, and we can not describe it satisfactorily by a simplifying model.

[6] Januschowski et al. (2020) provide a detailed discussion about ML methods for forecasting. The text discusses the way of distinction between "statistical" and "ML" methods adapted in time.

Overcoming the longstanding problem of the computational intensity of such a data-driven approach with advances in computer sciences adds to the temptation to use these methods for addressing new problems such as distribution predictions.

4.2 A Route Towards Probabilistic Forecasting via Deep Learning

Let us consider an economic time-series y_t collected over $t = 1 \ldots, T$. The main objective is to approximate the conditional cumulative distribution function $F(y_{t+h}|\mathcal{I}_t)$ as precisely as possible and use it for h-step-ahead probabilistic forecast made at time t with information $\mathcal{I}_t$ containing past values of y_t as well as, possibly, past values of other exogenous observable variables.

Consider a partition of the support of y_t by $p > 1$ fixed thresholds corresponding to set of empirical α_j-quantiles $\{q^{\alpha_j}\}_{j=1}^{p}$ where $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_p < 1$ are p regularly spaced probability levels on a unit interval $[0, 1]$. These partitions are further time-varying; thus, in general, the elements of the partition are indexed implicitly by t.

The main goal then is to approximate a collection of conditional probabilities corresponding to the empirical quantiles such as

$$\left\{F(q^{\alpha_1}), \ldots, F(q^{\alpha_p})\right\} = \left\{\Pr\left(y_{t+h} \leq q^{\alpha_1}|\mathcal{I}_t\right), \ldots, \Pr\left(y_{t+h} \leq q^{\alpha_p}|\mathcal{I}_t\right)\right\}$$

for the collection of thresholds $1, \ldots, p$. One convenient way of estimating such quantities is distributional regression. Foresi and Peracchi (1995) noted that several binary regressions serve as a good partial description of the conditional distribution. To estimate the conditional distribution, one can simply consider the distribution regression model

$$\Pr(y_{t+h} \leq q^{\alpha_j}|\mathcal{I}_t) = \Lambda(\beta_j), \tag{4.1}$$

where $\Lambda : z \to [0, 1]$ is a known (monotonically increasing) link function, such as logit, probit, linear, log-log functions[7] and $\beta(.)$ is an unknown function-valued parameter to be determined. In contrast to estimating separate models for separate thresholds, Chernozhukov et al. (2013) considered a

[7] As discussed by Chernozhukov et al. (2013), log-log link nests the Cox model making distribution regression important.

continuum of binary regressions and argued it provides a coherent and flexible model for the entire conditional distribution as well as useful alternative to Koenker and Bassett Jr (1978)'s quantile regression. Alternatively, Anatolyev and Baruník (2019) propose to tie the coefficients of predictors in an ordered logit model via smooth dependence on corresponding probability levels. While being able to forecast entire distribution and keeping $0 < F_j < 1$ and $0 < F_1(.) < F_2(.) < \cdots < F_p(.) < 1$, the approach still depends on heavy parametrization suited for a specific problem of the time-series considered making it an infeasible approach for larger number of variables.

4.2.1 (Deep) Machine Learning

Such probabilistic forecasts heavily depend on the model parametrization and, with a growing number of covariates, become quickly infeasible. Stationarity of data at hand is also requirement that complicates forecasts as it is hard to achieve in many cases. In sharp contrast to such approach, we propose more flexible and general way to the distribution regression via deep learning. We propose a novel multiple output neural network we refer to as a distribution neural network (DistrNN). Our approach aims to uncover non-linear and mostly complex relationship of time-series without specifying strict parametric structure and without requiring strict assumptions about data, while focusing on the out-of-sample predictive power of the model.

Machine learning has a long history in economics and finance (Baillie & Kapetanios, 2007; Hutchinson et al., 1994; Kuan & White, 1994; Racine, 2001). At its core, one may perceive machine learning as a general statistical analysis that economists can use to capture complex relationships that are hidden when using simple linear methods. As emphasized by Breiman et al. (2001), maximizing prediction accuracy in the face of an unknown model differentiates machine learning from the more traditional statistical objective of estimating a model assuming a data generating process. Building on this, machine learning seeks to choose the most preferable model from an unknown pool of models using innovative optimization techniques. As opposed to traditional measures of fit, machine learning focuses on the out-of-sample forecasting performance and understanding the bias-variance trade-off, as well as using data-driven techniques that concentrate on finding structures in large datasets. Further, if one dismisses the "black-box" view of machine learning as a misconception (Lopez de Prado, 2019), it seems

nothing should stop a researcher from exploring the power of these methods to solve problems like probabilistic forecasting. However, the problems in economics differ from typical machine learning applications in many aspects. In order to enjoy the benefits of machine learning, a user needs to understand key challenges brought by data.[8]

Deep feed-forward networks, also often called feed-forward neural networks or multilayer perceptrons, lie at the heart of deep learning models and are universal approximators that can learn any functional relationship between input and output variables with sufficient data Kuan and White (1994). As a class of supervised learning methods, these approaches are used for classification, recognition and prediction. While being increasingly popular in economics for solutions to particular problems Athey and Imbens (2019) and Mullainathan and Spiess (2017), probabilistic forecasting has not been explored by the literature yet.

This motivates us to reformulate distribution regression into a more general and flexible distributional neural network. The functional form of the new network is driven by data and we may relax assumptions on the distribution of the data, parametric model as well stationarity of data. The proposed distributional neural network is, as feed-forward network, a hierarchical chain of layers that represents high-dimensional and/or non-linear input variables with the aim to predict the target output variable. Importantly, we approximate the conditional distribution function with multiple outputs of the network as a set of probabilities jointly.

As a first step, we exchange a known link function from Eq. 4.1 for an unknown general function $\mathfrak{g}$ that will be approximated by a neural network:

$$\Pr\left(y_{t+h} \leq q^{\alpha_j} | \mathcal{I}_t\right) = \mathfrak{g}_j(\cdot). \tag{4.2}$$

Next, we consider a set of probabilities corresponding to $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_p < 1$ being p regularly spaced levels that characterize conditional distribution function using a set of predictors $z_t = (y_t, x_t^1, ..., x_t^n)^\top$, and model

[8]For example Israel et al. (2020) note that machine learning applied to finance is challenged by small sample sizes, naturally low signal-to-noise ratios making market behavior difficult to predict and the dynamic character of markets. Facing these critical issues, the benefits of machine learning are not so obvious as in other fields and research into understanding how impactful machine learning can be for asset management is just emerging. As the deep learning literature grows, machine learning applications in finance have begun to emerge (Bianchi et al., 2021; Bryzgalova et al., 2019; Feng et al., 2018; Gu et al., 2020; Heaton et al., 2017; L. Chen et al., 2020; Tobek & Hronec, 2020; Zhang et al., 2020).

them jointly as

$$\left\{ \Pr\left(y_{t+h} \leq q^{\alpha_1}|z_t\right), \ldots, \Pr\left(y_{t+h} \leq q^{\alpha_p}|z_t\right) \right\} = \mathfrak{g}_{W,b}(z_t), \tag{4.3}$$

where $\mathfrak{g}_{W,b}$ is a multiple output neural network with L hidden layers that we name as distributional neural network:

$$\mathfrak{g}_{W,b}(z_t) = g^{(L)}_{W^{(L)},b^{(L)}} \circ \ldots \circ g^{(1)}_{W^{(1)},b^{(1)}}(z_t), \tag{4.4}$$

where $W = \left(W^{(1)}, \ldots, W^{(L)}\right)$ and $b = \left(b^{(1)}, \ldots, b^{(L)}\right)$ are weight matrices and bias vector. Any weight matrix $W^{(\ell)} \in \mathbb{R}^{m \times n}$ contain m neurons as n column vectors $W^{(\ell)} = [w^{(\ell)}_{\cdot,1}, \ldots, w^{(\ell)}_{\cdot,n}]$, and $b^{(\ell)}$ are thresholds or activation levels which contribute to the output of a hidden layer allowing the function to be shifted.

It is important to note that in sharp contrast to the literature, we consider a multiple output (deep) neural network to characterize a collection of probabilities. Before discussing the details of estimation that allow us to keep the monotonicity of probabilities, we illustrate the framework. Figure 4.1 illustrates how $l \in 1, ..., L$ hidden layers transform input data into a chain using a collection of non-linear activation functions $g^{(1)}, \ldots, g^{(L)}$. A commonly used activation functions, $g^{(\ell)}_{W^{(\ell)},b^{(\ell)}}$, used as

$$g^{(\ell)}_{W^{(\ell)},b^{(\ell)}} := g_\ell\left(W^{(\ell)} z_t + b^{(\ell)}\right) = g_\ell\left(\sum_{i=1}^{m} W^{(\ell)}_i z_t + b^{(\ell)}_i\right)$$

are a sigmoid $g_\ell(u) = \sigma(u) = 1/(1 + \exp(-u))$, rectified linear units $g_\ell(u) = \max\{u, 0\}$, or $g_\ell(u) = \tanh(u)$. In case $\mathfrak{g}_{W,b}(z_t)$ is non-linear, neural network complexity grows with increasing number of neurons m, and with increasing number of hidden layers L and we build a deep neural network. We use activation function $g^{(L)}(\cdot) = \sigma(\cdot)$ to transform outputs to probabilities. Note that for $L = 1$, the neural network becomes a simple logistic regression.

4.2.2 (Deep) Recurrent Neural Networks

Predictors used by economists often evolve over time, and hence, traditional neural networks assuming independence of data may not approximate relationships sufficiently well. Instead, a Recurrent Neural Network (RNN) that takes into account time-series behavior may help in the prediction task.

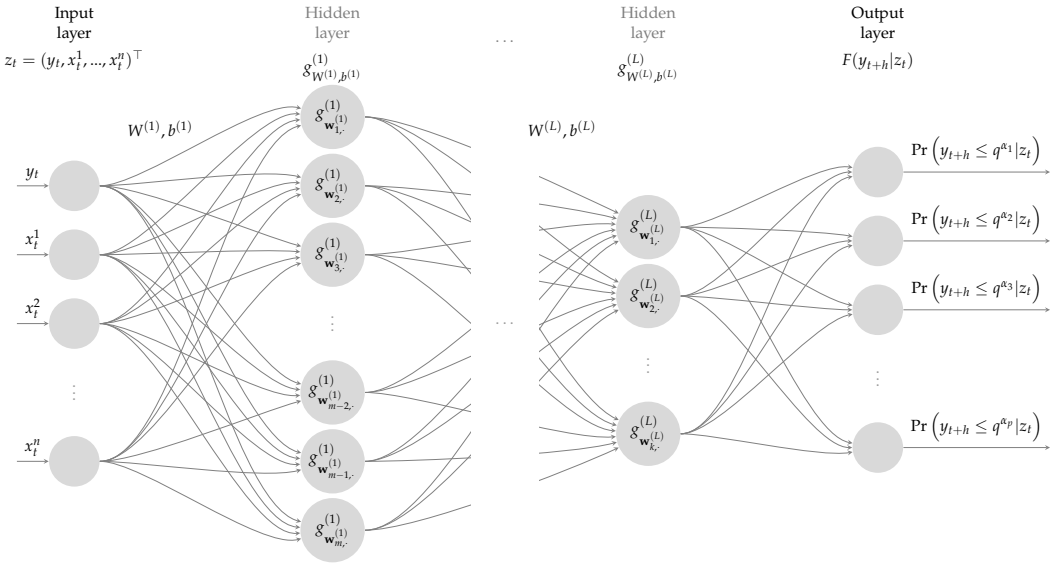

Figure 4.1: Distributional (Deep) Feed-forward Network.

An illustration of a multiple output (deep) neural network $\mathfrak{g}_{W,b}(z_t)$ to model the collection of conditional probabilities $\left\{ \Pr\left(y_{t+h} \leq q^{\alpha_1}|z_t\right), \ldots, \Pr\left(y_{t+h} \leq q^{\alpha_p}|z_t\right) \right\}$ with set of predictor variables $z_t = (y_t, x_t^1, ..., x_t^n)^\top$. And with large number of hidden layers L the network is deep.

Taking into account the sequential nature of data that evolve over time and possess an autocorrelation structure, RNNs are more suitable for many economic problems. In contrast to plain neural networks, hidden layers in recurrent networks are being updated in a recurrence for every time step of the sequence, meaning that the weights of the network are shared over the sequential data, and hidden states remember the time structure.

Formally, RNNs transform a sequence of input variables to another output sequence with lagged (memory) hidden states

$$h_t = g(W_h h_{t-1} + W_z z_t + b_0). \tag{4.5}$$

Figure 4.2 illustrates distinctions of weights where dashed lines correspond to W_h and solid lines to W_z. Intuitively, RNN is a non-linear generalization of an autoregressive process where lagged variables are transformations of the observed variables. Nevertheless, the structure is only useful when the immediate past is relevant. In case the dynamics are driven by events that are further back in the past, the nodes of the network require an even more complex structure.

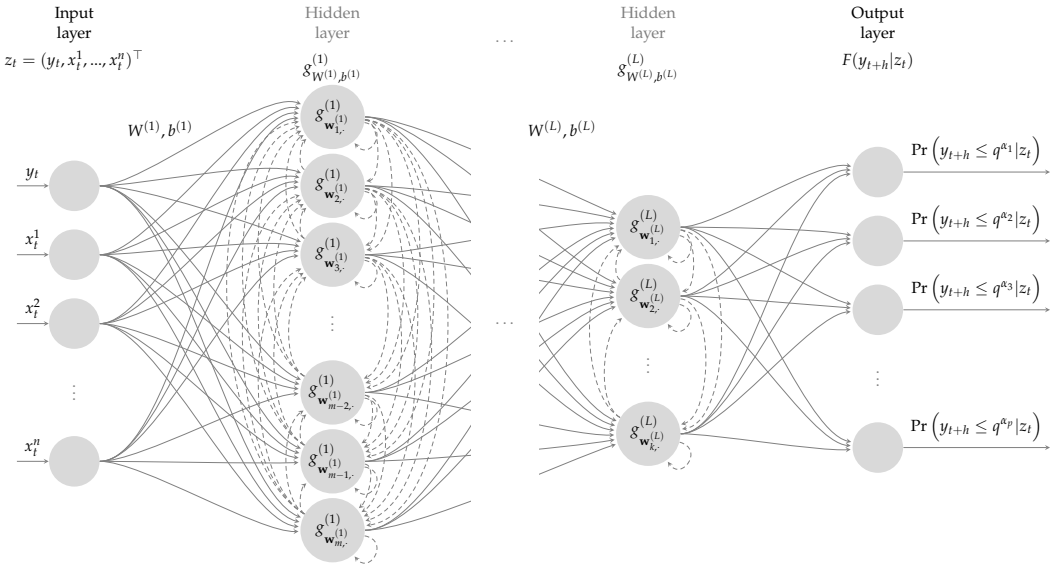

Figure 4.2: Distributional (Deep) Recurrent Network.

An illustration of a deep recurrent neural network $\mathfrak{g}_{W,b}(z_t)$ that captures relationship between all nodes (solid) and recurrent paths (dashed) in the network at time t to model the collection of conditional probabilities $\left\{ \Pr\left(y_{t+h} \leq q^{\alpha_1}|z_t\right), \ldots, \Pr\left(y_{t+h} \leq q^{\alpha_p}|z_t\right) \right\}$ with set of predictor variables $z_t = (y_t, x_t^1, ..., x_t^n)^\top$. And with a large number of hidden layers L, the network is deep.

Long Short-term Memory (LSTM)

As a particular form of recurrent networks, an LSTM provides a solution to the short memory problem by incorporating memory units into the structure (Hochreiter & Schmidhuber, 1997) and capturing potentially long-time dynamics in the time-series. Memory units allow the network to learn when to forget previous hidden states and when to update hidden states given new information. Specifically, the LSTM unit has five components: an input gate, a hidden state, a memory cell, a forget gate, and an output gate. The memory cell unit combines the previous time step memory cell unit, which is modulated by the forget and input modulation gates, together with the previous hidden state, modulated by the input gate. These components enable an LSTM to learn very complex long-term and temporal dynamics that a vanilla RNN is incapable of. Additional depth in capturing the complexity of a tie series can be added by stacking LSTM on top of each other.

Formally, at each time step, a new memory cell c_t is created taking current input z_t and previous hidden state h_{t-1}, and it is then combined with forget

gate that controls an amount of information kept in the hidden state as

$$
\begin{aligned}
h_t &= \sigma\left(\underbrace{W_h^{(o)} h_{t-1} + W_z^{(o)} = z_t + b_0^{(o)}}_{\text{output gate}}\right) \circ \tanh(c_t) \\
c_t &= \sigma\left(\underbrace{W_h^{(g)} h_{t-1} + W_z^{(g)} z_t + b_0^{(g)}}_{\text{forget gate}}\right) \circ c_{t-1} + \sigma\left(\underbrace{W_h^{(i)} h_{t-1} + W_z^{(i)} + b_0^{(i)}}_{\text{input gate}}\right) \circ \tanh(k_t).
\end{aligned}
$$

The term $\sigma(\cdot) \circ c_{t-1}$ introduces the long-range dependence, k_t is new information flow to the current cell. The forget gate and input gate states control weights of past memory and new information. In Figure 4.2, c_t is the memory pass through multiple hidden states in the recurrent network.

4.2.3 Loss Function

Since we aim to estimate the cumulative distribution function that is a non-decreasing function bounded on $[0, 1]$, we need to design an objective function that minimizes differences between targets and estimated distribution as well as imposes a non-decreasing property of the output. Since the problem is essentially a more complex classification problem closely related to logistic regression, we use a binary cross-entropy loss function. Moreover, to order the predicted probabilities, we introduce a penalty to the multiple output classification problem.

The loss function is then composed of two parts: traditional binary cross-entropy and a penalty imposing monotonicity of predicted output:

$$
\begin{aligned}
\mathcal{L} = & -\frac{1}{T}\sum_{t}^{T}\frac{1}{p}\sum_{j}^{p}[\mathbb{I}_{\{y_{t+h}\leq q^{\alpha_j}\}} \log\left\{\widehat{\mathfrak{g}}_{W,b,j}(z_t)\right\} \\
& \underbrace{+\left(1-\mathbb{I}_{\{y_{t+h}\leq q^{\alpha_j}\}}\right)\log\left\{1-\widehat{\mathfrak{g}}_{W,b,j}(z_t)\right\}]}_{\text{binary cross-entropy}} \\
& \underbrace{+\ \lambda_m \sum_{t}^{T}\sum_{j=1}^{p-1}\left(\widehat{\mathfrak{g}}_{W,b,j}(z_t) - \widehat{\mathfrak{g}}_{W,b,j+1}(z_t)\right)_+}_{\text{monotonicity penalty}} \qquad (4.6)
\end{aligned}
$$

where $(u)_+$ is a rectified linear units function, ReLU, $(u)_+ = \max\{u, 0\}$, which passes through only positive differences between two neighboring values, j and $j+1$, of Cumulative Distribution Function (CDF), those violating

the monotonicity condition, and $\mathbb{I}_{\{.\}}$ is an indicator function. This violation is controlled by the penalty parameter λ_m.[9] Note that in addition to its simplicity, ReLU is used for convenience reasons, allowing for general use.[10]

4.2.4 Networks Design and Estimation Steps

Due to the high dimensionality and nonlinearity of the problem, estimation of a deep neural network is a complex task. Estimation requires an optimal selection of parameters to provide good performance and avoid potential risks such as over-fitting or convergence problems. Moreover, each problem and data require specific careful choices to minimize the risks. Here, we provide a detailed summary of the model architectures and their estimations.

Learning, Regularization and Hyper-parameters

The selection of hyper-parameters, together with regularization methods, plays a crucial role in the reduction of risk from estimation. In particular, we use the ReLU activation function to introduce non-linearity to our problem and help the optimization algorithm converge faster. For the learning process, we use the adaptive gradient algorithm, Adam (Kingma & Ba, 2014) and its modification AdamW (Loshchilov & Hutter, 2019) that allows for regularization by decoupling the weight decay from the gradient-based update. The regularization is close to L_2-regularization with improved results.[11]

We use hyper-optimization algorithms based on random search over a grid/cube of parameter ranges, which are specific to a given experiment. For some experiments, one might search over the full grid of parameters, testing all possible combinations, which can be costly. Using hyper-optimization[12] we select the learning rate of the optimizer, η, the weight decay parameter of AdamW, λ_W, and a Dropout parameter regularizing models (Srivastava et al., 2014) that is an efficient way of performing model averaging with neural networks. Specifically, Dropout parameter turns-off a fraction $\phi \in (0.0, 1.0)$ of nodes in the layer of network at which it is applied. In an in-sample

[9] The choice of λ_m is specific to the problem and data we are working with.

[10] This choice allows the use of GPU and hence opens computational capacities for more complex problems. The use of own or not optimized functions for GPU is not desired, and $(u)_+$ is common to libraries working with GPUs.

[11] We keep decay of momentum parameters β_1, β_2 constant and at default values throughout all estimations, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

[12] We using Julia package HyperOpt.jl (https://github.com/baggepinnen/Hyperopt.jl).

training part, the model is given a number of epochs to learn on the data. The number of epochs depend on the size of the data and the batch size used in the estimation. However, we also use an early stopping technique that helps regularization to prevent from over-fitting. The early-stopping, or patience, criterion is the minimum number of epochs provided to the model to learn.

Code Implementation

We have estimated our models on 48 core Intel® Xeon® /i7 Gold 6126 CPU@ 2.60GHz, 128 GB of memory, and GeForce 3090 GPU. We implement the models using `Flux.jl` (Innes et al., 2018) package in JULIA 1.6.0. language.

Data Preparation and Information set

To predict the distribution function of a time-series y_t with observations $y_1, \ldots, y_T$, we split our time-series into several parts. The first partitioning creates, as known in time-series literature, in-sample $[1 : t_0]$ and out-of-sample $[t_0 + 1 : T]$ subsamples. Equivalently, in machine learning jargon, train and test sets. Test subsample is never available to the learning algorithm while training the model. We further divide the train subsample into training and validation sets, which are used to cross-validation of our model and the model's parameters selection. The model selection is based on the value of the loss function on the validation subsample(s), mainly the binary cross-entropy loss.

One of the crucial parts in the estimation of distributional neural networks is the information set. The information set $\mathcal{I}_{t_0}$ is based on the past observation available at time t_0. This is the maximal time-span providing historical information, in our case, up to the last observation of the validation subsample. The importance here lies in finding the empirical quantiles, q^{α}, corresponding to the set of probabilities $\{\alpha_1, \ldots, \alpha_p\}$, which are used to build the sequence of target values. Given the information about $\{y_t\}_{t=1}^{t_0}$ and the empirical quantiles, we are able to model the distribution conditional on the information set up to time t_0. Given the information set, we face the problem with non-variation or updating the conditional empirical quantiles for the future distributions. Although we do not assume the shape of the distribution, we assume, to some extent, small level of shift in the distribution. Further, the choice of empirical quantiles and probability levels faces

the same problem as quantile regression when it comes to small samples of data.

4.3 Empirical Application: Macroeconomic Fan-Charts in Era of Big Data

Macroeconomic fan charts are popular tools for communicating uncertainty surrounding economic forecasts. Recognizing the need to communicate uncertainty to public, the Bank of England started to publish fan charts in 1996, and quickly became a leader in communicating uncertainty. Yet the art and science of such an important tool for policy-making remains on shoulders of the methods chosen.

