Statistical and machine learning for credit and market risk management

A dissertation in partial fulfillment of the requirements for the degree of
Doktor der Wirtschaftswissenschaft (Dr. rer. pol.)

submitted to the

Faculty of Business, Economics,
and Management Information Systems

University of Regensburg

submitted by

Maximilian Nagl, M.Sc. in Business Administration

Advisors

Prof. Dr. Daniel Rösch

Prof. Dr. Rolf Tschernig

Date of disputation

January, 26th 2022

Acknowledgments

First and foremost, I would like to express my gratitude to Prof. Dr. Daniel Rösch for his continuous support, his valuable input, and enduring motivation to go one step further. I am very grateful to him for always having an open door and inspiring ideas. Furthermore, I would like to thank him for raising my enthusiasm for research and for encouraging me to continue on this path. I am grateful to Prof. Dr. Rolf Tschernig for being my second advisor.

I would like to thank all of my colleagues for uplifting discussions and, more than anything, for a wonderful time and cheerful evenings. In particular, I would like to thank my co-authors, Dr. Jennifer Betz, Prof. Dr. Ralf Kellner, and Dr. Michael Kratochwil for the challenging discussions and many hours spent together.

Last but not least, I would like to thank my wonderful family and awesome friends. Above all, I am grateful to my fantastic parents Lorenz and Kathrin for doing an amazing job as parents. *All I am, I am because of you.* I want to thank Matthias for being a wonderful brother and colleague and for reading and improving every paper I ever wrote.

I dedicate this thesis to my wonderful wife Cathrine. Thank you for your unconditional love, your care, and your continuous support during this journey. Thank you for your eternal belief in me, even in times I have lost it. *It was quite a journey - but the best is yet to come.*

Contents

List of Figures

List of Tables

Introduction

Motivation and area of research

Financial institutions play a major role in the stability of the financial sector. These institutions have a crucial role as intermediaries to support the supply of money and lending as well as the transfer of risk between entities. In general, the stability of modern financial systems is considered a building block of economic growth (Basel Committee on Banking Supervision, 2017). However, this intermediary function exposes financial institutions to several types of risk. Credit risk is defined as the risk that an obligor fails to meet its obligations (Basel Committee on Banking Supervision, 2000). This type of risk is characterized by three parameters. The Probability of Default (PD) quantifies the probability that an obligor will not fulfill his agreed obligations in a future period of time. The Loss Given Default (LGD) denotes the share of the outstanding amount that is lost due to the failure to comply with the obligation. The Exposure at Default (EAD) defines the outstanding amount at the time of failure. Market risk encompasses the potential financial losses due to movements in market prices (Basel Committee on Banking Supervision, 2019b). These markets include for example stock markets and derivative markets. Credit risk accounts for the largest share with roughly 84% of risk-weighted assets of 131 major EU banks as of June 2020, whereas market risk has a share of 4%. The latter type of risk increased by more than 22% compared to June 2019, which can be attributed to the turbulence caused by the COVID-19 crisis (European Banking Authority, 2020). Therefore, managing these risks is important for financial institutions, but also for the economy in general, as it can contribute to the ability of financial institutions to fulfill their role as intermediaries at any time. Given their systemic importance, regulatory requirements are imposed to give the financial institutions guidance on how to manage these risks and how to determine an adequate capital

buffer.

This capital should absorb potential losses from their business tasks (Basel Committee on Banking Supervision, 2017). This is especially important in difficult times, where a distressed financial sector can lead to a reduction of leading activities. Especially in economic downturns, the role of supplying liquidity and lending is more important than ever. In extreme cases, the reduction of intermediary activities can even cause a recession, see Ivashina and Scharfstein (2010). Therefore, the precise estimation of the determinants for various sources of risk is a highly important task for the economy in general, and for financial institutions in particular. During the last decades, the computational power and storage capacities increased substantially, whereas the costs declined sharply. This enables researchers and practitioners to use more advanced and computationally intensive algorithms (FERMA, 2019). This is especially important for machine learning models, but also for Bayesian statistical models.

During the last decades, the research on neural networks, in particular, has substantially increased and important results have been derived. Early works on the universal approximation theorem by Cybenko (1989) and Hornik (1991) prove the approximation capabilities of neural networks and paved the way for their successful application in various fields of research. Advanced statistical and machine learning models have been applied in risk management research early on, see, e.g., Odom and Sharda (1990); Tam (1991); Tam and Kiang (1992) or Hutchinson et al. (1994). With the ongoing development of new algorithms or inference methods, the potential for risk management is large. Surveys conducted by the Bank of Canada (2018), the Bank of England (2019), or the Deutsche Bundesbank (2020) reveals that machine learning applications are gradually applied and especially widespread in the financial industry. There are already real-life applications in some institutions, but the majority of potential use cases are expected in the upcoming years (Bank of England, 2019). Hence, many financial institutions are still in the early adoption phase. Moreover, if these methods should be used in a regulatory environment it is inherent to make the approximated relations and the driver of predictions transparent to regulators. Therefore, the field of Explainable Artificial Intelligence (XAI) becomes increasingly important for risk management applications, see, e.g., Fritz-Morgenthal et al. (2021). This is also apparent in publications of regulatory authorities, see, e.g., Basel Committee on Banking Supervision (2019a) or Deutsche Bundesbank (2020). At the time of writing, there is a consensus that XAI methods are a promising answer to regulatory concerns, as they make the machine learning models less opaque.

The thesis sheds light on the application of advanced statistical and machine learning methods

for credit and market risk management. These applications are handled in four independent research papers (see Chapter 1, 2, 3, and 4). The first deals with advanced Bayesian methods to address the challenging risk parameter EAD and its behavior in downturn periods. The second paper focuses on the combination of statistical and machine learning methods to cope with various aspects of LGD and has a special focus on XAI methods. The third research paper applies neural networks for the calibration of financial models with a special focus on their real-world benefits. The last research paper addresses in detail the non-linearity entailed in stock market movements. The following paragraphs give a brief introduction to the background and motivation for each research paper.

Research paper I — *Credit line exposure at default modeling using Bayesian mixed effect quantile regression*

For the majority of companies, credit lines are the dominant funding source, see Segura and Zeng (2020) and Lins et al. (2010). For example in the United States, 80 % of the small and medium-sized enterprises (SME) funding depends significantly on credit lines (Sufi, 2009). Following Colla et al. (2013), this type of credit is the second most important financing instrument for listed companies as well. Therefore, credit lines are a cornerstone of the financial strategy and needs for a vast majority of companies. These credit lines are not only important to support the growth of companies in expansion, but also in crisis periods. Agarwal et al. (2006) and Barraza and Civelli (2020) argue that credit lines are important for companies to sustain their investments, operations and their liquidity in economic downturns. Hence, they are crucial throughout the business cycle. However, there are two sides of the same coin. Especially in crisis periods, where credit lines are needed to soften economic shocks for companies, they expose banks to higher liquidity and credit risk as well. Ivashina and Scharfstein (2010) show that financial institutions faced bank runs in the Global Financial Crisis inducing high liquidity risk to them. Furthermore, Acharya and Mora (2015) shows that banks with many undrawn credit lines entail a higher risk in crisis periods. Although their importance for financial institutions and the economy, credit lines have found little attention in the academic literature, especially from the credit risk perspective. There exit only a few papers that study the determinants of credit line EAD. The Basel Accord (Basel Committee on Banking Supervision, 2017) require financial institutions to model the exposure at default using conversion factors. They relate the potential draw down on a credit line to the observed balance and limit one year prior to default. These conversion factors exhibit a challenging distribution and, thus, required special attention. The first research paper aims at investigating the crisis characteristics of credit lines based on

one of the world's largest international databases of defaulted credit lines.

Research paper II — *Opening the Black Box – Quantile Neural Networks for Loss Given Default Prediction*

The literature on Loss Given Default has attracted more and more attention in recent years. However, most of the literature focuses on market-based LGDs, which are available for traded debt such as bonds, see, e.g., Nazemi et al. (2021); Gambetti et al. (2019) or Sopitpongstorn et al. (2021). They are calculated as one minus the price of the bond 30 days after its default and exhibit values between 0 and 1. In contrast, bank loans are commonly not traded and, thus, no market price is available. For bank loans usually workout LGDs are considered, see Betz et al. (2020) or Bellotti et al. (2021). They are calculated based on actual recovery cash-flows during the resolution process. This kind of LGD entails some challenging characteristics. They exhibit a bimodal distribution with higher probability masses at the tails of the distribution. Due to the multitude of cash-flows during the resolution process, values lower than 0 and higher than 1 are frequently observed. Following Krüger and Rösch (2017), the drivers of workout LGDs have different (linear) impacts on the tails of the distribution compared to the middle. Furthermore, there is an open discussion on which macroeconomic variables drive workout LGDs and whether the impact differs for low and high LGDs. Addressing all these challenges and questions requires great flexibility of the underlying model, accompanied by an intuitive way to derive and interpret the main drivers. The second research paper tackles these issues by combining quantile regression with an artificial neural network. The aim is to model and identify all kinds of non-linearities between drivers and workout LGDs. As LGDs are one of the main risk parameters in credit risk, the predicted workout LGDs are commonly used for regulatory purposes. This requires a high level of transparency of the underlying dynamics. In order to meet this requirement, the paper applies novel and advanced explainable machine learning methods to quantify important marginal and joint drivers.

Research paper III — *Deep Calibration of Financial models: Turning theory into practice*

Asset pricing models are frequently used in financial institutions to calculate the value of derivatives or, as a preliminary step, to generate scenarios for subsequent Monte Carlo simulations. This is done by calibrating these models on current market prices. Calibration means that the input parameters of the asset pricing model are set, such that the difference between the current market and the model-implied prices are as small as possible. These models are usually complex and assume highly non-linear relationships between input parameters and

model-implied prices. Due to this complexity, the calibration of these models causes a large computational burden, especially since this calibration has to be done on a regular basis (e.g., daily). Therefore, the choice of an asset pricing model in financial institutions requires balancing the accuracy of the model and the time required for its calibration. These limitations have also led to the widespread use of local optimizers, see Liu et al. (2019). In recent years, applications of machine learning methods to reduce the computational burden have emerged. The acceleration is achieved by approximating the asset pricing model with machine learning methods, e.g., artificial neural networks. Subsequently, this approximation is used instead of the original asset pricing model in the calibration procedure. The third research paper aims at answer the question of whether currently employed calibration frameworks of financial institutions can be accelerated, maintaining similar calibration accuracy. This would make it possible to use more advanced financial models or/and optimizers for the calibration tasks. Furthermore, this could lead to more stable input parameters over time, which might contribute to less volatile Profit & Loss figures over time.

Research paper IV — *Does non-linearity in risk premiums vary over time?*

Following Gu et al. (2020), the risk premium is the difference between the conditional expected stock return and the risk-free rate. Therefore, it can be interpreted as the expected compensation of an investor for investing in risky stocks rather than in the risk-free asset. A primary goal of asset pricing is to investigate the main drivers of risk premiums, i.e., why different stocks earn different average returns. The empirical literature commonly uses cross-sectional linear regressions to predict future risk premiums and infer the important drivers via statistical tests, see, e.g., Fama and French (2008) or Lewellen et al. (2015). However, there is a growing body of literature that provides evidence of non-linear relationships between drivers and risk premiums, see, e.g., Gu et al. (2020); Chen et al. (2020) or Bryzgalova et al. (2020). The evidence is based on the observation that machine learning methods outperform linear regressions in forecasting exercises. This outperformance is attributed to the ability of machine learning models to allow for all kinds of non-linearities. Their flexibility may enable the models to better approximate the hidden and probably highly non-linear data generating process. However, there are still two open questions with respect to non-linearity in risk premiums. First, how much non-linearity is actually modeled? Second, does this non-linearity vary over time? The aim of the fourth research paper is at answering these important questions. By using a flexible machine learning method and novel explainable machine learning techniques, the paper seeks insights into the hidden dynamics of risk premiums and their evolution over time.

Literature

Focussing on the literature of Exposure at Default modelling, it can be stated that there are considerable fewer publications than for PD and LGD. There exist basically two strands of literature in EAD modelling. Direct approaches use the EAD as the dependent variable. These approaches usually involve multi-stage models, for example Hon and Bellotti (2016); Leow and Crook (2016); Tong et al. (2016) and Thackham and Ma (2019). In contrast, indirect approaches model the EAD of credit lines using so-called conversion factors. These factors relate the EAD to certain determinants one year prior to default. These are the Limit, i.e., how much can the obligor draw, and the Balance, i.e., how much has the obligor already drawn. This way to model EAD of credit lines is also the approach required by Basel regulations (see Basel Committee on Banking Supervision, 2017, §241, 242). However, these conversion factors are challenging. They exhibit a bimodal distribution with large probability masses at 0 and 1, accompanied by a high amount of outliers. However, the majority of studies employ classical linear regressions (see Araten and Jacobs Jr, 2001; Moral, 2006; Qi, 2009). Barakova and Parthasarathy (2013) use median regression to increase the robustness to outliers. First suggestions to consider the distributional features of conversion factors are multi-stage models (see Valvonis, 2008) or beta regression (see Jacobs Jr, 2010). Yang and Tkachenko (2012) find single layer neural networks to be superior, indicating that conversion factors might not be linearly related to covariates. There is an ongoing debate regarding the impact of macroeconomic variables, and whether credit line specific risk increases in economic downturns. Jiménez et al. (2009), Gatev and Strahan (2006), and Sufi (2009) conclude that firms tend to draw more in economic downturns, while Barakova and Parthasarathy (2013) report higher EADs in pre-crisis periods. Zhao et al. (2014) find statistically significant higher conversion factors during recession periods. However, Thackham and Ma (2019) state that the EAD of credit lines decrease in crisis periods. Nevertheless, estimates for (economic) downturns are mandatory for volatile segments in Basel regulations (see Basel Committee on Banking Supervision, 2017, §242) which is hampered by a lack of statistically evident systematic variables. Research paper I aims at the distributional challenges of these conversion factors and deeply analyses their behaviour in downturn periods.

Turning to research project II, the literature of LGD is more pronounced than for EAD, although the challenges faced with both risk parameters are comparable. Both distributions typically show a bimodal characteristic which makes the choice of a suitable model challenging as well. Most paper focus on market-based LGDs and compare a variety of different methods (see, e.g., Bastos, 2010; Grunert and Weber, 2009; Loterman et al., 2012; Qi and Yang, 2009; Qi and Zhao,

2011; Khieu et al., 2012; Gambetti et al., 2020; Bellotti et al., 2021; Sopitpongstorn et al., 2021). Summarizing the literature of market-based LGDs, machine learning models tend to perform best due to their ability of modelling non-linearities and interactions parsimoniously. The literature on workout LGDs is more focused on statistical models. Altman and Kalotay (2014) propose a Bayesian finite mixture model of normal distributions with an underlying ordered logit model. A frequentistic version of this model is used by Kalotay and Altman (2017) and a mixture of beta distributions by Calabrese (2014). Variants of mixture models are also used in Betz et al. (2018) and Tomarchio and Punzo (2019). The finite mixture models seem to be suitable to account for the more pronounced bimodality of workout LGDs and the potential non-linear dependence on their drivers. Krüger and Rösch (2017) use linear quantile regression to find varying impacts of the drivers over the full conditional distribution. Their empirical analysis show, that the linear quantile regression has the best distributional fit and outperforms finite mixture models. Research paper II aims at lifting the findings of both strands of literature by combining the linear quantile regression with an artificial neural network. Furthermore, the paper uses advanced explainable machine learning methods to evaluate the most important drivers and joint effects. A detailed literature overview of these models can be found in Chapter 2.

The application of machine learning for the valuation of derivative instruments started surprisingly long ago. One of the first paper by Hutchinson et al. (1994) aims at estimating the pricing function of derivatives using an artificial network. This can be seen as a non-parametric, model free way to value derivative instruments. In subsequent studies by Quek et al. (2008) and Culkin and Das (2017), this idea was resumed and extended to other applications, such as trading and hedging. The second strand of literature focuses on the approximation of advanced asset pricing models to accelerate the valuation of derivative instruments. Ferguson and Green (2018) approximate an asset pricing model for equity basket options. Hirsa et al. (2019) use neural networks for the valuation of European, American and Barrier options. Furthermore, Liu et al. (2019) focus on option valuations using the models of Heston (1993) and Bates (1996). With respect to interest rate models, Kienitz et al. (2020) is among the first to approximate the dynamic of swaptions using the model of Hull and White (1990) and Trolle and Schwartz (2009). In recent years, these approximations were not only used for valuation purposes, but also for calibration. Hernandez (2017) is the first to introduce this idea and apply a neural network to the calibration of a single-factor model based on Hull and White (1990). Subsequently, Dimitroff et al. (2018) use convolutional neural networks, which are commonly applied to image recognition and computer vision, to calibrate the stochastic volatility model of Heston (1993).

These papers show a substantial increase in performance, even for these more simplistic models. Therefore, neural networks are also applied to the more complex strand of rough volatility models, see Bayer and Stemper (2018); Bayer et al. (2019); Stone (2020) and Horvath et al. (2021). However, non of the previous studies investigated whether the accelerations can also be lifted in practical applications. Therefore, research paper III aims at turning the theoretical benefits into practice by comparing the neural network calibration framework to a real-life implementation at a large financial institution.

The final research paper of this dissertation focuses on the drivers of excess stock returns, i.e., risk premiums. Over the last decades, several hundred papers developed factors that explain the cross-section of excess stock returns, see Harvey et al. (2016). In addition, most of the variation in characteristic values and returns are in the extremes of the characteristic distribution and the dependence between characteristics and risk premiums seems to be non-linear, (Fama and French, 2008). Following the presidential address of Cochrane (2011), the identification of the variables, which provide independent (linear) explanatory power has gained large interest, see, e.g., Lewellen et al. (2015) and Green et al. (2017). As the form of dependence between the predictors and risk premiums seems to be non-linear, the application of machine learning methods has also increased. Gu et al. (2020) compare many statistical and machine learning models and show that neural networks and regression trees perform best. Bryzgalova et al. (2020) use decision trees to group similar stocks together and use this information for portfolio sorts. Their strategy triples the Sharpe ratios, compared to traditional portfolio sorts. Feng et al. (2020) use hidden states of a neural network to reduce the dimension of their portfolio sorts, which automatically allow non-linearities and interactions. They find again a superior performance compared to a traditional portfolio sort. Rossi (2018) use a machine learning algorithm to construct mean-variance efficient portfolios and document a superior performance. Chen et al. (2020) apply a combination of machine learning algorithms to estimate an asset pricing model for excess stock returns. Freyberger et al. (2020) applies adaptive group LASSO to select the variables with an independent (incremental) explanatory power for expected return predictions. They find that only a small number of predictors have an (time-varying) impact and non-linear relationships matter.[1] Recently, also risk premiums of bonds and hedge funds are targeted with machine learning. Bianchi et al. (2020) employ a battery of machine learning algorithms to forecast bond returns and find neural networks among the best performing methods. Wu et al. (2021) applies machine learning methods to forecast hedge fund returns

[1] Another strand of literature focusses on the application of machine learning methods to factor models, see Kelly et al. (2019), Pelger (2020) , Pelger and Xiong (2021) and Lettau and Pelger (2020).

and use them for selection. Again, neural networks are the best choice. Summarizing, there is broad evidence that the dependence of predictors and risk premiums is non-linear, which results in superior performance of machine learning methods. Furthermore, Freyberger et al. (2020) document a time-varying impact of some predictors. Research paper IV focuses on the non-linear relationships between predictors and the excess stock returns. Moreover, it aims at quantifying how much non-linearity is actually modelled in risk premium predictions.

Contributions

Related to research paper I, II, III, and IV, the main contributions of this thesis are structured by independent research papers which are presented in the individual chapters of this thesis (see Chapter 1, 2, 3, and 4).

Contribution I — *Credit line exposure at default modeling using Bayesian mixed effect quantile regression*

Although credit lines are important for companies and the economy in general, there exist only a small body of literature focusing on the EAD of this type of loan. Especially, with respect to the potential non-linear impact of drivers and the performance of these models to generate downturn estimates, this paper seeks out to give broad and in-depth evidence. To the best of my knowledge, this paper is the first to empirically compare the downturn characteristics of bank loan credit lines in two important regions, namely Europe and the USA. The empirical evidence is based on one of the world's largest databases on defaulted credit lines, provided by Global Credit Data (GCD). To account for the challenging shape of the EAD conversion factors, the paper applies a Bayesian quantile regression to allow for non-linear impacts over the conditional distribution and compares the results to a standard linear model. By using a quantile regression approach, this paper is - to the best of my knowledge- also the first paper to model the full conditional distribution of conversion factors. The empirical analysis reveals strong varying impacts over the distribution, which can be seen as evidence of diverging impacts of low and high conversion factors. Furthermore, the paper studies in-depth the impact of macroeconomic variables and their ability to generate sufficiently conservative downturn estimates, as required by Basel regulations (Basel Committee on Banking Supervision, 2017). The empirical results suggest that the evidence of macroeconomic variables seems to vanish in the tails of the distribution and for credit lines that are drawn heavily one year prior to their default. Therefore, the paper argues that credit lines with high risk (low utilization one year prior to default) are particularly affected by the economic shocks. Systematic variation which

cannot be measured by macroeconomic variables is modelled via time-specific random effects. This allows us to create adequate downturn estimates, even in settings where the identification of meaningful and evident macroeconomic variables is unfeasible. Furthermore, it offers banks and regulators an approach to incorporate their individual margin of conservatism for capital requirements of credit lines in stressed periods.

Contribution II — *Opening the Black Box – Quantile Neural Networks for Loss Given Default Prediction*

Modelling Loss Given Default has attracted more and more attention, accompanied with applications of up-to-date machine learning algorithms, see, e.g., Bellotti et al. (2021) or Gambetti et al. (2020). However, there are several papers that show that classical statistical models provide a reasonable distributional fit as well, see, e.g., Betz et al. (2018) or Krüger and Rösch (2017). The aim of research paper II is to take the next logical step by combining a powerful statistical model with a powerful machine learning method. This may lift the potential of both approaches and result in an overall superior model. The paper combines the well-known quantile regression, introduced by Koenker and Bassett (1978), with an artificial neural network. This allows the paper to perform the estimation of many quantiles in one single optimization step while controlling for monotonic increasing quantile estimates. Furthermore, this reduces the computational burden, as standard quantile regressions are fitted separately for every quantile. Moreover, the quantile regression neural network (QRNN) allows for any kind of non-linear relationships between input variables and every quantile estimate, without the need to specify the functional form of marginal and joint impacts in advance. The empirical analysis is performed on a sub-sample of US and European loans drawn from one of the largest loss databases in the world, provided by Global Credit Data (GCD). This database encompasses 55 globally acting financial institutions, several of which are systematically relevant. Therefore, the evidence is based on a broad sample of the banking sector. By comparing the distributional fit of the QRNN to a battery of challenger models, such as finite mixture regressions, beta regressions and the linear quantile regression, the paper finds a superior fit of the ORNN in-sample as well as out-of-time. The good performance of the QRNN may be traced back to its flexibility. With the application of advanced explainable machine learning approaches, the paper identifies the most important divers of each quantile. Opening up the black-box of neural networks is of major concern for financial institutions which aim at using these algorithms for regulatory purposes. With respect to the main drivers, the paper finds novel insights. First, the macroeconomy is two times more important in the US than in Europe. Second, the economic surrounding interacts

in Europe the most with the collateralization of the underlying loan. Interestingly, the level of seniority has in the United States large joint impacts with the economic variables. Overall, the paper finds large non-linearities especially in higher quantiles, which refer to higher losses. Furthermore, the paper shows that the QRNN can be easily used to generate downturn estimates, required by regulators (Basel Committee on Banking Supervision, 2017).

Contribution III — *Deep Calibration of Financial models: Turning theory into practice*

The third research paper builds on existing strategies to approximate asset pricing models for calibration, following Liu et al. (2019) and Horvath et al. (2021). Recent literature shows the benefits of neural networks on simulated data or stylized empirical applications. This paper is innovative by comparing the neural network approximations to a real-life implementation that is in action at a large financial institution. It sheds light on the practical benefits of machine learning applications in market risk management. The paper applies the calibration framework to an interest rate (IR) term structure model based on Trolle and Schwartz (2009). The empirical application entails historic market data for a consecutive series of trading days from January 2019 to September 2020. It contributes to the literature of calibration using neural networks in three ways. First, the study is innovative by approximating the model of Trolle and Schwartz (2009) based on a large set of swaptions using a wide range of historical market data. Second, the increase of calibration speed enables the use of a (slower) global optimizer instead of a local one, employed by the financial institution. Still, the neural network approximation is four times faster than the real-life calibration framework. The empirical results suggest that the calibrated model parameters of the TS model using the neural network approach are more stable over time, compared to the real-life implementation. This stability can have decisive managerial implications as more stable calibration results can contribute to less volatile Profit & Loss estimates over time. A positive side-effect of the TS model is that several more simplistic, but widely used in practice, IR term structure models can be derived, following Trolle and Schwartz (2009). Therefore, the empirical results are interesting for a wide range of financial institutions and market participants. Third, with respect to the regulatory requirements of calibration frameworks, lessons learned and practical guidelines are derived. This may ease the discussion with regulators to accept machine learning methods in real-life calibration frameworks.

Contribution IV — *Does non-linearity in risk premiums vary over time?*

Recently, a number of publications focus on the prediction of the risk premium using machine learning methods, see, e.g., Gu et al. (2020); Chen et al. (2020) or Bryzgalova et al. (2020). Overall, the machine learning algorithms outperform classical linear models. The majority of papers trace this superiority back to the flexibility of the algorithms to model non-linearity in almost any kind of functional form. Furthermore, Freyberger et al. (2020) document a time-varying impact of some drivers. However, there are two open questions. First, how much non-linearity is actually modelled? Second, does this non-linearity vary over time? The paper contributes to the literature by introducing a novel measure of non-linearity in machine learning predictions based on Apley and Zhu (2020). This measure is model agnostic and, thus, applicable to any machine learning method. To trace the non-linearity back to specific variables, the paper also extends approaches by Sadhwani et al. (2021). The combination of both extensions allows an in-depth quantification of non-linearity modelled from a very high perspective, i.e., how large is the overall non-linearity, to a very detailed perspective, i.e., which specific variables drive this non-linearity? With respect to risk premiums, the paper fit neural networks on subsequent time slices to quantify the non-linearity over time. Thereby, a time-varying behaviour of non-linearity in risk premium prediction is documented. Moreover, the analysis shows an inverse relationship of linearity in risk premium predictions and uncertainty measured by the VIX. In periods of high uncertainty, e.g., crisis periods, the non-linearity increases considerably. Overall, the paper documents non-linearity for many predictor variables, especially in crisis periods. For example, stock-level volatility measures show large non-linearities in uncertain times. Interestingly, in less uncertain times, the overall non-linearity decreases considerably. This indicates that classical linear asset pricing models, as employed by Fama and French (2008) or Lewellen et al. (2015), are suitable in normal times, but greater flexibility is required in economic downturns.

Structure

This thesis consists of four independent research papers with varying co-authors.[2] Chapter 1 presents the first paper (*Credit line exposure at default modeling using Bayesian mixed effect quantile regression*). In Chapter 2, the second paper (*Opening the Black Box – Quantile Neural Networks for Loss Given Default Prediction*) is propound. The third paper is subject to Chapter 3 (*Deep Calibration of Financial models: Turning theory into practice*). The fourth and last paper (*Does non-linearity in risk premiums vary over time?*) is comprised in Chapter 4. The Conclusion summarizes and provides an outlook.

[2] The co-authors and the current state of the research papers are mentioned at the beginning of each chapter.

Chapter 1

Credit line exposure at default modeling using Bayesian mixed effect quantile regression

This chapter is joint work with Jennifer Betz[1] and Daniel Rösch[2] and corresponds to a working paper with the same name (submitted to *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Revised and resubmitted).

For banks, credit lines play an important role exposing both liquidity and credit risk. In the advanced internal ratings based approach, banks are obliged to use their own estimates of exposure at default using credit conversion factors. For volatile segments, additional downturn estimates are required. Using the world's largest database of defaulted credit lines from the US and Europe and macroeconomic variables, we apply a Bayesian mixed effect quantile regression and find strongly varying covariate effects over the whole conditional distribution of credit conversion factors and especially between US and Europe. If macroeconomic variables do not provide adequate downturn estimates, the model is enhanced by random effects. Results from European credit lines suggest that high conversion factors are driven by random effects rather than observable covariates. We further show that the impact of the economic surrounding highly depends on the level of utilization one year prior default, suggesting that credit lines with high drawdown potential are most affected by economic downturns and hence bear the highest risk in crisis periods.

Keywords: Credit Risk, Credit Conversion Factor, Exposure at Default, Global Credit Data, Quantile Regression, Random Effects

JEL Classification: C23, G21, G33

[1] University Regensburg, Chair of Statistics and Risk Management, 93040 Regensburg, Germany, email: jenniefer.betz@ur.de.

[2] University Regensburg, Chair of Statistics and Risk Management, 93040 Regensburg, Germany, email: daniel.roesch@ur.de.