Here we aim to construct a data-driven macroeconomic fan chart from a best approximating model learned from hundreds of variables with deep learning. This is in sharp contrast to the literature providing uncertainty of macroeconomic variables using so-called prediction fan charts (Britton et al., 1998; Stock & Watson, 2017) using few variables with parametrized and structured model that requires a number of assumptions.

4.3.1 Data

To construct such a tool for measurement of uncertainty, we use a high-dimensional dataset of M. McCracken and Ng (2020) and M. W. McCracken and Ng (2016) that has been extensively used in the macroeconomic literature (Goulet Coulombe et al., 2022) and is available at the Federal Reserve of St-Louis's web site. From several alternatives, we opt for the harder to predict quarterly data FRED-QD. Our dataset contains 216 quarterly US macroeconomic and financial indicators observed from 1961Q1 until 2019Q4. Since a number of variables are non-stationary, we follow the transformation codes used by M. McCracken and Ng (2020). Using this dataset, we construct data-rich fan charts for real GDP growth (GDPC1), inflation (CPIAUCSL), and unemployment rate (UNRATE) that will, to the best of our knowledge, be the first of its kind to reflect high-dimensional information of 216 relevant variables. To contrast the data-driven fan charts, we use a state-of-the-art macroeconomic model based on Bayesian Vector Autoregression that con-

tains data-driven factors to incorporate the big data information (M. McCracken & Ng, 2020).

4.3.2 Deep-learning Based Fan-Charts

In order to obtain an h-step ahead forecasts that will form a fan chart, we consider a direct prediction scheme. Exploring the data structures, we form h distributional networks

$$\widehat{\mathfrak{g}}^{(1)}_{W,b}(z_t), \ldots, \widehat{\mathfrak{g}}^{(h)}_{W,b}(z_t), \tag{4.7}$$

where entire (continuous) h-step-ahead conditional distribution $\widehat{\mathfrak{g}}^{(h)}_{W,b}(z_t)$ is obtained by interpolation of cumulative distribution function preserving the monotonicity of the outcome. Here we apply Fritsch-Carlson monotonic cubic interpolation (Fritsch & Carlson, 1980), for details see Appendix 4.A.1 and use the predicted cumulative distribution function $\widehat{F}_{t+h}(\alpha|\mathcal{I}_t)$ to form $k-$*size* prediction intervals for a fan chart as

$$PI^k_{t+h} = \left[\widehat{F}^{-1}_{t+h}(\alpha_l|\mathcal{I}_t), \widehat{F}^{-1}_{t+h}(\alpha_u|\mathcal{I}_t)\right], \tag{4.8}$$

such that $k = \alpha_u - \alpha_l$ is size of the interval.

To show how useful our approach is, we contrast the predictions obtained from the distributional network with the Bayesian vector auto-regression estimated on the factors extracted from data as in M. McCracken and Ng (2020). This benchmark is a state-of-the-art approach in macroeconomics and at the same time uses information from the whole dataset, hence the forecasts are comparable. To obtain fan chart (prediction intervals), we use the best performing recursive (iterative) scheme $\widehat{y}_{t+h} = f(y_{t+h-1}|\mathcal{I}_t)$ where the prediction intervals are based on the distribution of residuals. Formally, when we assume normal distribution we obtain h-step ahead α prediction interval as $[\widehat{y}_{t+h} - \phi(1-\alpha/2)\widehat{\sigma}_h,\ \widehat{y}_{t+h} + \phi(1-\alpha/2)\widehat{\sigma}_h]$, where $\phi(1-\alpha/2)$ is corresponding quantile of the std. normal distribution, and, for example, for a naive forecast we have $\hat{\sigma}_h = \hat{\sigma}\sqrt{h}$ and $\hat{\sigma}_h$ is the residual standard deviation.[13]

We evaluate the h-step-ahead forecasts with quantile loss function (Clements et al., 2008)

$$L^h_{\alpha,m} = E(\alpha - \mathbb{I}\{e_{t+h,m} < 0\})e_{t+h,m}), \tag{4.9}$$

[13] Alternatively, one can obtain prediction intervals or fan charts using Bootstrap methods, Britton et al. (1998).

for a model m, horizon h, and α-quantile where $e_{t+h,m} = y_{t+h} - \widehat{F}^{-1}_{t+h,m}(\alpha|\mathcal{I}_t)$ is the difference between the original time-series and α-quantile forecast given the information set, $\mathcal{I}_t$. To compare the predictive accuracy of the models, we use the Diebold-Mariano test (Diebold & Mariano, 1995) with Newey-West variance for $h > 1$ cases and test the null hypothesis $H_0 : L_{\alpha,m_1} > L_{\alpha,m_2}$ against the alternative that m_2 is less accurate than m_1.

4.3.3 Setup

Working with quarterly data, we compute $h = 1, \ldots, 6$ horizon forecasts each quarter of the out-of-sample period starting at 2012:Q3 and ending with 2019:Q4. Conditional distribution is approximated with $j = 1, \ldots, 19$ empirical $\alpha_j = (0.01, ..., 0.99)$ probability levels. The learning explores 36 combinations of hyper-parameters to find the best approximating model for each h-step ahead forecast separately. The hyper-parameters space is optimized once on the training part of the data, and the training procedure performs growing-window forward-validation scheme on the training data using 3-folds. We split the training data while training each fold of validation on train and test parts by a ratio of 0.93. We present predictions for deep recurrent neural networks with two hidden LSTM layers of different numbers of neurons chosen in the hyper-optimization. Table 4.5 in the Appendix summarizes all parameters and details used in the estimation. To compare the deep-learning based fan charts, we perform the standard estimation procedure for the Bayesian Vector Autoregression (BVAR) model with four-factor components as in M. McCracken and Ng (2020). We use the information criteria[14] and choose the model with four lags. Further, we find the prediction intervals for GDP growth, inflation (Consumer Price Index (CPI)), and the unemployment rate (UNE). The data for both procedures are transformed according to M. McCracken and Ng (2020) codes and standardized to normal with zero mean and standard deviation one.

4.3.4 Discussion

We start the discussion by presenting the qualitative results of GDP growth, inflation and unemployment predictions in the form of fan charts. Figure 4.3 compares median as well as 50% and 90% prediction intervals made

[14] Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)

at four different periods by both recurrent distributional neural network and Bayesian vector auto-regression approaches and highlights the benefits of deep learning-based predictions.

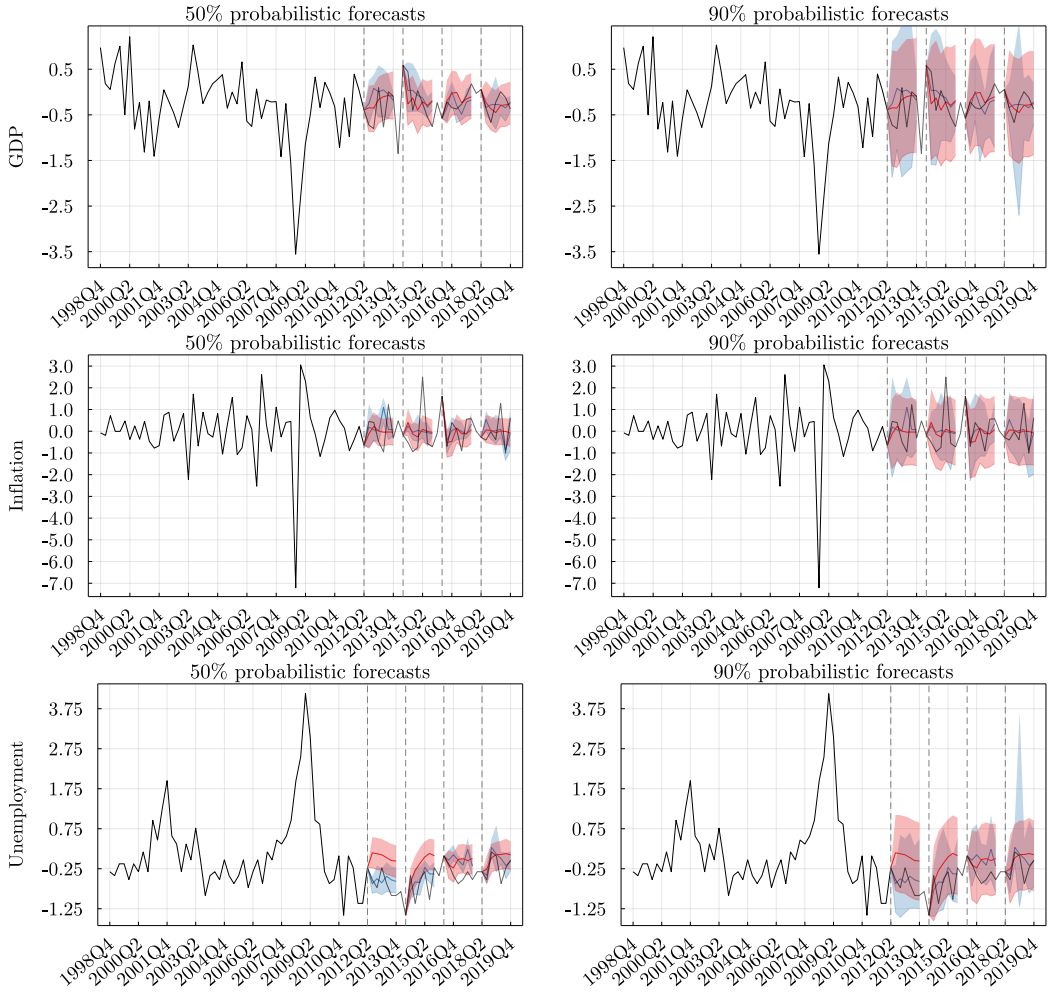

Figure 4.3: Deep-learning based (blue) and BVAR (red) fan charts
6-step-ahead quarterly forecasts of GDP growth (top), Inflation (middle), and Unemployment rate (bottom) with 50% (left column) and 90% (right column) fan charts obtained by the distributional network (blue) using 216 quarterly US macroeconomic and financial indicators from the FRED-QD database, and a factor three+four-variable BVAR (red). Forecasts are made at the end of the 2012:Q2, 2014:Q2, 2016:Q2 and 2018:Q2 depicted by dashed vertical lines. Train data is plotted by black solid line, and test data by grey solid line.

The prediction intervals from the distributional neural network are asymmetric and, in contrast to traditional time-series represented by BVAR, are not very smooth over the forecast horizon. Even with growing uncertainty when looking ahead into the future, deep learning still learns some structure from data, and probability intervals resemble less the form of "fans". The intervals are narrower in comparison to BVAR in most of the cases, especially

when looking at 50% intervals. In the case of the unemployment rate, the BVAR model is less capable of reducing the uncertainty about the future observation most probably because of strong spikes in previous years. In contrast, our DistrNN with LSTM units captures the uncertainty precisely.

While Figure 4.3 is illustrative, it only shows a few periods, and to support the gains of deep learning approach, we further quantify the prediction differences for the whole out-of-sample period. Tables 4.1 and 4.2 present the quantitative comparison of predictions from both models. We compare the forecasts at $h = 1, \dots, 6$ horizons using tick loss (Eq. 4.9) for selected $\alpha = \{0.1, 0.25, 0.5, 0.75, 0.9\}$ probability levels.

Table 4.1: Quantile loss of DistrNN and BVAR

	0.1		0.25		0.5		0.75		0.9	
	DistrNN	BVAR	DistrNN	BVAR	DistrNN	BVAR	DistrNN	BVAR	DistrNN	BVAR
GDP										
$h = 1$	**1.985**	2.146	**2.829**	3.645	**3.751**	4.280	3.118	3.060	**1.917**	2.381
$h = 2$	**2.253**	2.305	**3.338**	3.478	3.957	3.671	**2.800**	3.026	**1.826**	2.665
$h = 3$	2.317	2.193	**3.444**	3.510	**4.249**	4.307	3.524	3.203	**2.287**	2.363
$h = 4$	**1.977**	2.130	**3.437**	3.626	**4.013**	4.513	**2.931**	3.349	**1.710**	2.679
$h = 5$	**1.843**	2.199	**3.141**	3.193	**3.850**	4.222	**2.564**	3.188	**1.549**	2.546
$h = 6$	2.176	2.159	3.298	3.036	**3.776**	4.066	**2.682**	3.061	**1.909**	2.374
Inflation										
$h = 1$	**4.017**	4.051	**6.294**	6.884	**7.711**	9.509	**7.373**	8.552	**4.472**	7.925
$h = 2$	3.408	3.102	6.339	5.864	8.554	8.411	**7.325**	8.611	**3.959**	8.650
$h = 3$	**2.509**	3.171	5.319	5.295	8.394	7.937	8.291	8.252	**5.574**	8.453
$h = 4$	**2.899**	2.955	5.889	5.596	8.481	8.399	**7.485**	8.169	**5.023**	8.210
$h = 5$	3.454	3.037	5.973	5.261	8.004	7.944	**7.138**	8.102	**4.941**	8.101
$h = 6$	4.004	3.055	6.747	5.215	8.198	7.850	**6.941**	8.006	**4.645**	8.178
Unemployment										
$h = 1$	**1.950**	2.182	**3.060**	3.718	**3.586**	4.799	**2.928**	3.495	**1.639**	2.634
$h = 2$	**1.860**	2.359	**3.126**	4.022	**3.703**	4.814	**2.995**	3.155	**1.794**	2.168
$h = 3$	**1.902**	1.956	**3.407**	3.753	**4.004**	5.128	3.579	3.563	**2.486**	2.614
$h = 4$	2.605	1.841	3.712	3.666	**3.884**	5.600	**3.252**	3.769	**2.193**	2.735
$h = 5$	2.187	1.910	**3.851**	3.851	**4.383**	5.777	**3.263**	3.851	**1.816**	2.621
$h = 6$	2.113	1.916	**3.253**	3.751	**3.731**	5.776	**2.929**	3.728	**1.756**	2.518

Note: Quantiles losses of Distributional Recurrent Neural Network (DistrNN) and Bayesian VAR (BVAR), for variables GDP growth (GDP), Inflation, and Unemployment rate. The out-of-sample forecasts for 25 quarters are made at α-levels $\{0.1, 0.25, 0.5, 0.75, 0.9\}$, horizons $h = 1, \dots, 6$, starting at Q3/2013 and ending at Q4/2019. Cases with DistrNN forecast being smaller in comparison to BVAR are in blue.

In Table 4.1, we report quantile loss of both Distributional Recurrent Neural Network (DistrNN) and Bayesian VAR (BVAR) forecasts of GDP growth (GDP), inflation, and unemployment rate. Deep-learning based DistrNN approach provides forecasts with lower error (in blue) at most considered probability levels and horizons, and it brings larger improvement at shorter horizons. Notable gains are at 90% where DistrNN dominates BVAR greatly. The inflation is an exception, and DistrNN also improves losses for median and 25% level forecasts.

While deep-learning provides us with better forecasts in most of the cases,

Table 4.2: Relative out-of-sample performance of DistrNN and BVAR

	0.1	**0.25**	**0.5**	**0.75**	**0.9**
GDP					
$h = 1$	-0.4234	-2.0609 *	-0.8421	0.1479	-1.0619
$h = 2$	-0.1094	-0.4939	0.5027	-0.2906	-0.9199
$h = 3$	0.1516	-0.0821	-0.0757	0.5351	-0.0948
$h = 4$	-0.8808	-1.6944	-0.6268	-0.9966	-1.1122
$h = 5$	-1.4336	-0.4723	-0.4374	-2.5705 **	-1.3932
$h = 6$	0.0508	0.8273	-0.3862	-0.6077	-0.5314
Inflation					
$h = 1$	-0.0369	-0.5153	-1.6002	-0.8522	-1.718 *
$h = 2$	0.7832	0.5475	0.1647	-1.1302	-3.1631 ***
$h = 3$	-4.7111 ***	0.0436	1.0352	0.0345	-2.1701 **
$h = 4$	-0.3859	0.3249	0.151	-0.9863	-1.7402 *
$h = 5$	0.5635	3.9307 ***	0.101	-5.553 ***	-14.7935 ***
$h = 6$	0.8104	1.5175	0.2902	-1.0714	-2.1468 **
Unemployment					
$h = 1$	-0.3937	-0.8682	-1.7259 *	-1.0862	-1.2746
$h = 2$	-0.9097	-1.3733	-1.5014	-0.2501	-0.4865
$h = 3$	-0.1361	-0.4044	-1.4017	0.0155	-0.1076
$h = 4$	0.8105	0.0321	-1.3888	-0.4531	-0.5153
$h = 5$	1.479	-0.0008	-1.1089	-0.7159	-1.1653
$h = 6$	1.1521	-0.6429	-2.0862 **	-1.7029	-1.6384

Note: The values are Diebold-Mariano test statistics, with the null hypothesis H_0 : $L_{\alpha,DistrNN} > L_{\alpha,BVAR}$ against the alternative that Bayesian VAR is less accurate than the Distributional Recurrent Neural Network. Stars indicate statistical significance that ***, **, * correspond to 1%, 5%, 10% levels, accordingly. The negative sign of the DM statistics states that DistrNN has better OOS performance than BVAR, and the positive sign states the opposite. We report the out-of-sample forecasts for 25 quarters for three variables GDP growth (GDP), Inflation, and Unemployment rate, at α-levels $\{0.1, 0.25, 0.5, 0.75, 0.9\}$, horizons $h = 1, \ldots, 6$, starting at Q3/2013 and ending at Q4/2019.

a number of these cases is also significantly different from a traditional BVAR approach. Table 4.2 reports the results supporting the relative performance of the two methods. Deep learning delivers significantly better prediction in comparison to BVAR in 12 cases, while BVAR never outperforms the deep-learning approach significantly with the exception of the 5-step-ahead forecast of inflation at a 25% quantile level. We note that the results depend on 25 out-of-sample observations, which is a size that might be limiting for the test.

4.4 Empirical Application: Conditional Distributions of Asset Returns

Stock returns data are notoriously known to contain heavy tails and low signal-to-noise ratio (Fama, 1965; Israel et al., 2020). Despite the large literature uncovering these empirical properties, only a few studies attempt to forecast the distribution of returns.[15] Among the few, Anatolyev and Baruník (2019) parametrize a simple ordered logit to deliver the distribution forecasts.

Here, we aim to build a machine-learning based alternative that is capable of exploring a large number of informative variables. We compare the forecasts to the benchmark Anatolyev and Baruník (2019) (henceforth AB) model to see how well the machine learning approach approximates the parametrized model, but we mainly focus on using more variables that classical models can not explore due to its in-feasibility connected with large parameter space.

This application, hence, serves as a good complement to the previous one, where we used machine learning for multiple variable forecasts using big data. In contrast, liquid stock returns are known to be hardly predictable, and hence, even a small improvement is valuable.

4.4.1 Data and Estimation

Our dataset includes 29 most liquid U.S. stocks[16] of S&P500. The main reason for this particular choice is the comparability of the results with Anatolyev and Baruník (2019). The daily data covers the period from July 1, 2005 to August 31, 2018. We preprocessed the data to eliminate possible problems with liquidity or biases caused by weekends or bank holidays. The final sample period contains 3261 observations.

We start building the models using the same predictors as in Anatolyev and Baruník (2019) to make a direct comparison of the model forecasts. Specifically, they use $\text{Ind}_t = \mathbb{I}_{\{r_t \leq q^{\alpha_j}\}}$ and $\text{LogVol}_t = \ln(1 + |r_t|)$ as a proxy to a volatility measure. We will refer to this first choice as the *AB predictors*. Next, we prepare five realized measures from one-minute intra-day high-

[15] Literature focusing on Value-at-Risk forecasting has a special interest in a chosen quantile of the return distribution, mostly left tail (Engle & Manganelli, 2004)

[16] Assets selected in the sample: AAPL, AMZN, BAC, C, CMCSA, CSCO, CVX, DIS, GE, HD, IBM, INTC, JNJ, JPM, KO, MCD, MRK, MSFT, ORCL, PEP, PFE, PG, QCOM, SLB, T, VZ, WFC, WMT, XOM.

frequency data obtained from TickData.[17] The realized measures for each of 29 asset returns are realized volatility, skewness, kurtosis, and positive and negative semi-variances labelled as RVol_t, RSkew_t, RKurt_t, RSemiPos_t, and RSemiNeg_t. These are informative about returns distribution and should help the forecast. We will refer to this set of as *RM predictor*

Table 4.3: Sets of predictors used in the three models

AB predictors	RM predictors	AB+RM predictors
Ind_t	-	Ind_t
LogVol_t	-	LogVol_t
-	RVol_t	RVol_t
-	RSkew_t	RSkew_t
-	RKurt_t	RKurt_t
-	RSemiPos_t	RSemiPos_t
-	RSemiNeg_t	RSemiNeg_t

Note: The indicator Ind_t contains J columns of dummy variables.

In the third model, we combine both sets of predictors to estimate the conditional distribution of return since they are both informative. Since the realized measures contain information about higher moments of return distribution, these might improve predictions of the conditional distributions of returns. At the same time, the inclusion of those predictors into the original (benchmark) ordered logit model of Anatolyev and Baruník (2019) would result in an over-parametrized model that is not feasible. This is an important note since our approach provides a flexible and more general way of predicting distributions in data-rich environment and, at the same time, explores possible non-linearity in data. Table 4.3 summarizes the predictors used in the three models.

Prior to the estimation, we normalize the input data to an appropriate range, which eases the job for the algorithm to find a better optimum. This is a standard procedure in the learning process since the optimization operates on closer ranges while learning in the network structure. Further, we split the data into train and test parts with ratios 0.9 and 0.1, specifically to 2934 and 327 observations, respectively. First, we search for the best hyper-parameters set on the training window, at which we perform a four-fold rolling-window forward-validation scheme. The model is trained and validated during hyper-parameter search on each split composed from

[17] www.tickdata.com

90% and 10% partitions - training and validation. Using a rolling window of size 2934, we predict one step-ahead out-of-sample forecasts, H = 1. The window size equals to the size of the training sample, $t_0 = 2934$. On the first rolling window, the training part, we search the grid for the best parameters set using Random and Latin hypercube search algorithms. Table 4.6 in Appendix 4.A.2 details the ranges of parameters for the learning rate, η, dropout parameter, ϕ, and the weight decay penalizing parameter, and λ_W, on which the hyper-optimization algorithm is searching for the best model hyper-parameters in the space of 50 combinations. We employ the ensemble method for prediction, thus, for each rolling window step, the best model is trained three times given the best model hyper-parameters. Three forecasts are obtained, and an average distribution forecast is made for all t in out-of-sample, $t \in [2935 : 3261]$. In addition, we use additional regularization technique of early stopping and Table 4.6 in Appendix 4.A.2 also provides number of epochs allowed to train. This is a number of epochs the model is patient about the algorithm and waiting for improvement. Finally, we take the model with the best validation loss and use it for prediction of out-of-sample distributions. We also study the effect of complexity specifying the size of neural networks. We set the number of nodes from a shallow to a deeper DistrNN to $[128]$, $[128, 64]$, and $[128, 64, 32]$. The network's final layer outputs size is $p = 10$, whose values correspond to probability forecasts approximating the conditional distribution of excess returns given the information set $\mathcal{I}_t$.