1.1 Introduction

Credit lines are the dominant funding source for companies all around the world (see Segura and Zeng (2020) and Lins et al. (2010)). In the US – a traditionally rather market-oriented country – 80% of small and medium sized enterprises (SME) heavily rely on these funding instruments (see Sufi, 2009) and credit lines are the second most important debt financing category for listed companies (see Colla et al., 2013). Acharya et al. (2014), Acharya and Mora (2015) and Acharya et al. (2020) argue that credit lines are important for the economy in general as they provide (short-term) liquidity to corporations to sustain investments. Particularly in crisis periods when credit quality deteriorates, credit lines ensure that companies can maintain their operations and contribute to sustain investments and liquidity (see also Agarwal et al., 2006; Gatev and Strahan, 2006; Cornett et al., 2011; Berrospide and Meisenzahl, 2015; Barraza and Civelli, 2020). As a flip-side, they expose banks to both higher liquidity and credit risk. Ivashina and Scharfstein (2010) show that there was a bank run in the Global Financial Crisis (GFC) inducing high liquidity risk. Following Acharya et al. (2013) and Acharya and Mora (2015), banks with undrawn lines become riskier due to this additional risk in times of increased aggregated volatility.
In addition to the well documented liquidity risk, credit lines – such as loan contracts in general – also expose banks to credit risk. In this paper, we focus solely on defaulted credit lines, as we are interested in the dimensions of credit risk induced by the type of loan. In the advanced Internal Ratings Based (IRB) approach of the Basel regulations, banks are obliged to use their own estimates of the three central credit risk parameters – the Probability of Default (PD), the Loss Given Default (LGD), and the Exposure at Default (EAD) – to calculate their capital requirements for loans. For credit lines, the EAD is particularly important because a bank's credit risk exposure is increased when a credit line is drawn and volatile over time.

While the literature on PD and LGD modeling has widened considerably during the last two decades, less attention has been paid to EAD modeling. Literature on EAD modeling can roughly be divided into direct and indirect approaches. Direct modeling of EAD usually involves multi-stage models (Hon and Bellotti, 2016; Leow and Crook, 2016; Tong et al., 2016; Thackham and Ma, 2019). In contrast, indirect approaches are based on conversion factors which can be interpreted as additional drawdowns on the credit line in a specific time period, e.g., one year prior to default (see Section 1.2). As this is also the approach required by Basel regulations (see Basel Committee on Banking Supervision, 2017, §241, 242), we follow this

strand of literature. While indirect approaches allow for beneficial interpretations, they are challenging, i.e., conversion factors tend to exhibit extreme bimodal distributions – comparable to loss rate distributions – and are characterized by high amounts of outliers. Regardless, many studies use a classical linear OLS regression framework (see Araten and Jacobs Jr, 2001; Moral, 2006; Qi, 2009). Barakova and Parthasarathy (2013) additionally apply median regression which is more robust to outliers. Although not recommended by the Basel regulations (see Basel Committee on Banking Supervision, 2017, §247), several studies trim or winsorize the data (see Araten and Jacobs Jr, 2001; Moral, 2006; Jacobs Jr, 2010; Qi, 2009; Yang and Tkachenko, 2012; Barakova and Parthasarathy, 2013). First suggestions to consider the distributional features of conversion factors are multi-stage models (see Valvonis, 2008) or beta regression (see Jacobs Jr, 2010). Yang and Tkachenko (2012) find single layer neural networks to be superior, indicating that conversion factors might not be linearly related to covariates. However, neural networks lack economic interpretability and transparency which hampers application for regulatory purposes.

The risky position of a bank is not only increased by higher exposures when credit lines are drawn, but also through a link between credit line usage and default that was found by several studies (see Jiménez et al., 2009; Araten and Jacobs Jr, 2001; Valvonis, 2008; Qi, 2009; Jacobs Jr, 2010; Jacobs Jr and Bag, 2011; Zhao et al., 2014). Hence, obligors seem to draw heavier when tumbling towards default. In the literature, there is an ongoing debate regarding the impact of macroeconomic variables, and whether credit line specific risk increases in economic downturns. Jiménez et al. (2009), Gatev and Strahan (2006), and Sufi (2009) find statistical evidence that firms tend to draw more lines in economic downturns, while Barakova and Parthasarathy (2013) report higher EADs in contraction (pre-crises) periods compared to crises. Zhao et al. (2014) find statistically significant higher conversion factors during recession periods. Thackham and Ma (2019) even state weak evidence of counter-cyclic patterns in the Global Financial Crisis, i.e., a negative relation of EADs and default rates. In general, the identification of meaningful and statistically evident macroeconomic variables is of high relevance with respect to modeling EAD and conversion factors. In analogy to loss rates, estimates of conversion factors for (economic) downturns are also mandatory for volatile segments in Basel regulations (see Basel Committee on Banking Supervision, 2017, §242) which is hampered by a lack of statistically evident systematic variables. With respect to the literature, conversion factors are almost exclusively estimated with mean-related methods (such as OLS), although the distribution is highly bimodal. Therefore, conclusion with respect to the mean, which is rarely observed, may not be representative for the whole distribution. Furthermore, the bi-modality may lead to heterogeneous (varying) covariate

effects for the different parts of the distribution. This may also be an explanation of the lack of statistically evident systematic variables. For a detailed discussion of heterogeneous covariate effects, we refer to Koenker (2005). Therefore, we argue that using a quantile regression may be more representative for this challenging setting. Additionally, individual quantile functions enable financial institutions to better differentiate between loans and their inherent risk profile.

Given the importance of credit lines and their relation to the macroeconomy, as well as the lack of clear evidence in the literature, this paper provides the following contributions. First, this paper is innovative by investigating the downturn, i.e. crisis periods, characteristics of credit lines for the first time and comparing two important regions, namely Europe and US. Furthermore, our evidence is based on one of the world's largest international datasets with respect to defaulted credit lines. Second, we apply a novel approach to model conversion factors. Because of the regulatory requirements for conversion factors and their bimodal distribution which can hardly be tackled by linear OLS regression, we apply a Bayesian quantile regression (QR) approach. Therefore, this paper is – to the best of our knowledge – the first to model the full conditional distribution of credit conversion factors. We show that the QR approach yields an up to twice as good distributional fit, compared to the OLS regression in an out-of-time forecasting exercise. Additionally, we show that the impact of covariates strongly varies across quantiles, which cannot be captured by standard regression techniques. This suggests that there are severe differences in the determinants of low or high additional drawdowns and between regions, which is not documented in the literature so far. Third, we deeply investigate the impact of macroeconomic variables and their ability to generate sufficiently conservative downturn estimates, as required by Basel regulations. We find that evidence of macroeconomic variables seems to vanish in the tails of the distribution and for credit lines which exhibit high utilization, i.e., lines which are drawn heavily one year prior to default. Thus, credit lines with high risk (low utilization one year prior default) are particularly affected by the economic surrounding. Systematic variation which cannot be measured by macroeconomic variables is modeled via time-specific random effects. This allows us to create adequate downturn estimates, even in settings where the identification of meaningful and evident macroeconomic variables is unfeasible. Furthermore, it offers banks and regulators an approach to incorporate their individual margin of conservatism for capital requirements of credit lines in stressed periods.

The remainder of this paper is structured as follows. Section 1.2 presents the data of defaulted credit lines. In Section 1.3, Bayesian quantile regression – including the extension by time-specific random effects – is introduced. The main results are outlined in Section 1.4. Finally,

Section 1.5 concludes.

1.2 Data

Summarizing the literature reviewed in Section 1.1, EADs might be modeled directly or indirectly by means of conversion factors. The latter represent additional drawdowns with respect to an observed limit, balance or difference at a specific time t. Hereby, a more complete picture of the drawdown behavior of defaulted credit lines can be modeled. For example, (possible) different drivers for low and high additional drawdowns can be determined. Furthermore, the use of conversion factors is recommended by the Basel Accord (see Basel Committee on Banking Supervision, 2017, §241-§250).

Generally, conversion factors should be estimated with a fixed-horizon approach, i.e., all predictions should be linked to information 12 months prior to default (see Basel Committee on Banking Supervision, 2017, §245). Therefore, in the following the time stamp t refers to 12 months before the default in T. A rigorous discussion of advantages and disadvantages of various horizon approaches can be found in Gürtler et al. (2018). In general, the conversion factors consist of a composition of the following variables. Balance_t is the drawn amount of the credit line at time t, Limit_t is the available amount provided by the financial institution up to with the obligor can draw the line, and EAD_T is the drawn amount of the credit line at the time of default T. In the literature four common conversion factors can be found: The Loan Equivalent Exposure (LEQ, calculated by $\frac{\text{EAD}_T - \text{Balance}_t}{\text{Limit}_t - \text{Balance}_t}$), the Credit Conversion Factor (CCF, calculated by $\frac{\text{EAD}_T}{\text{Balance}_t}$), the Exposure at Default Factor (EADF, calculated by $\frac{\text{EAD}_T}{\text{Limit}_t}$) and the Additional Utilization Factor (AUF, calculated by $\frac{\text{EAD}_T - \text{Balance}_t}{\text{Limit}_t}$). As the nomenclature of these factors is not universally defined, we follow the definitions of Leow and Crook (2016). A discussion about the drawbacks of the first three conversion factors can be found in Leow and Crook (2016) and Thackham and Ma (2019). The AUF is suggested by Yang and Tkachenko (2012) and found to be suitible for corporate credit lines by Barakova and Parthasarathy (2013). While incorporating the limit as well as the balance at time t, it is stable for almost completely drawn lines. The AUF is undefined if the limit one year prior default is exactly zero. However, these credit lines are of minor concern in estimating credit risk due to their low potential of additional drawdowns. Furthermore, extreme values occur only if the limit one year prior default is extremely small compared to the additional drawdown[3]. Due to these benefits and

[3] Note that an AUF of one indicates that the additional drawdown is equal to the limit one year prior default. This can only occur if there is no balance one year prior default.

the limited drawbacks, we apply the AUF in the following analysis. For robustness, we also run our analysis using the EADF, but find no differences regarding our contributions[4].

We use access to the world's largest loss and exposure database which is collected by Global Credit Data (GCD)[5]. This cooperative consists of 55 globally acting member banks all around the world encompassing several systemically important institutions. The access to a unique sample of defaulted US American and European corporate credit lines provides exclusive insights accessing a large and important proportion of the banking universe. We use a sample from 2006 until the end of 2018. The database contains information about balance and limit at the time of default and one year prior to default. We use the fixed-horizon approach for calculating the AUF which is in line with the Basel Accord. Imposing a materiality threshold of 500 Euro[6] and using only credit lines where all independent variables are available, we have 14,382 credit lines in Europe and 4,432 credit lines in the US. To reduce the problem of extreme values, we restrict the range of AUF values to $[-0.5, 1.5]$. By including negative AUFs, variables which impact balance reduction until default can be identified, whereas AUFs greater than one enable us to look deeper into the drivers of extreme additional drawdowns beyond the prearranged limits. These are possible due to accumulated interest or banks allowing borrowers to draw beyond their limits, resulting in values greater than 1. With respect to the interval, we delete 3,466 credit lines in Europe and 390 in the US, corresponding to 24.10 % and 8.80 % of the sample. In Europe, 2,976 of the deleted credit lines have limits of zero one year prior default which implies a non-defined AUF.[7] As these credit lines have a low EAD potential, these observations are of minor economic concern. Values with limits greater than zero account for 3.34 % in Europe.

Table 1.1 compares descriptive statistics of the AUF and applied covariables in the two regions. For metric variables the means and a range of quantiles are displayed. For each level of categorical variables, the means and quantiles of the AUF are shown.

[4] Rerunning our analysis using CCF would be counterintuitive, as we would have to omit the most risky credit lines, which are especially important in crisis periods. Furthermore, as the LEQ has severe drawbacks and is only weakly defined in our sample, an additional analysis would not add any robustness.

[5] GCD is a non-profit organization aiming to support its member banks in understanding and modeling credit risk parameters such as LGD and EAD by, inter alia, collecting and pooling detailed loss and exposure information of defaulted loan contracts including credit lines (for further information see https://www.globalcreditdata.org/).

[6] This is in line with the materiality threshold of the European Banking Authority (2016).

[7] In the US, only 34 lines have a limit of zero one year prior default.

Table 1.1: Descriptive statistics

(a) USA

Variable	Level	Quantiles					Mean	STD	Obs.
		0.05	0.25	0.5	0.75	0.95			
AUF		-0.33	-0.07	0.00	0.03	0.59	0.03	0.26	4042
log(Limit)		9.69	11.63	12.96	14.51	16.72	12.86	2.92	4042
Age		0.10	0.84	1.99	3.79	7.92	2.73	2.63	4042
Utilization		0.15	0.80	1.00	1.00	1.00	0.84	0.28	4042
∆GDP		-0.04	0.00	0.02	0.02	0.03	0.01	0.02	4042
Facility type	Medium term revolver	-0.31	-0.06	0.00	0.04	0.61	0.04	0.26	3250
	Short term revolver	-0.37	-0.11	0.00	0.00	0.45	-0.01	0.24	792
Seniority	Pari-passu	-0.40	-0.20	-0.02	0.00	0.61	-0.03	0.28	1010
	Super senior	-0.28	-0.05	0.00	0.06	0.54	0.04	0.25	1550
	Non senior	-0.36	-0.10	-0.02	0.01	0.77	0.03	0.31	150
	Unknown	-0.23	-0.03	0.00	0.06	0.59	0.06	0.25	1332
Industry	Finance, insurance, real estate (FIRE)	-0.28	-0.05	-0.02	0.00	0.42	-0.01	0.21	754
	Agriculture, forestry, fishing (AFF)	-0.34	-0.12	-0.01	0.02	0.38	-0.01	0.24	133
	Mining (MIN)	-0.39	-0.15	0.00	0.21	0.80	0.06	0.34	133
	Construction (CON)	-0.38	-0.05	0.00	0.11	0.55	0.04	0.27	428
	Manufacturing (MAN)	-0.34	-0.10	0.00	0.10	0.75	0.06	0.31	528
	Transp., commu., sanitary serv. (TCEGS)	-0.26	-0.04	0.00	0.06	0.75	0.06	0.28	223
	Wholesale and retail trade (WRT)	-0.35	-0.09	0.00	0.05	0.60	0.02	0.27	541
	Services (SER)	-0.32	-0.05	0.00	0.05	0.61	0.04	0.27	830
	Other (OTH)	-0.30	-0.08	0.00	0.00	0.30	-0.01	0.19	472

(b) Europe

Variable	Level	Quantiles					Mean	STD	Obs.
		0.05	0.25	0.5	0.75	0.95			
AUF		-0.31	-0.01	0.03	0.36	1.04	0.20	0.41	10916
log(Limit)		8.13	9.90	11.38	12.90	15.58	11.51	2.25	10916
Age		0.00	1.21	3.43	6.48	19.26	5.25	6.80	10916
Utilization		0.00	0.48	0.97	1.00	1.00	0.72	0.39	10916
∆GDP		-0.05	-0.01	0.01	0.02	0.03	0.00	0.02	10916
Facility type	Medium term revolver	-0.32	-0.03	0.01	0.18	0.86	0.11	0.32	3206
	Short term revolver	-0.29	0.00	0.00	0.06	0.95	0.09	0.31	379
	Overdraft	-0.31	0.00	0.06	0.50	1.12	0.25	0.44	7331
Seniority	Pari-passu	-0.29	0.00	0.04	0.38	1.06	0.21	0.41	9835
	Super senior	-0.39	-0.08	0.00	0.17	0.94	0.08	0.34	981
	Non senior	-0.40	-0.16	0.04	0.46	0.88	0.14	0.41	100
Industry	Finance, insurance, real estate (FIRE)	-0.28	-0.01	0.01	0.26	1.03	0.18	0.40	2723
	Agriculture, forestry, fishing (AFF)	-0.29	-0.01	0.05	0.36	1.00	0.20	0.38	426
	Mining (MIN)	-0.20	-0.02	0.04	0.40	1.13	0.24	0.45	50
	Construction (CON)	-0.31	0.00	0.08	0.59	1.19	0.27	0.46	1138
	Manufacturing (MAN)	-0.36	-0.02	0.02	0.39	1.05	0.20	0.42	1069
	Transp., commu., sanitary serv. (TCEGS)	-0.34	-0.04	0.02	0.35	1.01	0.18	0.41	518
	Wholesale and retail trade (WRT)	-0.34	-0.03	0.05	0.37	1.06	0.20	0.42	2057
	Services (SER)	-0.31	0.00	0.13	0.62	1.18	0.30	0.46	979
	Other (OTH)	-0.24	0.00	0.03	0.22	0.96	0.15	0.34	1956

Note: The table shows means, standard deviations and quantiles for the AUF and the metric variables. For categorical variables, means, standard deviations and quantiles of the AUF for each level are displayed. The macro variable growth of the Gross Domestic Product (GDP) is lagged by one year. The variables log(Limit) and Utilization correspond to the logarithm of the limit respectively the utilization of the credit line one year prior to default.

Comparing the variable Age, which represents the number of years from origination of the credit line until one year prior default, European lines are on average more than twice as old. This may be attributed to the fact, that in Europe it is much more common to have tight and long-lasting business relationships to banks with respect to funding, whereas in the US, companies are

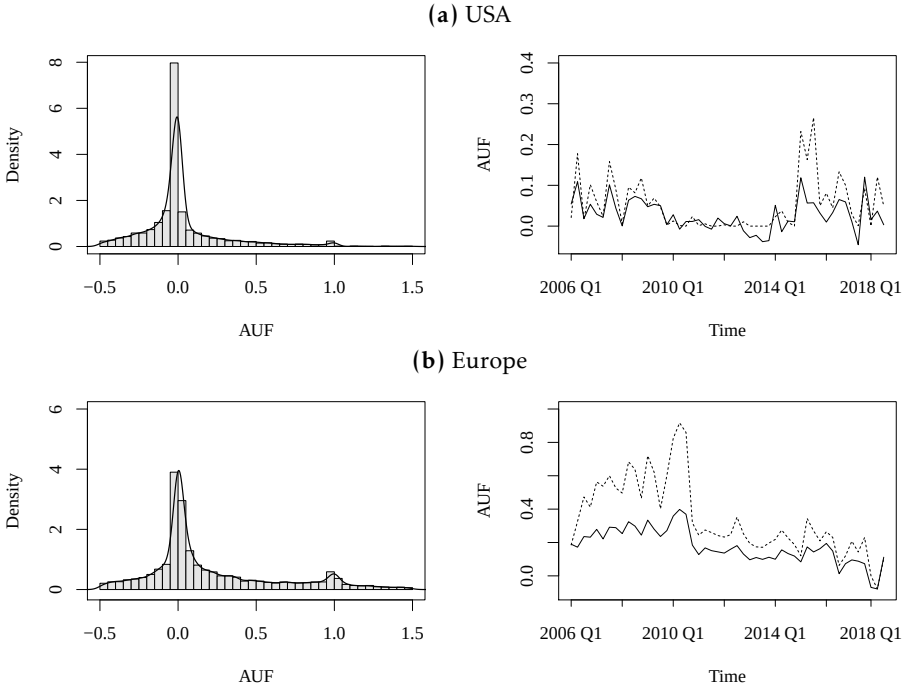
usually more often funded by capital markets (see Antoniou et al., 2008). Furthermore, it is apparent that the AUF differs among regions – especially in higher quantiles as (positive) additional drawdowns are much more common in Europe. This is in line with the observation that Utilization, which represents the percentage of how much is already drawn one year prior default, is higher in the US. In the first quartile, the lines are drawn up to 80%, whereas in Europe, only up to 48%. Due to the higher utilization in the US, the potential of additional drawdowns is limited which might result in a lower AUF. To control for the economic surrounding, we include the year-on-year growth of the Gross Domestic Product (GDP), labeled as ΔGDP in the final model. We also considered other macroeconomic variables, such as stock market growth, changes in house prices, volatility indexes, interest rate spreads, unemployment rates and overall liquidity. ΔGDP has the highest and most evident impact among all tested variables. Following Betz et al. (2018), we use one macro variable in the final model, as they are highly correlated and hence influence their statistical inference. Furthermore, our results in Section 1.4.2 show that the remaining systemic variation can be easily captured with the introduced random effect, avoiding all problems with highly correlated macroeconomic variables. We further include line-specific variables. Facility type controls for different revolving types of credit line and their maturity (overdraft[8], short & medium term revolver). Additionally, the order of claims in the resolution process is included via different levels of Seniority[9]. Log(Limit) controls for the size of the credit line with respect to the available limit one year prior default. We also tested whether the size of the company is a driver of the AUF, but found no evident effect. The impact of the company size may be absorbed by the log(Limit) as larger firms usually require larger credit lines[10]. Furthermore, in the literature the borrower rating is found to be suitable to model additional drawdowns for non-defaulted and defaulted credit lines. However, as we focus on *defaulted* credit lines using the fixed horizon approach, the ratings of the credit lines probably worsen for all defaulted lines one year prior default. To check this, we use a subsample of our data for which we have ratings, but find no difference between the rating categories in terms of the AUF distribution, and a very large part has a non-investmentgrade rating. This is similar to Thackham and Ma (2019), who do not include ratings in their final model for EAD prediction either.

[8] In general the Basel Accord does not require banks to estimate credit conversion factors for non-revolving lines, like overdrafts. Instead, a comparatively low CCF of 10 % is assigned. The descriptive statistics however show that these type of lines have a much greater potential of additional drawdowns. Hence, we include them in our sample to investigate their behavior as well.

[9] Super senior refers to a priority order where only one creditor has prior claims. If there is at least another claimant on the same rank, the seniority is defined as pari-passu.

[10] We also tested other credit-line-specific characteristics such as collateral, but did not find an evident impact, similar to Thackham and Ma (2019).

Figure 1.1: Distribution and time variation of AUF

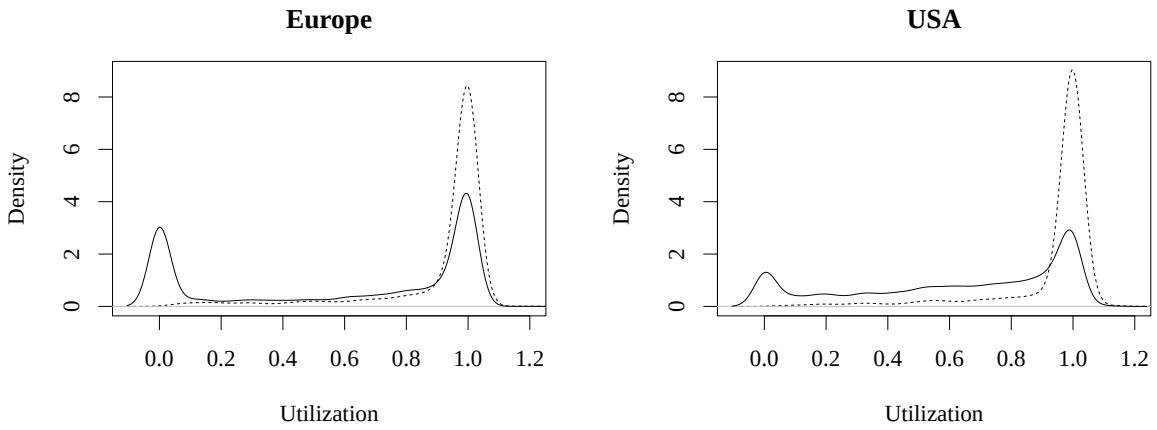


Note: The left panels of the figure show the distribution of the AUF separated by regions. The black lines represent the kernel density estimates, whereas the gray bars illustrate the histograms. The right panels illustrate the time patterns of the AUF divided by regions. The solid lines represents the mean in the quarter of default and the dotted line is the 75% quantile.

The left panels of Figure 1.1 illustrate the kernel density estimates of the AUF. The probability mass around zero is more pronounced in the US, whereas the probability mass around one is greater in Europe. The right panels of Figure 1.1 illustrate the time patterns of the average AUF (solid black line) and its 75 % quantile (black dotted line). Hereby, differences among the regions occur. The average AUF is lower in the US compared to Europe. The Global Financial Crisis and its aftermath is much more pronounced in Europe. This is especially true focusing on the 75 % quantile where the values increased considerably in the GFC and the subsequent quarters. Summarizing, time varying behavior is present in both regions, whereas it is more pronounced in Europe. This may be attributed to the fact of generally higher utilization one year prior default in the US American sample. To investigate this in more detail, we illustrate the distribution of Utilization depending on the realized AUF in Figure 1.2.

Lines with positive and negative AUFs seem to clearly differ in the level of utilization one year prior default. In Figure 1.2, the solid line illustrates the utilization of credit lines with positive AUFs and the dashed line represents credit lines with negative AUFs. Obligors with negative AUFs have more extensively drawn than obligors with positive AUFs. In Europe, there are many more credit lines with almost no and very high utilization one year prior to default, whereas in the US, there is a more equal level of utilization for positive AUFs.

Figure 1.2: Distribution of utilization level one year prior to default



Note: The figure shows the distribution of the level of Utilization separated by positive and negative AUFs. The solid line represents the density of the Utilization for lines with a positive additional draw-down (positive AUF) and the dashed line illustrates the density of the Utilization with exposure reduction (negative AUF).

Overall, there is also evidence that the time varying behavior is quantile-dependent. Usually, an explanation for different systematic behavior may be different default definitions. In this study, all loans have the same default definition according to Basel Committee on Banking Supervision (2017). Hence, we can eliminate the possibility that different systematic behaviors are attributed to different default definitions.

1.3 Methodology

With respect to the extreme bimodal distribution of the AUF (see left panels of Figure 1.1), analysis regarding the conditional mean of the distribution – such as a classical linear regression – may not be favorable as rigorously shown by Krüger and Rösch (2017). Modeling the entire distribution instead infers more comprehensive results. Furthermore, the impact of variables may differ over the distributional range. This is especially true in the existing setting as positive and negative AUFs are jointly modeled. Therefore, we analyze additional drawndowns using quantile regression introduced by Koenker and Bassett (1978) which allows us to model the full conditional distribution of the response variable. As each quantile is modeled separately by a linear regression, a more comprehensive picture of the distribution is obtained. Additionally, it allows for varying impacts of covariates over the entire distributional range. This enables us to detect the (different) drivers of low and high additional drawdowns. These implications are important to financial institutions as they can adjust their line management and, hence, distinguish between low and high drawdowns more exactly.

In the quantile regression approach, each quantile τ of the dependent variable Y is modeled based on a linear function. The corresponding regression function is

$$y_i = x_i\,\beta(\tau) + \epsilon_i(\tau), \tag{1.1}$$

where y_i represents the i-th observation of the response variable and x_i is the known covariate vector which includes a one for the τ-dependent intercept. The vector $\beta(\tau)$ contains the unknown parameters including the intercept and $\epsilon_i(\tau)$ is the quantile-specific error term. Assuming expectation $Q_\tau(\epsilon_i(\tau)) = 0$, the expected τ-quantile of the response variables is given by $Q_\tau(y_i\,|\,x_i) = x_i\,\beta(\tau)$ for $0 < \tau < 1$. The τ-specific estimates of $\beta(\tau)$ are obtained by minimizing the objective function with respect to $\beta(\tau)$:

$$\begin{aligned} &\sum_{i=1}^{n} \rho_\tau\left(y_i - x_i\,\beta(\tau)\right) \\ &\text{with } \rho_\tau(\omega) = \begin{cases} \tau\omega, & \text{if } \omega \geq 0, \\ (1-\tau)\,|\omega| & \text{else.} \end{cases} \end{aligned} \tag{1.2}$$

According to Koenker and Bassett (1978), the minimization problem of Equation (1.2) is solved with simplex algorithms . Yu and Moyeed (2001) and Yu and Zhang (2005) linked the minimization to the maximum likelihood theory via the asymmetric Laplace distribution (ALD). This distribution is parametrized by μ, σ, and τ. The random variable ϵ follows the asymmetric Laplace distribution as its probability density is:

$$\begin{aligned} &f(\epsilon\,|\,\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma}\exp\left\{-\rho_\tau\left(\frac{\epsilon-\mu}{\sigma}\right)\right\}, \\ &\text{with} -\infty < \mu < \infty,\ 0 < \tau < 1, \text{and},\ \sigma > 0, \end{aligned} \tag{1.3}$$

where ρ_τ is the objective function defined in Equation (1.2). The parameter μ determines the location, τ controls the skewness, and σ is the variance. In general, σ can be considered as a nuisance parameter and the skewness parameter τ corresponds to the desired quantile. Therefore, maximizing Equation (1.3) with respect to μ is equivalent to solving the minimization problem in Equation (1.2). Yu and Moyeed (2001) argue that the resulting prosterior is valid even if it is a misspecification of the true error and Sriram et al. (2013) provide a theoretical justification for posterior consistency under the ALD misspecification. The location parameter changes to $\mu_i = x_i\,\beta(\tau)$ and, for a fixed skewness parameter τ, the likelihood function – up to a

proportional constant (see Luo et al., 2012) – results in

$$L(\beta(\tau), \sigma \,|\, y, \tau) \propto \sigma^{-1} \exp\left\{ -\sum_{i=1}^{n} \rho_\tau \left(\frac{\epsilon - \mu_i}{\sigma} \right) \right\}. \tag{1.4}$$

Geraci and Bottai (2006) extended this approach to include a mixed effects model by including a random effect. In this setting, we implement a time-specific random effect F to account for clustering in the time line.[11] According to Geraci and Bottai (2006), the regression function of Equation (1.1) (and, thus, the location parameter) changes to

$$y_i = x_i \beta(\tau) + F(\tau) + \epsilon_i(\tau), \tag{1.5}$$

where $\epsilon(\tau) \sim AL(0, \sigma_\epsilon)$ and $F(\tau) \sim N(0, \sigma_F(\tau))$. The realization of $F(\tau)$ corresponds to the quarter of default, e.g., 2008 Q3, of the obligor. Therefore, obligors which default in the same quarter are exposed to the same τ-dependent realization of the random effect. The model in Equation (1.5) can be seen as a mixed effect model, where we treat impact of the covariates $\beta(\tau)$ as fixed and the impact of the time variation $F(\tau)$ as (additional) random intercept. Following Section 1.2, time patterns of AUFs vary among quantiles. Hence, it may be favorable to assign each quantile an individual impact of the random effect. Equation (1.2), (1.3), and (1.4) apply to the model with random effects by analogy. [12]

The models (with and without random effects) are estimated via Bayesian inference as the likelihood in Equation (1.4) cannot be maximized analytically. The posterior distribution is generated via Markov Chain Monte Carlo (MCMC) procedure. By constructing reversible Markov chains, the algorithm samples from the posterior distribution which corresponds to the

[11] Alternatively, one could use time-specific dummies to control for the remaining time variation. However, this might have at least two drawbacks. First, we want to use our model for predicting future conversion factors. Therefore, predicting an appropriate value for a future time-dummy is not straightforward. Second, with respect to the downturn estimates, the random effects structure gives financial institutions as well as prudential regulators a great flexibility to apply their margin of conservatism individually.