4.4.2 Statistical Evaluation Measures

We evaluate our probabilistic forecasts using several measures. First, we evaluate the precision of forecasts using the mean square prediction error calculated as

$$MPSE = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} \frac{1}{p} \sum_{j=1}^{p} \left(\mathbb{I}_{\{y_{t+h} \leq q^{\alpha_j}\}} - \widehat{\mathfrak{g}}_{W,b,j}(z_t) \right)^2, \tag{4.10}$$

where the out-of-sample predicted outputs $\widehat{\mathfrak{g}}_{W,b,j}(z_t)$ is a matrix keeping dimension of time $[t_0 + 1, \ldots, T]$ and conditional probability levels $\{1, \ldots, p\}$.

To evaluate the compatibility of a cumulative distribution function with an individual time-series observations, we use the continuous ranked prob-

ability score (CRPS, Matheson and Winkler (1976), Hersbach (2000)):

$$CRPS_t = -\int_{-\infty}^{\infty} \left(\widehat{\mathfrak{g}}_{W,b}(z_t) - \mathbb{I}_{\{y_{t+h} \leq y\}} \right)^2 dy, \tag{4.11}$$

where the conditional CDF $\widehat{\mathfrak{g}}_{W,b}(z_t)$ is obtained by CDF interpolation (see Appendix 4.A.1) while the integral is computed numerically using the Gauss-Chebyshev quadrature formulas (Judd (1998), section 7.2) with 300 Chebychev quadrature nodes on $[2y_{min}, 2y_{max}]$. CRPS score is of the highest value when the distributions are equal. We obtain an average CRPS score of the out-of-sample forecast as $CRPS_{OOS} = \frac{1}{T}\sum_{t=t_0+1}^{T} CRPS_t$.

Another measure for the distributional forecast accuracy is Brier score (Gneiting & Raftery, 2007). At time t, it calculates a squared difference of binary realization and the probability forecast,

$$B_t = -\sum_{j=1}^{p+1} \left(\mathbb{I}_{\{q^{\alpha_{j-1}} < y_t \leq q^{\alpha_j}\}} - \widehat{\Pr}\{q^{\alpha_{j-1}} < y_t \leq q^{\alpha_j}\} \right)^2. \tag{4.12}$$

We also compute the average value of the Brier score for the out-of-sample period.

We compare proposed models using the relative predictive performance of two models, M_1 / M_2, where M_i is a particular measure (MPSE, CRPS, or Brier) corresponding model i. The model M_1 performs better when the ratio is lower than one.

4.4.3 Discussion

Table 4.4, Figure 4.4 and Figure 4.5 in Appendix 4.A.2 provide all out-of-sample results with the horizon $h = 1$ comparing sizes of various machine learning models and different variables used as predictors. The performance for all measures is put relative to the maximum likelihood ordered logit model, an AB benchmark of Anatolyev and Baruník (2019).

Overall, we document that machine learning is capable of forecasting the conditional distribution of asset returns well and providing informative variables it delivers improving predictions. Particularly, we observe an average 1% out-of-sample improvements in MSPE for all studied assets in comparison to the parametric models. The improvement is even larger in terms of the

Data	Model	**MSPE**	**Bin.CE**	**CRPS**	**Brier**
AB	NN:128	14/29	14/29	20/29	11/29
AB	NN:128x64	13/29	11/29	15/29	11/29
AB	NN:128x64x32	8/29	7/29	8/29	11/29
RM	NN:128	25/29	25/29	27/29	25/29
RM	NN:128x64	25/29	24/29	27/29	25/29
RM	NN:128x64x32	25/29	24/29	26/29	25/29
AB+RM	NN:128	22/29	19/29	27/29	19/29
AB+RM	NN:128x64	22/29	21/29	26/29	21/29
AB+RM	NN:128x64x32	23/29	23/29	27/29	23/29

Table 4.4: Results according to different scores, assets
The table shows the performance of three models with different sizes and of three different input features. All results are benchmarked to Ordered Logit of Anatolyev and Baruník (2019). The ratio indicates the number of times out of 29 that a given model outperforms the benchmark.

Continuous rank probability score evaluating the compatibility of predicted and data distributions.

It is important to note here that in financial forecasting, the relationship between statistical and economic gains from predictions is non-trivial. Campbell and Thompson (2008) and Rapach et al. (2010) note that seemingly small statistical improvement could generate large benefits in practice, which has recently been confirmed on expected returns forecast by machine learning (Babiak & Baruník, 2020; Gu et al., 2020). Hence an average 1 % improvement in the out-of-sample predictions we document will most likely be interesting for practitioners forming their portfolios based on our forecasts. While it is tempting to explore such a strategy, it is far beyond the scope and space of this text.

More specifically, Table 4.4 shows that, on average, neural networks of all sizes deliver better performance than the parameterised AB model when using the same set of predictors in about half of the stocks tested. This result suggests that data does not contain any further non-linearities that are not captured by a parametric AB model, and since machine learning is more flexible in number of parameters to be estimated, it learns and approximates the AB parametrization with a small degree of error when considering CRPS measure.

The situation changes when data contains additional predictors and the AB approach becomes infeasible, the machine learning approach offers pos-

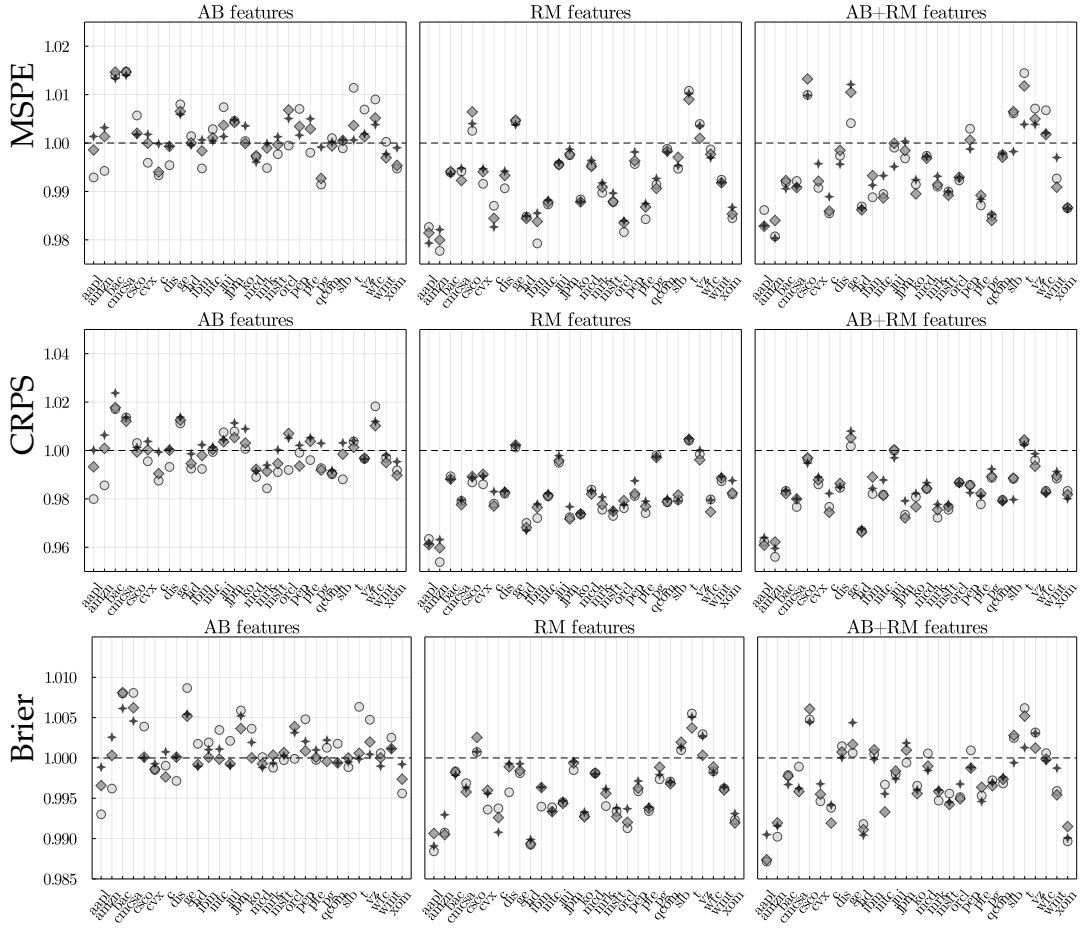

Figure 4.4: Comparison of the out-of-samples forecasts of 29 U.S. stocks
The three statistical measures, MSPE (top), CRPS (middle), and Brier score (bottom), are used for the three sets of predictors and three machine learning models depicted as *star* for 128, *diamond* for 128x64, and *circle* for 128x64x32. Anatolyev and Baruník (2019) ordered logit model is a benchmark with value 1. A value lower than 1 shows better better performance of a model in comparison to the benchmark.

sibility to explore how informative the predictors are for forecasts in terms of forecast errors. When additional five realised measures (RM) are used as predictors, the performance increases for all measures, see Table 4.4 rows with data RM and RM+AB. We observe that the NN approach outperforms the benchmark AB model for more than two thirds of the assets. With respect to the depth of networks, the shallow (NN 128) neural network shows the best results. This result is similar to Gu et al. (2020) who find that shallow network performs better than deeper structures one on asset returns data.

While Table 4.4 provides information on relative counts, Figures 4.5 and 4.4 complement it with all measures reported for individual assets in boxplots. The detailed look uncovers that machine learning improves the performance of individual stocks such as AAPL, AMZZN, GE, or WMT even

more. At the same time, Figure 4.5 shows that in most cases, deeper networks show lower variance for most of the stocks.

4.5 Conclusion

In this paper, we have proposed a new approach to modelling probability distributions of economic variables using state-of-the-art machine learning methods. The distributional neural network relaxes the assumption on the distribution family of time-series and lets the model explore the data fully. The approach is particularly beneficial for modelling data with non-Gaussian, non-linear and asymmetric structures. We show the usefulness of the approach in an economic and financial application. At the same time the approach is general and can be applied to any other dataset.

We have illustrated that our distributional neural network is useful in constructing big data-driven macroeconomic fan charts that are the first of their kind since they are learned from the structure between 216 relevant economic variables. Further, we illustrate how deep learning can be used to improve probabilistic forecasts of data notoriously known to contain low signal-to-noise ratios, heavy tails and asymmetries.

References

Anatolyev, S., & Baruník, J. (2019). Forecasting dynamic return distributions based on ordered binary choice. *International Journal of Forecasting*, *35*(3), 823–835.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Babiak, M., & Baruník, J. (2020). Deep learning, predictability, and optimal portfolio returns. *arXiv preprint arXiv:2009.03394*.

Baillie, R. T., & Kapetanios, G. (2007). Testing for neglected nonlinearity in long-memory models. *Journal of Business & Economic Statistics*, *25*(4), 447–461.

Baruník, J., & Hanus, L. (2024). Fan charts in era of big data and learning. *Finance Research Letters*, *61*, 105003.

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1046–1089.

Box, G. E., Draper, N. R., et al. (1987). *Empirical model-building and response surfaces* (Vol. 424). Wiley New York.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.

Breiman, L., et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Britton, E., Fisher, P., & Whitley, J. (1998). The inflation report projections: Understanding the fan chart. *Chart*, *8*(10).

Bryzgalova, S., Pelger, M., & Zhu, J. (2019). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458 (forthcoming in Journal of Finance)*.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Clements, M. P., Galvão, A. B., & Kim, J. H. (2008). Quantile forecasts of daily exchange rate returns from forecasts of realized volatility. *Journal of Empirical Finance*, *15*(4), 729–750.

Diebold, F. X. (2021). What's the big idea? "big data" and its origins. *Significance*, *18*(1), 36–37.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3).

Duan, T., Avati, A., Ding, D. Y., Basu, S., Ng, A. Y., & Schuler, A. (2019). Ngboost: Natural gradient boosting for probabilistic prediction. *arXiv preprint arXiv:1910.03225*.

Engle, R. F., & Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367–381.

Fama, E. F. (1965). Portfolio analysis in a stable paretian market. *Management Science*, *11*(3), 404–419.

Feng, G., He, J., & Polson, N. G. (2018). Deep learning for predicting asset returns. *arXiv preprint arXiv:1804.09314*.

Foresi, S., & Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, *90*(430), 451–466.

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition methods in economics. In *Handbook of labor economics* (pp. 1–102, Vol. 4). Elsevier.

Friedman, W. A. (2009). The harvard economic service and the problems of forecasting. *History of Political Economy*, *41*(1), 57–88.

Fritsch, F., & Carlson, R. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, *17*(2).

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., & Januschowski, T. (2019). Probabilistic forecasting with spline quantile function rnns. *The 22nd International Conference on Artificial Intelligence and Statistics*, 1901–1910.

Geweke, J., & Whiteman, C. (2006). Bayesian forecasting. *Handbook of economic forecasting*, *1*, 3–80.

Gneiting, T. (2008). Probabilistic forecasting. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 319–321.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5), 920–964.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*, *33*(1), 3–12.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hu, T., Guo, Q., Li, Z., Shen, X., & Sun, H. (2019). Distribution-free probability density forecast through deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(2), 612–625.

Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance*, *49*(3), 851–889.

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: The state space approach*. Springer Science & Business Media.

Chen, L., Pelger, M., & Zhu, J. (2020). Deep learning in asset pricing. *Available at SSRN 3350138*.

Chen, Y., Kang, Y., Chen, Y., & Wang, Z. (2020). Probabilistic forecasting with temporal convolutional neural network. *Neurocomputing*.

Chernozhukov, V., Fernández-Val, I., & Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, *81*(6), 2205–2268.

Innes, M., Saba, E., Fischer, K., Gandhi, D., Rudilosso, M. C., Joy, N. M., Karmali, T., Pal, A., & Shah, V. (2018). Fashionable modelling with flux. *CoRR*, *abs/1811.01457*.

Israel, R., Kelly, B. T., & Moskowitz, T. J. (2020). Can machines "learn" finance? *Journal of Investment Management*, *18*(2), 23–36.

Iworiso, J., & Vrontos, S. (2020). On the directional predictability of equity premium using machine learning techniques. *Journal of Forecasting*, *39*(3), 449–469.

Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., & Callot, L. (2020). Criteria for classifying forecasting methods. *International Journal of Forecasting*, *36*(1), 167–177.

Judd, K. (1998). *Numerical methods in economics*. MIT Press.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 33–50.

Kuan, C.-M., & White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews*, *13*(1), 1–91.

Lahiri, K., Martin, G., et al. (2010). Bayesian forecasting in economics. *International Journal of Forecasting*, *26*(2), 211–215.

Leorato, S., & Peracchi, F. (2015). Comparing distribution and quantile regression. *EIEF Working Papers Series*, (No. 1511).

Lerner, A. P. (1947). Measuring business cycles. by arthur f. burns and wesley c. mitchell. *The Journal of Economic History*, *7*(2), 222–226.

Lim, Y.-S., & Gorse, D. (2020). Deep probabilistic modelling of price movements for high-frequency trading. *arXiv preprint arXiv:2004.01498*.

Liu, C., & Maheu, J. M. (2009). Forecasting realized volatility: A Bayesian model-averaging approach. *Journal of Applied Econometrics*, *24*(5), 709–733.

Lopez de Prado, M. (2019). *Beyond econometrics: A roadmap towards financial machine learning*.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096.

McCracken, M., & Ng, S. (2020). *FRED-QD: A quarterly database for macroeconomic research* (tech. rep.). National Bureau of Economic Research NBER Working Paper No. 26872.

McCracken, M. W., & Ng, S. (2016). FRED-MD: a monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Racine, J. (2001). On the nonlinear predictability of stock returns using financial and economic variables. *Journal of Business & Economic Statistics*, *19*(3), 380–382.

Rapach, D., Strauss, J., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, *23*(2), 821–862.

Rothe, C. (2012). Partial distributional policy effects. *Econometrica*, *80*(5), 2269–2301.

Sadhwani, A., Giesecke, K., & Sirignano, J. (2020). Deep Learning for Mortgage Risk*. *Journal of Financial Econometrics*, *19*(2), 313–368.

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181–1191.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, *58*(3), 644–719.

Stock, J. H., & Watson, M. W. (2017). Twenty years of time series econometrics in ten pictures. *Journal of Economic Perspectives*, *31*(2), 59–86.

Timmermann, A. (2000). Density forecasting in economics and finance. *Journal of Forecasting*, *19*(4), 231.

Tobek, O., & Hronec, M. (2020). Does it pay to follow anomalies research? machine learning approach with international evidence. *Journal of Financial Markets*, 100588.

Wen, R., Torkkola, K., Narayanaswamy, B., & Madeka, D. (2017). A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*.

Wright, J. H. (2008). Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics*, *146*(2), 329–341.

Zhang, Z., Zohren, S., & Roberts, S. (2020). Deep learning for portfolio optimization. *The Journal of Financial Data Science*.

Žikeš, F., & Baruník, J. (2016). Semi-parametric conditional quantile models for financial returns and realized volatility. *Journal of Financial Econometrics*, *14*(1), 185–226.

4.A Appendix

4.A.1 CDF interpolation

The Fritsch–Carlson monotonic cubic interpolation (Fritsch & Carlson, 1980) provides a monotonically increasing CDF with range $[0, 1]$ when applied to CDF estimates on a finite grid.

Suppose we have CDF $F(y)$ defined at points $(y_k, F(y_k))$ for $k = 1, \dots, K$, where $F(y_0) = 0$ and $F(y_K) = 1$. We presume that $y_k < y_{k+1}$ and $F(y_k) < F(y_{k+1})$ for all $k = 0, \dots, K-1$, which is warranted by continuity of returns and construction of the estimated distribution. First, we compute slopes of the secant lines as $\Delta_k = (F(y_{k+1}) - F(y_k)))/(y_{k+1} - y_k)$ for $k = 1, \dots, K-1$, and then the tangents at every data point as $m_1 = \Delta_1$, $m_k = \frac{1}{2}(\Delta_{k-1} + \Delta_k)$ for $k = 2, \dots, K-1$, and $m_K = \Delta_{K-1}$. Let $\alpha_k = m_k/\Delta_k$ and $\beta_k = m_{k+1}/\Delta_k$ for $k = 1, \dots, K-1$. If $\alpha_k^2 + \beta_k^2 > 9$ for some $k = 1, \dots, K-1$, then we set $m_k = \tau_k \alpha_k \Delta_k$ and $m_{k+1} = \tau_k \beta_k \Delta_k$, with $\tau_k = 3(\alpha_k^2 + \beta_k^2)^{-1/2}$. Finally, the cubic Hermite spline is applied: for any $y \in [y_k, y_{k+1}]$ for some $k = 0, \dots, K-1$, we evaluate $F(y)$ as

$$F(y) = (2t^3 - 3t^2 + 1)F(y_k) + (t^3 - 2t^2 + t)hy_k + (-2t^3 + 3t^2)F(y_{k+1}) + (t^3 - t^2)hm_{k+1},$$

where $h = y_{k+1} - y_k$ and $t = (y - y_k)/h$.

4.A.2 Additional tables and figures

Hyper parameters	**Values**
Learning rate, η	0.0001, 0.001, 0.005
Dropout rate, ϕ	0.2, 0.4
L_2-decay regularization rate, λ_W	0.00001, 0.00005
Nodes dimensions	32x32, 64x64, 60x50
Fixed parameters	Value
Number of layers	2
Mini batch size	8
Epochs	350
Monotonicity parameter, λ_m	5.0
Cross-validation, k-folds	3
Train/test ratio	0.93
Ensembles	1

Table 4.5: Fan chart recurrent DistrNN parameters space for the empirical application, Sec. 4.3

The hyperoptimization algorithm searches through the whole hyperparameter space and tries all sets/combinations of hyperparameters to evaluate the model.

Hyper parameters	**Minimum value**	**Maximum value**
Learning rate, η	0.0001	0.02
Dropout rate, ϕ	0.0	0.5
L_2-decay regularization rate, λ_W	0.0	0.0018
Fixed parameters	Value	
Epochs	250	
Early stopping patience	25	
Monotonicity parameter, λ_m	0.2	
Mini batch size	32	
Ensembles	3	
Number of layers	1, 2, 3	
Nodes dimensions	128, 128x64, 128x64x32	

Table 4.6: DistrNN parameters space for the emprical application, Sec. 4.4

The hyperoptimization algorithm searches through the hyperparameter space and randomly tries sets of parameters to evaluate the model.

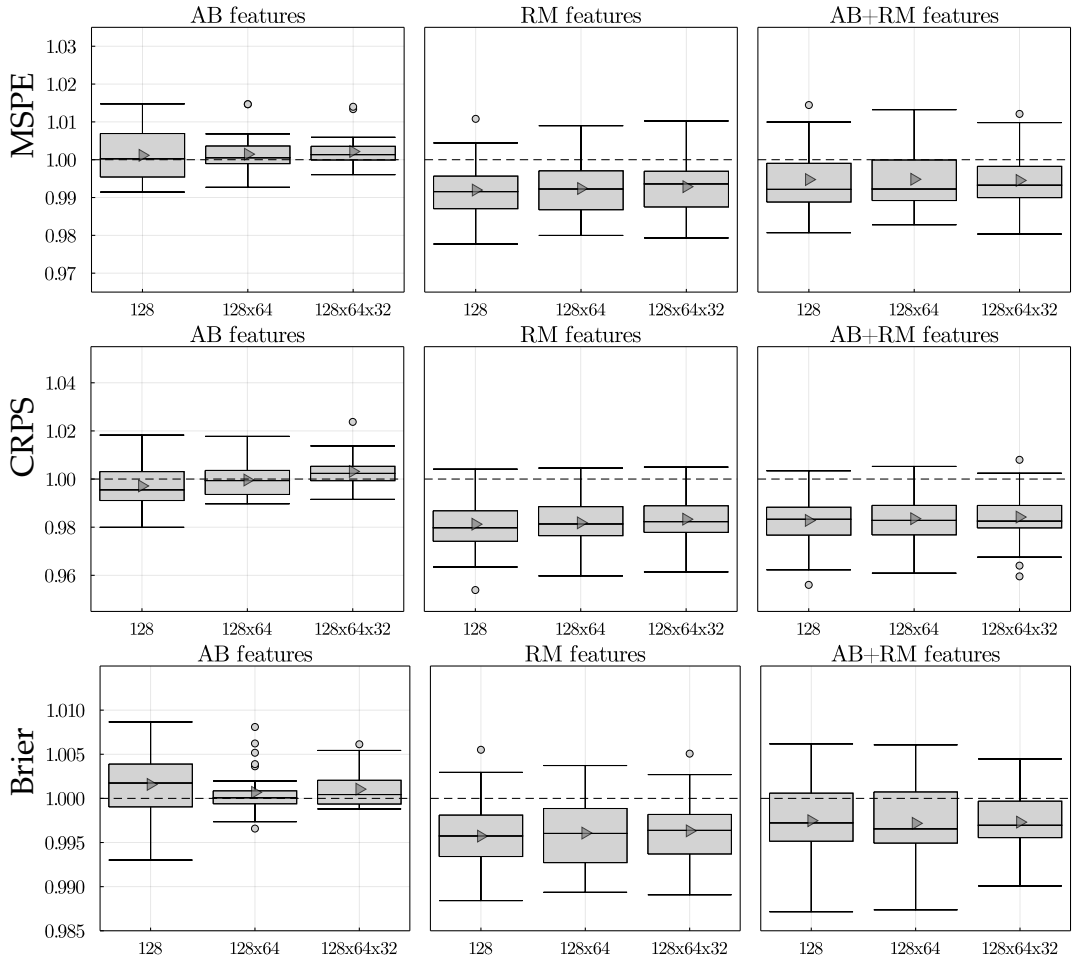

Figure 4.5: Boxplot comparison of the out-of-sample forecasts
The three statistical measures: MSPE (top), CRPS (middle), and Brier (bottom). Each boxplot depicts benchmark values of 29 assets of given NN model size. Anatolyev and Baruník (2019) ordered logit model is benchmark=1. Value lower than 1 states that the purposed model is better that benchmark.