[12] Alternatively, we could have used finite mixture models as in Calabrese (2014), Altman and Kalotay (2014), Kalotay and Altman (2017), Betz et al. (2018) or Betz et al. (2021) for Losses Given Default (LGDs). These models assume a latent variable which describes the affiliation to individual components of the mixture model and use observable and unobservable covariates to model this latent variable. Some of these studies include a time-specific random intercept, as we did, and evaluate the impact of this time variation on the latent variable. However, the ordered logit or probit does not allow a direct link between changes in the latent variable and the resulting affiliation probabilities to the mixture components. An increase of the latent variable results in a higher probability of the highest component and a lower probability for the lowest component. However, the impact on intermediate components can not be inferred directly. Therefore, we think that the interpretation in terms of quantiles and the impact of the random effect on each quantile allows for a more direct interpretation. Moreover, one can think of fitting an unconditional mixture model on the conversionfactor's distribution, following Tomarchio and Punzo (2019) for LGD estimation. As we observe different shapes of the conversion factor's distribution for different facility types or industries in our sample, we would have to redo the inference for many subsets of our data.

target distribution in the equilibrium. More details on the estimation and the specified prior distributions for every parameter in the model can be found in Appendix 1.A.

Alternatively, frequentistic approaches could be used following, e.g., Geraci and Bottai (2007), Chernozhukov et al. (2013), Galvao et al. (2013), Galvao and Kato (2017), Graham et al. (2018) or Galvao and Poirier (2019). However, the Bayesian framework has some favorable properties. Following the statements by Yu et al. (2005); Yue and Rue (2011) and Bernardi et al. (2015) the Bayesian quantile regression provides estimations and predictions which take into account parameter uncertainty. This is especially interesting if the sample size is not extensively large. Furthermore, inferring distributions instead of point estimates of the parameters contributes to a more comprehensive understanding, see, e.g., Bernardi et al. (2015), and the interpretation of credibility intervals, e.g., Highest Posterior Density Intervals (HPDIs), is quite intuitive. Additionally, the convergence and stability for extreme quantiles can easily be assessed using the standard tools of Bayesian inference. With rising computational power, the estimation of Bayesian models is fairly efficient using standard software. Moreover, recent literature suggests that Bayesian quantile regressions are especially suitable for tail risk estimations, see, e.g., Carriero et al. (2020); Clements et al. (2020) and Ferrara et al. (2021). Summarizing, we think that a Bayesian mixed effect quantile regression is a reasonable choice for modeling the challenging distribution of the AUF.

As we use a default database, there might be a concern regarding endogeneity in particular due to sample selection. Meaning, that our target variable is only observed after default and is censored otherwise. This could imply that the sample is not representative for the population. However, the endogeneity problem arises only if there is a dependence between the censoring event (i.e., the default) and the resulting AUF. This problem may be alleviated by including the time-to-default into the modeling framework. However, this metric is not known before default and, thus, it is difficult to estimate. An alternative solution might be the joint modeling of AUF and the probability of default and account for their dependencies via copulae, see, e.g. Krüger et al. (2018). More specifically regarding the methods employed in this article, Arellano and Bonhomme (2017) propose a correction method for (frequentistic) quantile regressions in the case of sample selection by ”rotating” the check function by an amount that depends on the strength of selection. However, one has to quantify the strength of selection a priori. There is some evidence for sample selection regarding LGD, see, e.g. Rösch and Scheule (2014) or Krüger et al. (2018). To the best of our knowledge, there is no study which focuses on the dependence between probability of default and conversion factors and, thus, it is difficult to determine the

potential impact of endogeneity in our empirical application. However, the question of sample selection in conversion factor models is certainly a interesting path of future research.[13]

We further include the ordinary-least-squares (OLS) regression as a benchmark for our novel approach. This model focuses on the conditional mean of the distribution by neglecting varying impacts through the bimodal distribution. However, it is the most common method in literature, see e.g. Barakova and Parthasarathy (2013); Jacobs Jr (2010); Jacobs Jr and Bag (2011); Qi (2009) and Zhao et al. (2014). We estimate this regression in a Bayesian framework using uninformed priors such that the posterior means coincide with the point estimates in the Frequentistic framework.

1.4 Empirical Results

In this section, we present the empirical results based on a subsample from 2006 to mid 2016. The remaining observations are used in an out-of-time validation at the end of this section. We start with the quantile regression without random effects – labeled as *Macro Only Model* (see Equation (1.1) and Section 1.4.1) – to investigate the impact of the independent variables on the AUF distribution in the US and Europe. Afterwards, we look deeper in crisis periods and evaluate the model's ability to provide an AUF downturn distribution comparable to the one observed in the GFC. As the Macro Only Model only provides a sufficiently conservative downturn distribution in the US, we include a time-specific random effect in the quantile regression for Europe. This model is labeled as *Random Effects Model* (see Equation (1.5) and Section 1.4.2). It provides sufficiently conservative downturn distributions for Europe.

To interpret the models in Bayesian terms, we follow two coherent concepts. The first is based on posterior odds which are used to quantify the statistical evidence of the posterior means' estimated signs. Posterior odds coincide with the Bayes factor if the prior odds are equal to one. This is true for any symmetric prior distribution with a mean of zero. Since we assume a normal distribution with a mean of zero as prior for each parameter in the β vector (see Appendix 1.A), the posterior odds are equal to the Bayes factor. Posterior odds are defined as the ratio of the probability mass favoring the sign of the posterior mean and the probability

[13] We would like to thank an anonymous associate editor for suggesting this discussion.

mass of the opposite sign:

$$\text{Posterior odds}_{\mathrm{E}[\beta_i]<0} = \frac{\mathbb{P}(\beta_i < 0 | \text{data})}{\mathbb{P}(\beta_i \geq 0 | \text{data})}$$
$$\text{Posterior odds}_{\mathrm{E}[\beta_i]>0} = \frac{\mathbb{P}(\beta_i > 0 | \text{data})}{\mathbb{P}(\beta_i \leq 0 | \text{data})}$$

Therefore, we can directly quantify the evidence favoring the sign of the posterior means, e.g., posterior odds of 10 indicate that it is ten times more likely that the sign of the posterior mean is true compared to the opposite sign. Based on Kass and Raftery (1995), posterior odds greater than 3.2 indicate substantial evidence, values exceeding 10 correspond to strong evidence and posterior odds larger than 100 to decisive evidence.

The second concept to evaluate the evidence of posterior means are Highest Posterior Density Intervals (HPDI). These intervals quantify a range of the posterior distribution in which the unobservable parameter is located with a given probability, e.g., 95%. If zero is not included in the HPDI, statistical evidence for the sign of the posterior mean is assigned. For all model parameters, we assume non-informative priors as we do not impose a direction of impact. Nevertheless, due to the two coherent concepts, we are able to learn about the relation of covariates and AUF in a consecutive step.

1.4.1 Macro Only Model

In this subsection, results of the Macro Only Model and OLS with all variables described in Table 1.1 plus an interaction between ΔGDP and Utilization, i.e., $\Delta\text{GDP} \cdot \text{Utilization}$, are presented. This interaction gives us insights, whether the impact of the macroeconomy depends on the level of Utilization. This could have important implications for risk management practice in general and for credit line exposure at default in particular. We choose for each categorical variable a reference category, which is indicated in brackets in the first column of Table 1.2. This table compares the posterior means of the parameter estimates for the 5 %, 50 %, 95 % quantile and the OLS regression in the US and Europe. Appendix D shows some conversion diagnostics of the estimated models.[14]

[14] The estimation of quantile regressions can be challenging in the tails of the distribution due to a very low number of observations, as for example outlined by Chernozhukov (2005). This is frequently the case if we think about distributions like normal, logit or Cauchy. However, considering the distribution of the conversion factors we can detect differences to the aforementioned distributions. We observer considerable more realizations in the tails of the distribution compared to the middle as both modes are at 0 and 1. Therefore, in our application, the tails of the distribution are well observed. Similar observations can be found in Krüger and Rösch (2017) and Kellner et al. (2022), who found no instability problems concerning LGD as target variable.Furthermore, we check for every estimated quantile regression the common convergence checks which were all satisfied as outlined in our

For interpretation please note that the AUF distribution is negative for quantiles lower than the median and positive for quantiles greater than the median. Therefore, a negative posterior mean indicates a higher amount of exposure reduction for the left part of the distribution and a lower additional drawdown in the right part of the distribution. As there is a direct link between AUF and EAD in terms of lower or higher values, we can interpret the posterior means interchangeably for EAD and AUF. An increase of AUF results in an increase of EAD and vice versa. In Table 1.2, the coefficients vary over the quantiles and (in many cases) change their signs. This underpins the assumption that credit lines which reduce exposure are differently impacted by the independent variables than credit lines with positive additional drawdowns. This observation cannot be accounted for in the OLS model, where impacts are related only to the conditional mean. Hence, conclusions regarding positive or negative impacts of covariates for all levels of AUF are not possible. The applied quantile regression approach is well suited to consider this quantile-varying influence. Furthermore, setting AUFs outside the tolerated range back to the limits, e.g. 0 or 1, which is common in the EAD literature, might distort the results gathered from these models. This can be seen by the different signs of coefficients for positive and negative additional drawdowns. Setting outliers back to the limits may also hamper the identification of significant drivers of credit conversion factors.

Table 1.2: Results | Macro Only Model & OLS

(a) USA

Variable	Level	$\tau = 0.05$	$\tau = 0.50$	$\tau = 0.95$	OLS
Intercept		$0.128^{\circ\circ\circ}$	$0.599^{\circ\circ\circ}$	$1.125^{\circ\circ\circ}$	$0.704^{\circ\circ\circ}$
Facility	Short term revolver	$-0.042^{\circ\circ\circ}$	$-0.010^{\circ\circ}$	$-0.019^{\circ\circ\circ}$	$-0.042^{\circ\circ\circ}$
Industry	Agriculture	$-0.120^{\circ\circ\circ}$	-0.008°	0.018°	-0.026
(FIRE)	Mining	$-0.115^{\circ\circ\circ}$	$-0.074^{\circ\circ\circ}$	$-0.024^{\circ\circ}$	$-0.087^{\circ\circ\circ}$
	Construction	$-0.075^{\circ\circ\circ}$	$-0.017^{\circ\circ\circ}$	$0.018^{\circ\circ}$	$-0.052^{\circ\circ\circ}$
	Manufacturing	$-0.072^{\circ\circ\circ}$	$-0.011^{\circ\circ}$	$0.115^{\circ\circ\circ}$	-0.023°
	Transportation	-0.008	0.001	$0.076^{\circ\circ\circ}$	-0.009
	Wholesale	$-0.088^{\circ\circ\circ}$	$-0.016^{\circ\circ\circ}$	$0.038^{\circ\circ\circ}$	$-0.046^{\circ\circ\circ}$
	Service	$-0.070^{\circ\circ\circ}$	$-0.010^{\circ\circ}$	$0.041^{\circ\circ\circ}$	-0.027°
	Other	$-0.070^{\circ\circ\circ}$	-0.008°	0.003	$-0.050^{\circ\circ\circ}$
Seniority	Super senior	$0.080^{\circ\circ\circ}$	$0.014^{\circ\circ\circ}$	$-0.098^{\circ\circ\circ}$	-0.003
(pari-passu)	Non senior	$-0.037^{\circ\circ\circ}$	0.008°	$-0.103^{\circ\circ\circ}$	-0.034°
	Unknown	$0.133^{\circ\circ\circ}$	$0.025^{\circ\circ\circ}$	$-0.125^{\circ\circ\circ}$	$-0.030^{\circ\circ}$
log(Limit)		$-0.016^{\circ\circ\circ}$	$-0.011^{\circ\circ\circ}$	$-0.007^{\circ\circ\circ}$	$-0.017^{\circ\circ\circ}$
Age		$-0.003^{\circ\circ}$	$-0.002^{\circ\circ\circ}$	$-0.004^{\circ\circ\circ}$	$-0.006^{\circ\circ\circ}$
ΔGDP		-0.129	$-2.922^{\circ\circ\circ}$	-0.207	$-1.100^{\circ\circ}$
Utilization		$-0.234^{\circ\circ\circ}$	$-0.480^{\circ\circ\circ}$	$-0.870^{\circ\circ\circ}$	$-0.475^{\circ\circ\circ}$
Interaction		0.179	$2.902^{\circ\circ\circ}$	-0.398°	1.076°

Appendices. Alternative approaches for extreme quantiles can be found in Alhamzawi (2016), Huang and Chen (2015), Tian et al. (2017) or Hu et al. (2021).

(b) Europe

Variable	Level	$\tau = 0.05$	$\tau = 0.50$	$\tau = 0.95$	OLS
Intercept		$0.132^{\circ\circ\circ}$	$0.815^{\circ\circ\circ}$	$1.099^{\circ\circ\circ}$	$0.731^{\circ\circ\circ}$
Facility	Short term revolver	0.017°	$0.015^{\circ\circ}$	$-0.013^{\circ\circ}$	0.027
(medium term)	Overdraft	$-0.029^{\circ\circ\circ}$	$0.012^{\circ\circ\circ}$	$0.220^{\circ\circ\circ}$	$0.045^{\circ\circ\circ}$
Industry	Agriculture	-0.013°	0.004	$0.117^{\circ\circ\circ}$	0.044°
(FIRE)	Mining	0.029°	0.007	$0.611^{\circ\circ\circ}$	0.110
	Construction	$-0.050^{\circ\circ\circ}$	-0.007°	$0.047^{\circ\circ\circ}$	-0.001
	Manufacturing	$-0.053^{\circ\circ\circ}$	$-0.019^{\circ\circ\circ}$	$0.056^{\circ\circ\circ}$	-0.014
	Transportation	$-0.065^{\circ\circ\circ}$	$-0.021^{\circ\circ\circ}$	$0.037^{\circ\circ}$	-0.001
	Wholesale	$-0.050^{\circ\circ\circ}$	$-0.020^{\circ\circ\circ}$	0.019°	$-0.027^{\circ\circ}$
	Service	$-0.043^{\circ\circ\circ}$	$-0.009^{\circ\circ}$	$0.121^{\circ\circ\circ}$	0.011
	Other	$-0.039^{\circ\circ\circ}$	$-0.027^{\circ\circ\circ}$	0.015°	$-0.054^{\circ\circ\circ}$
Seniority	Super senior	$-0.040^{\circ\circ\circ}$	0.001	$0.045^{\circ\circ\circ}$	-0.002
(pari-passu)	Non senior	$-0.045^{\circ\circ\circ}$	$0.060^{\circ\circ\circ}$	$0.461^{\circ\circ\circ}$	$0.143^{\circ\circ\circ}$
log(Limit)		$-0.013^{\circ\circ\circ}$	$-0.010^{\circ\circ\circ}$	$-0.037^{\circ\circ\circ}$	$-0.028^{\circ\circ\circ}$
Age		$-0.002^{\circ\circ\circ}$	0.000	$0.004^{\circ\circ\circ}$	0.000
ΔGDP		-0.114°	$-1.997^{\circ\circ\circ}$	-0.255°	$-0.869^{\circ\circ\circ}$
Utilization		$-0.269^{\circ\circ\circ}$	$-0.687^{\circ\circ\circ}$	$-0.295^{\circ\circ\circ}$	$-0.382^{\circ\circ\circ}$
Interaction		$0.393^{\circ\circ}$	$1.978^{\circ\circ\circ}$	$2.569^{\circ\circ\circ}$	$1.088^{\circ\circ\circ}$

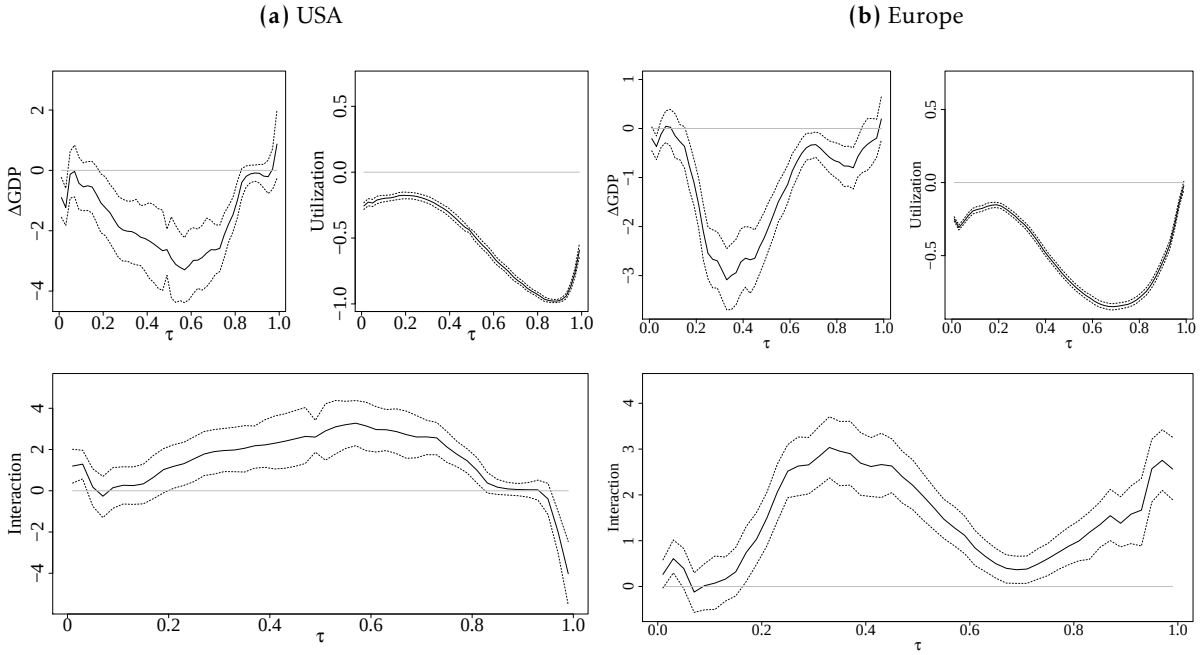
Note: This table shows the estimated posterior means for several selected quantiles. The first column inherits the name of the different independent variables. If they are categorical, the reference group is indicated in brackets. The second column illustrates the different levels of categorical variables. Statistical evidence is indicated by the following circles :° corresponds to substantial evidence (Odds >3.2), °° corresponds to strong evidence (Odds > 10),°°° corresponds to decisive evidence (Odds >100).

In the US, we find decisive evidence that short-term revolving lines have lower additional drawdowns and larger exposure reductions compared to medium term lines. These findings are valid in Europe for the positive part of the response distribution. Contrary, we find decisive evidence that another kind of short term lines – so called overdrafts – have higher additional drawdowns compared to medium term lines. To summarize, short term lines in the US have lower EADs, whereas in Europe it depends on the type of credit line. A possible explanation may be that overdrafts are less in the focus of monitoring processes as they are unconditionally revocable. With respect to the results, we may see that these lines, however, also expose credit risk to banks.

With respect to seniority, we find decisive evidence that non-senior credit lines draw less, respectively, reduce more exposure than pari-passu in the US. In Europe, we find decisive evidence that non-senior lines draw considerably more compared to pari-passu lines. The variable log(Limit) controls for the size of the credit line with respect to the limit one year prior to default. We find decisive evidence that larger lines reduce more or draw less additional exposure. This might be explained by the fact that banks monitor larger lines more tightly than smaller lines. The variable Age shows decisively evident negative signs for the quantiles

in the US. Thus, obligors with a short business relationship draw more, respectively, reduce less. Banks may not know theses obligors well and, hence, it is harder to foresee default and the drawdowns of the firm one year prior to default. In Europe, we find the same pattern for reductions, but the contrary sign for high additional draws. This might be explained by the fact that the overall business relationship is longer and, in some cases, longstanding obligors may be granted more financial leeway to draw their lines in the hope that default may be prevented.

Figure 1.3: Results | Macro Only Model (coefficient plots)



Note: The left three plots of the figure show the estimated coefficients for ΔGDP, Utilization and the interaction term over the whole distributional range in the US. The black lines represent the posterior means, whereas the dotted lines illustrate 95% HPDIs. The right three plots illustrate the estimated coefficients in Europe.

Figure 1.3 illustrates the impact of the variables ΔGDP, Utilization and their interaction term over the full response distribution, based on the Macro Only Model illustrated in Table 1.2. In Appendix C figures of all remaining independent variables are presented. We can clearly see that the posterior mean of all three variables varies considerably over the response distribution. The posterior mean (solid line) of ΔGDP is evidently negative for large parts of the distribution as the 95% HPDI (dotted line) does not include zero. The negative sign indicates an increase of the AUF in economic downturns, i.e., when ΔGDP is negative. This is in line with Figure 1.1 as quantiles of the AUF increase in the GFC. However, there is no statistically evident impact of the macroeconomic variable in the tails of the response distribution. This lack of evidence cannot be revealed by the OLS model, which underpins that our approach may be better suited to the non-linear impact of macroeconomic variables on the AUF and further reveals novel results to the literature of EAD modeling. This also suggests that the systematic of high additional drawdowns cannot be captured with the observable macrovariable and hence,

downturn estimates may be difficult to obtain.

Regarding Utilization, we find a throughout evidently negative impact on the AUF distribution indicating that the exposure reduction increases and, respectively, the additional drawndowns decrease with increasing Utilization. The latter effect may be explained by the fact that the potential of additional drawdowns is limited with higher utilization one year prior to default. Furthermore, credit lines with exposure reductions are heavily drawn one year prior default (see Figure 1.2).

We include an interaction term between ΔGDP and Utilization to control for a different impact of the macroeconomic environment with respect to the available limit. The interaction term has an evidently positive posterior mean in large parts of the response distribution. The total impact of the macroeconomic variable with respect to the level of Utilization is:

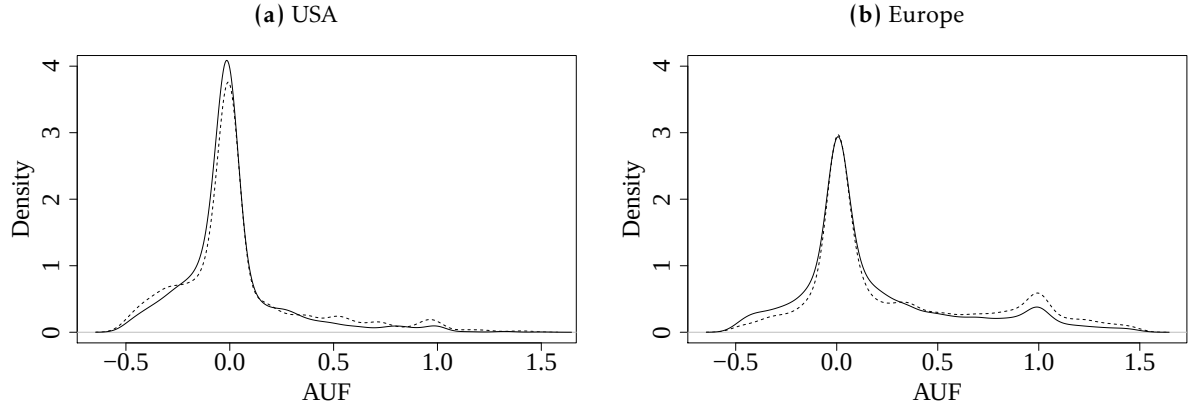
$$\text{total effect} = \beta^{(-)}_{\Delta\text{GDP}} + \beta^{(+)}_{\text{Interaction}} \cdot \text{Utilization}.$$

The overall negative impact of ΔGDP decreases with a higher Utilization as the interaction term is positive throughout the quantiles in both regions (see lower panel of Figure 1.3). For example, at the 50% quantile, the overall (negative) impact of the macroeconomic environment in Europe is reduced from -1.598 for 20% of utilization to -0.02 for 99% of utilization. Thus, the macroeconomic environment, especially in the inner quantiles, is more relevant for less drawn credit lines and less important for heavily drawn lines. This is plausible as less drawn lines have a higher drawdown potential which can be affected by economic downturns. Furthermore, the macroeconomic environment seems to be less important for credit lines with exposure reductions as they draw heavily one year prior default. This might have substantial consequences for credit risk management as crises affect those parts of the exposure distribution which bear higher risk – in terms of higher EADs.

Downturn estimation based on Macro Only Model

In this paragraph, we investigate the ability of the Macro Only Model to produce appropriate downturn distributions – comparable to the one observed in the GFC. Hereby, we assume an adverse realization of the macroeconomic variable ΔGDP to adopt an economic downturn. The adverse realization is set to - 5.5 % in Europe and -3.9 % in the US, corresponding to the 95% quantile of the observed growth rates in the sample period.

Figure 1.4: Distribution of AUF in the GFC



Note: The figure illustrates kernel density estimates of the AUF during the GFC (gray line) and the remaining periods in the sample (black line). With respect to the comparability of the density estimates, the same bandwidth was applied to both regions.

Figure 1.4 compares the density of the AUF during the GFC (crises distribution, dashed lines) and in the remaining time period (non-crises distribution, solid lines). According to the OECD[15], the GFC lasts from 2007 Q4 to 2009 Q2 in the US, whereas it is slightly shifted in Europe (2008 Q1 to 2009 Q3). In the US, the crises and non-crises distributions are very similar. This is in line with Figure 1.1 where only small variations of the AUF over time and slightly higher AUFs during the GFC arise. Contrary, there is less probability mass on exposure reduction (AUF < 0) and much more mass on higher additional drawdowns (AUF ≥ 1) in Europe, indicating a substantial impact of the GFC.

To evaluate the fit of the posterior predictive distribution and the empirical distribution, we use Probability-Probability (PP) plots following Michael (1983). Hereby, the empirical and theoretical quantiles are compared. The empirical quantiles $p_{\text{empirical},i}$ are generated via the posterior predictive distribution, whereas the theoretical quantiles $p_{\text{theoretical},i}$ are calculated from the data:

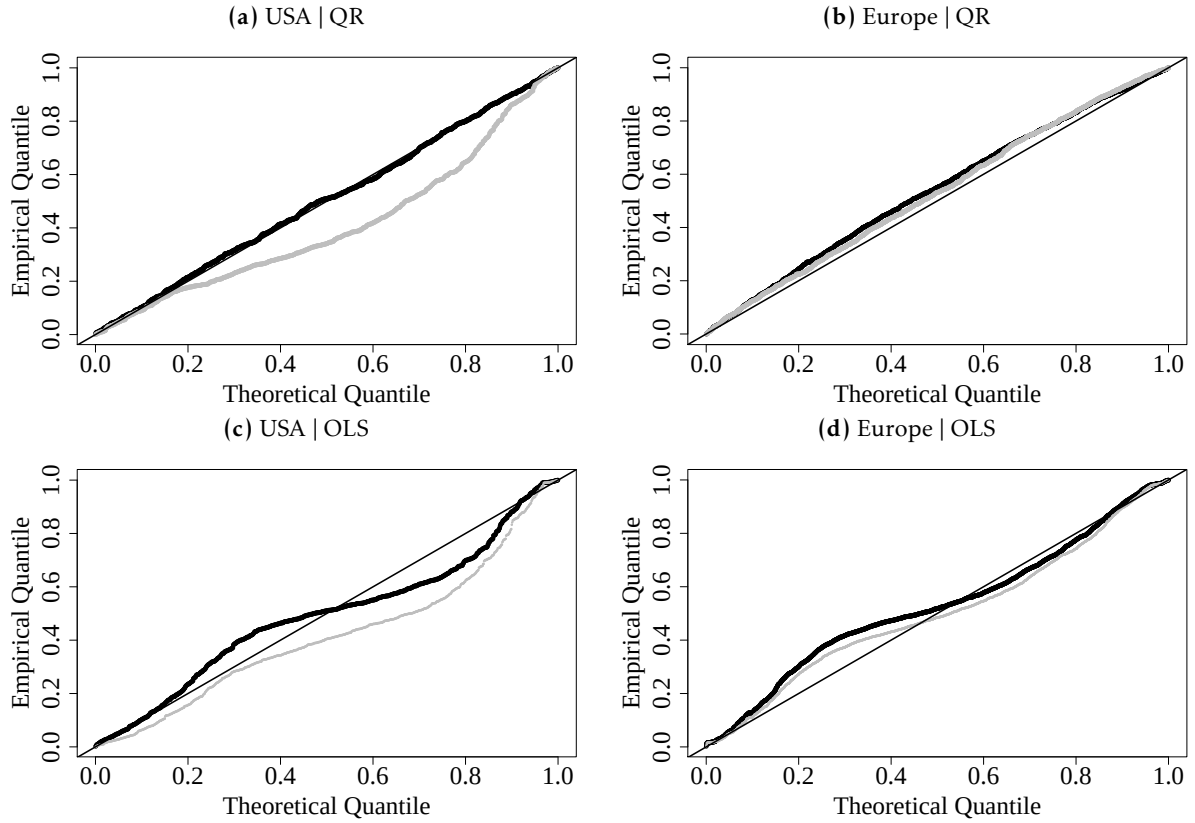
$$p_{\text{empirical},i} = \hat{F}(y_i), \qquad \text{and} \qquad p_{\text{theoretical},i} = \frac{i - 0.5}{n} \tag{1.6}$$

where the credit lines $i = 1, \ldots, n$ are ordered by y_i to ensure monotone increasing quantiles $\hat{F}(y_i)$. The compliance of all theoretical and empirical quantiles indicate perfect fit. Graphically, a perfect fit is obtained when the points in the PP plot lie on the bisecting line. If the points are above the bisecting line, the crisis distribution is underestimated, e.g. to little mass on high additional drawdowns, and vice versa. For the PP plot of the estimation sample, the points lie

[15] The recession indicators of the OECD are available at https://fred.stlouisfed.org/series/USARECDM for the US Area and available at https://fred.stlouisfed.org/series/EUROREC for the European Area.

on the bisection line perfectly, thus, in-sample perfect fit is achieved for the *Macro Only Model*. Contrary, the OLS shows considerable deviations.[16]

Figure 1.5: Distributional fit in downturn periods | Macro Only Model & OLS



Note: The figure shows the distributional fit in the Global Financial Crisis separated by regions. The black lines indicate the fit of the posterior predicted distribution, whereas the gray lines illustrate the fit using a stress scenario. The stress scenario is generated by considering an extreme value of the macro variable ΔGDP for each obligor defaulting during the crisis period. We used the 95% quantile of ΔGDP during the whole sample period. For the US, the extreme value corresponds to −3.9 % and to −5.5 % for Europe. An underestimation of the empirical crisis distribution is indicated by a PP-line above the bisecting line. Contrary, overestimation, i.e., a too conservative posterior predictive distribution, is indicated by a line below the bisecting line.

Figure 1.5 illustrates the distributional fit in a downturn period, i.e., the GFC, for the US (left panel) and Europe (right panel). The black points indicate the PP plot of the posterior predictive distribution. In the US, the Macro Only Model produces an almost perfect fit. This might be expected as the crises and non-crises distribution do not substantially differ (see Figure 1.4). However, the linear model deviates strongly from the bisecting line, showing a rather poor distributional fit. In Europe, the empirical distribution is underestimated in the GFC as the points are above the bisecting line. Hence, the posterior predictive distribution is not sufficiently conservative. Again, the OLS provides a considerably lower fit.

To generate a stressed posterior predictive distribution, an extreme realization of ΔGDP is applied. We use the 95 % quantile of ΔGDP which corresponds to −3.9 % in the US and −5.5 %

[16] The corresponding figures for the estimation sample are available from the authors upon request.

in Europe. According to the negative posterior mean of ΔGDP, a negative realization results in a higher AUF. In Figure 1.5, the gray dots correspond to the stressed predictive distribution. The stressed predictive distribution is too conservative in the US which might have been expected as the posterior predictive distribution already delivers a perfect fit. Contrary, the stressed predictive distribution is still not conservative enough in Europe. This might be due to two reasons. First, ΔGDP does not have an evident impact on the tails of the distribution. Second, there are more credit lines with positive AUF and high utilization in Europe as shown in Figure 1.2. As we have seen, the negative impact of the macroeconomic environment is reduced with higher utilization, and hence the ability to stress the distribution via macroeconomic variables is limited.

To summarize, the Macro Only Model provides a good distributional fit in crises and non-crises periods in the US, whereas the OLS does not. On the contrary, the macroeconomic variable does not seem to be able to capture the true systematic pattern in Europe. Therefore, we include a time-specific random effect in our quantile regression approach in the next step.

1.4.2 Random Effects Model

The model set-up for the Random Effects Model is similar to the Macro Only Model as the observable variables remain in the modeling framework. We extend the model by a time-specific random effect as stated in Equation (1.5). The realizations of the random effect refer to the quarter of default t. Obligors who default in the same quarter t, share the same realization of the random effect and, thus, their AUFs are either higher (positive realization of the random effect) or lower (negative realization of the random effect) on average.

This enables us to capture the co-movement in the time dimension. As the coefficients of the independent variables are very similar to the ones obtained by the Macro Only Model, we focus only on the extension of this model. The coefficients for selected quantiles can be found in Table B.1 in Appendix 1.B.

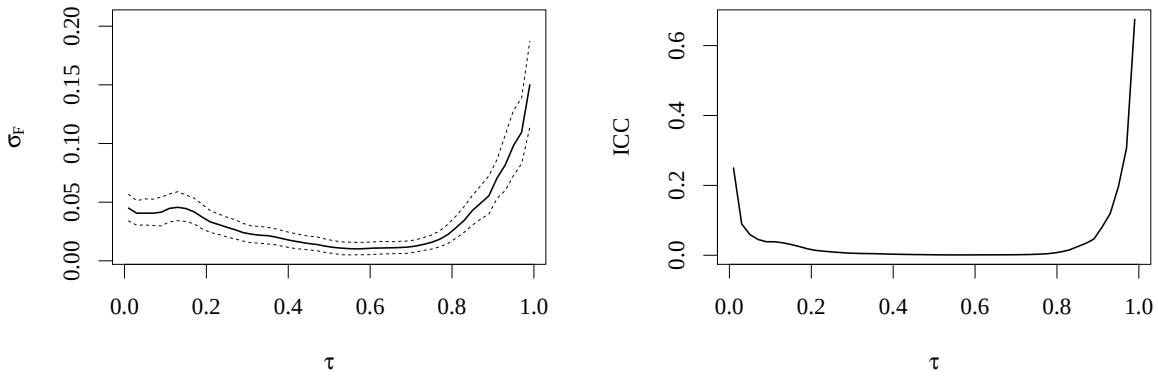
The main parameter of the random effect and, thus, the Random Effects Model, is the standard deviation σ_F. It can be interpreted in terms of magnitude of the random effect's impact. The higher the standard deviation, the larger the impact of the random effect on the specific quantile. As an additional measure we use the Inter Cohort Correlation (ICC) coefficient. It illustrates the proportion of variation in the quantile captured by the random effect.

According to Geraci and Bottai (2006), the ICC is defined as:

$$\text{ICC} = \frac{\sigma_F^2}{\sigma_F^2 + \sigma_\epsilon^2}, \tag{1.7}$$

where σ_F^2 is the variance of the random effect and σ_ϵ^2 is the variance of the error term in the quantile function (see Equation (1.5)). The higher the ICC, the more the random effect accounts for the variation in the quantile.

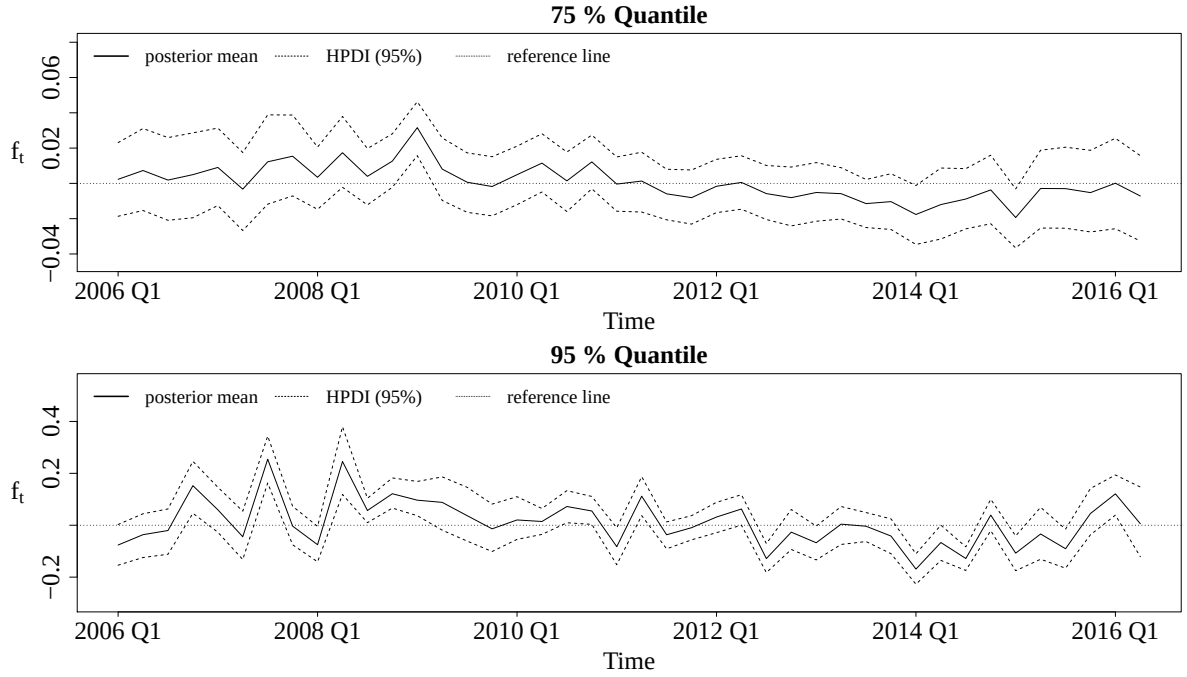
Figure 1.6: Results | Random Effects Model (coefficients plots of σ_F and ICC)



Note: The left panel of the figure illustrates the estimated posterior mean of σ_F in the Random Effects Model. The dashed lines indicate the 95 % HPDIs. The standard deviation σ_F can be interpreted as the impact strength of the random effect in the corresponding quantile. The right part of the figure displays the posterior mean of the ICC coefficient (see Equation (1.7)). It indicates how much of the variation in each quantile is due to the random effect compared to the fixed effects.

Figure 1.6 illustrates the standard deviation σ_F of the random effect (left panel) and the ICC coefficient (right panel) for each quantile. The random effect has the highest impact in the tails of the distribution. This coincides with the lack of statistical evidence for the macroeconomic variable in this range (see right panels of Figure 1.3). From a credit risk management perspective, it is noteworthy that the impact of the random effect is stronger in the right tail of the distribution. Thus, unobservable systematic patterns are crucial for extreme positive additional drawdowns. According to the ICC, the random effect accounts for more than 60 % of the variation in the far right tail. This has two major implications. First, modelling a quantile-dependent random effect is favourable as the impact differs along the response distribution. Second, the random effect accounts for the true systematic variation in a value range where macroeconomic variables lack statistical evidence.

Figure 1.7: Results | Random Effects Model (random effect realizations)



Note: The figure illustrates the posterior means (solid gray line) of the random effect realizations for the 75 % and 95 % quantile. The dashed lines correspond to the 95 % HPDIs. A positive posterior mean indicates a positive effect on the corresponding quantile function and, therefore, a higher AUF.

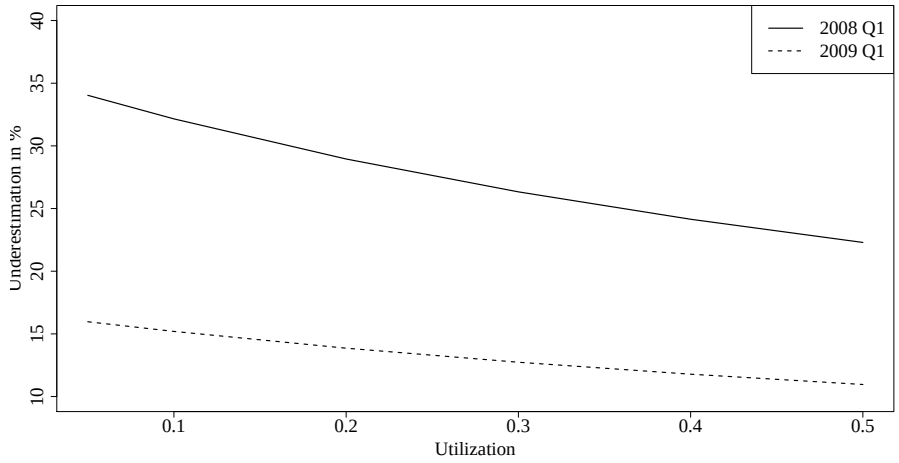
Figure 1.7 illustrates the posterior means (black solid line) and the HPDIs (black dashed line) of the random effect realizations for the 75 % and 95 % quantile. The dotted line marks the reference point of zero. As indicated by Figure 1.6, the magnitudes of the realizations substantially differ among the quantiles. Regarding the 95% quantile, the posterior means are up to ten times as high compared to the 75 % quantile. In the GFC, large positive realizations indicating higher AUFs occur. So the question arises why the random effect accounts for systematic variation, especially in the early stages of the financial crisis and for higher quantiles? One reason may be that credit lines in general are among the first financial instruments that companies use to sustain their liquidity and financing duties when the economic condition deteriorates. This is in line with findings of Barakova and Parthasarathy (2013) who find that EAD of syndicated credit lines is especially high in pre- and early stages of crisis periods, where defaults are hard to anticipate for banks. Hence, finding an observable variable for very early stages of crisis periods may be tedious and largely portfolio-dependent. The random effects approach provides a straightforward and tailor-made solution to this problem. Banks and regulators may use a *baseline* macroeconomic variable, like ΔGDP, to account for the overall economy and use the random effect to capture the remaining systematic variation of credit lines, as suitable variables are hard to find.

To underline the importance of the random effect, assume a short term revolver, located in the FIRE industry, pari-passu in seniority, one year history of credit line and an available limit of 250,000. To forecast an adverse realization of the EAD, a bank may use the posterior means, displayed in Table B.1, of the Random Effects Model for the 95% quantile:

$$\begin{aligned} Q_{95th}(y_i|x_i) = 1.094 - 0.015 - 0.037 \cdot 250,000 + 1 \cdot 0.0004 - \Delta\text{GDP} \cdot 0.319 \\ - 0.287 \cdot \text{Utilization} + 2.221 \cdot \Delta\text{GDP} \cdot \text{Utilization} \end{aligned} \tag{1.8}$$

We can calculate the AUF based on observable variables in Equation (1.8) and subsequently estimate the EAD. To calculate the EAD with the random effect, its realizations can simply be added to the AUF based on Equation (1.8). For covering downturn characteristics, we use the realization in 2008 Q1 of 0.22 and 2009 Q1 of 0.10 with the corresponding values of ΔGDP. To assess the importance of the random effect, the relative difference[17] between the EAD estimate with random effect and the EAD estimate based on Equation (1.8), depending on the level of Utilization is shown in Figure 1.8:

Figure 1.8: Results | Impact of the Random Effect



Note: The figure illustrates the relative difference of EAD estimates with and without considering the random effect. The black solid line represents the realization of 2008 Q1, whereas the dashed line illustrates the realization of 2009 Q1.

We can obtain two important insights from this stylized example. First, the comparison of the two lines indicates that the realization of the random effects has a large impact on the EAD estimates, underlining the importance of this approach. The estimated EAD with the realization of the random effect is up to 35% higher than when neglecting the realization. Furthermore, we can see that the random effect, again, is most important for less drawn lines, which entail

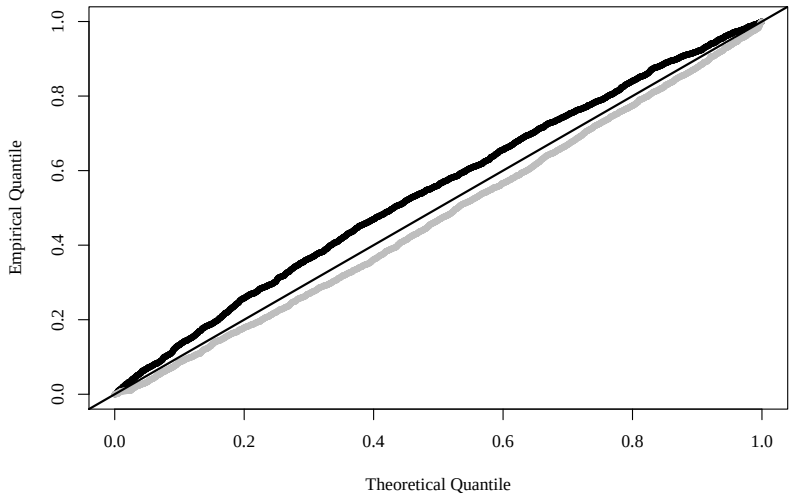
[17] The relative difference is calculated by $\left(\frac{EAD_{\text{with random effect}}}{EAD_{\text{without random effect}}} - 1\right)$. Hence, a value greater than zero indicates a larger EAD estimate by using the realization of the random effect.

the greatest risk to banks. This clearly shows that the random effect accounts for a large and important share of systemic variation of credit lines, especially for higher quantiles of the AUF distribution.

Downturn estimation based on Random Effects Model

In analogy to Section 1.4.1, we investigate the model's ability to produce sufficiently conservative downturn distributions. In Europe, the Macro Only Model underestimates the empirical AUF distribution – even if the macroeconomic variable is stressed to its 95 % quantile. This might be due to its lack of statistical evidence in the tails of the AUF distribution. The downturn AUF distribution in the Random Effects Model is generated by applying an adverse realization of the random effect. As an adverse realization, we use the 95 % quantile of each quantile-specific normal distribution with mean zero and standard deviation $\sigma_F(\tau)$. The posterior predictive distribution is generated by setting the random effect to its mean.

Figure 1.9: Distributional fit in downturn periods | Random Effects Model



Note: The figure shows the distributional fit during the GFC for the Random Effects Model. The black line indicates the fit of the posterior predictive distribution, whereas the gray line illustrates the fit using a stress scenario. The stress scenario is generated by considering an extreme realization of the random effect for each obligor defaulting during the GFC. Recall that the quantile-specific random effect follows a normal distribution with mean zero and standard deviation σ_F. The 95 % quantile of each quantile-specific random effect distribution is applied as extreme realization. An underestimation of the empirical crisis distribution is indicated by a PP-line above the bisecting line. Contrary, a too conservative posterior predictive distribution is indicated by a line below the bisecting line.

Figure 1.9 illustrates the PP plots of the posterior predictive distribution and downturn distribution based on the Random Effects Model in the GFC. The interpretation coincides to the one in Figure 1.5. The black points indicate the distributional fit of the posterior predictive distribution, whereas the gray dots illustrate the fit of the downturn distribution. The posterior predictive distribution underestimates the empirical AUF distribution as the black dots are

above the bisecting line. However, the downturn distribution via the random effect delivers a sufficiently conservative distribution. Summarizing, the random effect accounts for systematic variation in the tails of the distribution where macroeconomic variables lack impact and statistical evidence. Therefore, sufficiently conservative downturn distributions can be generated based on the random effect in Europe.

Out-of-time comparison[18]

The final part of this section focuses on the out-of-time performance of quantile regression and the benchmark model. In credit risk, we are usually interested in predicting the future. Hence, a model should be capable of predicting the EAD in unseen time periods. We use the hold-out sample ranging from mid 2016 to the end of 2018 to conduct this out-of-time validation. To provide a more broad picture, we sample 1,000 portfolios including 200 credit lines each of the hold-out sample instead of comparing both methods only once. As the comparison of all PP plots is tedious, we summarize them using the Harmonic Mass Index (HMI). This measure averages the absolute deviations of empirical and theoretical quantiles which are plotted in the PP plot (Wagenvoort, 2006). Formally, it is defined as:

$$HMI = \frac{2}{n} \sum_{i=1}^{n} |p_{\text{empirical},i} - p_{\text{theoretical},i}| \tag{1.9}$$

The lower the calculated HMI, the better the distributional fit. A perfect fit results in an HMI of zero. Table 1.3 reports mean and standard deviation over the 1,000 samples:

Table 1.3: Harmonic Mass Index

(a) USA

	Quantile Regression	**OLS**
Mean	0.0458	0.0823
Standard deviation	0.0080	0.0067

(b) Europe

	Quantile Regression	**OLS**
Mean	0.1216	0.1616
Standard deviation	0.0170	0.0130

Note: The table shows means, standard deviations of the HMI over the 1,000 sampled portfolios in each region. The HMI summarizes the absolute deviations from the perfect fit. Hence, the lower the value, the better the distributional fit. For the European Data set, the Random Effects Model is used, as it turned out to be superior. The random effects in the Random Effects Model are set to their expectation for prediction. The Macro Only Model is used in the US American data set.

[18] We thank discussants of the CFE 2019 for suggesting this comparison.

Regarding Table 1.3, the quantile regression performs much better over all samples and in both regions. In the US American sample, the HMI is almost cut by half and in Europe it decreased by 24.75%. The standard deviations across the 1,000 portfolios in each region are similar. To underline the superiority of the quantile regression in each and every portfolio we would like to stress the point that there is not a single portfolio in which our approach provides a worse fit than the linear model.

1.5 Conclusion

By using access to one of the world's largest loss and exposure data bases, this paper sheds light onto the topic of modeling EADs and conversion factors and, thus, the drawdown behavior of eventually defaulted credit lines. We apply Bayesian quantile regressions to model the full conditional distribution of conversion factors. If the identification of adequate (i.e., meaningful and statistically evident) macroeconomic variables is unfeasible, the quantile regression approach is extended by time-specific random effects to capture the unexplained systematic time patterns of conversion factors.

Quantile regression turns out to be a superior modeling technique in this setting as deviating effects among quantiles are captured. The most striking deviations throughout the quantile range refer to the impact of macroeconomic variables. We find statistically evident impacts on the inner quantiles, while evidence vanishes in the outer tails of the distribution. This is of special relevance in the light of the requirement for downturn estimates, i.e., estimates which reflect economic downturn conditions. Furthermore, macroeconomic effects on conversion factors vary for different utilization levels. Less drawn lines (low utilization) are affected to a higher extent by economic downturns. This entails tangible consequences for credit risk managements as these lines bear the highest risk in terms of an EAD increase. Credit lines which are already exhausted one year prior to default react less to economic decline.

With respect to downturn estimation, we reveal major differences among the two considered regions – the US and Europe. In the US, macroeconomic variables seem to capture wide parts of the systematic co-movement of conversion factors in the time line. Thus, sufficiently conservative downturn estimates are able to be generated via these observable systematic variables. This might be due to the fact that comovements are generally less pronounced compared to Europe. In contrast to the US, macroeconomic variables do not seem to be suitable

to produce adequate downturn estimates in Europe. Hence, time-specific random effects are included into the modeling framework. These unobservable systematic effects are able to capture the true systematic patterns in conversion factors. Indeed, the impact of the random effect is largest regarding the tails of the distribution where the impact of the macroeconomic variables vanishes. As a consequence, sufficiently conservative downturn estimations can be generated based on random effects for Europe. Comparing our approach with the most common method in literature, the OLS regression, we can provide evidence of superior fit and greater flexibility. Especially in the out-of-time forecasting exercise, our model provides an up to twice as good distributional fit compared to the benchmark model.

The results of this paper have three major implications for financial institutions and politics. First, less drawn credit lines not only bear the highest risk in terms of an EAD increase, but are also more severely affected by economic downturn. Second, systematic patterns in conversion factors might be of different kind and magnitude depending on the considered region. Thus, random effects might offer a reasonable option to generate sufficiently conservative downturn estimates if the identification of adequate macroeconomic variables is challenging. Furthermore, we can show that credit lines also induce higher credit risk besides the well documented liquidity risk in crisis periods, which is important for politics and regulators.

Acknowledgements

The authors would like to thank Global Credit Data for granting access to their database and participants of the Global Credit Data European Conference 2019 in Vienna for fruitful discussions and helpful comments. Furthermore, we gratefully acknowledge many useful comments and discussions of participants at the International Conference on Computational and Financial Econometrics (CFE) 2019 in London and the participants at the Finance Research Letters Annual Event 2021 in Valencia. We also thank participants at the International Risk Management Conference (IRMC) 2021 in Cagliari.

1.A Bayesian model specification

The quantile regression and its extensions are estimated using Bayesian inference. Hence, for each parameter prior distributions have to be specified. Furthermore, to ensure a more efficient estimation, this paper uses the decomposition of the asymmetric Laplace Distribution based on Yu and Stander (2007) and Luo et al. (2012). A random variable of the asymmetric Laplace Distribution can be expressed as a mixture of a standard normal and an exponential random variable. Therefore, Equation (1.5) changes to:

$$y_i = x_i\beta(\tau) + F(\tau) + c_1 e_i + \sqrt{c_2 \sigma e_i} z_i, \tag{1.10}$$

where $c_1 = \frac{1-2\tau}{\tau(1-\tau)}$, $c_2 = \frac{2}{\tau(1-\tau)}$, $z_i \sim N(0,1)$ and $e_i \sim \text{Exp}\left(\frac{1}{\sigma_\epsilon}\right)$.

The Bayesian quantile regression and its priors can be formulated as follows:

$$\begin{aligned} f(y_i|\beta(\tau), F(\tau), \sigma_\epsilon, e_i, z_i) &= (2\pi c_2 \sigma_\epsilon e_i)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\pi c_2 e_i}(y_i - x_i\beta(\tau) - F(\tau) - c_1 e_i)^2\right\} \\ F(\tau) &\sim N(0, \sigma_F(\tau)) \\ \sigma_F(\tau) &\sim N(0, 10^5)[0, \infty] \\ \beta(\tau) &\sim N(0, 10^5) \\ \sigma_\epsilon &\sim N(0, 10^5)[0, \infty] \\ z_i &\sim N(0,1) \\ e_i &\sim \text{Exp}\left(\frac{1}{\sigma_\epsilon}\right). \end{aligned} \tag{1.11}$$

The squared brackets in the model specifications of the dispersion parameters indicate truncation. The prior specifications of model parameters are set to be uninformative assuming large values of their dispersion parameters. The random effect follows a Normal distribution with mean zero and the random effect specific standard deviation $\sigma_F(\tau)$. In this hierarchical setting, we also specified a truncated Normal distribution for this dispersion parameter as the prior distribution. The models are sampled using two MCMC chains each. We use a chain length of 10,000 for the European sample and 20,000 for the US sample due to a smaller sample size. Furthermore, the burn-in length was set to 2,000 in Europe and 4,000 in the US.