Chapter 5

Learning probability distributions of day-ahead electricity prices

Abstract

We propose a novel machine learning approach to probabilistic forecasting of hourly day-ahead electricity prices. In contrast to recent advances in data-rich probabilistic forecasting that approximate the distributions with some features such as moments, our method is non-parametric and selects the best distribution from all possible empirical distributions learned from the data. The model we propose is a multiple output neural network with a monotonicity adjusting penalty. Such a distributional neural network can learn complex patterns in electricity prices from data-rich environments and it outperforms state-of-the-art benchmarks.

This chapter was co-authored with Jozef Baruník and has been submitted and is under review. We are grateful to Wolfgang Hardle, Lukáš Vácha, František Čech, and the participants at various conferences and research seminars for many useful comments, suggestions, and discussions. We gratefully acknowledge the support from the Czech Science Foundation under the EXPRO GX19-28231X project. We provide the computational package `DistrNNEnegry.jl` in JULIA available at https://github.com/luboshanus/DistrNNEnergy.jl that allows one to use our measures on time-series data.

5.1 Introduction

> "We will make electricity so cheap that only the rich will burn candles."
>
> —Thomas A.Edison (1880)

Electricity is essential to modern life. Its prices are inherently difficult to predict due to its complex, non-linear nature driven by the dynamics of supply and demand, incorporating uncertainty about future price movements is key to decision making in energy companies. While researchers have primarily focused on point forecasts over a long period of time, with weather-dependent renewable energy sources, turbulent times leading to increased imbalances between production and consumption, and higher price volatility (Maciejowska, 2020), research focusing on probabilistic price forecasting has recently gained importance (Nowotarski & Weron, 2018; Petropoulos et al., 2022). Probabilistic forecasting is rapidly becoming essential for producers, retailers and traders who need to assess uncertainty and improve optimal strategies for short-term operations, derivative pricing, value-at-risk, hedging and trading (Bunn et al., 2016).

Hand in hand with this surge in probabilistic electricity forecasting, researchers eager to use large numbers of series to understand fluctuations in electricity prices are collecting data unimaginable a few decades ago. Faced with an explosion in the volume, velocity and variety of data, the need to unlock the information hidden in big data has become a key issue not only in energy economics (Diebold, 2021). Challenged by the proliferation of parameters and the strong criticism of arbitrarily chosen restrictions in both reduced and structured models in recent decades, economists wishing to explore the potentially rich information content of new datasets have recently turned their hopes to machine learning (Mullainathan & Spiess, 2017). A key idea of (machine) learning, which can be thought of as the inference of plausible models to explain observed data, has recently attracted a number of researchers who document how learning patterns from data can be useful (Bianchi et al., 2021; Feng et al., 2018; Goulet Coulombe et al., 2022; Gu et al., 2020; Heaton et al., 2017; Israel et al., 2020; Iworiso & Vrontos, 2020; Mullainathan & Spiess, 2017; Sadhwani et al., 2020; Tobek & Hronec, 2020). A burgeoning literature and a growing number of applications in energy economics focus mostly on cross-sectional data and, ultimately, point forecasts. While machines can use such models to make predictions about future data,

shifting the focus from point forecasting to probabilistic forecasting using big data is an essential next step.

The contribution of this paper is that we propose a novel machine learning approach to probabilistic forecasting of hourly day-ahead electricity prices. Specifically, we propose a distributional neural network to provide probabilistic forecasts that reflect the time-series dynamics of large amounts of available information relevant to future prices. Such data-driven probabilistic forecasts aim to improve the current state of the art in forecasting and communicating uncertainty. Our approach provides data-rich forecasts that are not constrained by distributional or other model assumptions, hence fully allows the exploration of non-Gaussian, heavy-tailed and asymmetric data. Our distributional neural network significantly outperforms state-of-the-art methods. Finally, we provide an efficient computational package.

In the empirical exercise, we build a distributional network to forecast German hourly day-ahead electricity prices using the 221 characteristics, including lagged prices, total load, external variables such as EU allowance prices, fuel prices, in particular coal, gas and oil. Such data-rich uncertainty forecasts, which do not rely on distributional assumptions or model choice, are the first of their kind. To benchmark our framework, we use the naive model and the two quantile regression-based linear models with autoregressive and exogenous variables: quantile regression averaging (QRA) and quantile regression committee machine (QRM) Marcjasz et al. (2020) and Nowotarski and Weron (2015) estimated with lasso estimated autoregression. Our model significantly outperforms the benchmarks.

Why should we believe that machine learning can improve probabilistic forecasting? Classical time-series econometrics (Box et al., 2015; Hyndman et al., 2008) focuses mainly on predetermined autocorrelation or seasonality structures in data that are parameterised. When researchers have large amounts of time series at their disposal, these methods quickly become infeasible and unable to explore more complex data structures. Bearing in mind the famous adage that *"all models are wrong..., but some of them are useful."* (Box, Draper, et al., 1987), modern machine learning methods can easily overcome these problems. As a powerful tool for approximating complex and unknown data structures (Kuan & White, 1994), these methods can be useful in a range of application problems where data contain a rich information structure that cannot be satisfactorily described by a simplifying model. Overcoming the long-standing problem of computational intensity of such

data-driven approach with advances in computer science adds to the temptation to use these methods to address new problems such as distribution prediction.

The use of machine learning methods is also emerging in the literature on hourly electricity price forecasting. Lago et al. (2021b), Lehna et al. (2022), and F. Zhang et al. (2022) use a hybrid recurrent network for point forecasts. Nowotarski and Weron (2018) reviews recent advances in probabilistic forecasting, while Mashlakov et al. (2021) uses probabilistic forecasting with auto-regressive recurrent networks (DeepAR) developed by Amazon Research Germany (Salinas et al., 2020). Marcjasz et al. (2020) also uses nonlinear autoregressive networks with exogenous variables, Klein et al. (2023) constructs deep recurrent networks. Mashlakov et al. (2021) assesses the performance of deep learning models for multivariate probabilistic forecasting, and Marcjasz et al. (2023) uses a deep neural network with output from the normal and Johnson's SU distributions.

All these approaches rely on restrictive models and assumptions, particularly, the literature usually proposes to learn only some features of the distribution, such as moments. In contrast, our approach selects the best distribution from all possible empirical distributions learned from the data.

5.2 Probabilistic forecasting via distributional neural network

Consider hourly day-ahead electricity time-series $y_{t,h}$ collected over $t = 1\ldots,T$ days and $h = 1,\ldots,24$ hours. The main objective is to approximate as closely as possible the conditional cumulative distribution function $F(y_{t,h}|\mathcal{I}_{t-1})$ and use it for a 1-step-ahead probabilistic forecast made at time $t-1$ with information $\mathcal{I}_{t-1}$ containing past values of $y_{t,h}$ and possibly past values of other exogenous observable variables x_t.

The main goal is then to approximate a collection of conditional probabilities corresponding to the empirical quantiles, such as

$$\left\{F(q_h^{\alpha_1}),\ldots,F(q_h^{\alpha_p})\right\} = \left\{\Pr\left(y_{t,h} \leq q_h^{\alpha_1}|\mathcal{I}_{t-1}\right),\ldots,\Pr\left(y_{t,h} \leq q_h^{\alpha_p}|\mathcal{I}_{t-1}\right)\right\}$$

for the collection of thresholds $1,\ldots,p$. A convenient way to estimate such quantities is distribution regression. Foresi and Peracchi (1995) noted that

several binary regressions serve as good partial descriptions of the conditional distribution. To estimate the conditional distribution, one can simply consider a distributional regression model with a (monotonically increasing) link function, such as logit, probit, linear, log-log functions. In contrast to estimating separate models for separate thresholds, Chernozhukov et al. (2013) considered a continuum of binary regressions and argued that it provides a coherent and flexible model for the entire conditional distribution as well as a useful alternative to Koenker and Bassett Jr (1978)'s quantile regression. Alternatively, Anatolyev and Baruník (2019) suggest binding the coefficients of predictors in an ordered logit model via smooth dependence on corresponding probability levels. While this approach is able to predict the entire distribution, keeping $0 < F_j < 1$ and $0 < F_1(.) < F_2(.) < \dots F_p(.) < 1$, it still depends on a strong parameterisation suited to a specific problem of the time-series considered, making it an unfeasible approach for a larger number of variables.

5.2.1 Distributional neural network

Such probabilistic predictions are highly dependent on the model parameterisation and quickly become infeasible with increasing number of covariates. This motivates us to reformulate distributional regression into a more general and flexible distributional neural network. The functional form of the new network is driven by the data, and we can relax assumptions about the distribution of the data, the parametric model as well as the stationarity of the data. The proposed distributional neural network, as a feed-forward network, is a hierarchical chain of layers representing high-dimensional and/or non-linear input variables with the aim of predicting the target output variable. Importantly, we approximate the conditional distribution function with multiple outputs of the network as a set of joint probabilities.

As a first step, we replace a known link function with an unknown general function $\mathfrak{g}$, which is approximated by a neural network. Next, we consider a set of probabilities corresponding to $0 < \alpha_1 < \alpha_2 < \dots < \alpha_p < 1$ being p regularly spaced levels that characterise the conditional distribution function using a set of predictors z_t to be specified later, and model them jointly as

$$\left\{ \Pr\left(y_{t,h} \leq q_h^{\alpha_1} | z_{t-1}\right), \dots, \Pr\left(y_{t,h} \leq q_h^{\alpha_p} | z_{t-1}\right) \right\} = \mathfrak{g}_{W,b,h}(z_{t-1}), \qquad (5.1)$$

where $\mathfrak{g}_{W,b,h}$ is a multiple output neural network with L hidden layers that we name as distributional neural network:

$$\mathfrak{g}_{W,b,h}(\boldsymbol{z}_{t-1}) = g^{(L)}_{W^{(L)},b^{(L)}} \circ \ldots \circ g^{(1)}_{W^{(1)},b^{(1)}}(\boldsymbol{z}_{t-1}), \tag{5.2}$$

where $W = \left(W^{(1)}, \ldots, W^{(L)}\right)$ and $b = \left(b^{(1)}, \ldots, b^{(L)}\right)$ are weight matrices and bias vector. Any weight matrix $W^{(\ell)} \in \mathbb{R}^{m \times n}$ contain m neurons as n column vectors $W^{(\ell)} = [w^{(\ell)}_{\cdot,1}, \ldots, w^{(\ell)}_{\cdot,n}]$, and $b^{(\ell)}$ are thresholds or activation levels.

It is important to note that, in contrast to the literature, we consider a multi-output (deep) neural network to characterise the collection of probabilities. Before discussing the estimation details that allow us to preserve the monotonicity of the probabilities, we illustrate the framework. Figure 5.1 illustrates how $l \in 1, ..., L$ hidden layers transform input data into a chain using a collection of non-linear activation functions $g^{(1)}, \ldots, g^{(L)}$. A commonly used activation function, $g^{(\ell)}_{W^{(\ell)},b^{(\ell)}}$, is used as the

$$g^{(\ell)}_{W^{(\ell)},b^{(\ell)}} := g_\ell\left(W^{(\ell)}\boldsymbol{z}_{t-1} + b^{(\ell)}\right) = g_\ell\left(\sum_{i=1}^{m} W^{(\ell)}_i \boldsymbol{z}_{t-1} + b^{(\ell)}_i\right)$$

are a sigmoid $g_\ell(u) = \sigma(u) = 1/(1+\exp(-u))$, rectified linear units $g_\ell(u) = \max\{u, 0\}$, or $g_\ell(u) = \tanh(u)$. In case $\mathfrak{g}_{W,b,h}(u)$ is non-linear, neural network complexity grows with increasing number of neurons m, and with increasing number of hidden layers L and we build a deep neural network. We use activation function $g^{(L)}(\cdot) = \sigma(\cdot)$ to transform outputs to probabilities. Note that for $L = 1$, neural network becomes a simple logistic regression.

5.2.2 Loss Function

Since we want to estimate the cumulative distribution function (CDF), which is a non-decreasing function bounded on $[0, 1]$, we need to design an objective function that minimises the differences between the target and the estimated distribution, as well as imposing a non-decreasing property on the output. As the problem is essentially a more complex classification problem, logistic regression, we use a binary cross-entropy loss function. In addition, we introduce a penalty to the multiple output classification problem to order the predicted probabilities.

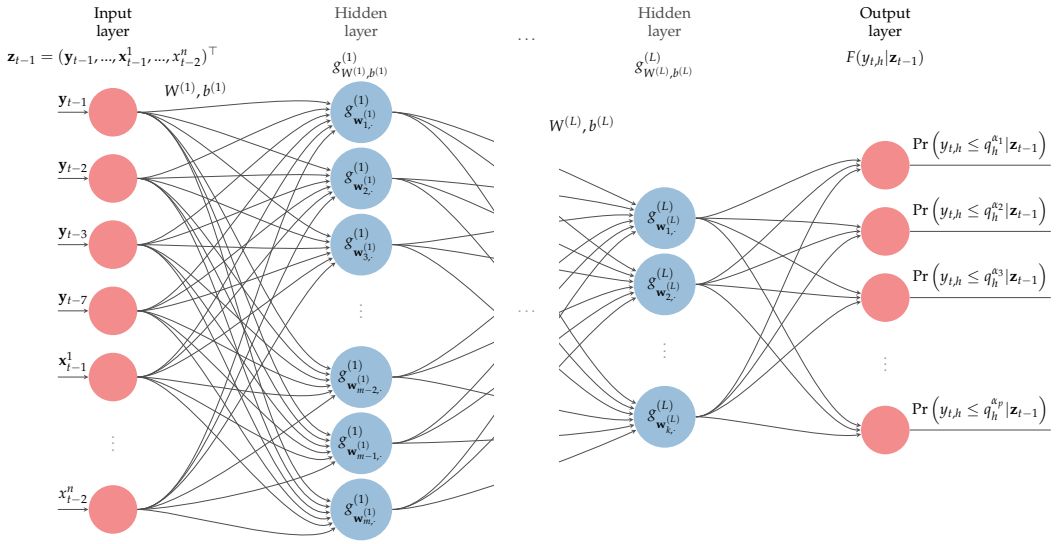

Figure 5.1: Distributional (Deep) Feed-forward Network.

An illustration of a multiple output (deep) neural network $\mathfrak{g}_{W,b,h}(\mathbf{z}_{t-1})$ to model the collection of conditional probabilities $\left\{\Pr\left(y_{t,h} \leq q_h^{\alpha_1}|\mathbf{z}_{t-1}\right), \ldots, \Pr\left(y_{t,h} \leq q_h^{\alpha_p}|\mathbf{z}_{t-1}\right)\right\}$ with set of predictor variables $\mathbf{z}_t = (\boldsymbol{y}_t, \boldsymbol{x}_t^1, ..., \boldsymbol{x}_t^n)^\top$. Large number of hidden layers L makes the network deep.

The loss function is then composed of two parts: traditional binary cross-entropy and a penalty adjusting for monotonicity of predicted output:

$$\begin{aligned}\mathcal{L} \;=\; & -\frac{1}{T}\sum_t^T \frac{1}{p}\sum_j^p \Big(\mathbb{I}\{y_{t,h} \leq q_h^{\alpha_j}\}\log\left\{\widehat{\mathfrak{g}}_{W,b,h,j}(\mathbf{z}_{t-1})\right\} \\ & \underbrace{+\left(1-\mathbb{I}\{y_{t,h} \leq q_h^{\alpha_j}\}\right)\log\left\{1-\widehat{\mathfrak{g}}_{W,b,h,j}(\mathbf{z}_{t-1})\right\}\Big)}_{\text{binary cross-entropy}} \qquad (5.3) \\ & + \underbrace{\lambda_m \sum_t^T \sum_{j=1}^{p-1}\left(\widehat{\mathfrak{g}}_{W,b,h,j}(\mathbf{z}_{t-1}) - \widehat{\mathfrak{g}}_{W,b,h,j+1}(\mathbf{z}_{t-1})\right)_+}_{\text{monotonicity penalty}} \qquad (5.4)\end{aligned}$$

where $(u)_+$ is a rectified linear units function, ReLU, $(u)_+ = \max\{u, 0\}$, which passes through only positive differences between two neighbouring values, j and $j+1$, of CDF, those violating the monotonicity condition, and $\mathbb{I}\{.\}$ is an indicator function. This violation is controlled by the penalty parameter λ_m. Note that in addition to its simplicity, ReLU is used for convenience reasons allowing for general use.[1]

[1]This choice allows to use GPU and hence opens computational capacities for more

5.3 Data

We use the spot market price, which is important for day-ahead auctions. In the day-ahead electricity market, the day-ahead forecast is used to formulate bids for 24 hours, which generally means that on day $t-1$ participants submit bids for the 24 hours of the day-ahead. These are executed up to a certain hour (deadline), after which the market clears and participants receive energy allocations at the clearing price. We consider the hourly day-ahead electricity market in Germany. The data cover the period from 7 January 2015 to 31 December 2020.[2]

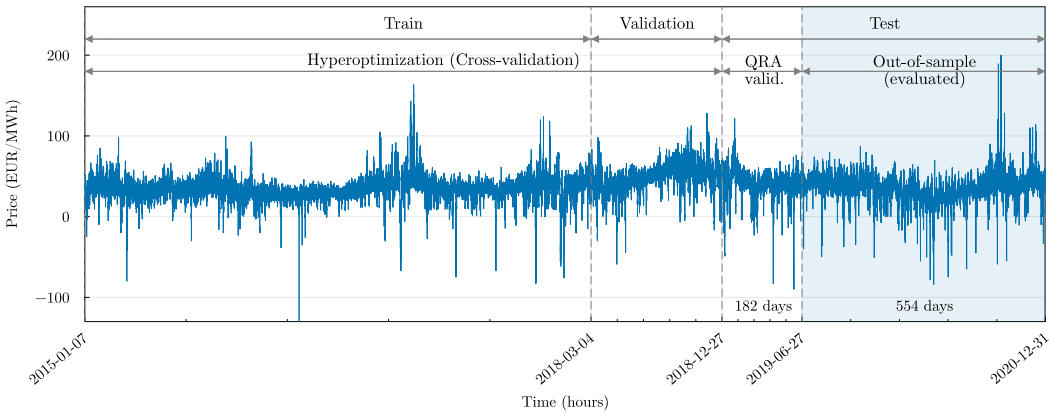

Figure 5.2: Electricity price data with distinguished and depicted periods for estimation. Train and Validation subsamples show how much of data is used for training the model at both stages, hyper-optimization training and rolling window learning.

To carry out the forecasting exercise of day-ahead electricity prices $y_{t,h}$ with hourly observations, we partition the data, in machine learning jargon, into train, validation, and test sets. As illustrated in Figure 5.2, we consider the last 736 days, period as the test subsample (out-of-sample, OOS). The days prior to OOS are used as training and estimation, which we split into train and validation parts. The block QRA validation is the first 182 days of OOS partition are used as a calibration window for the quantile regression as calibration window, thus we do not consider this part when we evaluate OOS results of the distributional neural network. This leaves us with last

complex problems. The use of own or not optimized functions for GPU is not desired and $(u)_+$ is common to libraries working with GPUs.

[2]Data accompany the text of Marcjasz et al. (2023) and are available online. Also available at https://transparency.entsoe.eu/.

554 days period between 2019-06-07 and 2020-12-31, which is a sufficient number of days considered for good practice to evaluate techniques in the EPF literature. Test subsample (out-of-sample, OOS) is never available to the learning algorithm while training the model. We further divide the train subsample into training and validation sets, which are used to cross-validation of neural network and to find its parameters selection.

5.3.1 Data transformation

Prior to the estimation procedure, we transform the data, as is common in the EPF literature, in order to stabilise its variance and make the distribution more symmetrical. Since German electricity prices are allowed to be negative, we cannot use a logarithmic transformation. We adopt the variance-stabilising transformation of Uniejewski et al. (2018) in its simpler form, as also discussed in Narajewski and Ziel (2020). We do the median *normalisation* of the price, $p_{t,h}$, as $p^n_{t,h} = 1/b(p_{t,h} - \widehat{a})$, where $\widehat{b} = \text{MAD}(p_{t,h|\mathcal{I}_{t-1}})1/z_{0.75}$, $\widehat{a} = \text{median}(p_{t,h|\mathcal{I}_{t-1}})$, and $1/z_{0.75} = 1.4826$ is the 75% quantile of $\mathcal{N}(0,1)$. After this first step, we apply *inverse hyperbolic transformation* such that $y^n_{t,h} = \text{asinh}(p^n_{t,h})$ to obtain more variance stable and symmetric price data.

To get the data back to the original scale, we do the transformation in the inverse order such that $\widehat{\mathbb{E}(p_{t,h})} = \widehat{\mathbb{E}(y_{t,h})} \cdot \widehat{b} + \widehat{a} \approx \sinh(\widehat{\mathbb{E}(y_{t,h})}) \cdot \widehat{b} + \widehat{a}$, where $\widehat{\mathbb{E}(y_{t,h})})$ is a model forecast for the price (quantiles), sinh is the hyperbolic sin function, and $\widehat{b}$ and $\widehat{a}$ are sample parameters of the transformation, MAD and median, respectively. In Figure 5.3, we plot histograms of day-ahead price prior and after the transformation showing changes in scale and shape.

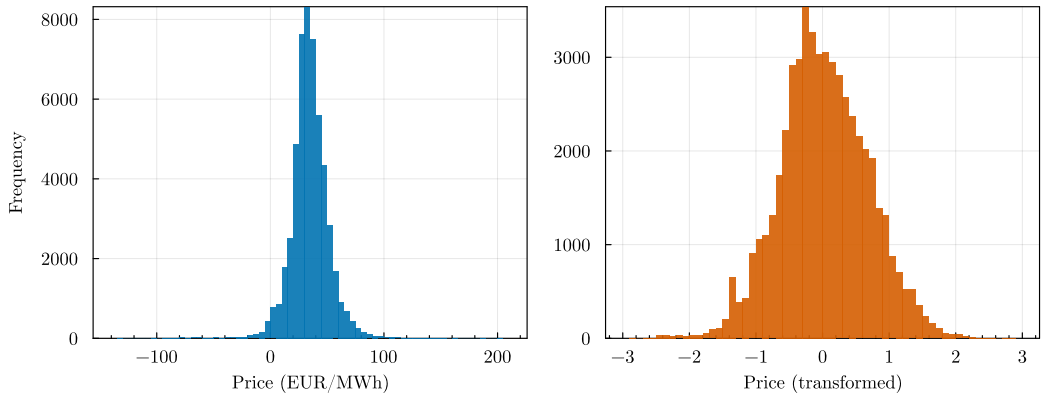

Figure 5.3: Histogram of original and transformed price data.

5.3.2 Input variables

Based on the data provided by Marcjasz et al. (2023), a consistent approach is taken for constructing the inputs for all models. The input features correspond to the day-ahead price data at their respective time points, resulting in 24 hour-ahead prices denoted by $\boldsymbol{y}_t = [y_{t,1}, \dots y_{t,24}]$. The inputs $\boldsymbol{z}_{t-1}$ constitute the information set $\mathcal{I}_{t-1}$, which includes historical price data and other exogenous variables.

However, as neural networks estimate time-dependent variables with intricate and non-linear relationships, it remains necessary to provide time-series lagged inputs, since the data are autocorrelated and seasonal patterns such as daily and weekly are present. Therefore, we begin by incorporating previous *day-ahead prices* as lags, specifically $\boldsymbol{y}_{t-1}$, $\boldsymbol{y}_{t-2}$, $\boldsymbol{y}_{t-3}$, and $\boldsymbol{y}_{t-7}$. Next, the variable *total load* is significant in the EPF studies as it's a targeted variable. We incorporate all hours of the day-ahead forecast of total load for the previous two days, including $\boldsymbol{x}_t^1$, $\boldsymbol{x}_{t-1}^1$, and $\boldsymbol{x}_{t-7}^1$. The final variable to be incorporated in the 24-hour size pertains to a day-ahead prediction of *renewable energy sources*. For this, we include data for the day ahead and the prior day, $\boldsymbol{x}_t^2$ and $\boldsymbol{x}_{t-1}^2$. Other external variables to be included are the closing prices of EU allowances, x_{t-2}^3, and the prices of fuels, in particular *coal, gas and oil*, x_{t-2}^4, x_{t-2}^5 and x_{t-2}^6. Since we forecast (t) today, these costs reflect the most recent data available, from two days ago, following the standard practice of the day-ahead auction market. Finally, to address the weekly pattern in the data, we incorporate a vector of weekday dummies, $x_{t,weekday}^7$ for the specific day of the week. Total number of columns in the input matrix is 221 features. We consider inputs $\boldsymbol{z}_t = [\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \boldsymbol{y}_{t-3}, \boldsymbol{y}_{t-7}, \boldsymbol{x}_t^1, \boldsymbol{x}_{t-1}^1, \boldsymbol{x}_{t-7}^1, \boldsymbol{x}_t^2, \boldsymbol{x}_{t-1}^2, x_{t-2}^3, x_{t-2}^4, x_{t-2}^5, x_{t-2}^6, x_{t,weekday}^7]$ for all models except the naive one.

5.3.3 Target variable and the information set

To forecast the probability of the day-ahead electricity price being below specific quantile levels, we model it as a set of probabilities based on the conditional price information set, i.e. $\Pr\left(y_{t,h} \leq q_h^{\alpha_j} | \mathcal{I}_{t-1}\right)$. When forecasting, it is crucial to ensure the information set is set accurately. The previous observations prior to the day on which the forecast is made make up the

information set $\mathcal{I}_{t-1}$. In our setup, we consider the data of the training and validation subsamples available to the information set.

The accuracy of the forecast outcomes is largely dependent on the precisely defined empirical quantiles, q_h^{α}, which correspond to a set of probabilities $\{\alpha\}$. Due to different location, size, and shape that influence hourly price distributions, target variable is on an hourly basis. The predicted variable is a set of hourly indicators related to p equidistant probability levels $\alpha_j = \{0.01, \ldots, 0.99\}$, where $p = 31$.[3] As a result, the target variable is

$$\boldsymbol{y}_{t,h,\alpha_j} = \mathbb{I}\{y_{t,h} \leq q_h^{\alpha_j} | \mathcal{I}_{t-1}\}, \text{ for } h = 1, ..., 24, \forall \alpha_j. \tag{5.5}$$

The unconditional quantiles defined by hours allow us to assume that the distribution within the information set is hour specific.

Last, the data for the target variable are subjected to *winsorisation*, we use a proportion of 0.1% to deal with the extreme minimum and maximum values within the information set. Eq. 5.5 shows that the cumulative distribution function approximation approach uses unconditional quantiles q^{α_j} for the indicator of the target variable. Winsorisation with a small fraction has no effect because the lowest and highest α values are less than 0.1%. The handling of extreme outliers, e.g. negative prices, is beneficial in the post-estimation inverse transformation of $\widehat{F}(y_{t,h})$ into $\widehat{F}^{-1}(\alpha)$, where we follow Fritsch and Carlson (1980).

5.4 Estimation

We begin by presenting the forecasting setup and outlining our approach of non-parametric distributional neural network procedure. Next, we propose benchmark models, such as the naive model, and then consider two versions of quantile regression based on a linear model with autoregressive and exogenous variables (Marcjasz et al., 2023; Nowotarski & Weron, 2018; Serafin et al., 2019), which are stable benchmarks in the probabilistic electricity price forecasting literature. The bids in the auctions are posted once a day for all hours, we follow the literature (Liu et al., 2017; Maciejowska et al., 2016) and predict the distributions of day-ahead prices for each hour given the same information for all hours.

[3] We have also experimented with different number of probability levels p and while the results not change we have used $p = 31$ as a sufficient approximation.

5.4.1 Distributional neural network

In line with the saying, *"a picture is worth a thousand words"*, we present cumulative distribution functions and quantile functions in Figure 5.4, which encapsulate the idea of how we obtain distributional forecast, from left to right. The left panel illustrates the approximation of CDFs for every day and hour using p equidistant points that correspond to α_j probabilities. In the plot on the right, we show the predicted 99 quantiles of day-ahead prices for given hours, obtained by inversion and interpolation. Both figures are based on the unconditional values of the complete data set. Nonetheless, they demonstrate how our results look for a single forecasted day of 24 hours.

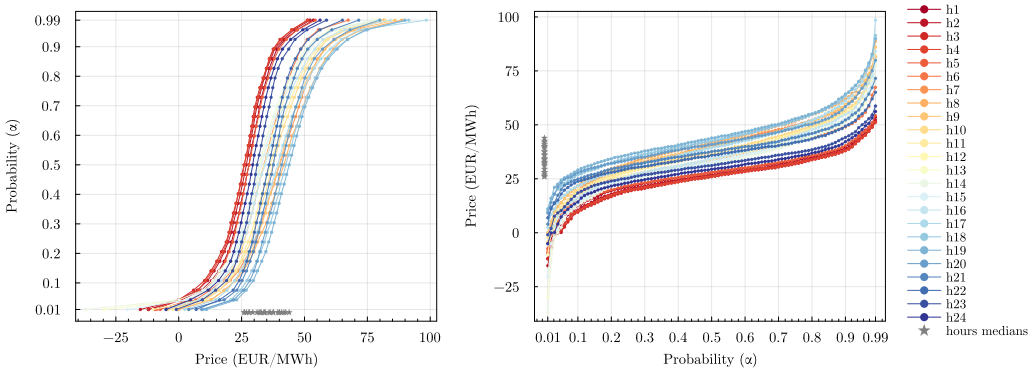

Figure 5.4: Unconditional cumulative distribution, $F_h(y_{y,h})$, and quantile functions by hours (depicted by warm to cold colors), $Q_h(\alpha)$, for whole data sample. Left: Illustrates 31 α probability levels used to provide our target variable and similar to these unconditional CDF we train the conditional one. Right: Unconditional quantiles for 99 α probability levels mimic the final result of distributional forecasts.

Our distributional neural network is a multilayer perceptron, it has two (possibly more) hidden layers, with each layer containing different numbers of neurons, which are fine-tuned by hyper-optimization. The DistrNN's input size is 221 features, which is taken in. We make no assumptions regarding the shape of distributions, as DistrNN directly outputs the vector of probabilities, which are 31 values approximating the CDF. We allow for a different distribution for each hour and as a result, we have twenty-four distributional neural networks to train.

To develop and analyse the model, we utilise the Julia programming language, specifically using the Flux.jl package (Innes et al., 2018) for neural network training. Most neural networks have essential components, such as

optimisation algorithms and techniques to prevent over-fitting. We implement AdamW (Loshchilov & Hutter, 2019) as our optimisation algorithm, which includes regularisation techniques. AdamW mimics $L_2 - norm$ regularisation through its weight decay as learning occurs. Then, we apply dropout regularization method (Srivastava et al., 2014) and batch normalization to the first layer's weights in the model. The learning rate for stochastic gradient descent (Adam, Kingma and Ba (2014)), denoted as η, the weight decay regularising parameter, denoted as λ_W, and the proportion parameter for dropout, indicating how many neurons to turn off in each layer, denoted as ϕ, are subjected to hyper-optimisation.

Training and hyper-optimization tuning

Our forecasting procedure is similar to other forecasting studies that make use of a daily data forward rolling scheme. We use data from a training and validation period consisting of four and a half years (shown in Figure 5.2) to perform a hyper-optimization search for parameters. To reduce computational costs, the hyper-optimization is conducted before implementing the rolling window scheme. We use k-folds cross validation on randomly shuffled data to increase the possibility of model generalisation rather than data memorisation. The dataset is divided into seven cross-validation sets, with data being separated into a 1:7 train-validation ratio. We must determine the ideal parameters for the distributional neural network for every hour of the day-ahead prices. For hyper-optimization,[4] the algorithm considers 60 parameter combinations in a grid fashion based on the parameter ranges and sets provided in Table 5.1. The top parameters set is that with the smallest mean of validation losses from cross-validation. We run our neural network for a maximum of 1000 epochs. Additionally, we employ early stopping with a patience of 15 epochs when the validation loss does not show improvement. Furthermore, we utilise batches of 64 data points. The input data for DistrNN are augmented with noise from $\mathcal{N}(0, 0.1)$.

Forward rolling forecasting (recalibration)

We evaluate the models on the out-of-sample period of 736 days, shown as a test in Figure 5.2, with focus on the shaded area of the last 554 days. Starting from 12 December 2018, we train DistrNN using a tuned parameter set for

[4]We use the Julia package Hyperopt.jl

Table 5.1: Parameter values used to train distributional neural network.

Hyper parameters	Values	Fixed parameters	Value
Learning rate, η	$Range_1(0.0001, 0.003)$	Number of HPO combinations	60
Dropout rate, ϕ	$Range_2(0.0, 1.0)$	Epochs	1000
L_2-decay rate, λ_W	$Range_2(0.000001, 0.01)$	Early stopping patience	15
Hidden neurons in each layer	$Range_2(32, 384)$	Monotonicity, λ_m	1.5
Mini batch size	$\{64\}$	Number of layers	2
Activation functions	{relu, tanh, sigmoid, softmax}	α levels	31
		CV k-folds	7
		Ensembles	8

Note: The hyper-optimization algorithm searches through the space of hyperparameters and randomly tries a number of parameters sets to train a network. $Range_1$ is evenly spaced log range, $Range_2$ is evenly spaced linear range.

each hour of the day in a rolling window fashion. To train DistrNN, we minimise the loss given by the binary cross entropy function (Eq. 5.3). In this part, we keep four and a half years of data available for training, and the split is with a ratio of 80% for training and 20% for validation subsamples of shuffled data. To reduce the forecast variance, we train the model several times with different initialisation of weights and biases. Taking into account the validation loss, we only consider the first better half of the results, i.e. with an ensemble size of $n = 8$ forecasts, we consider the first four, $\{\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_{n/2}\}$.

The distributional neural network predicts a CDF, $\widehat{F}_{t,h}(\cdot|\mathcal{I}_{t-1})$, which we use to find its inverse, $\widehat{F}^{-1}(y_{t,h}) = \widehat{Q}(\alpha)$, which is the quantile function, more precisely a collection of quantiles for a given $\alpha_j = \{0.01, \ldots, 0.99\}$. Before the inversion, we use the monotone cubic interpolation of (Fritsch & Carlson, 1980). On the interval [0,1] we obtain a monotonically increasing CDF on a finite grid of 400 points. We find the inverse function before aggregating the predicted DistrNN CDFs $\widehat{F}^i_{t,h}$. Thus, in the case of neural networks, the aggregated ensemble mean is the mean of the predictions. We average the predictions over the quantiles $\bar{Q}(\alpha) = \frac{1}{N_{ens}} \sum_i^{N_{ens}} \widehat{F}_i^{-1}(y)$, and evaluate $\bar{Q}_{t,h}(\alpha)$ as our result. Note that we consider ensembles over quantiles rather than probabilities, although both quantile and probability averaging of out-of-sample ensembles give similar results in Marcjasz et al. (2023).

We face the inverse problem of quantile crossing: possible violation of monotonicity of the cumulative distribution function. To solve this, we propose a loss function (Eq. 5.3) that penalises for such occurrences. During learning, the algorithm only retains models with parameters that satisfy the monotonicity condition during learning.

5.4.2 Naive benchmark

To be consistent with the literature, we use the naive model in this paper. The model, as the name suggests, is a simple way of predicting the next day's price distribution using the previous day's or week's prices. Once the price point forecasts are available, one can bootstrap the price distribution from the errors of a given day between the predicted price and the true price (Marcjasz et al., 2023; Nowotarski & Weron, 2015; Weron, 2014; Ziel & Weron, 2018). The expected price for day t and hour h is

$$\widehat{\mathbb{E}\left(y_{t,h}\right)} = \begin{cases} y_{t-7,h} & \text{for Monday, Saturday, and Sunday,} \\ y_{t-1,h} & \text{for Tuesday, Wednesday, Thursday, and Friday.} \end{cases} \tag{5.6}$$

Then the errors for one day, $\widehat{\varepsilon}_t = y_t - \widehat{y}_t$, are bootstrapped and added to the predicted prices from Eq. 5.6 such that

$$\widehat{y}^i_{t,h} = \widehat{\mathbb{E}(y_{t,h})} + \widehat{\varepsilon}^i_t, \text{ for } i \in 1, ..., M, \tag{5.7}$$

which gives the naive distributional forecasts from the sampled prices.

5.4.3 QRA and QRM benchmarks

For parametric distributional forecasting, we consider two approaches popular in the literature. We use the quantile regression averaging (QRA) introduced at EPF by Nowotarski and Weron (2015) and the quantile regression committee machine (QRM) of Marcjasz et al. (2020). Both quantile regression models require point forecasts of the price to estimate the probabilistic forecasts. It is argued that the Lasso Estimated Auto-Regressive (LEAR) model of Uniejewski et al. (2016) may be the most accurate linear model (Lago et al., 2021b) for point forecasting. We consider this prominent state-of-the-art model as a sufficient parametric benchmark model (Marcjasz et al., 2023; Mpfumali et al., 2019; Uniejewski et al., 2019; W. Zhang et al., 2018).

The LEAR model of Lago et al. (2021b) is a linear regression model with numerous parameters, both autoregressive and exogenous, estimated using LASSO regularisation (Tibshirani, 1996). To ensure consistency with the existing literature and reproducibility, we follow the specifications of (Lago et al., 2021b; Marcjasz et al., 2023) and use calibration windows of identical length to forecast electricity prices. The LEAR procedure uses a forward

rolling window scheme, with each window based on an information set of 56, 84, 1092 and 1456 days. The LEAR model requires the selection of a hyperparameter - the regularisation parameter λ. It uses a cross-validation scheme with 7-fold search on a grid of 100 values and chooses to use the least angle regression (Efron et al., 2004).[5] This produces four-point OOS forecasts of size 736 days, which are used in the QRA scheme to obtain probabilistic forecasts. The model is estimated independently for each hour h, while the information set is the same for each day t.

Both QRA and QRM are estimated using quantile regression (Koenker & Bassett Jr, 1978), which is used to predict the conditional α-quantile of $y_{t,h}$) with a set of regressors. For QRA, the regressors consist of the intercept and four LEAR price forecasts, [$\mathbf{1}$, $\widehat{\boldsymbol{y}}_t^{56}$, $\widehat{\boldsymbol{y}}_t^{84}$, $\widehat{\boldsymbol{y}}_t^{1092}$, $\widehat{\boldsymbol{y}}_t^{1456}$] to perform quantile averaging. For QRM, we compute the average of the LEAR forecasts (LEAR-Avg), referring to the name of the *"committee machine"* that is taken as input. The estimation is done by minimising the quantile loss function (Eq. 5.10) for each α quantile. We estimate 99 quantiles to approximate the future distribution of prices as closely as possible. In the forward rolling scheme, we use the in-sample (calibration) window of 6 months (182 days) to obtain 554 days of out-of-sample results. According to Serafin et al. (2019), the performance of QRM is better than QRA, which is not necessarily true for every valuation metric, e.g. Marcjasz et al. (2023).

5.4.4 Evaluation criteria

We assess the quality of the probabilistic forecast using two measures. First, the empirical analysis focuses on the reliability and uncertainty of the forecasts, in other words, the prediction intervals. We evaluate forecast intervals of size $(1-\alpha)$ using the unconditional coverage score, or α coverage, which measures whether or not the price occurs within such an interval. The occurrence rate should be close to the nominal value of the interval, i.e. if the prediction interval is $(1-\alpha) = 90\%$, the occurrence or coverage should be as close as possible to 90%-coverage.

Further, to focus on sharpness of the probabilistic forecasts of all models we follow Gneiting and Raftery (2007) and use the Continuous Rank

[5]Other options could be to use the Akaike information criterion or the Bayesian information criterion.

Probability Score (CRPS) measure

$$CRPS_{t,h}(\widehat{F}_{y_{t,h}}, y_{t,h}) = \int_{\mathbb{R}} (\widehat{F}_{y_{t,h}}(z) - \mathbb{I}\{y_{t,h} \leq z\})^2 dz, \tag{5.8}$$

where $\mathbb{I}\{y_{t,h} \leq z\}$ is the indicator function. As is common in the EPF literature, we use the discrete approximation of the CRPS as

$$CRPS_{t,h} = \frac{1}{N_\alpha} \sum_{\alpha=0.01}^{0.99} QL_\alpha(\widehat{q}^\alpha_{t,h}, y_{t,h}), \tag{5.9}$$

where QL_α is the α quantile loss function, or pinball loss, which we state as

$$QL_\alpha(\widehat{q}^\alpha_{t,h}, y_{t,h}) = (\mathbb{I}\{y_{t,h} \leq \widehat{q}^\alpha_{t,h}\} - \alpha)(\widehat{q}^\alpha_{t,h} - y_{t,h}), \tag{5.10}$$

where α is probability, $\widehat{q}^\alpha_{t,h}$ is the quantile prediction obtained from $\widehat{F}_{t,h}$, $y_{t,h}$ is the original time-series, and N_α is the number of quantile probability levels we approximate the quantile function from CDF. In this way we approximate the CRPS sum of pinball scores over the discrete set of $\alpha = \{0.01, 0.02, ..., 0.99\}$ for all out-of-sample.

To complement the measures of distributional accuracy, we also report standard metrics for assessing median forecasts. Two criteria for median accuracy are mean absolute error and root mean squared error accuracy measures, where the lower the criteria, the better the accuracy, but this does not guarantee the quality of the model. First, mean absolute error (MAE)

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{H} \sum_{h=1}^{H} |y_{t,h} - \widehat{y}_{t,h}|, \tag{5.11}$$

and second, the Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \frac{1}{H} \sum_{h=1}^{H} (y_{t,h} - \widehat{y}_{t,h})^2}, \tag{5.12}$$

where $y_{t,h}$ is the electricity price and $\widehat{y}_{t,h}$ is the predicted median price.

To assess the significance of forecast accuracy and performance between models, we use the Diebold-Mariano test (Diebold & Mariano, 1995), the version with adjusted Newey-West variance. Using the DM test, we take two approaches to testing. First, we compare the errors of the models on a day-ahead basis, and second, we test the disaggregated accuracy for each

of the 24 hours. In our evaluation, we have a loss for model m and hour h denoted as $L_m^{t,h}$, i.e. the vector of CRPS loss. In the overall test, we aggregate losses to days where $L_m^t = \sum_{h=1}^{24} L_m^{t,h}$ and measure statistical significance between all pairs between models. We specify the null hypothesis about two models that the difference of the models' $L_1 - norm$ is lower-equal to zero as $\mathcal{H}_0 : \mathbb{E}[\Delta_{m_1,m_2}^t] \leq 0$, where formally $\Delta_{m_1,m_2}^t = ||L_{m_1}^t||_1 - ||L_{m_2}^t||_1$. Let us consider the disaggregated differences between the accuracy of the models, so that for $h = 1, ..., 24$ with losses $L_m^{t,h}$ we test the null hypothesis $\mathcal{H}_0^h : \mathbb{E}[\Delta_{m_1,m_2}^{t,h}] \leq 0$, where $\Delta_{m_1,m_2}^{t,h} = ||L_{m_1}^{t,h}||_1 - ||L_{m_2}^{t,h}||$. The null hypotheses of both tests are against the alternative that m_2 is more accurate than m_1 (Clements et al., 2008; Nowotarski & Weron, 2018).[6]

5.5 Results

This section presents the results of all methods for German day-ahead prices. It provides results of accuracy and quality measures as well as results of statistical tests. We start with the Figure 5.5, which shows the example of the probabilistic forecast of electricity prices from our distribution networks together with the actual realised price. We can see that the model provides asymmetric probability forecasts that respond precisely to the data.

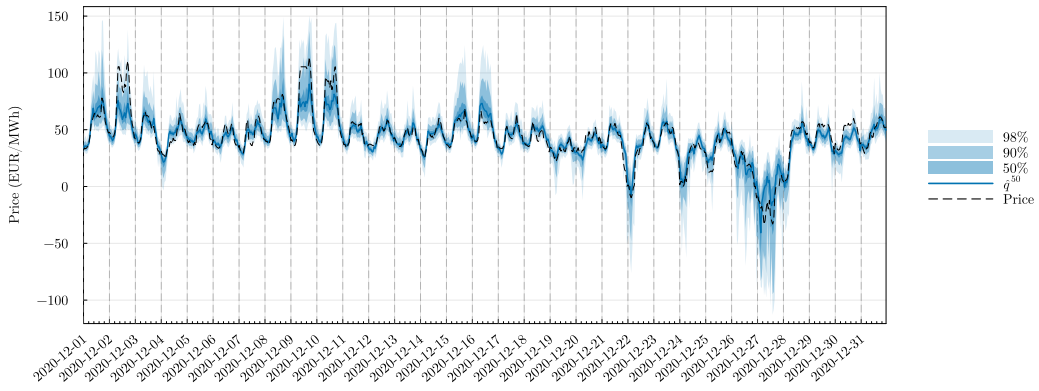

Figure 5.5: Example of electricity price probabilistic forecasts. Figure plots hourly forecasts for a month of December 2020, predicted by DistrNN.

We provide the hyper-optimisation results that precede the following out-of-sample evaluation in the Appendix 5.A in Figures 5.2 and 5.3. The figures

[6]To perform Diebold-Mariano test we use https://github.com/JuliaStats/HypothesisTests.jl.

show the values of the validation losses of the forward rolling scheme and the corresponding CRPS values of the out-of-sample predictions, and other figures show the values of the hyperparameters selected for the forward rolling scheme specifically for each hour.

5.5.1 Ouf-of-sample evaluation

Table 5.2 gives an overview of the results for all models considered, both in point and probabilistic angles. In terms of CRPS loss, the lowest loss value of DistrNN, 1.3669, is about 17% less than the second lowest LEAR-QRA, 1.6497. The accuracy of the LEAR-based quantile regression models, QRA and QRM, is similar. We also observe evidence of better accuracy of DistrNN than LEAR-QRA(QRM) models in the columns of unconditional coverage of prediction intervals. For all three interval sizes, 50%, 90% and 98%, DistrNN reports occurrence rates closest to the nominal values, i.e. the most reliable coverage of prediction intervals. Both LEAR-QRA and LEAR-QRM show larger distances to the nominal value and the occurrence rates are all lower than the nominal αs, meaning that the quantile regression methods underestimate the size of the prediction intervals. For DistrNN, this is true for 50% and 90%, although the difference for the latter is less than 1%. The 98%-coverage is matches by DistrNN almost ideally.