1.B Random effects model

Table 1.B.1: Results | Macro Only Model (MOM) and Random Effects model (REM) for Europe

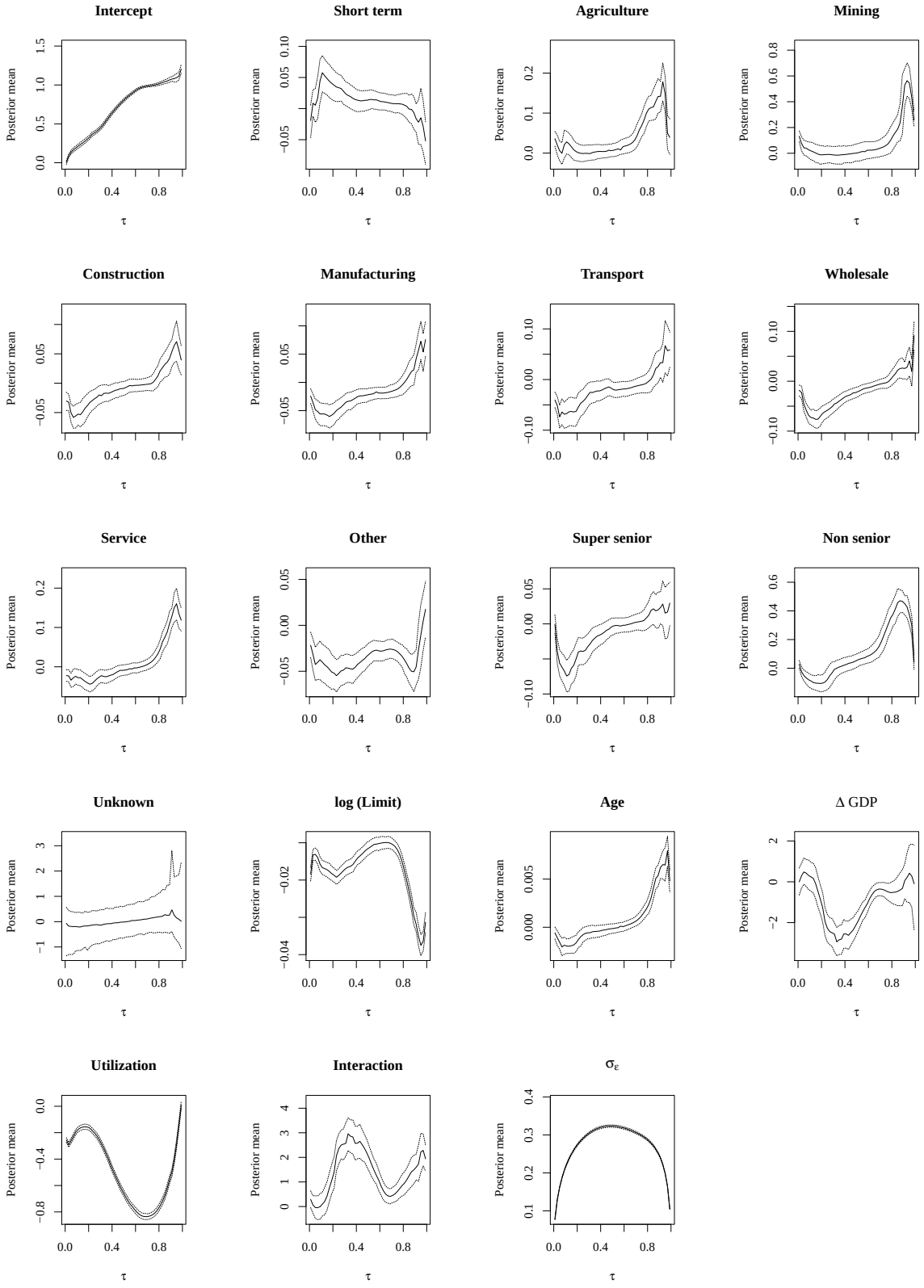
Variable	Level	$\tau = 0.05$		$\tau = 0.50$		$\tau = 0.95$	
		MOM	REM	MOM	REM	MOM	REM
Intercept		$0.132^{\circ\circ\circ}$	$0.136^{\circ\circ\circ}$	$0.815^{\circ\circ\circ}$	$0.818^{\circ\circ\circ}$	$1.099^{\circ\circ\circ}$	$1.094^{\circ\circ\circ}$
Facility Type	Short term	0.017°	0.005°	$0.015^{\circ\circ}$	$0.014^{\circ\circ}$	-0.013	-0.015
(medium term)	Overdraft	$-0.029^{\circ\circ\circ}$	$-0.035^{\circ\circ\circ}$	$0.012^{\circ\circ\circ}$	$0.008^{\circ\circ\circ}$	$0.220^{\circ\circ\circ}$	$0.189^{\circ\circ\circ}$
Industry	Agricult.	-0.013°	0.006	0.004	0.007	$0.117^{\circ\circ\circ}$	$0.148^{\circ\circ\circ}$
(FIRE)	Mining	0.029°	$0.043^{\circ\circ}$	0.007	0.004°	$0.611^{\circ\circ\circ}$	$0.543^{\circ\circ\circ}$
	Construct.	$-0.050^{\circ\circ\circ}$	$-0.051^{\circ\circ\circ}$	-0.007°	-0.008°	$0.047^{\circ\circ\circ}$	$0.071^{\circ\circ\circ}$
	Manufact.	$-0.053^{\circ\circ\circ}$	$-0.048^{\circ\circ\circ}$	$-0.019^{\circ\circ\circ}$	$-0.021^{\circ\circ\circ}$	$0.056^{\circ\circ\circ}$	$0.073^{\circ\circ\circ}$
	Transport	$-0.065^{\circ\circ\circ}$	$-0.074^{\circ\circ\circ}$	$-0.021^{\circ\circ\circ}$	$-0.020^{\circ\circ\circ}$	$0.037^{\circ\circ}$	$0.067^{\circ\circ\circ}$
	Wholesale	$-0.050^{\circ\circ\circ}$	$-0.046^{\circ\circ\circ}$	$-0.020^{\circ\circ\circ}$	$-0.021^{\circ\circ\circ}$	0.019°	$0.041^{\circ\circ\circ}$
	Service	$-0.043^{\circ\circ\circ}$	$-0.034^{\circ\circ\circ}$	$-0.009^{\circ\circ}$	-0.008°	$0.121^{\circ\circ\circ}$	$0.160^{\circ\circ\circ}$
	Other	$-0.039^{\circ\circ\circ}$	$-0.043^{\circ\circ\circ}$	$-0.027^{\circ\circ\circ}$	$-0.031^{\circ\circ\circ}$	0.015°	-0.010
Seniority	Super sen.	$-0.040^{\circ\circ\circ}$	$-0.056^{\circ\circ\circ}$	0.001	-0.003	$0.045^{\circ\circ\circ}$	$0.015^{\circ\circ\circ}$
(pari-passu)	Non sen.	$-0.045^{\circ\circ\circ}$	$-0.052^{\circ\circ\circ}$	$0.060^{\circ\circ\circ}$	$0.058^{\circ\circ\circ}$	$0.461^{\circ\circ\circ}$	$0.371^{\circ\circ\circ}$
log(Limit)		$-0.013^{\circ\circ\circ}$	$-0.013^{\circ\circ\circ}$	$-0.010^{\circ\circ\circ}$	$-0.011^{\circ\circ\circ}$	$-0.037^{\circ\circ\circ}$	$-0.037^{\circ\circ\circ}$
Age		$-0.002^{\circ\circ\circ}$	$-0.002^{\circ\circ\circ}$	0.000	0.000	$0.004^{\circ\circ\circ}$	$0.006^{\circ\circ\circ}$
ΔGDP		-0.114°	0.045	$-1.997^{\circ\circ\circ}$	$-1.952^{\circ\circ\circ}$	-0.255°	0.319
Utilization		$-0.269^{\circ\circ\circ}$	$-0.261^{\circ\circ\circ}$	$-0.687^{\circ\circ\circ}$	$-0.677^{\circ\circ\circ}$	$-0.295^{\circ\circ\circ}$	$-0.287^{\circ\circ\circ}$
Interaction		$0.393^{\circ\circ}$	0.483°	$1.978^{\circ\circ\circ}$	$1.960^{\circ\circ\circ}$	$2.569^{\circ\circ\circ}$	$2.221^{\circ\circ\circ}$
σ_F			$0.041^{\circ\circ\circ}$		$0.011^{\circ\circ\circ}$		$0.098^{\circ\circ\circ}$

Note: This table shows the estimated posterior means for several selected quantiles and compares the Macro Only with the Random Effects Model. As one can see, the estimated posterior means do not differ much. The first column inherits the name of the different independent variables. If they are categorical, the reference group is indicated in brackets. The second column illustrates the different levels of categorical variables. Statistical evidence is indicated by the following circles :° corresponds to sub- stantial evidence (Odds > 3.2), °° corresponds to strong evidence (Odds > 10),°°° corresponds to decisive evidence (Odds >100).

1.C Coefficient Plots

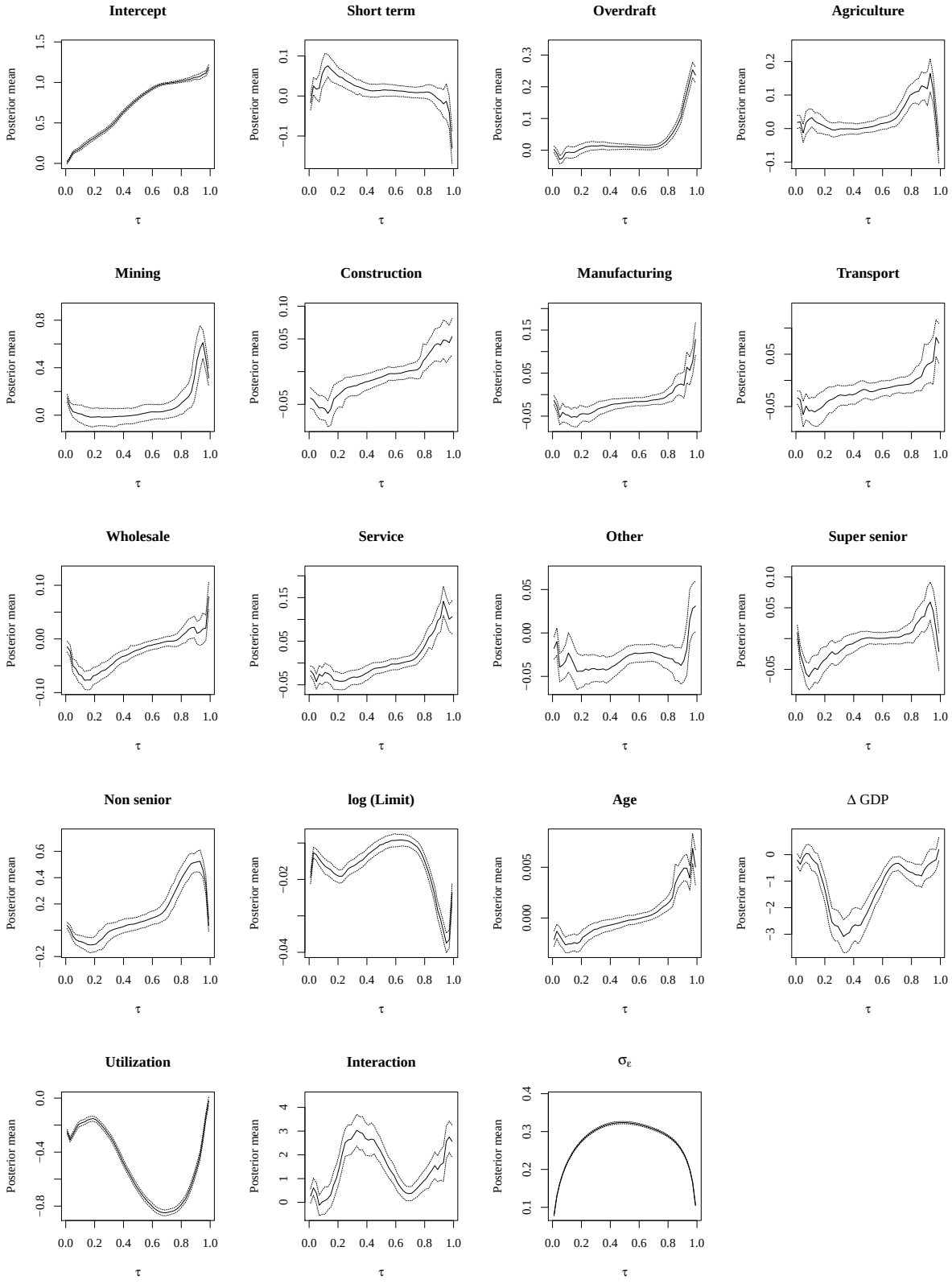
The following figures show the estimated posterior means and the 95% HPDI for each parameter in the three different quantile regressions. Statistical evidence is indicated if zero is not included in the 95% HPDI.

Figure 1.C.1: Coefficients USA | Macro Only Model



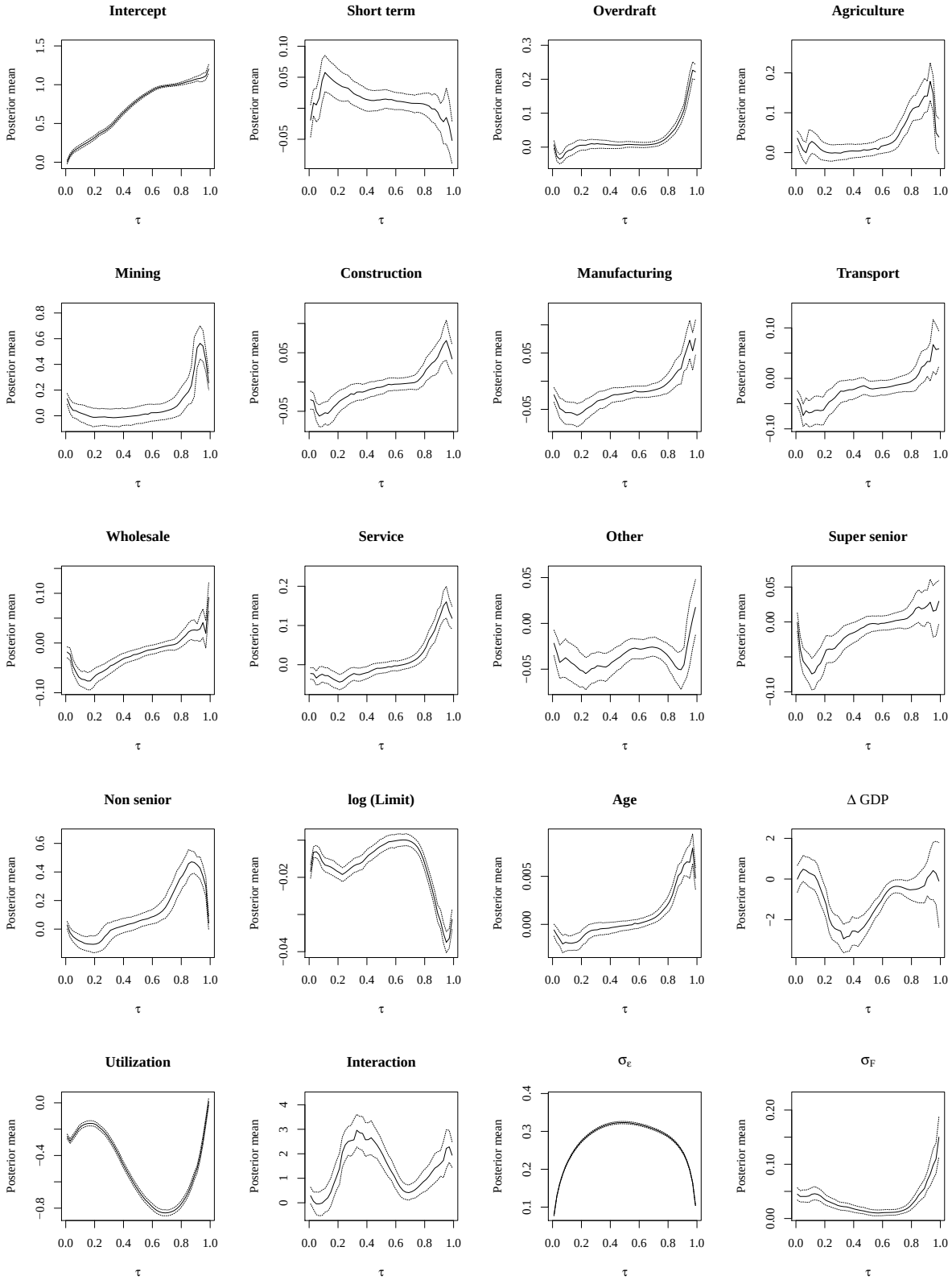
Note: The figure shows the estimated coefficients and their 95% HPDI for all parameters in the whole distributional range in the US. The black lines represent the posterior means, whereas the dotted lines illustrate 95% HPDIs.

Figure 1.C.2: Coefficients Europe | Macro Only Model



Note: The figure shows the estimated coefficients and their 95% HPDI for all parameters in the whole distributional range in the European sample. The black lines represent the posterior means, whereas the dotted lines illustrate 95% HPDIs.

Figure 1.C.3: Coefficients Europe | Random Effects Model



Note: The figure shows the estimated coefficients and their 95% HPDI for all parameters in the whole distributional range in the European sample. The black lines represent the posterior means, whereas the dotted lines illustrate 95% HPDIs.

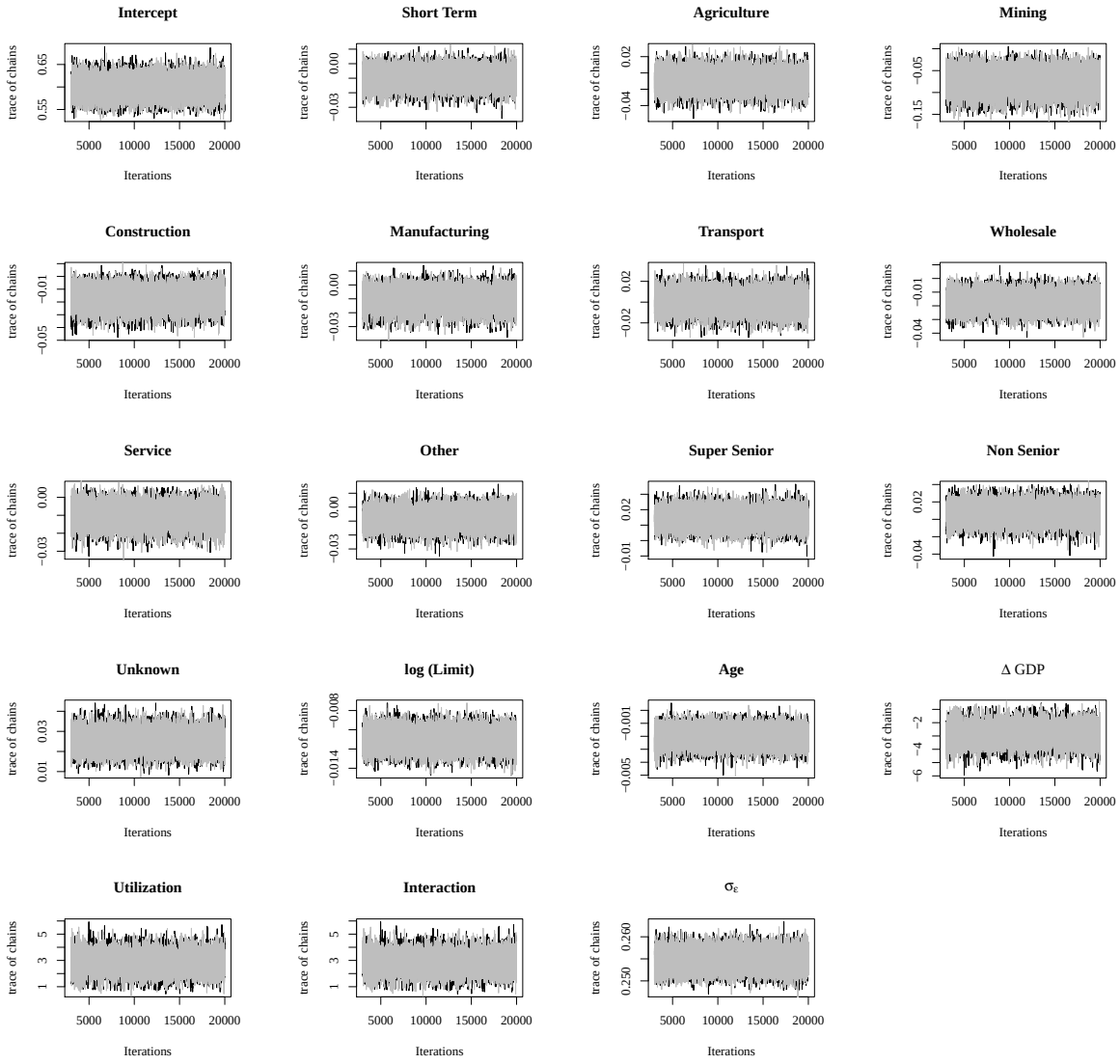
1.D Convergence Diagnostics

To evaluate the convergence of the estimated models, trace plots are the primary source of convergence diagnostics. Stable trace plots indicate that the chains converge to a steady state. Hence, priors are well calibrated and the burn-in is sufficient. Furthermore, we examine two well-known figures in Bayesian inference – the Gelman-Rubin and Heidelberger-Welch diagnostic. Both are hypotheses tests in frequentist terms, however, applied widely to evaluate the length of burn-in (Gelman-Rubin) and the length of chains (Heidelberger-Welch). Furthermore, we display the diagnostic only for the median ($\tau = 0.5$). Please note that for all quantiles convergence is achieved.[19]

[19] Traceplots, Gelman-Rubin and Heidelberger-Welch diagnostics for all quantiles are available from the authors upon request.

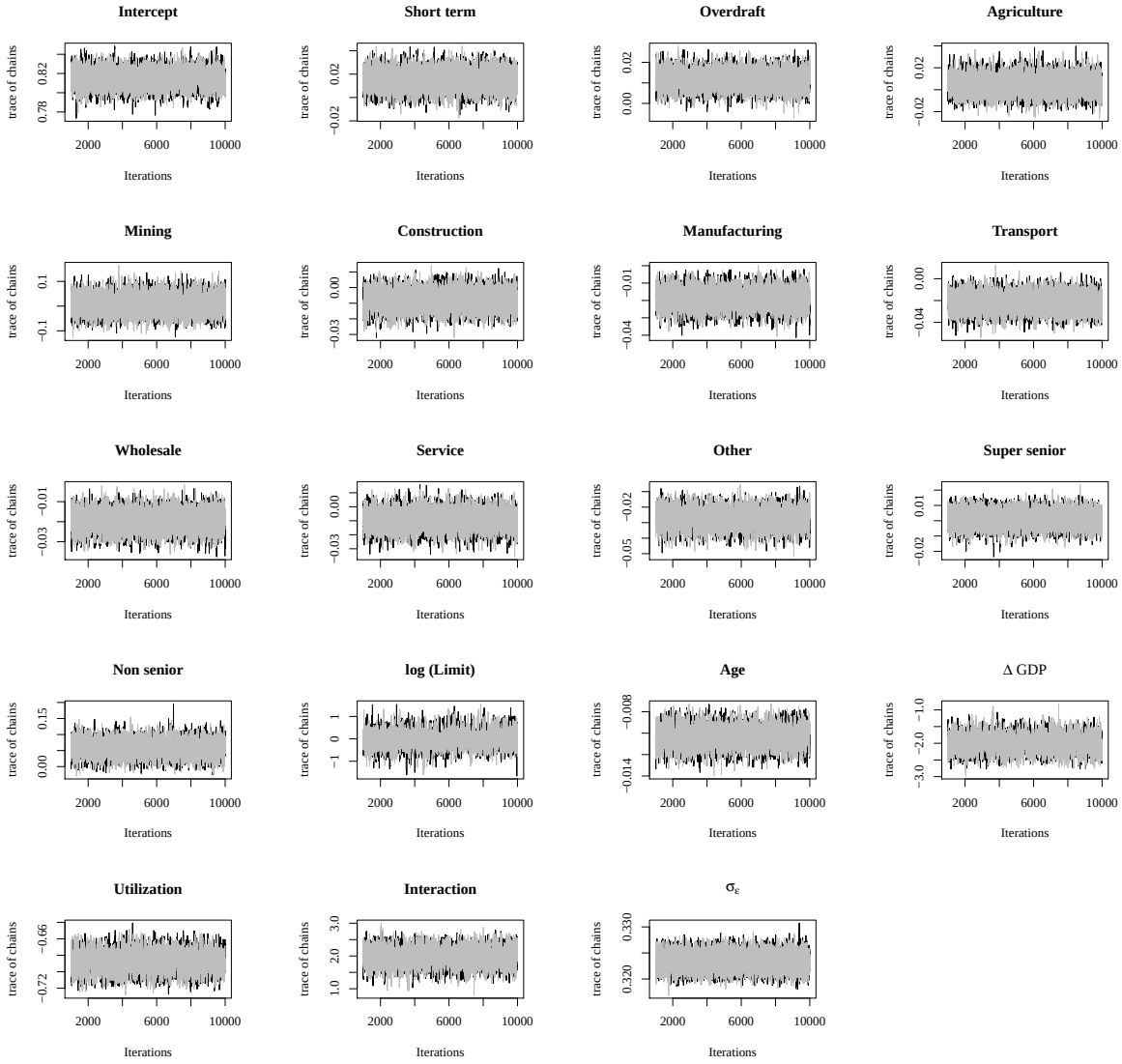
1.D.1 Traceplots

Figure 1.D.1: Traceplot USA | Macro Only Model | $\tau = 0.5$



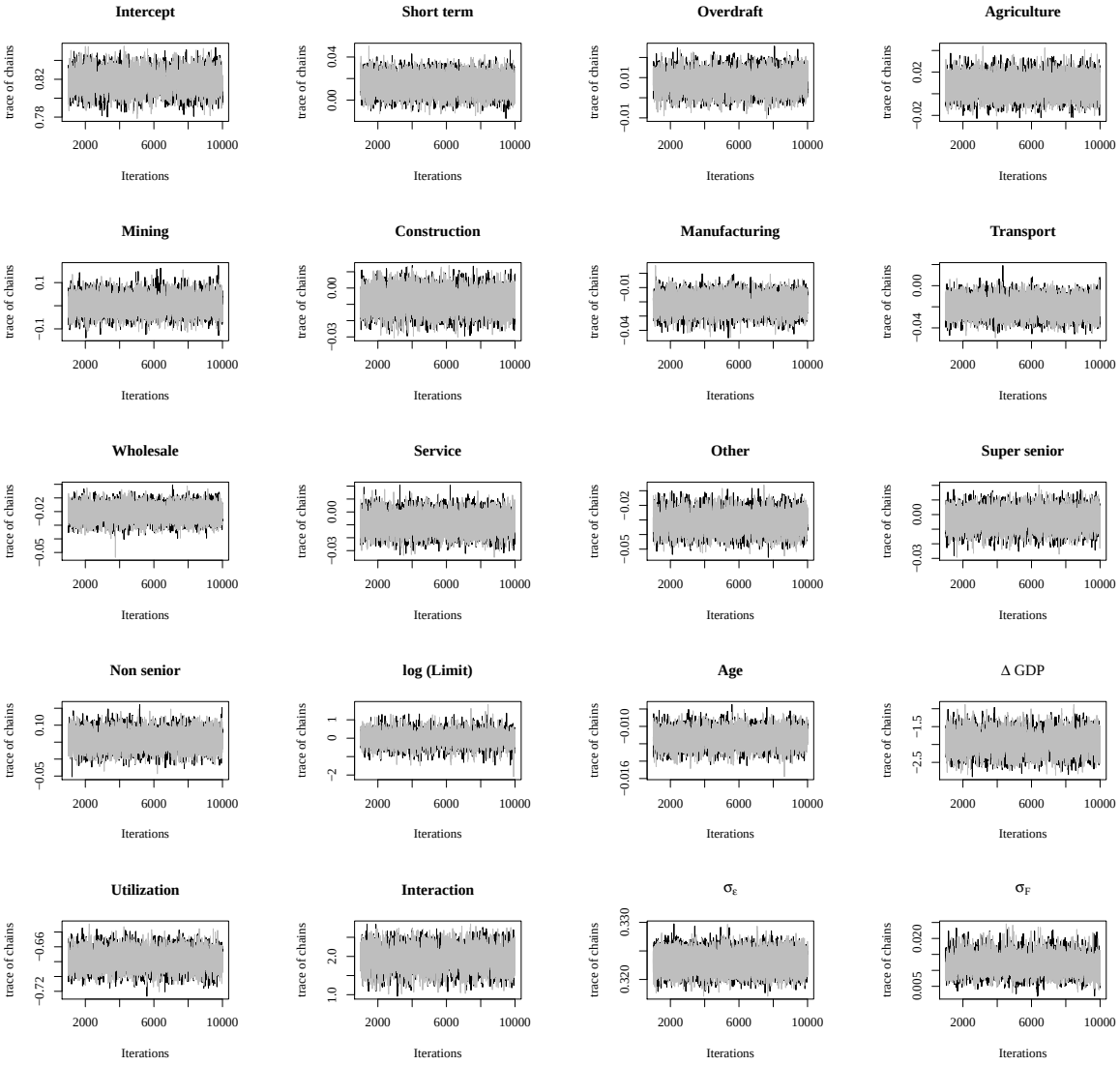
Note: The figure illustrates the MCMC chains for the Macro Only Model in the US American sample. The first chain is colored in black, whereas the second one in gray.

Figure 1.D.2: Traceplot Europe | Macro Only Model | $\tau = 0.5$



Note: The figure illustrates the MCMC chains for the Macro Only Model in the European sample. The first chain is colored in black, whereas the second one in gray.

Figure 1.D.3: Traceplot Europe | Random Effects Model | $\tau = 0.5$



Note: The figure illustrates the MCMC chains for the Macro Only Model in the European sample. The first chain is colored in black, whereas the second one in gray

1.D.2 Gelman Rubin Diagnostic

Table 1.D.1: Results | Macro Only Model (MOM) and Random Effects Model (REM) for Europe | $\tau = 0.50$

Level	MOM— Europe		MOM — USA		REM Model — Europe	
	Point estimate	Upper confid. limits (90%)	Point estimate	Upper confid. limits (90%)	Point estimate	Upper confid. limits (90%)
$\beta_{Intercept}$	1.0016	1.0016	1.0008	1.0028	1.0007	1.0019
$\beta_{Shortterm}$	1.0000	1.0001	1.0010	1.0039	1.0000	1.0001
$\beta_{Overdraft}$	1.0011	1.0040			1.0009	1.0034
$\beta_{Agriculture}$	1.0003	1.0011	1.0027	1.0102	1.0003	1.0008
β_{Mining}	1.0003	1.0004	1.0010	1.0010	1.0001	1.0005
$\beta_{Construction}$	1.0014	1.0053	1.0036	1.0061	1.0003	1.0011
$\beta_{Manufact.}$	1.0015	1.0037	1.0005	1.0005	0.9999	1.0000
$\beta_{Transport}$	1.0011	1.0045	1.0014	1.0037	1.0001	1.0004
$\beta_{Wholesale}$	1.0002	1.0011	1.0000	1.0000	1.0001	1.0004
$\beta_{Service}$	1.0035	1.0136	1.0005	1.0016	1.0023	1.0074
β_{Other}	1.0005	1.0007	1.0003	1.0009	1.0014	1.0057
$\beta_{SuperSenior}$	1.0013	1.0028	1.0000	1.0001	1.0004	1.0012
$\beta_{NonSenior}$	1.0002	1.0005	1.0008	1.0032	1.0002	1.0004
$\beta_{Unknown}$			1.0004	1.0017		
$\beta_{log(Limit)}$	1.0010	1.0037	1.0060	1.0165	0.9999	0.9999
β_{Age}	1.0008	1.0032	1.0008	1.0033	1.0012	1.0046
$\beta_{\Delta GDP}$	1.0011	1.0026	1.0007	1.0022	1.0002	1.0006
$\beta_{Utilization}$	1.0012	1.0021	1.0001	1.0002	0.9999	0.9999
$\beta_{Interaction}$	1.0012	1.0034	1.0007	1.0020	1.0003	1.0007
σ_ϵ	1.0000	1.0000	1.0006	1.0025	1.0002	1.0006
σ_F					1.0020	1.0073

Notes: The table summarizes the Gelman Rubin diagnostic for the different quantile regressions with $\tau = 0.5$. The first column indicates the estimated parameters. The Gelman-Rubin diagnostic examines the length of burn-in. The potential reduction factor and the upper confidence limit are displayed in this table. Convergence is achieved if chains do not depend on their initial values, i.e., for upper limits close to one (Gelman et al. (1992)). A rule of thumb assumes 1.1 as the critical value.

1.D.3 Heidelberger Welch Diagnostic

Table 1.D.2: Results | Macro Only Model (MOM) and Random Effects Model (REM) for Europe | $\tau = 0.50$

Level	MOM— Europe			MOM — USA			REM Model — Europe		
	Stationary test	Start	p-value	Stationary test	Start	p-value	Stationary test	Start	p-value
$\beta_{Intercept}$	passed	1	0.8105	passed	1	0.1476	passed	1	0.1537
$\beta_{Shortterm}$	passed	1	0.3552	passed	1	0.2930	passed	1	0.5847
$\beta_{Overdraft}$	passed	1	0.1478				passed	1	0.2819
$\beta_{Agriculture}$	passed	1	0.6500	passed	1	0.2812	passed	1	0.1539
β_{Mining}	passed	1	0.5665	passed	1	0.1009	passed	1	0.8425
$\beta_{Construction}$	passed	1	0.6427	passed	1	0.4143	passed	1	0.7893
$\beta_{Manufact.}$	passed	1	0.5964	passed	8001	0.0791	passed	1	0.8941
$\beta_{Transport}$	passed	1	0.2271	passed	1	0.5938	passed	1	0.7341
$\beta_{Wholesale}$	passed	1	0.1283	passed	1	0.4641	passed	1	0.3796
$\beta_{Service}$	passed	5401	0.0705	passed	1	0.5843	passed	1	0.1254
β_{Other}	passed	1	0.5231	passed	1	0.5648	passed	1	0.2908
$\beta_{SuperSenior}$	passed	1	0.3019	passed	1	0.2010	passed	1	0.3966
$\beta_{NonSenior}$	passed	1	0.3736	passed	1	0.4174	passed	1	0.6930
$\beta_{Unknown}$				passed	1	0.2013			
$\beta_{log(Limit)}$	passed	1	0.6185	passed	1	0.0766	passed	1	0.3555
β_{Age}	passed	1	0.7987	passed	1	0.7029	passed	1	0.8754
$\beta_{\Delta GDP}$	passed	1	0.3652	passed	1	0.3879	passed	1	0.2158
$\beta_{Utilization}$	passed	1	0.1887	passed	1	0.6711	passed	1	0.4300
$\beta_{Interaction}$	passed	1	0.5972	passed	1	0.3807	passed	1	0.5506
σ_ϵ	passed	1	0.5997	passed	1	0.1964	passed	1	0.4853
σ_F							passed	1	0.2112

Notes: The table summarizes the results of the Heidelberger-Welch diagnostic for the different quantile regression in the two samples. To evaluate whether the chain length is sufficiently long, both chains in each model are combined. In the Heidelberger-Welch diagnostic, a criterion of relative accuracy for the posterior means is calculated. The frequentistic stationary test uses the Cramer-von-Mises statistic to test the null hypotheses that the sampled values originate from a stationary process (see Heidelberger and Welch (1981, 1983)).

Chapter 2

Opening the Black Box – Quantile Neural Networks for Loss Given Default Prediction

This chapter is joint work with Ralf Kellner[1] and Daniel Rösch[2] and published as:

Kellner, R., Nagl, M., Rösch, D (2022). Opening the Black Box – Quantile Neural Networks for Loss Given Default Prediction. ***Journal of Banking & Finance 134*****, 106334**

https://doi.org/10.1016/j.jbankfin.2021.106334

We extend the linear quantile regression with a neural network structure to enable more flexibility in every quantile of the bank loan loss given default distribution. This allows us to model interactions and non-linear impacts of any kind without the need of specifying the exact form beforehand. The precision of the quantile forecasts increases up to 30% compared to the benchmark, especially for higher quantiles which are most important in credit risk. By using a novel feature importance measure, we calculate the strength, direction, interactions and other non-linear impacts for every conditional quantile and every variable. This enables us to explain why our extension exhibits superior performance over the benchmark. Moreover, we find that the macroeconomy is up to two times more important in USA than in Europe and has large joint impacts in both regions. The macroeconomy is most important in the US, whereas in Europe collateralization is essential.

Keywords: Quantile Regression, Black Box, Neural Networks, Explainable Machine Learning, Global Credit Data

JEL Classification: C21, G21, G33

[1] University Regensburg, Chair of Statistics and Risk Management, 93040 Regensburg, Germany, email: ralf.kellner@ur.de.

[2] University Regensburg, Chair of Statistics and Risk Management, 93040 Regensburg, Germany, email: daniel.roesch@ur.de.

2.1 Introduction

Estimation and prediction of loss given default (LGD) is an important and challenging task for financial institutions. The LGD is the fraction of loss from the exposure at default and depending on the instrument, it can be divided into market-based and workout LGDs. The former is relevant for publicly traded instruments such as bonds and defined as one minus the ratio of the market price 30 days after default over the outstanding amount. The latter is commonly used for loan contracts and is determined by subsuming discounted payments from debtors during the process of default resolution. LGD distributions exhibit extreme and versatile shapes, typically with high probability masses centered around zero and one. Furthermore, predicting LGDs is a challenging task due to long-lasting and complex resolution processes. In this paper, we use access to a unique database of workout LGDs provided by Global Credit Data (GCD), which is a non profit initiative that supports banks by collecting and analyzing historical loss data from a multitude of member banks worldwide, encompassing several systematically relevant institutions (www.globalcreditdata.org). Moreover, our analysis is separated into a data set of American and European loans, as previous studies detected profound differences between those regions which can likely be ascribed to differences in the legal and regulatory environment.

After years of developing models for the probability of default (PD), LGD modeling has attracted more and more attention. In general, common drivers for LGDs are identified in different studies which also compare a variety of LGD models (see, e.g., Bastos, 2010; Grunert and Weber, 2009; Loterman et al., 2012; Qi and Yang, 2009; Qi and Zhao, 2011). Khieu et al. (2012) identified loan characteristics as more important for recovery rates of bonds, but macroeconomic variables also play an important role. Hence, the discussion whether and which economic variables are important is still active, (see, e.g., Leow et al., 2014; Krüger and Rösch, 2017; Betz et al., 2018; Nazemi et al., 2021, 2017; Nazemi and Fabozzi, 2018). The latter two use principal components gathered from many different macroeconomic variables, which we also follow in this study. Furthermore, Krüger and Rösch (2017) find a varying (linear) impact over the entire conditional LGD distribution, which points to a complex and potentially non-linear relationship between LGDs and the economy. This is in line with findings of Sopitpongstorn et al. (2021), who find non-linearities between covariates and market based recovery rates by using a local logit regression.

However, a high amount of publications is dedicated to market-based or expected LGDs which exhibit significant differences to workout LGDs.[3] In contrast to the latter, market-based LGDs are bound in the interval $[0,1]$ and are characterized by short resolution processes. Especially for workout LGDs, the distributional form is extreme due to values below zero, above one, along with high probability masses at zero and one. This is probably the reason why mixture distributions and other models with flexible distributional forms best capture the workout LGD distribution. Altman and Kalotay (2014) develop a Bayesian finite mixture model of normal distributions with an underlying ordered logit model which links debtor features to mixture component affiliation. A frequentistic version of this model is presented by Kalotay and Altman (2017) and a mixture of beta distributions is applied by Calabrese (2014). Variants of mixture models are also used in Betz et al. (2018) and Tomarchio and Punzo (2019). An alternative approach is shown by Krüger and Rösch (2017) who use linear quantile regression to predict different parts of the LGD distribution. Even though this approach is able to capture a varying impact of predictors over the distribution, it is restricted to a linear relationship between predictors and the variable of interest, and the evaluation of (non-linear) interactions would be computationally burdensome.[4]

Neural networks have previously been applied to the estimation of LGDs (see, Qi and Zhao, 2011; Loterman et al., 2012). However, in comparison to our approach, the network is calibrated to predict the mean value. In contrast, we calibrate the network for a discrete set of quantiles of the LGD distribution using the quantile specific loss function and control for strict monotony of quantile estimates. As shown by Krüger and Rösch (2017), it is important to account for varying impact for different quantile levels. This may explain why the application of neural networks to LGD modeling in previous studies has often led to worse results in comparison to other flexible approaches like regression trees and support vector regression. However, neural networks already exhibit promising results in comparison to less flexible approaches like transformed regression type models. Yet, a disadvantage is the alleged incapability of identifying (relevant) predictors for LGDs (see, Qi and Zhao, 2011). However, this disadvantage vanishes in the light of recent techniques. In this paper, we use a feature importance measure based on gradient information as shown in Horel et al. (2018) and Nagl (2021). It enables us to decompose the prediction of neural networks into their relative feature importance and interactions with all other features. This gives us a broad and detailed description of the underlying relations.

[3] A comprehensive study on expected LGDs based on Credit Default Swaps can be found in Doshi et al. (2018).

[4] Similar approaches focusing on the estimation of the distribution of LGD can be found in Hwang and Chu (2018) and Hwang et al. (2020). Both rely on inverse-probability-transformations of the true LGD values which is not feasible for workout LGDs as they are not bounded between zero and one.

Furthermore, the computational burden to model and test joint effects and interactions between independent variables can be considerably reduced. As an example, assume that we want to estimate 100 quantiles using 26 different predictor variables. To test and model every pairwise interaction, e.g. $x_1 \cdot x_2$, one has to fit $26 \cdot \frac{26-1}{2} \cdot 100 = 32,500$ different models. If we think about non-linear interactions as well, e.g. $\exp(x_1 \cdot x_2)$ or $x_1^2 \cdot x_2$, this number rises fast and results in a computationally expensive and tedious task. By using our approach we have to fit only one model to capture all these possible forms at once. Hence, it is not necessary to assume a structural relationship beforehand, but we can quantify the strength of joint effects and non-linearities for every quantile afterwards.

Quantiles are important not only for the overall distribution, but also to differentiate between loans and their inherent risk profile. Consider the following stylized example of different quantiles of the Loss Given Default:

	$\hat{Q}_{0.05}$	$\hat{Q}_{0.25}$	$\hat{Q}_{0.50}$	$\hat{Q}_{0.75}$	$\hat{Q}_{0.95}$
loan 1	0.03	0.08	0.20	0.59	0.70
loan 2	0.03	0.08	0.20	0.67	0.90
loan 3	0.02	0.05	0.20	0.59	0.70

If we focus only on median estimates ($\hat{Q}_{0.50}$), one might come to the conclusion that every loan exposes the bank to the same risk in terms of LGD. But focusing on higher quantiles, e.g. unfavourable scenarios for the bank, we observe that loan 2 entails a considerably higher risk. On the contrary, loan 3 contains the least risk as the lower quantiles are smaller. Moreover, our empirical analysis finds that higher quantiles are driven by a higher sensitivity to the macroeconomy and overall higher non-linearity and interactions compared to the median, which underlines the importance of quantiles. In summary, mean-related methods may not be representative for bimodal distributions, which are characterized rather by their tails than their expectation and that quantiles can enhance the bank's ability to differentiate between risk profiles of obligors.

Estimating conditional quantiles has emerged as a powerful method and widespread potential application. The most common method is to minimize the expectation of the so-called check-function leading to the linear quantile regression introduced by Koenker and Bassett (1978) and for example used by Krüger and Rösch (2017). For a timely and comprehensive overview of various extensions and applications we refer to Koenker et al. (2017). In general, non-linearity can be allowed by additive quantile regressions using splines, see e.g. Koenker et al. (1994), Horowitz and Lee (2005) and Hoshino (2014). However, one has to choose which variables

should expressed as splines or tensors if non-linear interactions should be allowed.

Fully non-parametric quantile regressions were introduced by Koenker (2005), Li and Racine (2008) and Li et al. (2013). Along with the rise of quantile regressions, also the discussion of non-monotone quantile estimates appeared. For example, Takeuchi et al. (2006) shows that monotonicity can be included directly in the estimation procedure. A more general post-hoc method to ensure monotone quantile function is introduced by Chernozhukov et al. (2010) who argue to simply rearrange quantile estimates. This approach is for example used by Wu and Yan (2019).

We contribute to the literature by developing a quantile neural network regression model with a sound estimation procedure. The calibration to each quantile is subsumed in a single optimization step. This considerably reduces the computational burden, especially with respect to possible forms of interactions between variables. Furthermore, we find a superior in- and out-of-time performance in both out-of-time periods (Great Financial Crisis (GFC) and post GFC) compared to the linear counterpart. This can be traced back to non-linear relationships between predictors and LGDs especially in higher quantile levels. By using the feature importance measure, we can attribute the superior performance of the QRNN approach to non-linearity in specific quantiles and the variables driving it. This offers insights into risk drivers that have not been detected in earlier literature. Furthermore, explainability is of great importance for financial institutions, as neural networks are often falsely accused of being black boxes in the financial community and regulators strictly allow LGD predictions if their derivation is transparent to them. This has prohibited the use of superior models for regulatory and internal risk management purposes. From an economic perspective, we find that the impact of the macroeconomy is up to two times more important in the US than in Europe. The results suggest that in Europe the economic surrounding interacts most with variables describing the different forms of collateralization. On the contrary, in the United States the level of seniority has a large joint impact with the macroeconomy on the LGD prediction, especially for higher quantiles.

The remainder of this paper is structured as follows. In Section 2.2, we give a short summary of relevant literature with the use of machine learning for credit risk modeling. Data is presented in Section 2.3, while the methodology is described in Section 2.4. Our empirical results are discussed in Section 2.5 and Section 2.6 concludes.

2.2 Literature Review

Over the last decade, there is an increasing attention of machine and deep learning algorithms in credit risk. We can name two possible reasons. First, computational power increased massively in recent years and open-source solutions have been widely developed. This makes highly complex algorithms available to a very broad audience, which covers academics, but also practitioners and regulators. Second, superior performance of these algorithms is well documented. The following section provides a brief review of recent studies using machine or deep learning in credit risk.

Overall, the literature concerning PD is much wider than for LGD. There are many studies which compare supervised machine learning algorithms with respect to their predictive power (see, e.g. Cowden et al., 2019; Li and Chen, 2019; Chen et al., 2020; Petropoulos et al., 2020; Luo et al., 2020; Dumitrescu et al., 2021). A general consensus exists that more flexible models outperform the linear logit regression. Bakoben et al. (2020) employ unsupervised learning approaches to detect different credit card account behaviours, increasing predictive power. A promising part of machine learning are deep learning models. Kvamme et al. (2018) utilize deep convolutional networks to predict probability of default of mortgage loans, showing a superior performance. Mai et al. (2019) also use deep convolutional networks and ensemble techniques to predict corporate defaults. They further use textual disclosures into to enhance the discrimination power of their models. Another perspective of deep learning is shown by Sariev and Germano (2019). They utilize Bayesian regularized neural networks to automatically determine the regularization in the network. Furthermore, the combination of machine learning and classical statistical models seems to be promising as well. Li and Chen (2019) combine logistic regression and neural networks to enhance discrimination power. Sigrist and Hirnschall (2019) unite Tobit regression with regression trees. Our paper also follows this line, as we combine statistical models with machine learning methods. For a detailed overview of machine learning in PD modelling, we refer to Mai et al. (2019). Jing et al. (2021) are among the first to incorporate the evolution of PDs over time in a long short-term memory network.

The application of machine learning for Loss Given Defaults has become very popular in the last decade. Early studies were conducted by Matuszyk et al. (2010) and Bastos (2010) using tree based methods. Comprehensive benchmark studies were conducted by Qi and Zhao (2011) , Bellotti and Crook (2012) and Loterman et al. (2012).

The latter one may be the most comprehensive by testing 24 different regression algorithms in total, based on six real world datasets and finding evidence that non-linear techniques perform best. Some other studies put more emphasis on the comparison of two stage and single models, such as for example Tobback et al. (2014), Sun and Jin (2016) or Tanoue and Yamashita (2019). Yao et al. (2017) propose two-stage approaches involving support vector machines for LGD prediction. Contrary, Nazemi et al. (2017) find that fuzzy decision methods perform best. Nazemi et al. (2018) show that using principal components, an unsupervised learning approach, derived from a wide range of macroeconomic variables enhances the prediction performance. Most recent studies focus on the comparison of a very wide range of models, macroeconomic variables and LGD types. Bellotti et al. (2021) conduct an exhaustive analysis with various different models and find tree based methods to be superior. Kaposty et al. (2020) conduct a horse race of different models, in which random forests turn out to be the best ones. Gambetti et al. (2020) also use a vast selection of machine learning models and introduce meta-learning strategies, providing evidence that the macroeconomic surrounding is important for all methods if it is incorporated via uncertainty measures. This complements findings of Gambetti et al. (2019) that uncertainty is the most important macroeconomic variable for (market) LGD prediction.

In recent years, the body of literature focusing on explanation methods has grown fast. One can divide this body into local explanations, i.e. explain the individual prediction of an observation, and global explanations, i.e. explain the learned relation of the black-box model. For an excellent and detailed review regarding these methods, we refer to Horel and Giesecke (2020). Partial Dependence Plots (PDP) are introduced by Friedman (2001). They plot the importance of a feature by varying over its marginal distribution and calculate the (global) effect on the resulting prediction. Goldstein et al. (2015) extend this idea to individual predictions by introducing the Individual Conditional Expectation (ICE) Plots. Apley and Zhu (2020) introduce Accumulated Local Effects (ALE) Plots and focus on the conditional distribution of features instead, solving problems of PDP and ICE plots. One of the most prominent method is Local Interpretable Model-agnostic Explanations (LIME) introduced by Ribeiro et al. (2016). To explain any black-box model LIME perturbs the data for a given observation and get the black box predictions for these new points. Afterwards, a white-box model, such as a linear regression, is fitted to the permuted data and predictions. Ribeiro et al. (2018) introduced anchors, which decompose black-box predictions into highly interpretable if-else rules, e.g. if x_{i1} is greater than threshold z_{i1}, predict y_i. Recently, many applications of SHapley Additive exPlanations (SHAP), introduced by Lundberg and Lee (2017), can be found in the literature.

This feature importance measure is the only one backed by an economic theory called coalitional game theory. SHAP values can be calculated for individual predictions and for global explanations. Horel and Giesecke (2020) continues the work by Horel et al. (2018) and derive a way to statistically test the impact of the feature importances. Their test statistics are valid for single-layer networks using mean-squared error loss and sigmoid activation functions. Nagl (2021) builds on the work by Horel et al. (2018) and addresses the quantification of non-linearity and interactions entailed in black-box predictions.

Summarizing, recent studies focusing on machine learning applied LGD are typically conducted in the spirit of "horse races" in which various methods are compared in their performance. We rather show how to extend the economically useful and meaningful method of quantile regression with non-linearity, feature importance of independent variables and with the identification of their interactions. Moreover, other studies commonly focus on mean predictions and not on the entire distribution (or quantiles). The approach that we present explicitly focusses on quantiles and, hence, delivers a much broader picture of LGDs. Additionally, our measure for feature importance can especially and easily be applied to the quantile regression neural network and sheds light on risk drivers and joint effects that have not been documented in the literature before. While the general idea to extend quantile regression with stacked layers of neural network architectures has been applied to fields other than credit[5] – to the best of our knowledge – this paper is the first using this approach in credit risk. Hereby, domain specific adjustments such as a unified estimation procedure with monotonicity regularization are developed as a new unique contribution to the credit risk literature.

2.3 Data

We use access to Global Credit Data (GCD), one of the world's largest loss data bases. This consortium consists of 55 globally acting banks, encompassing several systemically relevant institutions. The data offer an unique and broad perspective of the banking universe.[6] The information is based on transactions, providing a detailed view on occurred losses and their determinants. We focus on workout Recovery Rates, including post-default cash flows. Recovery Rates are the difference between discounted positive cash flows (CF^+) and discounted direct as well as indirect costs (CF^-), divided by the exposure at default (EAD).

[5] See, e.g., applications for time series models in Xu et al. (2016), Salinas et al. (2019) and Wu and Yan (2019).
[6] For recent information about GCD we refer to Brumma et al. (2020a,b).

The LGD is defined as one minus the Recovery Rate[7]:

$$LGD = 1 - \frac{\sum_{i=1}^{n} CF^{+} - \sum_{i=1}^{n} CF^{-}}{EAD}. \tag{2.1}$$

To check for appropriateness of the calculated LGD values, we use the same procedure as Betz et al. (2018). We impose a materiality restriction of 500 USD and only use resolved loans to avoid the well-known default resolution bias, see e.g. Betz et al. (2018) and Betz et al. (2020). Considering this, we restrict our data sample to the time period of 2000-2016, as the resolution of loan contracts can last several years. Workout LGDs are not compulsorily restricted to the interval of 0% and 100%. Hence, we cut values outside of the range [-25%, 125%].[8] Actually, this restriction is only of minor concern using quantile-based methods, because they are generally less sensitive to outliers. Nevertheless, it insures a rather homogeneous sample. To compare different economic and geographic regions, we use a European and US American sample.

Table 2.1 shows descriptive statistics for both considered regions. For metric dependent variables, selected quantiles, the mean and standard deviation are presented. For categorical variables, quantiles of the LGD distribution in these categorical subgroups are shown. The first level of each categorical variable is used as the reference level.

For example, facility types are divided into two subcategories, namely term loans and credit lines. Regarding the quantiles in the US American sample, we observe that the distribution is very similar for quantiles lower than the median, whereas credit lines have considerably more mass at the 75% quantile. The difference between credit line and term loan is even more pronounced in the European sample. This may indicate that high quantiles are quite different regarding the facility type. Considerably deviating effects can also be observed in other categories like asset class, industry and collateral. This may imply that one has to account for different impacts on different parts of the LGD distribution, underlining our quantile based approach. To account for systematic effects, we conduct a Principal Component Analysis (PCA) of the most common macroeconomic variables regarding workout LGD estimation, following Nazemi et al. (2018).[9] To account for roughly 95% of their variance, we derive eight components in Europe and eleven components in the United States.[10]

[7] For a detailed overview of positive and negative cash flows, see Betz et al. (2020).
[8] In total, we cut off 3.48% of the available resolved loans due to all restrictions.
[9] For a detailed methodical overview of the PCA, we refer to Nazemi et al. (2018).
[10] We also conducted all analyses with 90% and 99,5% accounted variance as a robustness test. All conclusions and results remain the same as outlined in the following sections. The evaluations with the alternative settings are available upon request.

Table 2.1: Descriptive statistics

(a) US sample

Variable	Level	Quantiles					Mean	STD	Obs.
		0.05	0.25	0.50	0.75	0.95			
LGD		0.00	0.37	4.51	51.34	100.00	27.89	37.29	9649
log(EAD)		9.99	12.10	13.56	15.24	17.00	13.59	2.18	9649
PC 1		37.97	59.46	78.39	94.04	128.34	78.89	27.40	9649
PC 2		−194.10	−165.38	−127.17	−107.14	−74.34	−134.68	38.89	9649
PC 3		98.01	126.82	154.96	202.31	260.96	164.30	47.94	9649
PC 4		10.52	28.96	38.94	58.29	82.48	43.04	21.07	9649
PC 5		−44.24	−27.30	−17.93	−5.76	38.45	−14.72	21.75	9649
PC 6		−49.21	−14.80	−4.72	3.29	8.89	−7.39	15.44	9649
PC 7		−5.56	7.41	21.53	34.49	112.00	27.79	31.62	9649
PC 8		−27.16	−15.26	−1.80	13.52	35.46	0.34	18.47	9649
PC 9		23.48	30.01	37.84	55.61	75.20	42.58	15.31	9649
PC 10		−30.02	−4.65	7.90	14.70	30.74	5.58	15.96	9649
PC 11		−61.04	−20.00	−15.41	−12.76	−2.12	−19.52	16.52	9649
Facility type	Term loan	0.00	0.49	5.00	44.47	100.00	25.56	34.79	5331
	Credit line	−0.01	0.25	3.77	63.80	100.00	30.77	39.99	4318
Seniority	Pari passu	−0.14	0.17	6.62	100.00	100.00	38.33	44.35	3859
	Senior	0.00	0.54	4.06	33.87	93.43	20.89	29.61	5623
	Non senior	0.00	0.06	1.01	38.60	100.00	22.33	34.51	167
Industry	FIRE	−0.03	0.42	5.56	65.55	100.00	31.64	40.51	1181
	Agriculture	0.00	0.03	2.39	40.18	100.00	23.99	34.82	663
	Mining	−0.08	0.00	1.09	29.15	96.85	18.21	28.98	167
	Construction	0.00	0.34	4.38	42.36	100.00	24.63	34.04	1853
	Manufacturing	0.00	1.60	7.51	26.42	79.96	19.49	25.96	279
	Transport	0.00	0.69	6.59	66.63	100.00	32.24	39.94	1021
	Wholesale	0.00	0.42	4.91	52.96	100.00	28.06	36.86	2189
	Services	0.00	0.33	3.85	59.24	100.00	29.34	39.05	2296
Asset Class	SME	0.00	0.26	3.54	68.79	100.00	31.19	40.82	5870
	Large corporates	0.00	0.63	6.11	37.96	93.26	22.76	30.33	3779
Collateral	No	−0.05	0.54	6.28	69.27	100.00	31.73	40.29	1919
	Real Estate	−0.02	0.42	5.19	97.47	100.00	33.92	42.36	1160
	Yes	0.00	0.33	4.04	47.00	100.00	25.70	35.16	6570
Guaranteee	No	0.00	0.51	5.02	53.47	100.00	28.42	37.51	6864
	Yes	0.00	0.19	3.26	48.34	100.00	26.58	36.72	2785

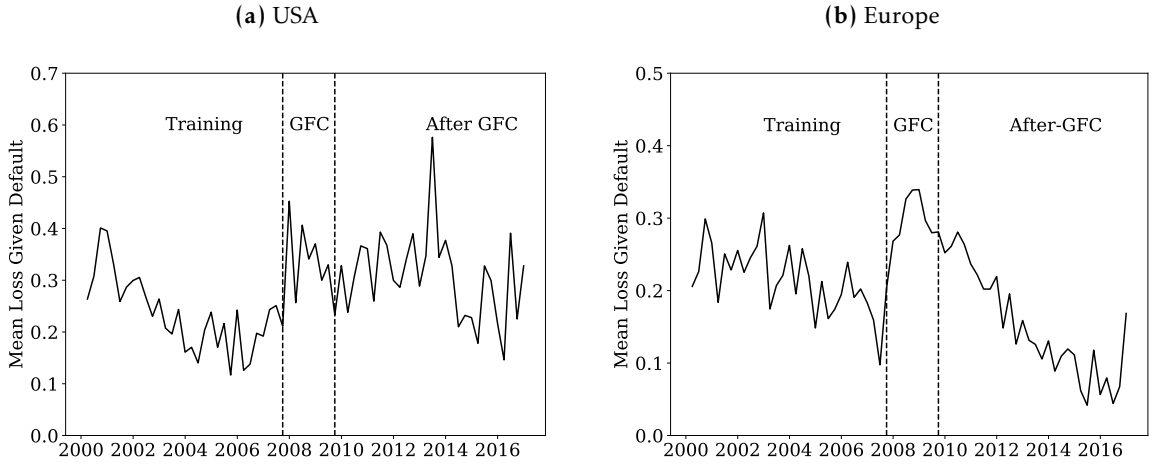
(b) European sample

Variable	Level	Quantiles					Mean	STD	Obs.
		0.05	0.25	0.50	0.75	0.95			
LGD		0.00	0.00	1.18	26.25	100.00	21.18	34.89	44480
log(EAD)		8.52	10.52	11.83	13.15	15.26	11.85	2.04	44480
PC 1		−380.13	−253.38	−81.99	280.37	879.00	66.46	409.94	44480
PC 2		−5288.11	−1994.81	−91.59	831.00	1938.78	−721.75	2308.33	44480
PC 3		−1419.16	−570.62	141.55	1601.34	4117.37	614.83	1778.93	44480
PC 4		−5057.54	−1946.21	−189.13	710.03	1928.93	−699.37	2232.38	44480
PC 5		−899.02	−363.31	97.23	1032.14	2646.32	403.21	1145.24	44480
PC 6		−765.30	−280.80	124.80	826.23	2103.84	322.12	916.72	44480
PC 7		−399.50	−103.94	160.36	576.27	1457.96	292.55	588.54	44480
PC 8		−3109.94	−1169.66	−37.36	525.33	1250.03	−375.62	1399.83	44480
Facility type	Term loan	0.00	0.00	0.91	15.92	100.00	17.26	31.24	26705
	Credit line	0.00	0.00	1.94	57.00	100.00	27.08	39.02	17775
Seniority	Pari passu	0.00	0.00	1.66	31.94	100.00	21.72	34.50	6553
	Senior	0.00	0.00	1.18	26.38	100.00	21.30	35.09	36264
	Non senior	0.00	0.00	0.28	10.52	98.43	16.57	31.53	1663
Industry	FIRE	0.00	0.00	0.69	14.07	100.00	17.76	32.65	7039
	Agriclulture	0.00	0.00	0.94	27.00	100.00	21.01	35.19	1382
	Mining	0.00	0.00	1.72	14.29	94.86	15.80	28.70	239
	Construction	0.00	0.02	1.72	24.90	100.00	20.92	34.28	7841
	Manufacturing	0.00	0.00	1.62	18.08	99.66	19.59	33.66	262
	Transport	0.00	0.00	1.49	26.36	100.00	21.42	34.88	5901
	Wholesale	0.00	0.00	1.01	31.84	100.00	22.40	36.08	10592
	Services	0.00	0.00	1.37	31.93	100.00	22.41	35.50	11224
Asset Class	SME	0.00	0.00	1.03	26.77	100.00	21.37	35.28	39563
	Large corporates	0.00	0.02	3.76	22.68	100.00	19.71	31.56	4917
Collateral	No	0.00	0.03	1.95	50.43	100.00	26.28	39.15	15964
	Real Estate	0.00	0.00	0.47	7.83	94.06	13.81	27.88	19420
	Yes	0.00	0.03	3.69	55.96	100.00	27.98	37.37	9096
Guarantee	No	0.00	0.00	1.51	38.19	100.00	23.58	36.57	34936
	Yes	0.00	0.00	0.44	7.24	88.64	12.40	26.08	9544

Note: The table shows means, standard deviations and quantiles for the LGD and the metric variables. For categorical variables, means, standard deviations and quantiles of the LGD for each level are displayed.The PCs capture roughly 95% of the macroeconomic variable's variance outlined in Table 2.A.1.The first level of each category is used as the reference level.