We further observe that DistrNN has the lowest MAE and RMSE for point forecasts. Even, the distributional models provide medians, $\widehat{q}^{0.50}$, of probability forecasts, the results are better than the LEAR-Avg optimizing for the mean.[7]

	point		*probabilistic*			
	MAE	RMSE	CRPS	50%-cov	90%-cov	98%-cov
Naive	9.2559	14.2027	3.3409	0.3509	0.6965	0.7915
LEAR-Avg	4.4655	6.7939	-	-	-	-
LEAR-QRM	4.3848	6.7547	1.7048	0.4272	0.8318	0.9348
LEAR-QRA	4.3230	6.6908	1.6497	0.4329	0.8432	0.9577
DistrNN	3.7507	6.3119	1.3669	0.4558	0.8800	0.9792

Table 5.2: Quantitative results. Point (median) and probabilistic accuracy results. For MAE, RMSE, and CRPS measures lower $\Longrightarrow$ better, for α-coverage measure closer to nominal % coverage $\Longrightarrow$ better. Colours highlight differences in values from red to green are from the worst to the best.

[7]Diebold-Mariano test results are in Appendix 5.1.

Figure 5.6 breaks down the CRPS results from Table 5.2 into individual quantile losses corresponding to the average of QL_α for the OOS period. We see that the DistrNN quantile loss is the lowest for all probability levels. The difference between the losses is negative for all αs. This supports that the loss of DistrNN is lower than that of LEAR-QRA.

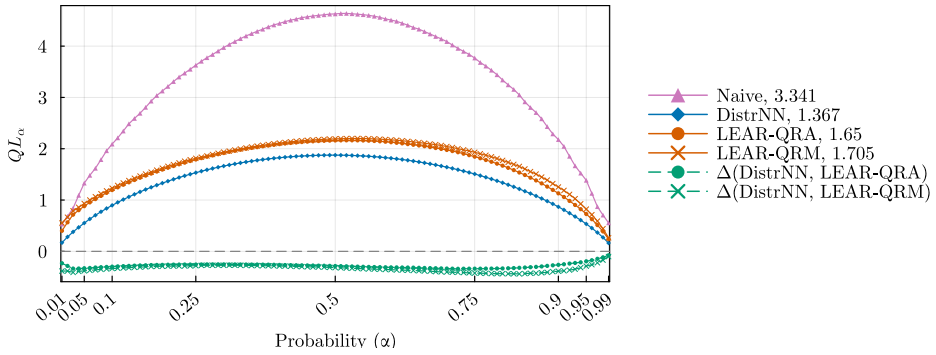

Figure 5.6: Continuous probability score between models for $\alpha = \{0.01, \ldots, 0.99\}$ probability levels. Labels in the legend provide the average CRPS of models' results. Differences between DistrNN and LEAR-QRA(M) are dashed.

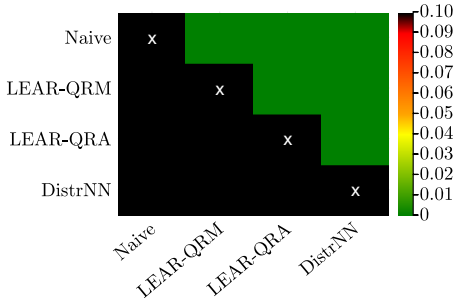

Figure 5.7: P-values of Diebold-Mariano tests with $\mathcal{H}_0$: Loss of model on y-axis is better than loss of model on x-axis. It depicts p-values for each pair of models. In a cell where the colour is other than black, the model on the x-axis is significantly better than one on the y-axis. When cell is black the p-value is greater than 10%.

In Figure 5.7 we show the DM test results for the multivariate loss between models. In other words, the overall accuracy between pairs of all models is tested as the sum of the absolute CRPS loss over 24 hours, L_m, within for the OOS period. The DM test suggests that we reject the null hypothesis that LEAR-QRA is statically better than DistrNN and accept the alternative that DistrNN has significantly better accuracy of probabilistic prediction. In

parallel, we do not reject the null that DistrNN has better accuracy than LEAR-QRA. Furthermore, as expected, we see that both LEAR-QRA and DistrNN have statistically better accuracy than the naive predictions.

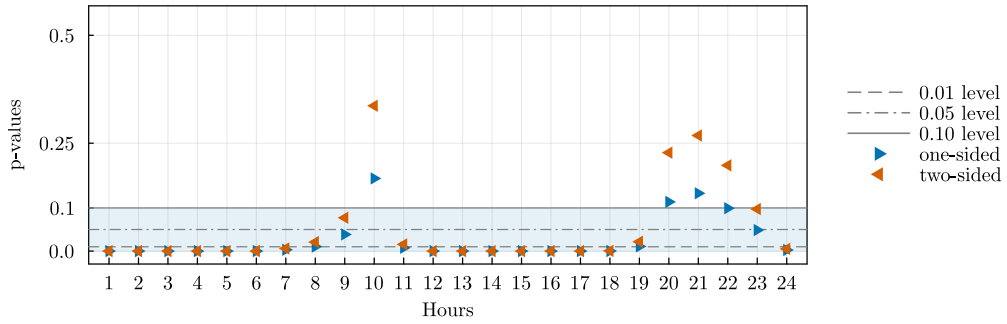

Figure 5.8: P-values of two Diebold-Mariano tests disaggregated by hours. One-sided test $\mathcal{H}_0$: Loss of LEAR-QRA is better than loss of DistrNN. Two-sided test $\mathcal{H}_0$: Losses of both models are not different.

Finally, in Figure 5.8 we show the p-values of the Diebold-Mariano test for each hour separately. Above, we have shown that DistrNN provides better accuracy than other models when considering total daily losses. This next result provides an insight into the importance of performance disaggregated by hour. Figure 5.8 shows that DistrNN has significantly better accuracies than LEAR-QRA for most hours (21 hours) at the 10% probability level, with the three insignificant hours not far from significance. Furthermore, for 16 hours, we reject both one- and two-sided null hypotheses at the 1% significance level.

Note that our CRPS values differ from Marcjasz et al. (2023). To begin with, we do not utilise an ensemble method that combines OOS runs across various hyper-optimization customisations. In contrast, Marcjasz et al. (2023) conducted 4 such runs for both Normal and Johnson's SU distributions. We obtain single (hyper-optimised) models for the above setup, which exhibit a lower CRPS compared to any of the 8 single hyperparameter set-based models in that study. Since their study confirms that an ensemble of models results in performance gains, we assume that this is also applicable to DistrNN.

The authors use a larger hyperparameter set of 2048 compared to 60 in our setting. The number of searches is also increased by the fact that Marcjasz et al. (2023) manually turn on and off input features, and they also use parameter regularisation in training as dropout and L_2-norm. In this sense,

we use dropout and AdamW learning algorithms to regularise, where the latter mimics L_2 norm regularisation. In addition, for computational reasons, we limit the number of neurons in the network layers to be between 32 and 384, as opposed to the maximum size of 1024. Comparing the computational cost of the approaches is not straightforward, although the times of both are similar, DistrNN is narrower in this paper and has fewer values in the output layer, it needs to run separately for all 24 hours of a day. Compared to Marcjasz et al. (2023) whose output layer is for one day and all 24 hours using a wider neural network.

Berrisch and Ziel (2023) further improve the results of Marcjasz et al. (2023) with their technique of CRPS learning on already provided OOS results of different models. They provide how to average such results to get a more accurate ensemble average. We do not compete with these results, as the technique can be applied equally well to our DistrNN results.

We do not restrict the reader to taking these results as definitive or to using our approach only as a feed-forward neural network. There may be potential accuracy and performance benefits if the distributional network is recurrent, convolutional, temporal-attentional, and many others. This also opens up space for further analysis, taking into account parsimony and computational cost. Figure 5.3 in the appendix shows a comparison of the most chosen number of hidden nodes and the most preferred activation function.

5.5.2 Software and computational time

In recent years, the use of software, particularly in econometrics, has developed rapidly and enormously. We provide a Julia (Bezanson et al., 2012) code that replicates our results and also serves as an example of how to use environments other than languages, such as python or R. The exercise uses the Flux.jl package (Innes et al., 2018) and the results can be replicated using examples at https://github.com/luboshanus/DistrNNEnergy.jl, which may make the process easier for those using Julia to predict (energy) time-series.[8]

The complete estimation process of DistrNN, involving the hyper-optimisation search and a forward rolling window scheme across 24 hours, 736 OOS observations, 60 hyperparameter sets, 7 folds, 8 ensembles, using 1000 epochs and 64 mini-batch size, entails the estimation of 10080 (24*60*7) and 141312 (24*736*8) networks. Therefore, obtaining an out-of-sample prediction can

[8]The code uses several Julia packages provided in Project.toml file.

take approximately 24 to 48 hours, depending on the number of ensembles (2-8). We distribute the hyper-optimisation and rolling estimation tasks over 60 CPU cores.[9] The complete estimation of LEAR-QRA(QRM) in Julia takes about 15 minutes when distributed over 15 CPU cores.[10]

5.6 Conclusion

This paper proposes a novel machine learning approach to probabilistic forecasting of hourly day-ahead electricity prices. Compared to the state-of-the-art frameworks in the (probabilistic) electricity price forecasting literature, our model provides more accurate forecasts. This is mainly due to the fact that it does not rely on restrictive model assumptions and allows for non-Gaussian, heavy-tailed data and their non-linear interactions. By relaxing the assumption on the distribution family of the time-series, our distributional neural network explores the data fully. We also provide an efficient computational package that can be used by researchers.

References

Anatolyev, S., & Baruník, J. (2019). Forecasting dynamic return distributions based on ordered binary choice. *International Journal of Forecasting*, *35*(3), 823–835.

Berrisch, J., & Ziel, F. (2023). Multivariate probabilistic CRPS learning with an application to day-ahead electricity prices. *arXiv preprint arXiv:2303.10019*.

Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*.

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1046–1089.

Box, G. E., Draper, N. R., et al. (1987). *Empirical model-building and response surfaces* (Vol. 424). Wiley New York.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.

[9] We used 60 CPU cores of the AMD Ryzen Threadripper 3990X 64-core processor.

[10] We rewrote parts of the authors' open access toolboxes in Julia (Lago et al., 2021a; Marcjasz et al., 2023).

Bunn, D., Andresen, A., Chen, D., & Westgaard, S. (2016). Analysis and forecasting of electricty price risks with quantile factor models. *The Energy Journal*, *37*(1).

Clements, M. P., Galvão, A. B., & Kim, J. H. (2008). Quantile forecasts of daily exchange rate returns from forecasts of realized volatility. *Journal of Empirical Finance*, *15*(4), 729–750.

Diebold, F. X. (2021). What's the big idea? "big data" and its origins. *Significance*, *18*(1), 36–37.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3).

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.

Feng, G., He, J., & Polson, N. G. (2018). Deep learning for predicting asset returns. *arXiv preprint arXiv:1804.09314*.

Foresi, S., & Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, *90*(430), 451–466.

Fritsch, F., & Carlson, R. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, *17*(2).

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5), 920–964.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*, *33*(1), 3–12.

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: The state space approach*. Springer Science & Business Media.

Chernozhukov, V., Fernández-Val, I., & Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, *81*(6), 2205–2268.

Innes, M., Saba, E., Fischer, K., Gandhi, D., Rudilosso, M. C., Joy, N. M., Karmali, T., Pal, A., & Shah, V. (2018). Fashionable modelling with flux. *CoRR*, *abs/1811.01457*.

Israel, R., Kelly, B. T., & Moskowitz, T. J. (2020). Can machines "learn" finance? *Journal of Investment Management*, *18*(2), 23–36.

Iworiso, J., & Vrontos, S. (2020). On the directional predictability of equity premium using machine learning techniques. *Journal of Forecasting*, *39*(3), 449–469.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, N., Smith, M. S., & Nott, D. J. (2023). Deep distributional time series models and the probabilistic forecasting of intraday electricity prices. *Journal of Applied Econometrics*.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 33–50.

Kuan, C.-M., & White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews*, *13*(1), 1–91.

Lago, J., Marcjasz, G., De Schutter, B., & Weron, R. (2021a). EPFTOOLBOX: The first open-access PYTHON library for driving research in electricity price forecasting (EPF).

Lago, J., Marcjasz, G., De Schutter, B., & Weron, R. (2021b). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, *293*, 116983.

Lehna, M., Scheller, F., & Herwartz, H. (2022). Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account. *Energy Economics*, *106*, 105742.

Liu, B., Nowotarski, J., Hong, T., & Weron, R. (2017). Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid*, *8*(2), 730–737.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Maciejowska, K. (2020). Assessing the impact of renewable energy sources on the electricity price level and variability–a quantile regression approach. *Energy Economics*, *85*, 104532.

Maciejowska, K., Nowotarski, J., & Weron, R. (2016). Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. *International Journal of Forecasting*, *32*(3), 957–965.

Marcjasz, G., Narajewski, M., Weron, R., & Ziel, F. (2023). Distributional neural networks for electricity price forecasting. *Energy Economics*, *125*, 106843.

Marcjasz, G., Uniejewski, B., & Weron, R. (2020). Probabilistic electricity price forecasting with narx networks: Combine point or probabilistic forecasts? *International Journal of Forecasting*, *36*(2), 466–479.

Mashlakov, A., Kuronen, T., Lensu, L., Kaarna, A., & Honkapuro, S. (2021). Assessing the performance of deep learning models for multivariate probabilistic energy forecasting. *Applied Energy*, *285*, 116405.

Mpfumali, P., Sigauke, C., Bere, A., & Mulaudzi, S. (2019). Day ahead hourly global horizontal irradiance forecasting—application to south african data. *Energies*, *12*(18).

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Narajewski, M., & Ziel, F. (2020). Econometric modelling and forecasting of intraday electricity prices. *Journal of Commodity Markets*, *19*, 100107.

Nowotarski, J., & Weron, R. (2015). Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, *30*(3), 791–803.

Nowotarski, J., & Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, *81*, 1548–1568.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., et al. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*(3), 705–871.

Sadhwani, A., Giesecke, K., & Sirignano, J. (2020). Deep Learning for Mortgage Risk*. *Journal of Financial Econometrics*, *19*(2), 313–368.

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181–1191.

Serafin, T., Uniejewski, B., & Weron, R. (2019). Averaging predictive distributions across calibration windows for day-ahead electricity price forecasting. *Energies*, *12*(13), 2561.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Tobek, O., & Hronec, M. (2020). Does it pay to follow anomalies research? machine learning approach with international evidence. *Journal of Financial Markets*, 100588.

Uniejewski, B., Marcjasz, G., & Weron, R. (2019). On the importance of the long-term seasonal component in day-ahead electricity price forecasting: Part ii — probabilistic forecasting [Energy Markets Dynamics in a Changing Environment]. *Energy Economics*, *79*, 171–182.

Uniejewski, B., Nowotarski, J., & Weron, R. (2016). Automated variable selection and shrinkage for day-ahead electricity price forecasting. *Energies*, *9*(8).

Uniejewski, B., Weron, R., & Ziel, F. (2018). Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, *33*(2), 2219–2229.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, *30*(4), 1030–1081.

Zhang, F., Fleyeh, H., & Bales, C. (2022). A hybrid model based on bidirectional long short-term memory neural network and catboost for short-term electricity spot price forecasting. *Journal of the Operational Research Society*, *73*(2), 301–325.

Zhang, W., Quan, H., & Srinivasan, D. (2018). Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination. *Energy*, *160*, 810–819.

Ziel, F., & Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, *70*, 396–420.

5.A Appendix

5.A.1 Point forecasts

If taken into accounts means and medians from the models and distribution, we provide results of DM test of mean absolute errors between models. The average of point forecasts of different calibration windows performs best in this case, then it is the DistrNN median, see Figure 5.1.

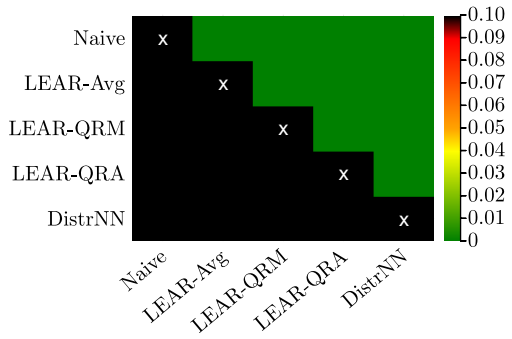

Figure 5.1: P-values of Diebold-Mariano tests for MAE of $\widehat{q}^{0.50}$ or $\bar{y}_t$ in L_m with $\mathcal{H}_0$: Loss of model on y-axis is better than loss of model on x-axis. Where the colour is other than black, the model on the x-axis is significantly better than one on the y-axis.

5.A.2 Hyper-optimization results

Here we provide figures documenting training and validation process, as results of all validation losses related to forward rolling training scheme, Figure 5.2, , and the selection of best parameters sets, see Figure 5.3.

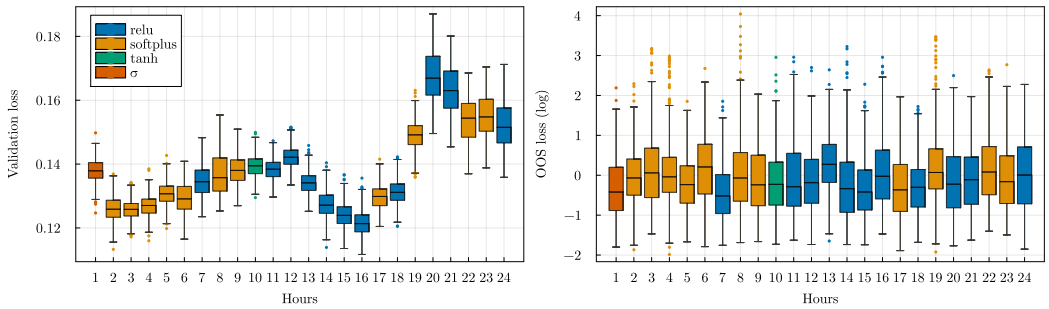

Figure 5.2: Rolling scheme results of validation losses and out-of-sample losses for 24 hours. Figure (left) shows evolution of validation losses (binary cross-entropy) in comparison to out-of-sample loss (CRPS) values for each hour. The colours show which activation function has been selected by the hyper-optimization for given hour. Every histogram contains 554 (+182 discarded) values *times* 4 of those considered for ensembles.

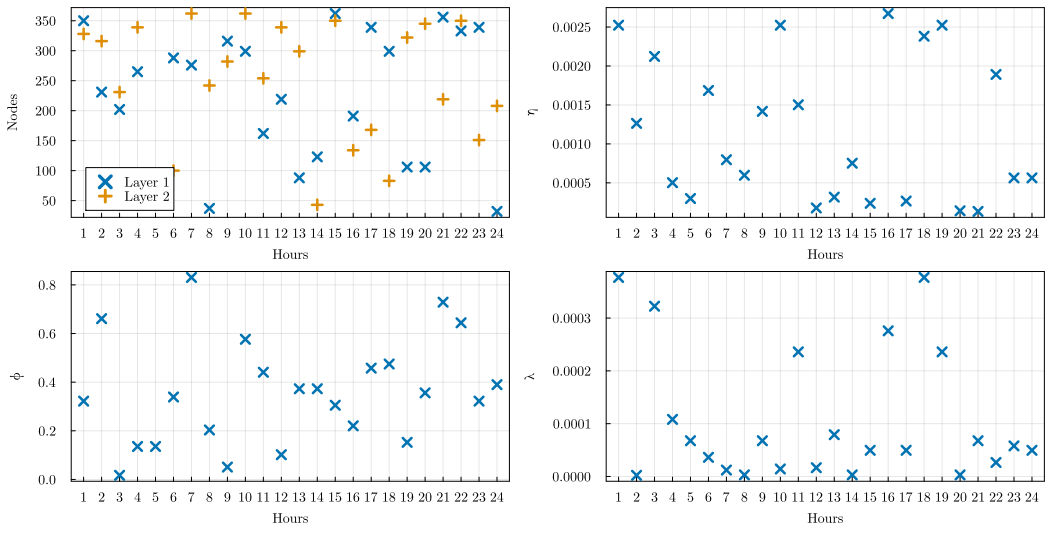

Figure 5.3: Hyper-optimisation results. The figures show the values of the parameters that make up the best sets used to re-learn/re-calibrate the neural networks for each hour. Together with these parameters, the hyper-optimisation selects the activation function, and the selected ones are shown in colour in the figure 5.2.

5.A.3 CDF interpolation

The Fritsch–Carlson monotonic cubic interpolation (Fritsch & Carlson, 1980) provides a monotonically increasing CDF with range $[0, 1]$ when applied to CDF estimates on a finite grid.

Suppose we have CDF $F(y)$ defined at points $(y_k, F(y_k))$ for $k = 1, \ldots, K$, where $F(y_0) = 0$ and $F(y_K) = 1$. We presume that $y_k < y_{k+1}$ and $F(y_k) < F(y_{k+1})$ for all $k = 0, \ldots, K-1$, which is warranted by continuity of returns and construction of the estimated distribution. First, we compute slopes of the secant lines as $\Delta_k = (F(y_{k+1}) - F(y_k)))/(y_{k+1} - y_k)$ for $k = 1, \ldots, K-1$, and then the tangents at every data point as $m_1 = \Delta_1$, $m_k = \frac{1}{2}(\Delta_{k-1} + \Delta_k)$ for $k = 2, \ldots, K-1$, and $m_K = \Delta_{K-1}$. Let $\alpha_k = m_k/\Delta_k$ and $\beta_k = m_{k+1}/\Delta_k$ for $k = 1, \ldots, K-1$. If $\alpha_k^2 + \beta_k^2 > 9$ for some $k = 1, \ldots, K-1$, then we set $m_k = \alpha_k \alpha_k \Delta_k$ and $m_{k+1} = \alpha_k \beta_k \Delta_k$, with $\alpha_k = 3(\alpha_k^2 + \beta_k^2)^{-1/2}$. Finally, the cubic Hermite spline is applied: for any $y \in [y_k, y_{k+1}]$ for some $k = 0, \ldots, K-1$, we evaluate $F(y)$ as

$$F(y) = (2t^3 - 3t^2 + 1)F(y_k) + (t^3 - 2t^2 + t)hy_k + (-2t^3 + 3t^2)F(y_{k+1}) + (t^3 - t^2)hm_{k+1},$$

where $h = y_{k+1} - y_k$ and $t = (y - y_k)/h$.

Chapter 6

Conclusion

The thesis comprises four articles that contribute to non-parametric modelling of economic time-series. The papers could be arranged into two parts or themes, while the first two articles focus on frequency domain methods modelling macroeconomic data, the other two articles deal with forecasting of probabilistic distributions of time-series using neural networks.

The chapter 2 studies dynamics of the economic integration of the Visegrad Four countries with the European Union. We propose the wavelet cohesion measure as a means to assess a multivariate comovement with time-varying weights. The wavelet cohesion measure is a weighted average of pairwise comovement, where the weights are attached to each pair of time-series. From results localised in time and frequency (cycles), we suggest that economic integration may lead to increased comovement of business cycles, which may reflect the benefits of convergence and coordination of economic policies. However, it is vital to have policy coordination and flexibility to ensure that all countries in the region share the advantages of economic integration.