This reduces the problem of selecting the most suitable macroeconomic variables, which is especially difficult for workout LGDs due to their very long resolution time, see for example Betz et al. (2020). The considered variables can be found in Table 2.A.1 in Appendix 2.A. In general, we include variables to capture the economic uncertainty, following Gambetti et al. (2019), the general economic situation and monetary and inflation related measures. Furthermore, our analysis includes several survey based variables.

Figure 2.1: Time variation of average LGDs

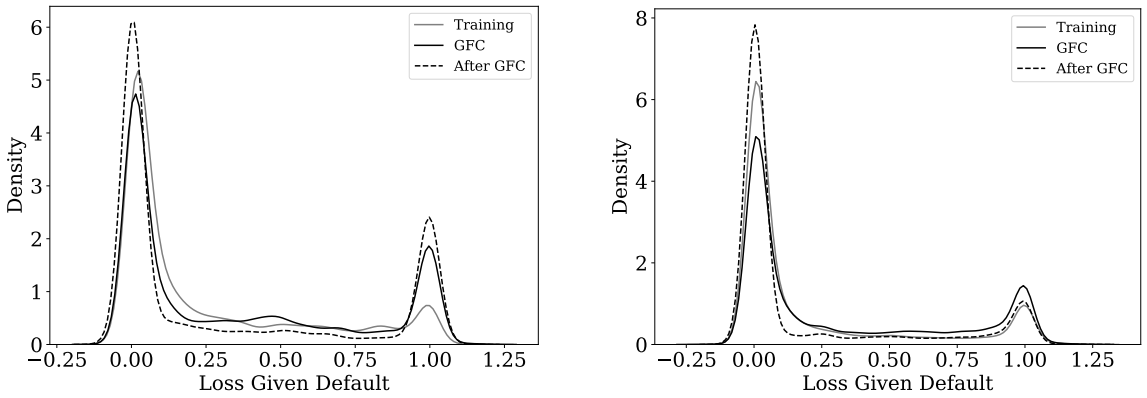


Note: This figure shows the time variation of the LGD means over time. The left panel illustrates mean values for the US American sample, whereas the right panel refers to the European sample. Furthermore, the time span of different samples is indicated via vertical dashed lines. We divide our sample in training, GFC and the time after the GFC.

For the rest of the paper, we divide the subsamples further, namely into a training sample starting in the year 2000, another one that contains LGDs during the GFC and the final one consisting of LGDs after the GFC until 2016. We follow the OECD which specifies the crisis period in the US from 2007 Q4 until 2009 Q2 and from 2008 Q1 until 2009 Q3 in Europe. We do not additionally split the training sample into in- and out-of-sample on a cross-sectional basis, as predicting the future is of major concern in credit risk. For example, Kalotay and Altman (2017) put emphasis on the fact that it is crucial to predict future and not only contemporary LGDs. Figure 2.1 illustrates the different behaviour of average LGDs over time. In the training sample, the mean values follow a very similar path over time, whereas in early years the LGDs in USA are slightly higher. Furthermore, in both regions, the GFC is observable and characterized by considerably higher LGDs. The main difference between both regions can be deduced after the GFC. In the European sample, the mean LGD values deteriorate towards low levels. On the contrary, the LGD values in the US American sample remain high and even increase in the time period around 2013-2014. Both different time patterns may be challenging for any model, due to signals of non-stationary behaviour requiring great flexibility. The different behaviour also points towards different systematic effects in both regions.

Figure 2.2: Estimated densities of LGD distributions

(a) USA **(b)** Europe



Note: This figure shows kernel density estimates of different time periods in different regions. The left panel illustrates the shape of the LGD distributions in the US American sample, whereas the right panel refers to the European sample. We used the same bandwidth for all density estimates to allow for comparison between regions as well as time periods.

Figure 2.2 shows the density estimates for both regions and different subsamples. The bandwidth for the density plot is held equal to allow a comparison. The LGD values used in this paper show a bimodal shape with large masses around zero and one. With respect to the US American sample, we can see that the mass on extreme high losses (LGD $\geq$1) is lowest in the training sample. Furthermore, the probability mass of high losses is even higher in the sample period after the GFC than in the crisis period itself. This may mainly be driven by the peak around 2014. Overall, we can see that the probability mass is shifted from low to high losses along the time line. This is contrary to Europe, where we observe the largest mass on high losses during the GFC and very equal masses in the other periods. This analysis gives another empirical indication why modeling quantiles is important for Loss Given Defaults. Figure 2.2 shows that for the 25% to 75% quantiles, i.e. the middle of the distribution, we have only very little observations as the very large part of all LGD realizations lie on the left and right tail. Hence, focusing on the expectation of this distribution would imply that one would focus on regions of the distribution which are rarely observed in practice.

With respect to the descriptive analysis of both regions, some differences and modelling implications are revealed. Both regions may be shaped by different systematic behaviours, which implies different models for both regions. Furthermore, the very different out-of-time behaviour requires a very flexible modelling approach. Additionally, we see deviating impacts on different parts of the LGD distribution, which suggests the need for increasing the model flexibility in this direction as well.

2.4 Methods

Quantile Regression (QR)

Krüger and Rösch (2017) use a linear quantile regression approach to tackle the challenging distributional form. Following Koenker and Bassett (1978), the τ-th conditional quantile of the response y_i given $\boldsymbol{x}_i$ is:

$$Q_\tau(y_i|\boldsymbol{x}_i) = \beta_{0,\tau} + \boldsymbol{\beta}_\tau^T \boldsymbol{x}_i, \tag{2.2}$$

with $\boldsymbol{x}_i \in \mathbb{R}^p$ as a vector of p covariates for any observation $i = 1,...,N$ and $\beta_{0,\tau} \in \mathbb{R}$, $\boldsymbol{\beta}_\tau \in \mathbb{R}^p$ as model parameters. The so-called check function $\rho_\tau(\omega)$ is defined as:

$$\rho_\tau(\omega) = \begin{cases} \tau\omega & \text{, if } \omega \geq 0, \\ (1-\tau)|\omega| & \text{, else.} \end{cases} \tag{2.3}$$

Parameter estimates are derived by minimizing the sum over all data points:

$$\sum_{i=1}^{N} \rho_\tau\left(y_i - \beta_{0,\tau} - \boldsymbol{\beta}_\tau^T \boldsymbol{x}_i\right). \tag{2.4}$$

This approach has at least two shortcomings: First, the quantile functions are estimated separately, leading to non-monotone quantiles. Hence, the estimated values of the quantile function may not increase with τ. Second, the impact of covariates is linear in the quantiles which may not be flexible enough, especially for extreme quantiles.

Quantile Regression Neural Networks (QRNN)

We extend this approach by using an Artificial Neural Network (ANN) to approximate the LGD distribution. Figure 2.B.1 in Appendix 2.B shows a graphical comparison of both approaches for the interested reader. To generate predictions for a set of τ-th LGD quantiles, the neural network starts with covariate matrix $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ as inputs in the input neurons. The network consists of stacked hidden layers $l = 1,...,L$ whereby each layer consists of K_l neurons $\boldsymbol{h}_l \in \mathbb{R}^{K_l}$ that are determined by an affine combination of neurons in the previous layer which is composed of an arbitrary (non-linear) activation function σ.

$$\boldsymbol{h}_l = \sigma\left(\boldsymbol{W}_l \boldsymbol{h}_{(l-1)} + \boldsymbol{b}_l\right)$$

with $\boldsymbol{W}_l \in \mathbb{R}^{K_l \times K_{l-1}}, \boldsymbol{b}_l \in \mathbb{R}^{K_l}$ as parameters which are usually called weights and biases. Quantiles are derived from the last layer, the so-called output layer $L+1$ and are given by choosing the identity function for σ, resulting in:

$$Q_\tau(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{W}_{\tau,L+1}\boldsymbol{h}_L + \boldsymbol{b}_{\tau,L+1}$$

with $\boldsymbol{y} \in \mathbb{R}^N$ is a vector of all response realizations y_i. It should be noted that weights and biases are shared among different levels of τ, except in the output layer, which is highlighted by the subscript τ for weights and biases in the output layer, only. That is different to traditional linear quantile regression and motivated to keep the model as parsimonious as possible. A graphical illustration can be found in Figure 2.B.1 in Appendix 2.B. The weights and biases are estimated via a backpropagation algorithm based on Rumelhart et al. (1986). This requires a loss function to be differentiable at any point. However, $\rho_\tau(\omega)$ cannot be differentiated at the origin. Therefore, we approximate this region following Huber (1964). This approach approximates this region quadratically, ensuring differentiation at any point. Moreover, as we estimate several quantile functions simultaneously, we have to ensure monotonicity using a penalty similar to Takeuchi et al. (2006). The penalty increases if there are non-monotone quantiles in any estimated quantile function for any different LGD observation.

Formally, the new quantile loss $\rho_\tau^{QRNN}(\omega)$, the Huber loss $h(\omega)$ and the monotonicity penalty $m(\boldsymbol{X})$ are defined as:

$$\begin{aligned}
\rho_\tau^{QRNN}(\omega) &= \begin{cases} \tau h(\omega) & \text{, if } \omega \geq 0, \\ (1-\tau)\,h(\omega) & \text{, else.} \end{cases} && \text{(Quantile Loss)} \\
h(\omega) &= \begin{cases} \frac{1}{2}\omega^2 & \text{, if } -\varepsilon \leq \omega \leq \varepsilon, \\ \varepsilon\left(|\omega| - \frac{1}{2}\varepsilon\right) & \text{, else.} \end{cases} && \text{(Huber Loss)} \\
m(\boldsymbol{X}) &= \sum_{i=1}^{N}\sum_{t=1}^{\theta-2} \max\left(0; Q_{\frac{t}{\theta}}(y_i|\boldsymbol{x}_i) - Q_{\frac{t+1}{\theta}}(y_i|\boldsymbol{x}_i)\right). && \text{(Monotonicity Penalty)}
\end{aligned} \tag{2.5}$$

where $\theta - 1$ is the number of quantiles which are estimated and $\frac{1}{\theta}, \frac{2}{\theta}, ..., \frac{\theta-1}{\theta}$ are corresponding quantile levels. The target function to minimize via the QRNN is defined as the sum of quantile and loss for all data points $i = 1, ..., N$ plus the monotonicity penalty which punishes non-(strict) monotonic behaviour for every data point:

$$L = \sum_{t=1}^{\theta-1}\sum_{i=1}^{N} \rho_{\frac{t}{\theta}}^{QRNN}(y_i - Q_{\frac{t}{\theta}}(y_i|\boldsymbol{x}_i)) + m(\boldsymbol{X}) \tag{2.6}$$

For Equation (2.6) we use $\varepsilon = 0.0001$. Overall, we find that the fit does not depend on ε.

Feature Importance

Neural networks have become widespread in finance over the past years. However, one main issue is still the lack of interpretability. We use approaches based on Horel et al. (2018) and the extensions in Nagl (2021) to open up these black boxes. The approaches focus on the "learned" relations of the neural network and are therefore estimated using the training data. Overall, we use three different measures to explain the QRNN. The first order feature importance $FI_\tau^{First}(x_r)$ quantifies the overall importance of an input variable $r = 1,\dots,p$. For our purpose, the first order feature importance is given by:

$$FI_\tau^{First}(x_r) = \frac{1}{C}\operatorname{sgn}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \hat{Q}_\tau(y_i|\boldsymbol{x}_i)}{\partial x_{ir}}\right)\right)\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \hat{Q}_\tau(y_i|\boldsymbol{x}_i)}{\partial x_{ir}}\right)^2} \tag{2.7}$$

$FI_\tau^{First}(x_r)$ is the feature importance of covariate x_r at quantile level τ, $\hat{Q}_\tau(y_i|\boldsymbol{x}_i)$ is the conditional quantile estimate and C is a normalizing constant that ensures $\sum_{r=1}^{p}|FI_\tau(x_r)| = 1$. The $\operatorname{sgn}(\cdot)$ operator defines the direction in which the feature drives the prediction. This feature importance employs the gradient for every covariate x_r in relation to $\hat{Q}_\tau(y_i|\boldsymbol{x}_i)$. All variables must be standardized, e.g. mean-scaling, to allow for comparison. The gradients are squared to avoid cancellations of positive and negative values. Furthermore, it sums up to 1, allowing an easy interpretation of "relative" importance. The extension in Nagl (2021) also quantifies the direction of the feature importance. This is achieved by taking the mean values of each gradient.

Additionally, we may argue that some input features have a joint impact, i.e. interacting with each other. For example, the importance of a collateral may also depend on the state of the business cycle, as in downturns bankruptcies may become more widespread. The feature importance can be extended to quantify joint impacts of features, see Nagl (2021). Additionally, we calculate the second partial derivative with respect to the same input feature to find the quantity of (single) non-linear impact. The second order feature importance $FI_\tau^{Second}(x_r)$ measures the extent of non-linear relationships of an input variable r and $FI_\tau^{Joint}(x_{rs})$ quantifies the strength of joint effects of two variables r and $s = 1,\dots,p$ (interactions).

We do not calculate the direction of impact, as the direction of joint impacts are tedious to disentangle. Rather, we are more interested in the question whether there is a joint-impact and it's potential strength.[11]

$$FI_{\tau}^{Second}(x_r) = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\frac{\partial^2 \hat{Q}_{\tau}(y_i|\boldsymbol{x}_i)}{\partial x_{ir} \partial x_{ir}} \right)^2}, \tag{2.8}$$

$$FI_{\tau}^{Joint}(x_{rs}) = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\frac{\partial^2 \hat{Q}_{\tau}(y_i|\boldsymbol{x}_i)}{\partial x_{ir} \partial x_{is}} \right)^2}. \tag{2.9}$$

If $FI_{\tau}^{Joint}(x_{rs})$ and $FI_{\tau}^{Second}(x_r)$ are close to zero in a quantile, we can negate single non-linear and joint impacts of the input variables. This leaves only a linear impact, which corresponds to the linear quantile regression.[12] Hence, we expect the quantile loss of QRNN and QR to be very similar. This allows us to explain differences in performance, as we can trace them back to $FI_{\tau}^{Joint}(x_{rs})$ or $FI_{\tau}^{Second}(x_r)$ and answer why this approach is superior and which effects (joint or second) and variables are responsible.

At the end of this section, we briefly discuss why we choose the explainability methods in Horel et al. (2018) and Nagl (2021). Regarding PDP, ICE and ALE, it is very tedious for interpret the QRNN using plots. For example, as our output consists of 99 quantiles and we use 26 variables, we would have to interpret $99 \cdot 26 = 2,574$ plots for the main importances and $\frac{26 \cdot (26-1)}{2} \cdot 99 = 32,175$ plots for interactions. The LIME and anchor approaches cannot be aggregated to global explanations and, thus, we cannot identify the (overall) main drivers of workout LGDs. The SHAP approach is from a computational perspective unfeasible in our application. Lastly, the approach in Horel and Giesecke (2020) is not valid for our quantile based definition of a neural network.

[11] The standardization with the constant C is also neglected for illustration purposes.

[12] Of course it is possible that LGD realizations are driven by higher order impacts (third or fourth order) or joint effects of more than two variables. However, our empirical results in Section 2.5 confirm that these effects are negligible.

2.5 Empirical Results

This paper compares the ability of forecasting the conditional distribution of LGDs using the QRNN. The main focus is a comparison with the linear quantile regression, for example used by Krüger and Rösch (2017). However, to give a very broad picture we include additional models. We compare our approach to beta regression, e.g., used by Yashkir and Yashkir (2013) and Gambetti et al. (2019), and fractional logit regression, e.g., employed in Bastos (2010) and Qi and Zhao (2011). Furthermore, we use regression trees used by Altman and Kalotay (2014) and Bellotti et al. (2021). Mixture models are known for their flexibility, especially for bimodal distributions. They are frequently used for LGD prediction, see Altman and Kalotay (2014) or Betz et al. (2018). Hence, they are solid challengers for the QRNN.[13]

One of the main challenges using neural networks is to find a suitable architecture. In various applications, a cross-validation strategy is used to find the optimal hyper parameters. This is very common in most machine learning applications in credit risk, see e.g. Bastos (2010); Hartmann-Wendels et al. (2014); Gambetti et al. (2020). However, in credit risk not only the out-of-sample (cross-sectional), but rather the out-of-time prediction is of major concern, as emphasized by Kalotay and Altman (2017). Hence, we also address this issue in the model calibration by splitting our training data into $k = 5$ time buckets. We call this procedure „k-fold Time Validation“.[14] Furthermore, we use so-called Dropout Layers based on Srivastava et al. (2014) and L1-Regularization of parameter weights to avoid overfitting and increase the robustness of our model.

Another issue which arises for most machine learning methods is the suspicion that the hyperparameters are tuned until the method beats the benchmark model. This „optimization“ is analogous to the p-hacking problem in classical statistics. To tackle this issue, we calculate the Spearman correlation coefficient between the estimated in-sample target function, defined in Equation (2.6) and for the two out-of-time samples in each k-fold Time Validation. If this coefficient is statistically different from zero and positive, it means that a reduction of in-sample target function implies a reduction of the out-of-time target function as well.

[13] We would like to thank the two anonymous referees for suggesting this comparison, which has substantially improved our paper.

[14] We also evaluate another validating strategy by subsequently filling up the training set and validating on the next time bucket. For example in the first run, time bucket 1 is used for training and the model is validated on time bucket 2. Next, time bucket 1 and 2 are used for training and the model is evaluated on time bucket 3 and so on. The final model and all evaluation metrics are very similar.

We choose $\tau = 0.01, 0.02, ..., 0.98, 0.99$ leading to 99 quantile estimates for each observation in our dataset.[15]

The QRNN network is special as the architecture has more output neurons (quantiles) than input neurons (features). Therefore, we choose a so-called "baseline" structure, which ensures that the number of nodes in the hidden layers increases from one to the other, adopting the strategy in Gu et al. (2020). The baseline of the first hidden layer contains eight neurons, whereas we use 16 neurons for the second hidden layer. The largest configuration with a multiple of eight would result in 64 neurons in the first and 128 neurons in the second hidden layer. We opt against using more than two layers to avoid the vanishing gradient problem, see Glorot and Bengio (2010). It is well known that from the universal approximation theorem, following Cybenko (1989) or Hornik (1991) among others, that a single layer neural network can approximate any continuous function. However, Rolnick and Tegmark (2018) show exemplarily that adding more layers is more efficient with the same approximation property. We use sigmoid and tanh as activation functions.

Table 2.B.1 in Appendix 2.B provides the selected ranges for the hyperparameters of the QRNN and a more detailed description of the hyperparameter.

Table 2.2: 5-fold Time Validation | Final values

Parameter	USA	Europe
Learning Rate	0.001	0.001
Dropout	0.20	0.20
Multiple	1	2
L1 Loss	0.005	0.005
Hidden Layer	2	1
Activation	tanh	sigmoid
Epochs	150	100
Loss based on Equation (2.6)| In Sample	7.2451	6.9115
Loss based on Equation (2.6)| GFC	10.8921	10.3597
Loss based on Equation (2.6)| After GFC	11.3331	7.2016

Note: The table shows the final values of the hyperparameter search. For each sample, an independent gird-search is employed. The results are comparable, although we observe differences in the number of hidden layers and the multiple.

Table 2.2 shows the resulting architecture in both regions which are very similar. For example, in both subsamples a *learning rate* of 0.001 provides the best alternative. A more interesting determinant is the *multiple*, which controls the shallowness and therefore to some extent, the complexity of the neural network. In both regions, a rather small value is selected.

[15] We also tried finer splits such as $0.001, 0.002, ..., 0.998, 0.999$, but did not find any substantial differences to the outcome of this analysis. If readers are interested in these results, please contact the corresponding author.

This may indicate two things. First, the 5-fold Time Validation ensures that the complexity does not go off the rails and second, that there is no need for extremely broad networks.

In the US American sample, a network with two *hidden layers* seems to be best, whereas a less deep and more shallow network fits better to the European sample. Comparing the in-sample with the out-of-sample target functions, we recover the discrepancies in the distributions. The value in the US American sample increases along the time line. On the contrary, the highest value in the European sample appears in the GFC, whereas the in-sample and After-GFC numbers are very similar. To investigate whether our validation strategy is robust, we provide the Spearman's ρ between in-sample and the two corresponding out-of-sample target functions for the 100 best specifications, see Table 2.3. For each 5-fold Time Validation subsample, a model is fitted based on four folds and the target function of the remaining fold is calculated. Subsequently, we calculate the target function values for the two out-of-time periods. All three values are stored for each parameter combination and averaged over the five repetitions. A positive value indicates that a very good model based on our validation strategy probably performs well in the out-of-time sample. For both out-of-time periods, we obtain ρ statistically different from zero with a positive sign.

Table 2.3: Spearman correlation coefficient of the top 100 target functions

(a) USA

	IS	GFC	After GFC
IS	1	0.28^{**}	0.34^{***}
GFC		1	0.33^{**}
After GFC			1

(b) Europe

	IS	GFC	After GFC
IS	1	0.34^{***}	0.40^{***}
GFC		1	0.24^{***}
After GFC			1

Note: The table shows the estimated Spearman's ρ. We test the hypothesis $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$. $***,**,*$ means statistically significant at the 1%, 5%, and 10% levels, respectively. We use a rank-based correlation metric as we are more interested in the similarity of the hierarchical structure of the validation task than the correlation in losses.

The majority of the literature focuses on mean predictions. For example, the standard coefficient of determination R^2 and the Pearson ρ are common measures to evaluate the fit of a model. They only consider the mean, not capturing the distributional characteristics of workout LGDs. However, to be in line with the literature, we report R^2 and $\rho_{Pearson}$ only for the sake of completeness and evaluate the goodness of fit for the entire distribution using more appropriate measures.

Table 2.4: Goodness of fit based on mean predictions

(a) European Sample

Method	R^2			$\rho_{Pearson}$		
	In Sample	GFC	After GFC	In Sample	GFC	After GFC
Quantile Regression	0.109	0.036	0.048	0.282	0.096	0.164
Quantile Neural Network	**0.145**	**0.067**	0.061	**0.363**	0.151	0.227
Gaussian Mixture Model	0.139	0.029	0.041	0.238	0.124	0.233
Regression Tree	0.114	0.051	0.048	0.227	**0.164**	0.262
Beta Regression	0.099	0.052	**0.096**	0.293	0.081	**0.278**
Fractional Logit Regression	0.050	0.019	0.022	0.297	0.089	0.272

(b) US Sample

Method	R^2			$\rho_{Pearson}$		
	In Sample	GFC	After GFC	In Sample	GFC	After GFC
Quantile Regression	0.068	0.032	0.024	0.236	0.045	0.081
Quantile Neural Network	0.070	0.038	0.043	0.245	**0.096**	0.086
Gaussian Mixture Model	**0.079**	0.037	**0.080**	**0.285**	0.073	**0.091**
Regression Tree	0.040	0.040	0.031	0.181	0.012	0.075
Beta Regression	0.077	**0.041**	0.030	0.233	0.066	0.057
Fractional Logit Regression	0.022	0.015	0.017	0.234	0.041	0.065

Note: This table shows on the left hand side the R^2 of the mean predictions for each method. Furthermore, on the right hand side the Pearson correlation coefficient between mean estimates and the true realizations of the LGD is displayed. The best value for each metric in each sample is indicated in bold. The mean predictions of the quantile regression and quantile neural network are calculated by taking the expectation over the estimated quantiles for every obligor. The number of components for the Gaussian Mixture model is chosen according to the lowest AIC on the training data, following Altman and Kalotay (2014). For both regions, three components fit best, which is in line with the literature, see e.g. Krüger and Rösch (2017). We optimize the hyperparamter (maximum depth, minimum samples required for the split and the minimum number of samples in a leaf node) of the regression tree using our Time Validation approach. To apply the beta regression and factorial logit regression, we transform the LGD values outside $[0,1]$ with $LGD_{[0,1]} = \frac{LGD - min(LGD)}{max(LGD) - min(LGD)}$, following Krüger and Rösch (2017) and Altman and Kalotay (2014). In general, the QRNN shows competitive results in Europe, but the Gaussian Mixture Model seems to be more appropriate in the US if one focuses only on mean estimates, neglecting the fit over the full conditional distribution.

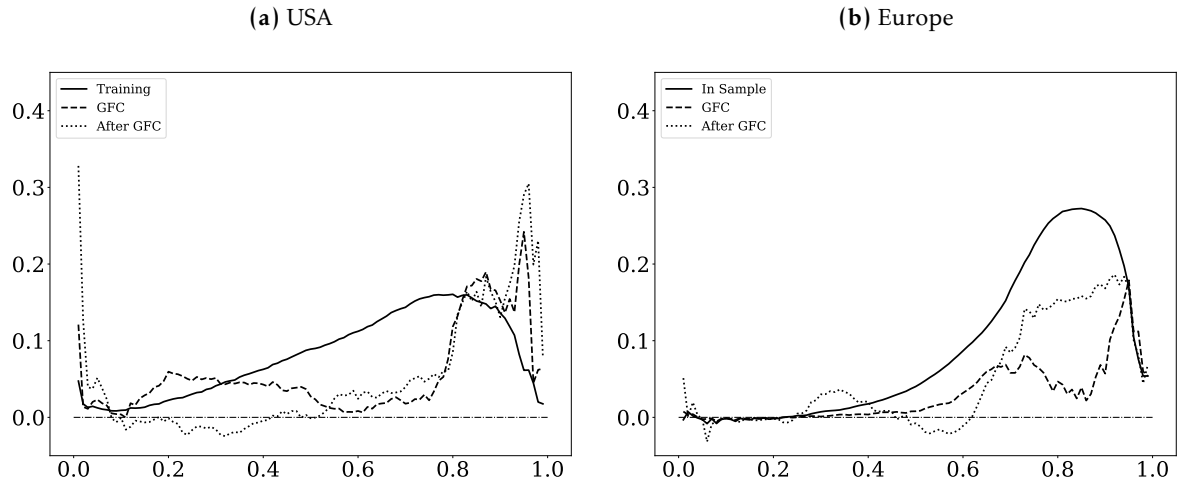
From Table 2.4, we can argue that the QRNN, although not constructed for mean predictions, provides competitive results in both regions. In the European dataset, the QRNN achieves the highest R^2 in sample and in the GFC, while the second best value after the GFC can be obtained. With respect to the $\rho_{Pearson}$ values, we achieve the highest correlation only in sample. In the US, the QRNN provides competitive results, but not the overall best performance. The results in the US American sample are in line with Krüger and Rösch (2017), that the Gaussian Mixture Model is a solid challenger and for some samples even outperforms the linear quantile regression. Overall, we can summarize that the QRNN provides also reasonable mean estimates. However, we would like to stress the point that a proper evaluation should be based on the whole distributional fit and not on one location parameter only.

To compare the distributional fit, we evaluate values from the quantile loss function in Equation (2.4) to determine which quantile predictions fit best. We propose this as the best way to compare model performance, as the lower the loss is, the better the conditional quantile estimates are. To quantify the improvement compared to the linear quantile regression, we focus on the relative values of Equation (2.4):

$$1 - \frac{\sum_{i=1}^{N} \rho_\tau \left(y_i - \hat{y}_{\tau,i}^{Alternative} \right)}{\sum_{i=1}^{N} \rho_\tau \left(y_i - \hat{y}_{\tau,i}^{QR} \right)}. \tag{2.10}$$

We estimate Equation (2.10) for every quantile, illustrated in Figure 2.3. The performance of the alternative model is better, if the value is greater than 0, e.g. a value of 0.3 means that precision of the quantile forecasts is 30% higher compared to the linear quantile regression. Focusing on the comparison of the QRNN and QR, values greater than zero can directly be ascribed to existing non-linear relationships and joint impacts between LGDs and covariates at this quantile level. This interpretation seems reasonable, as non-linearity and interactions are the only differences between both models. To exclude the chance that better results stem from the monotonicity constraint, we also estimated a linear QRNN model using no hidden layers, obtaining similar results.

Figure 2.3: Loss over quantiles



Note: These figures show the relative quantile-specific loss based on Koenker and Bassett (1978) for each quantile. The solid line refers to the training sample, the dashed line GFC and the dotted line to the after GFC sample. A value greater than zero indicates a better distributional fit of the QRNN.

The solid line represents the in-sample comparison, whereas the dashed line refers to the GFC and the dotted line to the After GFC period. In the US American sample, the precision improvement increases almost monotonically up to the 80% quantile. Hence, for higher LGD realizations more non-linearity and joint effects are present. The difference between QR an QRNN is up to 15%. If we focus on the out-of-time periods, we observe two interesting things.

First, we see steady values up to the 80% quantile, but afterwards the superiority of the QRNN yields over 30% improvement. This means, that more extreme realizations in the out-of-time perspective can be better predicted by the QRNN. For the lowest quantile we have a superior performance of the QRNN as well. On the contrary, for lower quantiles from 20% to 40% in the After GFC sample, the QR is slightly better with relative values of about -2.0%. In the European sample, we observe slightly different results. Both methods perform similar in-sample up to the median. Afterwards, the improvement of the QRNN increases sharply to values of around 25%. From the out-of-time perspective, the improvements start from the 60% quantile with values around 12%. Similar to the US sample, in the middle quantile in the After GFC sample the linear quantile regression has a small edge of -1.5% over the QRNN. In summary, Figure 2.3 shows that the QRNN clearly outperforms in-sample the linear quantile regression, especially for high LGD realizations. In both out-of-time samples, we observe the superiority for high LGD realizations as well. Yet, there are some quantiles in the After GFC sample, in which we do not outperform the QR. Nevertheless, for the vast majority of quantiles and most importantly for high realizations, the QRNN substantially outperforms the QR. To provide evidence, that the QRNN is indeed a reasonable extension of the linear quantile regression, we compare our results to other common methods in the literature. Table 2.5 shows the average improvement over the full conditional distribution compared to the linear quantile regression. Overall, we can summarize that the QRNN provides the largest improvement, which may be attributed to non-linear and joint effects. More interestingly, we recover the evidence of Figure 2.3 that especially on the right tail, i.e. for $\tau \in (0.75, 0.99)$, the QRNN performs very well. We also recover findings of previous studies, that the Gaussian Mixture Model performs reasonably well and in some cases outperforms the linear quantile regression, see e.g. Krüger and Rösch (2017).