In Chapter 3, we applied a frequency-specific methodology to further examine the dynamics of monetary policy transmission. We utilize frequency response functions to deconstruct the consistency of monetary policy within a framework that allows for time-varying coefficients and covariance structure. We estimated a conventional TVP-VAR model by analyzing data from the United States and its infinite moving average representation, which was necessary to identify the persistence. We observe a significant fluctuation in both output and inflation in response to temporal shocks, consistent with the existing literature. The persistence and transmission of monetary policy has

the greatest positive impacts on output during economic cycles longer than eight years. The results also show that the price puzzle phenomenon may have persistence effects and be propagated at different cycles, which is not observed using only impulse response functions.

In Chapters 4 and 5, we propose a new approach to probability distribution forecasting. Specifically, we introduce a multiple output neural network with a monotonicity adjusting penalty. The method is non-parametric, meaning that it does not assume a particular distributional form for the data but rather selects the best distribution from all possible empirical distributions learned from the data. This approach allows for the capture of complex patterns in data-rich environments. The distributional neural network proves useful for modelling data with non-Gaussian, non-linear and asymmetric structures. By allowing the model to fully explore the data, the proposed approach provides a more accurate and informative probabilistic forecasts.

In Chapter 4, we examine two distinct data sets. The first involves macroeconomic fan charts within the era of big data. Here, we rely upon 216 predictors to determine an approximation of the future distribution of various macroeconomic variables such as GDP growth, inflation, and unemployment. We compared our results to the macroeconomist's state-of-the-art Bayesian VAR model. Our findings show that, at various probability levels, our proposed method generates better predictions. In the second example, we focused on asset price distributions. We compared our approach to the ordered logistic regression model, which served as both a starting point and an inspiration, and acted as a benchmark. The benchmark model is a parametric counterpart to our neural networks and was specifically designed for forecasting asset prices. We report that a data-driven approach can yield benefits in predicting the distributions of heavy-tailed and asymmetric stock returns.

In Chapter 5, we present a novel machine learning approach to probabilistic forecasting of electricity prices. Our proposed method is applied to forecast hourly day-ahead electricity prices probabilistically. Our approach delivers data-intensive predictions without the restrictions of distributional or other model assumptions, thereby enabling the exploration of non-Gaussian, heavy-tailed and asymmetric data. We build a distributional network to forecast German hourly day-ahead electricity prices using the 221 characteristics. To compare the effectiveness of our framework, we employed the naive model and the two quantile regression-based linear models with

autoregressive and exogenous variables. After comparing our outcomes to other models considered state-of-the-art, our results were better than the outer models, and the difference was statistically significant. Additionally, we have included an effective computational package used to implement our proposed method.

Appendix A

Response to opponents

for the Dissertation Defense
January 17, 2024

on manuscript
Essays on Data-driven, Non-parametric Modelling of Time-Series

by Luboš Hanus

Dear Prof. Witzany, Dr. Ellington, Dr. Trimborn,

at first, let me greatly thank you for your opponent reports to my pre-defense version of this dissertation. I do appreciate your thorough reading of the thesis and providing many detailed comments. I have taken into account your suggestions and remarks, which have helped me improve the quality of the thesis, as well as the future work. In the following section, please find my explanation and discussion to your comments and suggestions.

Note: In the following part, the original Opponent's Reports and comments are in *italics*. My response and comments follow each comment or a set of comments.

A.1 Response to comments of Prof. Jiří Witzany, Ph.D.

A.1.1 Opponent's report

The thesis comprises of four papers which can be split into two topics. The first two papers apply frequency domain analysis methods to model macroeconomic data, namely synchronization of business cycles of EU countries, and effects of monetary shocks on selected US macroeconomic variables. The second two papers focus on machine learning applications for distributional predictions of macroeconomic and financial variables. All the papers are well motivated and deal with relevant policy-oriented or financial topics. The applied techniques are methodologically and computational on very high level. The questions stated above can be answered positively, that is:

a) Yes, there are significant original contributions of the author.

b) The thesis is based on an extensive set of relevant references that are used throughout the text.

c) The thesis would be definitely defendable at my home institution (FFA VSE).

d) The technical complexity of the papers might be related to the fact that only the first paper is reported to be published in the Empirical Economics journal (JCR Q3 in F5.2). The second paper has been published in the IES WP series (not submitted to a respected journal yet?), and the remaining two papers have been submitted to non- disclosed journals. All the three papers, in my opinion, allow their publication in a respected economic journal.

e) The presentation on methodologies and results might need some improvements in terms of clarification and readability. Otherwise, I do not have any major comments or recommendations for revisions.

f) The conclusion of my pre-defense review is to ***recommend the thesis for defense without substantial changes.***

More detailed comments to the four papers are as follows:

A.1.2 Comments to the first paper on *"Growth cycle synchronization of the Visegrad Four and the European Union"*

The first paper analyses the growth cycle synchronization of the four Visegrad countries with respect to synchronization of the other EU countries. The co-movement between two specific countries is decomposed into time and frequency dimensions using the wavelet transform methodology. The multivariate relationship is characterized by a weighted average of the wavelet cohesion measures of the individual pairs. The pair-based and multivariate cohesion analysis provides interesting results that might have important policy implications discussed in the text. I have the following minor comments:

Comment 1: *The multivariate cohesion measure defined in Eq. 2.4 uses weights $\omega_{ij}(\tau)$ attached to pairs of countries (i, j). However, the pair-dependent weights are not explicitly specified, it is only stated that the nominal GDP is used for time-varying weights. Therefore, the reader must speculate that $\omega_{ij}(\tau) = w_i(\tau)w_j(\tau)$ where $w_i(\tau)$ are the GDP based country specific weights? This should be clarified in the text, in particular since the multivariate measure is presented as the main methodological novelty.*

Response: Thank you for pointing this out. It is true the weights were not properly specified. I have clarified the notion in the final text.

Comment 2: *Why is the nominal rather than real GDP used? There might be an undesirable effect of exchange rate fluctuations on the nominal GDP of the Visegrad countries.*

Response: Thank you for your comment. It is a valid point that real GDP may be a better proxy for county size. However, since the weights are normalized and the measure is between -1 and 1, the difference between nominal and real GDP as a weight is likely to have a small impact on the final cohesion values.

Comment 3: *I assume that Figure 2.4 shows the multivariate cohesion of the set of all V4 and EU countries. Did you consider a multivariate measure where the pairs would be selected so that one country is in the first set (V4) and the other country in the second set (EU core)?*

Response: Thank you for your contribution. We have not yet explored this approach. This approach would provide a weighted measure of cohesion,

stripped of the internal pairwise synchronisations that occur within each group, namely the V4 and the EU core. One question may be whether or not we want to take into account the internal pairwise synchronisations of these groups, or whether we want to capture the possible full or future cohesion of the Union. The results we present opt for the latter. The proposed approach could provide additional information about the relationship between the groups. Furthermore, we provide coherence results of individual countries with a proxy for the EU, which contains similar values.

Comment 4: This comment was related to Chapter 4 (Paper 3), see responses to this chapter – Comment 1.

A.1.3 Comments to the second paper on *"Identification Persistence in Macroeconomic Responses"*

Comment 1: *The second paper aims to analyze the effects a monetary policy shock on the US GDP, inflation, and interest rate. The employed methodology includes the TVP-VAR model, time-frequency decomposition, and the impulse response analysis. The main results are presented in figures 3.3-3.8. My rather formal recommendation is to reconsider the format of the figures. Figure 3.3 presenting 3D time-varying impulse responses of the three variables over different impulse horizons is hardly legible in printed form. Figures 3.4-3.6 show responses and their confidence intervals of the three variables over the short, medium, and long-term horizons. My recommendation is to unify the scale and indicate the zero level by a horizontal line to make the visual comparison of the effects over different horizons easier.*

Response: Thank you for your suggestion. I agree that the coloured Figure 3.3 may be difficult to read when printed. I have tried to keep the shape of the figure according to the literature standards. Also, for the time-varying frequency response functions, I have not made the scale uniform because some time-variations might become unobservable. I appreciate the idea of having the zero level at the same position for the frequency responses.

A.1.4 Comments to the third paper on *"Taming data-driven probability distributions"*

The third paper proposes a machine learning approach to probabilistic (distributional) forecasting of macroeconomic and financial-time series. The key idea is to train a neural network with target being a vector of probabilities estimating the cumulative distribution function on a grid of values. This can be viewed as an analogy to the classical ordinal regression model. The input would characterize the information at time t (including the history in the LSTM framework), and the output should estimate the distribution of a variable at time $t + h$. The empirical results based on a dataset containing 216 quarterly US macroeconomic and financial variables demonstrate superiority of the ML approach compared to selected benchmark approaches such as BVAR. I have the following minor comments:

Comment 1: *The introduction to machine learning (4.2.1) includes, in my opinion, a few over-optimistic statements such as "...machine learning seeks to choose the most preferable model from an unknown pool of models using innovative optimization techniques. As opposed to traditional measures of fit, machine learning focuses on the out-of-sample forecasting performance and understanding the bias-variance trade-off; as well as using data driven techniques that concentrate on finding structures in large datasets. Further, if one dismisses the "black-box" view of machine learning as a misconception ..." I think that all of those statements could be opposed. For example, what is the most preferable model selected from a pool of models? Can we really dismiss the "black-box" view?*

Response: Thank you for your comment. I agree that these statements can be opposed and debated, but they are used to motivate the approach because there is a wide literature showing them as benefits (Israel et al., 2020). The most preferable model is selected based on statistics for a given problem. In my view, we can dismiss the "black-box" view since the neural network is just a more complex non-linear regression, and nowadays we do not perceive those as "black-boxes". Furthermore, efforts are being made to improve the interpretability and transparency of potential "black-box" models. The balance between model complexity and interpretability depends on a specific use case. While some applications may prioritize accuracy over interpretability, others require a balance between the two.

Regarding the pools of models and in terms of the optimization algorithm, there appear to be two pools. The initialisation of the neural network is

random in the first pool, and the learning process uses a criterial function to obtain the best model. The second pool is used for hyper-optimisation, where the same criterial function helps to find a suitable set of parameters for a model at a lower level. Finally, we perform model averaging of results from multiple models to reduce the variance of the final prediction.

Comment 2: *The loss function contains the monotonicity penalty term 4.7 including a meta-parameter λ_m. In spite of that, the trained NN outputs might have a number of monotonicity violations. A larger λ_m value will reduce the number of violations, but probably worsen the other part of the loss function performance. The choice of the meta-parameter should be better explained in the text.*

Response: It was not clear in the text, so I have added a footnote to the definition of the loss function. To reiterate, the choice of λ_m is specific to the problem and data we are working with.

Comment 3: *As mentioned above, there will be some monotonicity exceptions. How are those dealt with in the construction of the distribution function?*

Response: Thank you for pointing this out. Although it is rare, this situation can occur. To satisfy the monotonicity requirement, the probabilities are sorted before being fed into the Fritsch-Carlson monotonic cubic interpolation algorithm (Fritsch & Carlson, 1980).

Comment 4: *It is emphasized that the importance lays in finding of empirical quantiles q^α corresponding to a set of probabilities. Are the empirical quantiles based on the in-sample period only, or out-of-sample, or on the full sample? Some more detail on the explicit empirical quantiles' setup would be appropriate.*

Response: As the normalisation/standardisation is done only on the train part, the analysis does not contaminate the data with future information. It follows a common practice and does pure out-of-sample without looking into the future. This also applies to empirical quantiles. The quantiles can be found as constants, q^α, or one can use the Exponential Weighted Moving Average (EWMA) to obtain time-varying quantiles for the desired alphas, q_t^α.

A.1.5 Comments to the fourth paper on *"Learning probability distributions of day-ahead electricity prices"*

Comment 1: *The last paper applies basically that same ML methodology as the previous paper, but in this case applied to electricity prices with the goal to forecast distributions of day ahead hourly electricity prices. The empirical results are in this case compared to a naïve estimation, quantile regression averaging, or quantile regression committee machine models again demonstrating superiority of the ML approach. On the other hand, the computational time analysis indicates a high computational cost of the ML approach.*

Response: Thank you for bringing this to the attention. The computational cost is an acknowledgement of a potential limitation of the ML approach. However, it is important to note that the linear programming algorithm of penalised quantile regression also has its own convergence problems, which can result in lengthy computations.

A.2 Response to comments of Michael Ellington, Ph.D.

A.2.1 Opponent's report

Synopsis of the thesis

This thesis contains four stand alone empirical chapters that contribute to the literature modelling time-series. The theme connecting these chapters is a non-parametric approach to modelling economic and financial data. The first two chapters focus on frequency domain methods to analyse macroeconomic data. Meanwhile, the focal point of the latter two chapters utilizes contemporary Machine Learning (M-L) methods to generate predictive probability distributions for forecasting purposes.

The thesis is well presented and is structured in a coherent manner. Tables and Figures are all self-contained and are readable. In general the thesis contains high quality writing and properly motivates each paper. While there is room for improvement, it is not essential to successfully defend the thesis.

Overall, I can confirm that there are original contributions from the author that are relevant to the literature from each of the four chapters. The candidate utilizes contemporary methods and justifies their approach to each paper within the thesis using pertinent references. I am extremely confident that this thesis would pass

a defence at my home institution, and indeed any other respected international institution. The results, some of which are already published in journals with impact factor, are publishable in select general interest and/or many top field journals.

The comments I make below are minor editorial comments, which should be changed in advance of the defence. For each Chapter, I also provide comments and advice on how to improve the quality of work; particularly when preparing for publication. These are likely comments one would receive from a referee report at a reputable journal. Therefore, I recommend the thesis for defence without substantial changes.

A.2.2 Minor editorial comments

Comment 1: *Please make wherever "time series" is written to "time-series"so as to be consistent with the title. Do the same for "co-movement" change to "comovement" to be consistent with your references.*

Response: Thank you for pointing this out. I have corrected both terms to ensure consistency throughout the text.

Comment 2: *Please update your reference list. Many papers are published now, yet the citations are working papers; e.g. Ellington (2018).*

Response: Thank you for bringing this to my attention. I have reviewed the full list of references and updated several citations, such as those for Ellington (2018), Goulet Coulombe et al. (2022), and Sadhwani et al. (2020), and other.

Comment 3: *Reference list again, capital letters where they are meant to be: e.g. page 9, Hanus and Vácha (2020) in the thesis' reference list "European Union", and "Visegrad Four"read european union and visegrad four respectively.*

Response: Thank you for the correction. I have checked the names and capital letters and corrected them accordingly.

Comment 4: *Check references from Econometrica, they should all be "Econometrica"or "Econometrica: Journal of the Econometric Society", not a mix of both.*

Response: I edited and unified the names of the journals based on information from their official websites.

Comment 5: *Check references from Review of Financial Studies, should be either "Review of Financial Studies", or "The Review of Financial Studies", not a mixture.*

Response: As for the comment above, I edited names of the journals based on information from their official websites.

Comment 6: *Make all Journal titles begin with capital letters. In some cases, there are inconsistencies. For example, "Review of Economics and Statistics", sometimes reads "Review of economics and statistics"*

Response: Thank you for conducting a thorough check. These have also been corrected.

Comment 7: *The FRED-QD database stems from the publication M. W. McCracken and Ng (2016). Please include this in your reference list as oppose to the current technical report. Alternatively please use the correct citation of the NBER working paper M. McCracken and Ng (2020).*

Response: The citation of the FRED-QD database have been updated to the NBER working paper, and I have added the citation of M. W. McCracken and Ng (2016) in the final text.

A.2.3 Comments to the first paper on *"Growth cycle synchronization of the Visegrad Four and the European Union"*

Comment 1: *This paper is already published in Empirical Economics. I have no comments or suggestions for this chapter since it has already been through the peer review process. I enjoyed reading the paper and like the execution of the empirical analysis. On the whole, I like that the paper uses coherence and cohesion to assess growth cycle synchronization of the Visegrad Four. I think that the findings that cohesion is strong after 2005 is unsurprising due to the turbulence in the 90s. It is also reassuring, for policymakers, that cohesion is strong at business cycle frequencies. The paper suggests that with higher cohesion, the more efficient policies may be.*

Response: Thank you for your comment. The text has undergone minor grammatical changes and small revisions in response to other opponents' reports.

A.2.4 Comments to the second paper on *"Identification Persistence in Macroeconomic Responses"*

This paper proposes a frequency domain approach to analysing impulse response functions of variables within a vector autoregression (VAR). The focus is on the monetary policy transmission mechanism and the paper uses popular time-varying parameter VAR (TVP VAR) models (see e.g. Primiceri, 2005). The results of the paper reveal that using US data, low frequency cycles of output are prevalent and have positive cycles. However, the paper documents a price puzzle that many others show when analyzing how monetary policy shocks affect inflation.

I have some comments and suggestions for this paper that I outline below. I think the introduction of the paper should be re-written along the lines I specify below prior to the defence. The comments I have on the empirical application and suggestions for exploring the robustness of the findings are what I would suspect a referee would ask for and are not necessary to address in order to successfully defend the thesis.

Comment 1: *What is the non-parametric aspect of this study? I would be careful on selling this as a non-parametric approach to time-series modelling. The impulse transfer function is manipulations of the VAR's coefficient matrices (in this case time-varying). I note that you state the identification of persistence is non-parametric in the introduction to the thesis itself. I think you may need to specify precisely what is non-parametric about this chapter so as not to confuse the reader, or state that this chapter is concerned with data-driven modelling. I am not convinced the identification is non-parametric because it depends on a parametric model. It is fine to have in the thesis because you are modelling economic data in Bayesian manner which is data driven.*

Response: Thank you for bringing this to my attention. I agree that this was confusing in the original text. I also agree that the identification of persistence is not non-parametric and that it is data-driven. To discuss, the localised window in Bayesian estimation can be considered a non-parametric aspect, as well as the estimation of spectral densities. Nonetheless, I consent with the comment and I changed the term "non-parametric" to "data-driven" in the introduction.

Comment 2: *The introduction should be re-written. Currently we do not learn the contribution of the paper until the third page of the introduction. I think the empirical contribution of the paper is the analysis of the transmission mechanism of shocks*

across economic-cycles. This is new in a sense that I fail to see any economics paper that looks at the transmission mechanism from a frequency domain perspective; as the paper correctly cites, Dew-Becker and Giglio (2016) analyse frequency responses for asset pricing – but this is a peripheral contribution relative to the literature this paper contributes to. The paper could begin at a high level regarding how economists analyse responses of variables to shocks.

The next paragraph could narrow it down to monetary policy, citing the pertinent literature that they do. The problem is that traditional impulse response analysis fails to quantify the transmission mechanism across horizons of interest. This is particularly important for monetary policy and inflation targeting nations like the US who use a short-term interest rate to control inflation over the medium-term (i.e., two-years).

Following from this, is your contribution in paragraph 3. You can then spell out your main results and how you connect to the literature that does use frequency domain techniques in empirical macroeconomic models. I think an introduction written in this manner refines the contribution of your paper and also enhances readability.

Response: Thank you for your comments and for your attention to the matter. I agree that the introduction required reordering. I have attempted to rewrite it to enhance its readability and ensure that the reader can easily understand our contribution.

Comment 3: *In Figure 3.1, it would be great to see a couple of shocks with persistence in the time and frequency domain, perhaps as an additional two subplots below the current ones. This would further justify the importance of looking at shocks in the frequency domain.*

Response: I agree with the comment that only transitory shocks and responses were depicted in Figure 3.1. Therefore, I have added two subplots with persistent shocks to provide complete picture.

Comment 4: *A journal would expect the data to be updated to the most recent quarter. Not essential for the PhD thesis, but for publication in a journal I imagine they will ask for this.*

Response: Thank you for your comment. The dataset will be updated accordingly, and I will also incorporate the shadow rates suggested in the comment below.

Comment 5: *Why do you only identify the monetary policy shock? It is simple*

to identify the demand and supply shocks using sign restrictions (for signs see e.g. Ellington, 2022). On this note, I would also like to see the results from a Cholesky decomposition, you might also want to consider sign restrictions with a maximum FEVD on the monetary policy shock (see e.g. Uhlig, 2005; Volpicella, 2022). This would enable you to show the benefits of your contribution for different identification schemes and increase the quality of your empirical analysis. I would suggest exploring other identification schemes for the robustness of your approach when preparing the paper for publication.

Response: Thank you for this valid comment. The paper's original goal is to focus on the "price puzzle" phenomenon and its potential explanation through time and persistent structures. As the text evolved, it becomes clearer that the idea of expanding the persistence concept to demand and supply shocks is worth exploring. As stated, different identification schemes may benefit the paper during the revision process for publication. I will also elaborate on the restrictions via a maximum FEVD that have not been considered. These additional details will improve the quality of the paper.

Comment 6: *Please re-word sentence 1, paragraph 3, page 48. You impose sign restrictions every quarter throughout your estimation sample. It currently can be confused with imposing sign restrictions for every horizon you compute the impulse response for, whereas you only require the sign restrictions to be satisfied on impact.*

Response: Thank you for your comment. I have re-worded the sentence and provided additional explanation to clarify the sign restrictions.

Comment 7: *Check your measure of GDP deflator inflation. The plot in Figure 3.2 plots the change in the index from a year before (I download the index GDPDEF from FRED and replicate your figure). It does not plot the annual rate of GDP deflator inflation. It should be the latter in your models and not the plot in Figure 3.2.*

Response: Thank you for pointing this out. Unfortunately, it was an error when preparing the picture for the thesis. The figure now depicts percentage changes from year ago.

Comment 8: *Section 3.4.1, paragraph 2 final sentence, please change "3.5, 3.4"to "3.4 and 3.5".*

Response: The change has been made.

Comment 9: *Please move paragraph 3, Section 3.4.1 to after the discussion of Figures 3.3–3.5. You can bring the discussion of the frequency response of the*

interest rate in after this for completeness. I would expect the result for the long-run. Intuitively rising the Federal Funds rate to control inflation should be met with declines as the monetary transmission mechanism begins to work.

Response: Moving the paragraph improves the comprehensiveness of the comments on the results.

Comment 10: *Please specify how you compute the frequency responses, I presume they stem from a sum of the ωs within the defined frequency bands; please specify explicitly in the thesis.*

Response: The frequency responses in Figures 3.4, 3.5 and 3.6 are defined as average frequency response over given frequency band. I have extended the paragraph on page 44, where frequency response is defined. Further, the exact frequency bands corresponding to economic cycles are provided in the footnote 9 on page 49.

Comment 11: *Large posterior credible intervals when looking at correlations and impulse response/transfer functions at the end of the sample stem from the Federal Funds rate approaching its zero lower bound. I urge you to consider shadow rates to account for this (Wu & Xia, 2016). This should reduce the estimation uncertainty around your metrics throughout these periods. They are shown to capture unconventional monetary policy properly in New Keynesian Macroeconomic Models whilst retaining the key features of demand and supply shocks in Wu and Zhang (2019); empirical confirmation is in Ellington (2022).*

Response: Thank you for your comment. I agree that the shadow rates should be considered, not only due to the large confidence intervals. I will replace the reported Federal Funds rate with with the shadow rate of (Wu & Xia, 2016), as supported by the literature.

Comment 12: *Why not show the time-frequency-varying correlations of interest rates and GDP growth instead of the correlations between output and inflation. I think the correlations between gdp growth and inflation are not interesting in that they look stable (from posterior median estimates); adding to this.*

Response: Thank you for your suggestion. I agree that the correlation interest rates and GDP growth might be more interesting for monetary policy than the correlation output and inflation.

Comment 13: *It would be good to see the frequency specific forecast error variance*

decompositions that stem from manipulations of the impulse transfer functions to understand the economic importance of monetary policy shocks across these horizons.

Response: Thank you for this suggestion. It is a good idea and in an ideal situation, for data generating processes with different levels of persistence, the study could show a complete analysis of the responses and their decompositions in the time and frequency domain. The FEVDs and their spectral counterparts would also enhance the information on shock propagation. I choose to focus the reader's attention on our contribution in identifying the persistence of monetary transmission.

Comment 14: *A suggestion when preparing for publication: I wonder how the results change if you impose sign restrictions on the impulse response of the monetary policy shock for four quarters as opposed to one? Does this resolve the price puzzle you document?*

Response: This comment raises an interesting point. It is possible that imposing a sign restriction on four quarters could change the results in both time and frequency domains, as the size and direction of impulse responses with the sign restriction imposed only on the impact period vary over time and are not negative for shorter periods/horizons. This could serve as an additional robustness check and is worth further exploration.

Comment 15: *A suggestion when preparing for publication: it would be great to replace GDP growth with the output gap.*

Response: Thank you for the suggestion. I will certainly consider it as it is a good alternative.

Comment 16: *A suggestion for publication: would be to look at the model implied Taylor rules in the frequency domain to see how the reaction of the Federal Reserve to inflation and GDP differs over time and frequencies see e.g. Belongia and Ireland, 2016; Ellington, 2022, for mapping the Taylor rule from the TVP VAR. This would allow for a much more thorough analysis of the monetary transmission mechanism high lighting the benefits of your approach and idea of looking at this in the frequency domain.*

Response: In relation to the Taylor rule and monetary policy, it is interesting to investigate the cyclical effects that arise as central banks aim to react over different time horizons. The use of frequency inputs in estimation could enhance the discussion of dependence across different horizons and cycles.

A.2.5 Comments to the third paper on *"Taming data-driven probability distributions"*

This paper introduces a deep learning approach to generate predictive distributions for macroeconomic and financial time-series. The approach rests on a deep learning recurrent neural network which alleviates the need to specify a single model to generate predictive distributions. The main problem that underpins the idea is that practitioners care about uncertainty shrouding their forecast. I agree with this notion and believe this is an important contribution; particularly to the finance literature where there is a habit of focusing on measures of central tendency (i.e. Root Mean Squared Errors etc.).

The paper offers two empirical applications. The first is a data rich environment using a quarterly version of the FRED-MD database in M. W. McCracken and Ng, 2016. The first application benchmarks against a Bayesian VAR. The second is an application using daily data. The main results use the ordered logit model of Anatolyev and Baruník, 2019. In both cases there is evidence that the approaches within are at least as good as benchmark models.

Overall, I like the paper and enjoyed reading it. It is relatively well written and easy to follow. My specific comments are below. Unless specified otherwise, I feel the below need revision for the defence, those suggestions that are not required for defence are given to help improve the quality of the empirical applications. Noting that the paper is already at revise and resubmit status at a reputable journal, I am not recommending that further empirical analysis is required for defence.

Comment 1: *Please check sentence structure and wording throughout. For example in the Introduction paragraph 1 final two sentences: "We develop a distributional machine learning methods" should be "We develop a distributional machine learning method". Also "Such data-driven probabilistic forecasts are not possible with classical methods without set of restricting assumptions" should be "Such data-driven probabilistic forecasts are not possible with classical methods without a set of restricting assumptions". There are minor editorial issues like this throughout the chapter.*

Response: Thank you for your attention paid to the editorial issues. I have reviewed the chapter and made a significant number of corrections to the final text.

Comment 2: *Please do not start sentences with the word "with" or "because". I*

suggest going through the entire thesis for this as there are other chapters that this occurs.

Response: Thank you for the comment. Thoroughly, I have checked the entire thesis to correct or rephrase sentences in question.

Comment 3: *Please split the first paragraph in the introduction into two paragraphs. This is currently too long. I strongly encourage re-wording the first paragraph to make it substantially shorter. I recommend beginning the second paragraph with the sentence "In this paper...".*

Response: I have split the first paragraph into two in the final text. In addition, I have also shortened the first part with minor adjustments so as not to lose the points the texts make about data availability and the importance of linking uncertainty and big data.

Comment 4: *On this note: "Challenged by **the** proliferation of parameters...". Also, please re-word this sentence as it currently does not read well.*

Response: I have rewritten the sentence.

Comment 5: *Check paragraph 2 in the introduction. The first sentence is too long and requires a comma after "retail banking".*

Response: This sentence has been adjusted to be more readable.

Comment 6: *I would also like to see the final two paragraphs stating the main contribution and results on pages 63–64, sooner. Perhaps after the new second paragraph starting with "In this paper...". You should tell the reader what is this paper is doing and the main results. Then you can connect to the literature as you do.*

Response: The way how information flows in the introduction is important. Thus, with respect of the comment, I have changed the order of paragraphs and adjusted them to fit the new position. The literature links well too.

Comment 7: *Footnote 5 on page 63 is far too long. Consider intertwining it within the main body of the introduction to really speak to the literature you are contributing to. These methodological papers seem to coincide with your work so it would be great to see what is new about your paper relative to these methodological papers. If you are borrowing methods from them, or extending, then say so.*

Response: I have incorporated part of the footnote into the literature review. Our approach contributes to the literature presented. Furthermore, one of

the approaches is used a as benchmark in the publication that is based on this chapter (Baruník & Hanus, 2024). We are not borrowing or extending these methods in any sense, we try to answer the same question "to forecast time-series distribution".

Comment 8: *I like that you spell out clearly the M-L approaches relevant to the paper. A small typo in Section 4.2.3 is that the loss function $\mathcal{L}$ has two equation numbers associated to it since it lies on multiple lines. Please rectify. Also, page 72, does AdamW (weight decay parameter) need a citation?*

Response: Thanks for pointing this out. I have removed one number from the equation numbering. On the page 72, in the first paragraph, there is a citation of AdamW (Loshchilov & Hutter, 2019).

Comment 9: *Section 4.3: I think there is a large literature on uncertainty around forecasts, be it uncertainty around the forecast itself or indeed model uncertainty. One way that the literature addresses this problem is (Bayesian) model averaging or combination forecasts (weighted using some optimization function or equal weighting). The paper is silent on these issues. I think there should be a footnote at least identifying some work that utilizes this (see e.g. Liu & Maheu, 2009; Wright, 2008). Alternatively, Steel (2020) provide a good overview of this literature. Therefore I would expect to see a footnote citing the latter and some brief statement about why you are not speaking to this literature (as it can be utilized to make fan charts).*

Response: Thank you, I have listed and added the references in the literature review. Bayesian model averaging are relevant for addressing a model uncertainty and the resulting forecast uncertainty. BMA can be used as a benchmark. The approach of averaging is also used in the neural network procedure, where the model is initialized and trained multiple times, and the individual forecasts are averaged. This is known as "ensembling" in machine learning jargon. The ensembles can be used in various ways, such as model averaging, when models differ not only in parameter initializations but also in structure. I agree that the previous version was missing this information and it should be included in the final text.

Comment 10: *Is the BVAR a standard BVAR (with what priors?) that utilizes the factors estimated by the Expectations-Maximization Algorithm in M. McCracken and Ng (2020)? Please specify this and the number of factors used (I note you specify in the figures, but not the main text, unless I missed it).*

Response: I have used the R package of Kuschnig and Vashold (2021) which implements the hierarchical priors in the fashion of Giannone et al. (2015). The procedure is automatic. The factors used in the BVAR are four and it was only specified in the figure, thus, I have put the information in the main text as well. I have obtained the four factors using Principal component analysis and those resemble ones from M. McCracken and Ng (2020).

Comment 11: *On this note, why not estimate the factor and the forecasting model in the same step using a Bayesian Factor Augmented VAR model (FAVAR)? Code is readily available from Gary Koop's website for this that enables you to estimate the number of factors within the procedure.*

Response: Thank you for the suggestion. The Factor-Augmented Vector Autoregression (FAVAR) is a natural alternative to the Bayesian Vector Autoregression (BVAR) model. We chose the vector autoregression model due to its simplicity and validity as a state-of-the-art model in macroeconomic literature and practice. Further comparisons are possible. During the review process of Baruník and Hanus (2024), an opponent requested a comparison of our results with a state-of-the-art machine learning model, DeepAR, proposed by Salinas et al. (2020). Hence, we opted for the ML benchmark.

Comment 12: *This comment is not necessary to address for defence: I suspect a referee would like to see how your M-L approach compares against a Bayesian Model Averaging approach to generating predictive distributions; at least in my opinion this approach would be a more consistent experimental design.*

Response: Thank you for your comment. We agree that BMA would be a good benchmark approach, we chose to focus on the requests of the referees, as outlined in previous comments.

Comment 13: *Section 4.4 makes sense to benchmark forecasts against Anatolyev and Baruník (2019) given the frequency you observe the data. I also agree with the variables chosen as inputs. However one question I do have, that is not necessary to address in the thesis for defence is the following: How does the addition of common factors help prediction (at a daily frequency). Adding to this, there is no mechanism that allows for covariation among stocks. Is this problematic in a sense that a conditional predictive of an asset's return may be influenced by covariances among returns or some common systematic exposure to the market?*

Response: Thank you, this is an interesting and valid question. I agree that testing different factors could provide information about their influence

on stocks. We have not included any additional common features for stock forecasting, as we built the model to improve the forecasting of already established and published work of Anatolyev and Baruník (2019) in *International Journal of Forecasting*. Our model is proposed in the exercise of asset returns as a local model, in other words, a univariate forecasting approach. However, and most likely, adding more data or common features can improve the prediction when using a distributional neural network.

Comment 14: *I would expect to see a justification for the choice of the 29 stocks. All stocks listed on the S&P500 are liquid so what informed the choice of these 29?*

Response: The selection of 29 stocks and their asset prices was based on data availability. During the time we were working on the problem, we had access to a smaller sample of high frequency data (minutes intra-days) from which we could obtain daily realised measures. In addition, as mentioned in the previous answer, we wanted to see the improvement over the already published results.

Comment 15: *When referring to Appendices please use capital "A" throughout the thesis there are a couple of times on page 82 where it is not capitalized, but then it is on page 83.*

Response: All references to the Appendix have been corrected.

Comment 16: *Table 4.4: for example 22/29 implies that for 22 of the 29 assets, NN:128 performs better than the benchmark using MSPE. Is this correct? If so, can this please be made more explicit in the main text and in the Table notes.*

Response: The table's description has been edited. And the interpretation of the table has been added in the main text.

Comment 17: *Why not use the Diebold Mariano (DM) test to infer how many asset distribution forecasts are statistically different to the benchmark (in favour of and against the M-L method) as a complement to those results you report in Table 4.4? If they are not statistically different fine, since as you say economically the meaning can be significant for portfolio formation and trading strategies. I know Figure 4.5 goes some way to showing this, but it would be nice to know what DM tests say.*

Response: Thank you for your comment. In this case, we evaluate distributional forecasts using statistical measures commonly employed in the literature of probabilistic forecasting, such as the CRPS or Brier scores.

Comment 18: *On this note, not necessary for the defence: I imagine a referee would like to know what the bottom line for an investor in using these approaches might be to highlight what Campbell and others state about economic significance of marginal improvements in predictability. This would certainly showcase the approach in a more convincing manner.*

Response: Economic and statistical significance are closely related. As shown in Campbell and Thompson (2008) and Rapach et al. (2010), an improvement in statistical performance and predictability leads to increased economic gains. For a simple strategy to demonstrate economic gains, readers are referred to Anatolyev and Baruník (2019). As the DistrNN model is benchmarked against their model, economic performance should also improve, allowing an investor to benefit from a more accurate trading strategy.

Comment 19: *typo page 85 "information" not "informaiton".*

Response: This typo has been corrected.

A.2.6 Comments to the fourth paper on *"Learning probability distributions of day-ahead electricity prices"*

This paper presents distribution forecasts of electricity prices using contemporary M-L methods. This is a non-parametric approach to selecting the best predictive distribution from all possible distributions that the machine learns from the data. A multiple output neural network with a monotonicity adjusting penalty is able to learn complex patterns in electricity prices and outperforms a variety of M-L benchmarks and a naive forecasting model.

I learned a lot from this paper since electricity prices/markets is not my area of expertise. Overall, I like the paper and can see a clear contribution to existing knowledge. Below I list my comments, some comments are suggestions to help the candidate and are specified as not necessary to address for defence of the thesis. I note that this paper is under review at a reputable journal and therefore hope that the below recommended changes below help with a revision of the paper at a journal.

Comment 1: *Introduction: I would like to see the final two paragraphs of the introduction on page 99 on the first page of the introduction. Then, a summary of results. Following this you can place the paragraphs that connect to the literature.*

One minor revision here would be to explicitly state the value added for your paper against those (or a few most relevant to your study) that you cite here.

Response: Thank you for your contribution. The final two paragraphs of the introduction have been included to bring the contribution to the reader sooner and to connect it well with the literature. The last paragraph now states the added value, which is that the above-mentioned studies are restrictive with their parametric approaches.

Comment 2: *Please define acronyms QRA and QRM on page 99, you use one of the acronyms on page 104 before you define it on page 110.*

Response: Thank you for pointing this out. The acronyms were in the introduction (Chapter 1) and not in the Chapter 5, thus I have added them on page 99.

Comment 3: *page 105, $y_{t-1}, y_{t-2}, y_{t-2}$ and y_{t-7}. Please correct the second y_{t-2} to the lag you mean.*

Response: I have corrected the typo to the lag used in the analysis, y_{t-3}.

Comment 4: *What exogenous variables are you referring to on page 105? Where are they from? For instance what is total load? it seems to not be defined properly. I know you mention the variables in the introduction, but there is nothing on where they come from. Are they from the same place as the spot price data?*

Response: The exogenous variables mentioned on page 105 are those that are not directly related to the price-target variable. These include the day-ahead forecast of the total load, day-ahead forecast of renewable energy sources, EU emission allowance price, fuel prices, and dummy variables.

The source of the data is from the transparency platform ENTSOE.[1] The text is to fit the electricity forecasting literature and it employs data of Marcjasz et al. (2023) obtained via *Energy Economics* journal website.

Comment 5: *top of page 109 "with" is repeated twice. Please remove one of them.*

Response: I have removed the duplicate in the final text.

Comment 6: *I like the way the results are presented in this paper. They are clear and contain lots of information in a concise and easily digestible manner. It is also great that the candidate makes their package and replication code available freely for researchers.*

[1] ENTSOE, 2022. ENTSOE Transparency. https://transparency.entsoe.eu/.

Response: Thank you for your words and diligence in examining the thesis.

A.3 Response to comments of Simon Trimborn, Ph.D.

The following text is the original opponent's report, in which my surname has been corrected to *Hanus*.

A.3.1 Opponent's report

It is my pleasure to provide a review for the doctoral thesis handed in by Lubos Hanus at Charles University. Mr. Hanus thesis consists of 4 Chapters spanning topics of data-driven model construction and data analysis. He analysed the business cycle co-movements between Visegrad countries and the EU, time-varying effects between macroeconomic variables, probabilistic forecasting and energy price forecasting.

Mr. Hanus added in all 4 chapters to the academic literature. Chapters 3 and 4 stand out as they contain methodological advancements whereas empirical contributions were made in Chapters 2 and 5. In Chapter 3 Mr. Hanus suggests to use a time-varying parameter model in a VAR framework based on which to conduct impulse response analysis but in the frequency domain, not the time domain. This allows for differentiating between impulse response effects as by their occurrence. In Chapter 4, Mr. Hanus suggests to utilize deep learning to conduct probabilistic forecasting. Indeed the complex structures of density functions are difficult to estimate, in particular when only little data are available. Turning towards deep learning techniques for such an estimation is a novel and smart way of conducting such an analysis. Based on the methods suggested, I can confidently say that Mr. Hanus made original methodological contributions in his thesis.

Further in the Chapters 2 and 5, he conducted an empirical investigation of business cycles and electricity prices. Utilizing the method developed in Chapter 4, Mr. Hanus contributes to the literature on electricity price forecasting by showing that his method excels on the task Given the recent excessive price changes in electricity prices, accurate estimation of prices became more important to ensure stable pricing for the consumers. By this, Mr. Hanus makes an important contribution to recent societal issues.

Mr. Hanus extends with his methodological contributions studies which are pub-

lished in respected journals such as Review of Economic Statistics, Review of Financial Studies, Journal of Business and Economic Statistics, among others. Consequently he makes contributions to method which are relevant to the wider academic community. With his empirical work he adds to the literature published in Journal of Macroeconomics, Journal of Political Economy, International Journal of Forecasting, among others. These empirical studies are published in some of the most respected journals in their field.

Given the contributions made, I am certain all chapters will be published in respected topical journals. In fact the first chapter is already published, one has an invitation for revise and resubmit and another one is currently submitted. Based on what I saw, I could imagine that Chapter 4 stands a chance of publication in the Journal of Financial Econometrics and Chapter 5 in the International Journal of Forecasting. Also Journal of Empirical Finance or Quantitative Finance are possibilities.

In my opinion Mr. Hanus thesis is a fine piece of work which would also be eligible for thesis defense at other institutions. I do have one comment: Chapters 3 and 4 only compared the suggested method against one other method. This raises the question why that particular method was chosen and how the methods of Mr. Hanus would compare against other methods. It would be good to comment on this in the thesis. Apart from this, I do not have any major comments but a suggestion for future research. I observed that the analysis in Chapter 5 focused on the overall prediction accuracy. Given the recent excessive price changes in the electricity markets, it would be interesting to develop a method based on deep learning for probability forecasts to predict accurately prices during periods of market excesses as we saw them recently. This would be accompanied by sparsity in the data, resulting in a research problem which is interesting from a methodological and societal standpoint.

Lastly, I would like to recommend the thesis of Mr. Hanus for defense without substantial changes.

A.3.2 Comments that appear in the opponent's report

Comment 1: *Chapters 3 and 4 only compared the suggested method against one other method. This raises the question why that particular method was chosen and how the methods of Mr. Hanus would compare against other methods. It would be good to comment on this in the thesis.*

Response: Thank you for your comment. The main objective of this work

is to contribute to the literature by presenting an approach to distributional modelling using a neural network.

In the third paper, *"Taming data-driven probability distributions"*, and in the case of Macroeconomic Fan-chart predictions, we use the BVAR model as a representative and state-of-the-art model in the banking industry and among monetary policy practitioners. As mentioned in our response to Comment 11 from Dr. Ellington, a referee asked us to benchmark our model against the DeepAR model (Salinas et al., 2020), which is a state-of-the-art recurrent neural network model. For asset return forecasting, we use the Ordered Logit model proposed by Anatolyev and Baruník (2019) as a benchmark due to the availability of the same dataset, allowing for a direct comparison. Additionally, the AB model is well-suited for forecasting asset return distributions and is a parsimonious linear model.

Regarding the fourth paper titled *"Learning probability distributions of day-ahead electricity prices"*, the selection of a benchmark in the literature on electricity price forecasting can be challenging due to its vastness. However, the quantile regression models and the naive model are widely recognized as state-of-the-art benchmarks, making them a suitable choice. Additionally, the data and test sample used in this study are the same as those in Marcjasz et al. (2023), allowing for a direct comparison.

It has not been our intention to produce a horse race between models, as there are many different models that could be analysed and compared, including linear models, non-parametric models, and models from the neural networks library.

Comment 2: *Apart from this, I do not have any major comments but a suggestion for future research. I observed that the analysis in Chapter 5 focused on the overall prediction accuracy. Given the recent excessive price changes in the electricity markets, it would be interesting to develop a method based on deep learning for probability forecasts to predict accurately prices during periods of market excesses as we saw them recently. This would be accompanied by sparsity in the data, resulting in a research problem which is interesting from a methodological and societal standpoint.*

Response: Thank you for your suggestion. It is an interesting idea and may be worth considering for future research.

References

Anatolyev, S., & Baruník, J. (2019). Forecasting dynamic return distributions based on ordered binary choice. *International Journal of Forecasting*, *35*(3), 823–835.

Baruník, J., & Hanus, L. (2024). Fan charts in era of big data and learning. *Finance Research Letters*, *61*, 105003.

Belongia, M. T., & Ireland, P. N. (2016). The evolution of US monetary policy: 2000–2007. *Journal of Economic Dynamics and Control*, *73*, 78–93.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Dew-Becker, I., & Giglio, S. (2016). Asset pricing in the frequency domain: Theory and empirics. *Review of Financial Studies*, *29*(8), 2029–2068.

Ellington, M. (2018). The case for Divisia monetary statistics: A Bayesian time-varying approach. *Journal of Economic Dynamics and Control*, *96*, 26–41.

Ellington, M. (2022). The empirical relevance of the shadow rate and the zero lower bound. *Journal of Money, Credit and Banking*, *54*(6), 1605–1635.

Fritsch, F., & Carlson, R. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, *17*(2).

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior Selection for Vector Autoregressions. *The Review of Economics and Statistics*, *97*(2), 436–451.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5), 920–964.

Hanus, L., & Vácha, L. (2020). Growth cycle synchronization of the Visegrad Four and the European Union. *Empirical Economics*, *58*(4), 1779–1795.

Israel, R., Kelly, B. T., & Moskowitz, T. J. (2020). Can machines "learn" finance? *Journal of Investment Management*, *18*(2), 23–36.

Kuschnig, N., & Vashold, L. (2021). Bvar: Bayesian vector autoregressions with hierarchical prior selection in r. *Journal of Statistical Software*, *100*(14), 1–27.

Liu, C., & Maheu, J. M. (2009). Forecasting realized volatility: A Bayesian model-averaging approach. *Journal of Applied Econometrics*, *24*(5), 709–733.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Marcjasz, G., Narajewski, M., Weron, R., & Ziel, F. (2023). Distributional neural networks for electricity price forecasting. *Energy Economics*, *125*, 106843.

McCracken, M., & Ng, S. (2020). *FRED-QD: A quarterly database for macroeconomic research* (tech. rep.). National Bureau of Economic Research NBER Working Paper No. 26872.

McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, *72*(3), 821–852.

Rapach, D., Strauss, J., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, *23*(2), 821–862.

Sadhwani, A., Giesecke, K., & Sirignano, J. (2020). Deep Learning for Mortgage Risk*. *Journal of Financial Econometrics*, *19*(2), 313–368.

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181–1191.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, *58*(3), 644–719.

Uhlig, H. (2005). What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics*, *52*(2), 381–419.

Volpicella, A. (2022). Svars identification through bounds on the forecast error variance. *Journal of Business & Economic Statistics*, *40*(3), 1291–1301.

Wright, J. H. (2008). Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics*, *146*(2), 329–341.

Wu, J. C., & Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *Journal of Money, Credit and Banking*, *48*(2-3), 253–291.

Wu, J. C., & Zhang, J. (2019). A shadow rate new keynesian model. *Journal of Economic Dynamics and Control*, *107*, 103728.

,

The last page left empty.