Figure 2.4 illustrates a random sample of 100 different LGD distributions of the training sample. The left-hand side shows the QRNN and on the right-hand side the QR. Comparing both methods in the US, we see the impact of the monotonicity penalty in Equation (2.5). The QRNN approach yields monotonic quantiles, whereas the linear model shows heavily deviating quantiles. This is also true for the European dataset. However, for the outer right tail, e.g. from 95% to 98%, the monotonicity is not necessarily met in the QRNN approach.[16] An easy solution to the small range of non-monotonic quantiles could be the approach by Chernozhukov et al. (2010), i.e. simply rearrange the quantile to ensure monotonicity.

[16] We tested various other specifications of the penalty and increased the weight of the monotonicity loss in the optimizing problem and increased the number of estimated quantiles. Nevertheless, compared to our benchmark model, the estimated distributions look much more reasonable.

Table 2.5: Loss over quantiles in comparison to other methods

(a) European Sample

Method	Full range			Right tail		
	In Sample	GFC	After GFC	In Sample	GFC	After GFC
Quantile Neural Network	**0.087**	**0.029**	**0.049**	**0.221**	**0.070**	**0.146**
Gaussian Mixture Model	0.038	-0.008	0.036	0.138	-0.016	0.092
Regression Tree	-0.279	-0.182	-0.296	0.058	0.067	0.070
Beta Regression	-0.122	-0.092	-0.151	0.085	-0.009	0.088
Fractional Logit Regression	-0.325	-0.526	-0.261	0.010	-1.393	-0.500

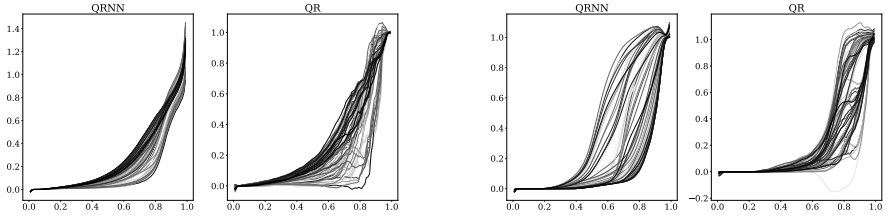
(b) US Sample

Method	Full range			Right tail		
	In Sample	GFC	After GFC	In Sample	GFC	After GFC
Quantile Neural Network	**0.080**	**0.056**	**0.049**	**0.123**	**0.134**	**0.156**
Gaussian Mixture Model	0.051	-0.010	-0.051	0.076	-0.068	-0.249
Regression Tree	-0.157	-0.125	-0.158	0.008	0.012	0.075
Beta Regression	-0.025	-0.082	-0.094	0.044	0.050	-0.177
Fractional Logit Regression	-0.208	-0.407	-0.377	-0.014	-0.526	-1.024

Note: This table shows the average values of Equation (2.10) for each method. A value larger than 0 indicates a better distributional fit compared to the linear quantile regression. The best value in each sample is indicated in bold. To put more emphasis on larger LGD realization, the quantile fit is also calculated for $\tau \in (0.75, 0.99)$, labelled as "right tail". The result shows that the QRNN outperforms all benchmark models, especially for the right tail. The quantiles of mean-focussed methods are calculated following Krüger and Rösch (2017).

Figure 2.4: CFD-Plot

(a) USA **(b)** Europe



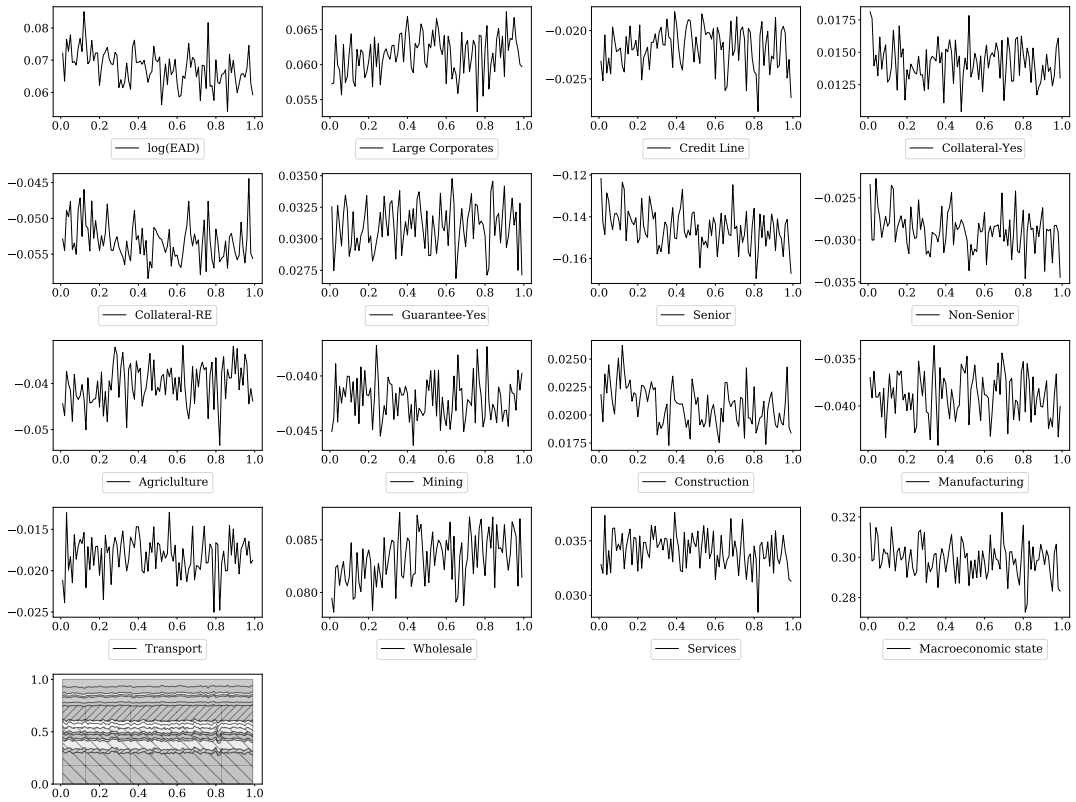
Note: These figures show a sample of 100 different estimated LGD cumulative distribution functions. The left panels show the results of the QRNN approach, whereas the right panel the linear method. Within every region, we select the same loans to allow for comparison. By definition, a cumulative distribution function must be monotonic in quantiles, which is guaranteed by the QRNN to a large extent. On the contrary, the linear quantile regression shows a non-monotonic behaviour for a wide range.

Figure 2.5 displays the estimated values for $FI_{\tau}^{First}(x_r)$. To the best of our knowledge, this paper is the first to disentangle the great flexibility of neural networks in single and joint impacts. We estimate the feature importance $FI_{\tau}^{First}(x_r)$ (vertical axis) for each quantile τ (horizontal axis) to discover the impact on each quantile. The sum of all absolute feature importances $\sum_{r=1}^{p} |FI_{\tau}^{First}(x_r)|$ sums up to 1 in each quantile. For both regions, all feature importances are aggregated in a stack plot to give a comprehensive overview. The ordering coincides with the

sequence of the displayed plots. We find that the impact of all feature importances is stable, but for some we find differences especially in the tails of the conditional distribution.

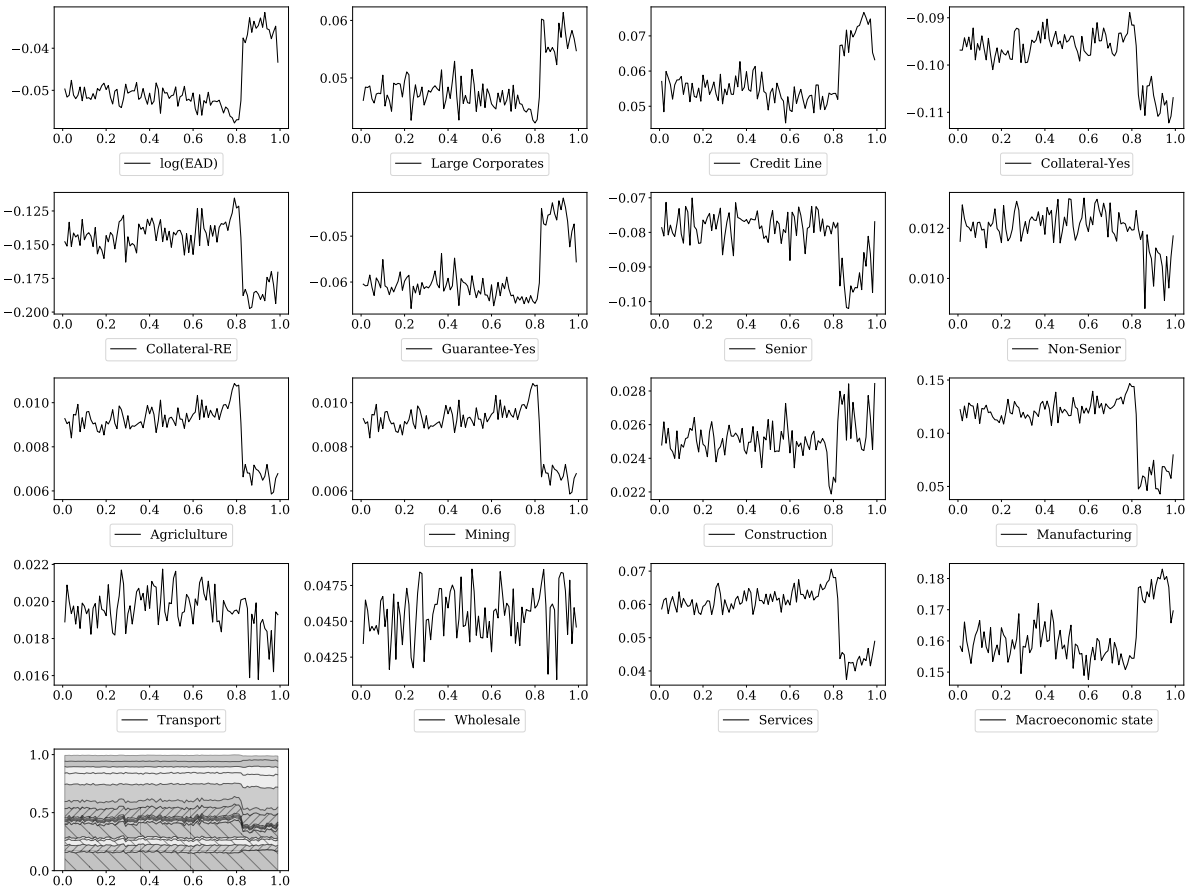
Figure 2.5: Feature Importance

(a) USA



Concerning the US, a positive value for log(EAD) means that the larger the log(EAD) of the loan, the higher the resulting quantile forecasts. The values around 0.07 indicate that the log (EAD) has a share of 7% of the total importance. In general, the higher the absolute value, the more important the feature is for quantile forecasts. Economically, the positive values of $FI_{\tau}^{Frist}(x_r)$ for log(EAD) indicate that banks potentially give credit to companies who cannot finance on capital markets and, hence, need larger loans from banks. As these companies may have a higher tendency to default, we observe higher LGDs. This coincides with the positive feature importance of large corporates. Another interesting feature is Collateral- RE with a negative effect. This means if a loan is protected by a real estate collateral, the LGD is lower. The single most important loan specific feature is the dummy variable Senior. The reference group is pari-passu, hence we can infer that if a bank is senior in the resolution process, the LGD realizations are lower. Abstracting from the estimated value, the dummy variable Senior has a share of 14% of the total importance. As mentioned in Section 2.3, we capture the systematic variation via principal components. As the direction and the exact meaning of every single

(b) Europe



Note: These plots show the estimated feature importance of every variable for every quantile. Please note that the sum of all feature importances for each quantile sums up to 1. The last plot illustrates the importance of all variables in a stacked fashion. This allows us to evaluate the overall importance for LGD prediction at a glance.

component are hard to interpret, we focus on the overall impact. Hence, the plot labelled as "macroeconomic state" displays the absolute sum of all principal components. From the absolute values of around 30% we can see that the macroeconomy is the most important determinant of workout LGDs in the United States. From these values we can infer that the macroeconomy has a quadruple impact on the LGD distribution compared to the log(EAD) (30% vs. 7%). Hence, the economic surrounding is four times more important than the loan value itself and twice as important as a senior rank in the resolution process (30% vs. 14%). As robustness, we also used a PCA with 90% and 99,5% of variance, but the overall effect is stable around 30% in any setting.[17] This result coincides with evidence that US American LGDs are very cyclical and the macroeconomy is an important driver, see Tobback et al. (2014); Betz et al. (2018); Nazemi and Fabozzi (2018).

[17] The conclusions about the importance of the macroeconomy in Europe and USA are similar when using less (more) principal components. Hence, the high importance comes not from the number of components used in the analysis. This may also seen as a robustness of our employed feature importance measure, as the overall importance does not seem to be sensitive to the number of components. The results with less (more) components are available upon request.

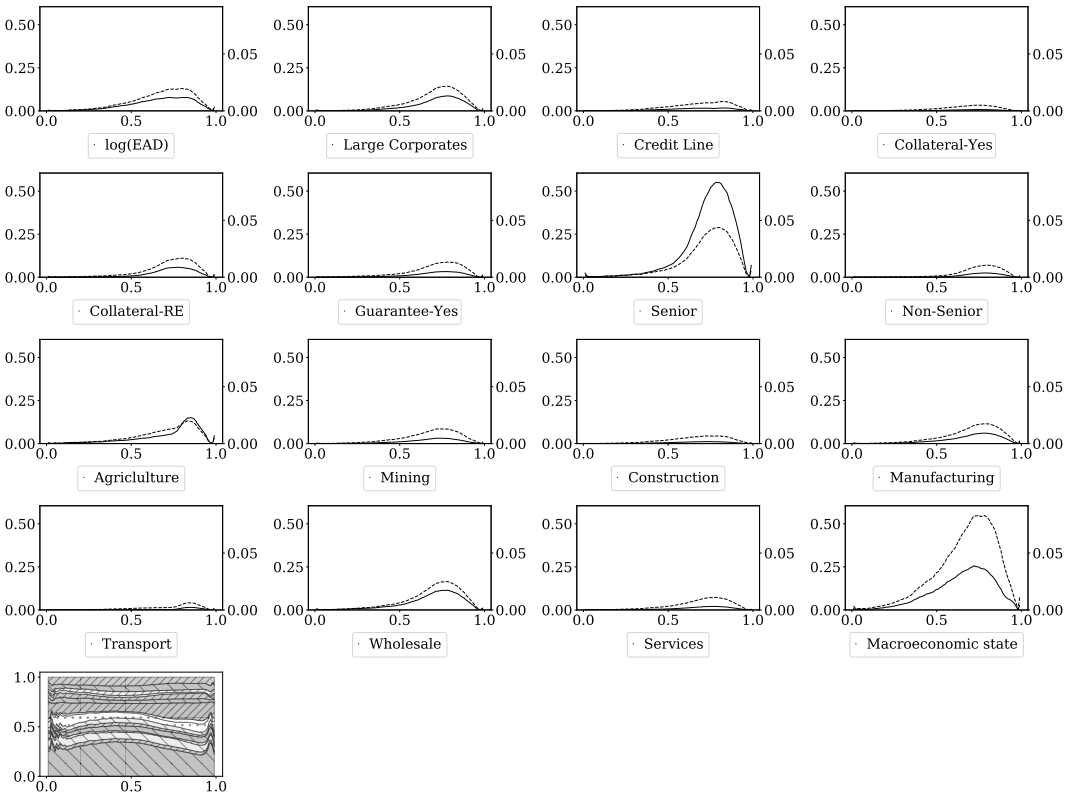
The results are somewhat different in the European sample, in which the single most important determinant is the protection of the loan, i.e. the combined impact of the collaterals of around 30%. Both feature importances have a negative sign, indicating that a protection decreases the losses of banks. If the loan is protected by real estate, the resulting LGD is decreased considerably, especially in high quantiles. Regarding the impact of the macroeconomy we find a considerable impact on the LGD prediction. Interestingly, the impact increases in quantiles, which indicates that especially for higher LGD values, the economy is more important. We find that the collateralization has a importance twice as high compared to the macroeconomy (30% vs 15%) and roughly four times higher than the senior rank in the resolution process (30% vs 7%). Moreover, we observe a shift in importance for quantiles above 80% to the macroeconomy, facility asset class, collaterals, high seniority and credit lines. This can be seen as evidence that high LGD realizations are driven by these five key drivers. Their relative importance sums up to roughly 65% over these quantiles.

Besides the first order importance, the second order and joint impacts are also interesting. The following analysis helps us to understand why the QRNN is superior in some quantiles and which variables drive this superiority. Figures 2.6 and 2.7 show the absolute sum of all pairwise interactions with all other variables for every input variable, $FI_{\tau}^{Joint}(x_{rs})$, illustrated by the dashed line and the second order effect, $FI_{\tau}^{Second}(x_r)$, via the solid black line. The last subplots in both figures show the relative importance of both, second order and joint impact, for every variable. The left y-axis corresponds to the values of $FI_{\tau}^{Joint}(x_{rs})$ and the right y-axis to $FI_{\tau}^{Second}(x_r)$.

From the last plot in Figure 2.6 for the United States, we can see that the principal components account for roughly 35% of these effects. The dummy variable Senior also shows considerable joint effects with all other variables. However, the macroeconomic state has a value roughly twice as high for joint effects and seems to be the overall most interacting determinant.[18] Regarding the second order effects, we observe the highest value with the variable Senior. This indicates that the rank in the resolution process has a strong non-linear relation to higher quantiles. The remaining variables show no substantial effects. Overall, we can see that the joint effects are clustered into two important drivers. Furthermore, important information can be obtained from quantiles, where these effects are close to zero. If we compare these ranges with the relative target functions in Figure 2.3, we can see that these coincide with the quantiles where QR and QRNN have a similar target function. Hence, we can see why the QRNN does not outperform the QR in these quantiles. There are simply no joint or second order effects for

[18] Again, the drawn conclusions for Europe and USA about second order and joint effects are similar when using less (more) components in the macroeconomic state. The results are available upon request.

Figure 2.6: $FI_{\tau}^{Joint}(x_{rs})$ and $FI_{\tau}^{Second}(x_r)$ | US sample

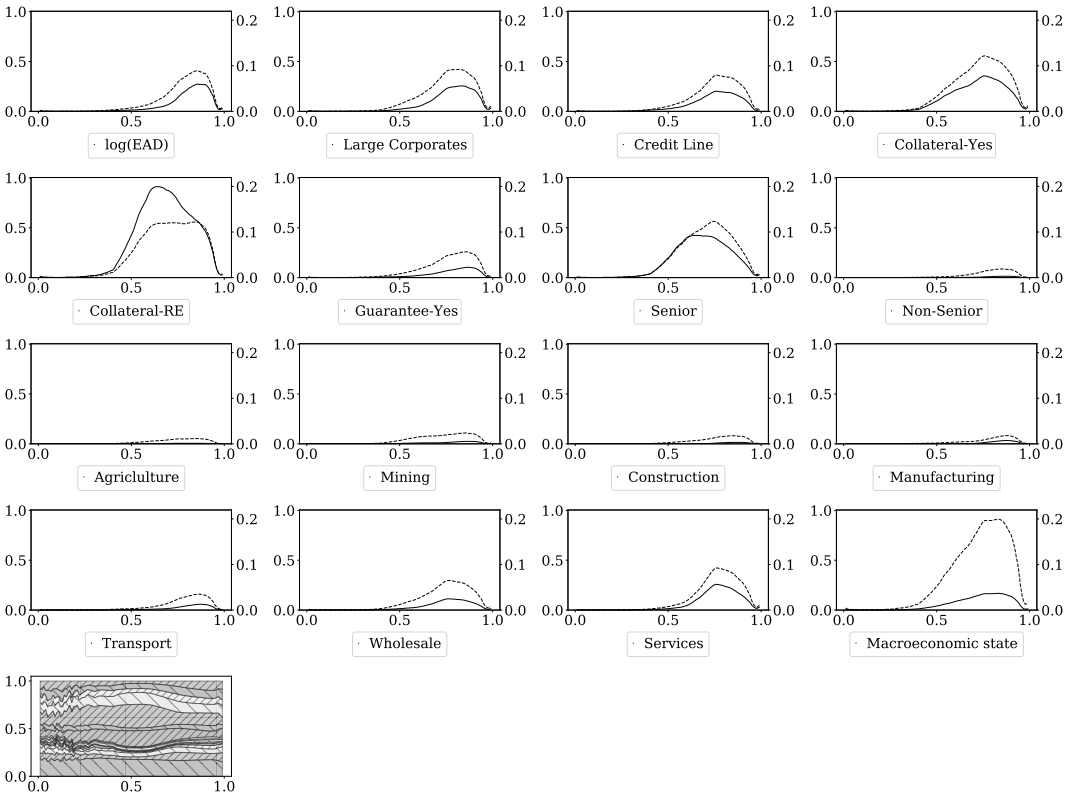


Note: These plots show the estimated importance of every variable for every quantile. The horizontal line shows the different quantiles, whereas the vertical line illustrates the importances. The left y-axis and the dashed line refers to $FI_{\tau}^{Joint}(x_{rs})$, whereas the right y-axis and the solid line refers to $FI_{\tau}^{Second}(x_r)$. This allows us the evaluation of the overall joint effects for LGD prediction at a glance.

any input variable. Hence, we conjecture that there is probably only a linear impact, very well described by the linear quantile regression.

A similar picture can be seen in Europe in Figure 2.7, where we have almost zero values for joint and second order effects for quantiles up to the median. This coincides with the similar performance regarding the target function, outlined in Figure 2.3. Regarding joint effects, we find two important things. First, the overall magnitude is higher in Europe, illustrated by the lager range of the vertical axes. This means that there are overall higher joint effects in Europe than in the United States. Second, the joint effects are far more distributed across the different input variables. Hence, we observe more than two main drivers of joint effects. Similar to the United States, the macroeconomy shows the highest value of joint effects, but in comparison to other variables the relative magnitude is not as big as in the United States. Especially, if we sum up the values for the collateralization, we get very similar numbers. With respect to second order effects, the collateralization shows the largest non-linear impact on the LGD distribution for higher quantiles.

Figure 2.7: $FI_{\tau}^{Joint}(x_{rs})$ and $FI_{\tau}^{Second}(x_r)$ | European sample



Note: These plots show the estimated importance of every variable for every quantile. The horizontal line shows the different quantiles, whereas the vertical line illustrates the importances. The left y-axis and the dashed line refers to $FI_{\tau}^{Joint}(x_{rs})$, whereas the right y-axis and the solid line refers to $FI_{\tau}^{Second}(x_r)$. This allows us the evaluation of the overall joint effects for LGD prediction at a glance.

Concluding, we find that the macroeconomy has the highest joint effects with other variables, closely followed by the collateralization. Finally, we analyse which variables have the largest joint effects with the macroeconomic state. Therefore, Figures 2.C.1 and 2.C.2 in Appendix 2.C show the absolute sum of interactions for all eight, respectively eleven, components for every remaining variable. Thus, we can disentangle the dashed line of the macroeconomic state in Figures 2.6 and 2.7 into its elements. For the sake of clarity, we moved these figures to the Appendix. In the United States, the variable Senior has the largest joint effects with the macroeconomy, especially for higher quantiles. This may be plausible as in economic downturns the rank in the resolution process is very important as the overall proceeds from bankruptcies may decrease and, thus, there is nothing left for lower ranks in the resolution process. In the European dataset, we observe that the largest joint effects are between the macroeconomy and the collateral of the loan. From an economic perspective, we can argue in the same way as for the seniority in the United States. In downturns it is probably more important to have a collateral for the loan as the proceeds from bankruptcies, excluding collaterals, might be lower and, thus, it probably reduces the losses faced by the bank. Overall, we find that collateralization and the

macroeconomy have the highest joint effects on the LGD distribution.

An interesting and actively debated topic is whether machine learning methods can be used for regulatory purposes. This question cannot be answered yet, as no final regulations are published. However, numerous workshops and discussion paper have emerged over the last years, see e.g. Basel Committee on Banking Supervision (2019a); Deutsche Bundesbank (2020); Paulsen et al. (2021). For a detailed discussion of various aspects of applying machine learning fo regulatory purposes, we refer to Fritz-Morgenthal et al. (2021). In general, explainability methods play a major role in this discussion. For example, Deutsche Bundesbank (2020) state that explainability methods are a promising answer to black-box models, making them less opaque and more comprehensible. From this standpoint, the presented feature importance in this paper is a step into the right direction, as we do not only obtain the main (first-order) importances, but also second-order and joint importances. This may be helpful in the discussion with regulators, as we are able to explain why the machine learning method is superior (non-linearity and joint effects) and we can attribute this to specific variables (macroeconomy and collateralization).

Another important regulatory aspect is the ability to generate downturn estimates, which is extensively researched, see e.g. Calabrese (2014); Krüger and Rösch (2017); Betz et al. (2018). We argue to use realizations of the PCA components during an economic downturn as a reasonable adverse scenario. To do so, we randomly sample 10,000 portfolios containing 500 obligors each, defaulted from 2010 to 2016. We use the QRNN to predict the conditional distribution of every obligor, calculate their expected LGD using the mean over the conditional distribution and finally aggregate this on portfolio level using the mean over all obligor's expected LGDs. Formally defined as:

$$\widehat{LGD}^{Portfolio} = \frac{1}{500} \sum_{i=1}^{500} \left(\frac{1}{99} \sum_{t=1}^{99} Q_{\frac{t}{100}}(y_i | \boldsymbol{x}_i) \right) \tag{2.11}$$

In the baseline scenario we use the true PCA component realization, whereas in the adverse scenario we use the PCA component realization during a quarter of economic downturn. Table 2.6 shows the mean of Equation (2.11) over all 10,000 simulated portfolios.

Table 2.6: Downturn estimates

Quarter	USA			Europe		
	Baseline	Adverse	Increase	Baseline	Adverse	Increase
	0.271			0.201		
2007 Q4		0.338	24.78%			
2008 Q1		0.325	19.93%		0.330	64.17%
2008 Q2		0.346	27.68%		0.313	55.72%
2008 Q3		0.345	27.31%		0.340	69.15%
2008 Q4		0.323	19.19%		0.343	70.65%
2009 Q1		0.344	26.94%		0.401	99.50%
2009 Q2		0.346	27.68%		0.385	91.54%
2009 Q3					0.362	80.10%

Note: This table shows the expected LGD value assuming the PCA realization in the given quarters. The quarters are chosen according to the OECD crisis indicator, which indicates that the GFC lasts from 2007 Q4 to 2009 Q2 in the USA, whereas it is slightly shifted in Europe (2008 Q1 to 2009 Q3). The values are obtained by randomly drawing 10,000 portfolios containing 500 obligors each. The column *Baseline* represents the average of the predicted LGD value using the true PCA realizations for each obligor and, thus, is identical over the quarters. The column *Adverse* reports the average predicted LGD value assuming the PCA realizations of the given quarter for all obligors. The column *Increase* shows the percent increase of the mean LGD value of the adverse scenario compared to the baseline values. A positive value indicates larger LGDs due to the adverse PCA realizations.

Although we did not calculate the direction of impact for the PCA components in Figures 2.6 and 2.7, we can now clearly see how the macroeconomic state impacts the predicted LGDs. In the US American sample, the estimated LGDs over all 10,000 portfolios increase by roughly 28% reflecting downturn conditions. The increase in the adverse scenario is even more pronounced in the European dataset. This can be attributed to the fact, that the difference between average LGDs in the GFC and after the GFC is much more pronounced than in USA. Hence, we see a downturn LGD that almost doubles the baseline LGD. Admittedly, our strategy relies on historical economic downturns and is difficult to extend to unprecedented scenarios. Nevertheless, this analysis shows that the QRNN is able to provide reasonably high downturn estimates, reflecting the impact of the macroeconomic state during economic downturns.

Summarizing, the contribution of this paper is a straight forward extension of the economically useful method of quantile regression into several important directions. Furthermore, we compare the QRNN to various other methods and obtain superior results regarding quantile specific forecasts. The neural network approach shows a considerably better distributional fit for the whole LGD distribution, especially in the out-of-time perspective with an improved precision of the quantile forecasts of up to 30%. We find that classical quantile regressions provide non-monotone estimates of the LGD's conditional distribution, contrary to the QRNN. With respect to the important drivers of LGDs, we see diverging determinants in USA and

Europe. In the US, the macroeconomic state is the single most important determinant. Seniority and the economic surrounding have the largest joint effects. The European LGD predictions rely more on loan-specific characteristics, such as different types of collateral. However, the largest joint effects are driven by the macroeconomy and the collateralization.

2.6 Conclusion

Recent literature shows that modelling the entire LGD distribution or quantiles thereof is more adequate than focussing just on means, and linear quantile regression outperforms ordinary least squares, fractional response regressions, beta regressions, regression trees and finite mixture models (see Krüger and Rösch (2017)). We extend this approach by allowing non-linearities and interactions in quantiles accomplished by the Quantile Regression Neural Network. This approach considerably enhances the modelling flexibility. The additional flexibility pays off in terms of a better distributional fit for in- and out-of-time samples with an improvement of up to 30% in quantile forecast precision compared. Machine learning models are prone to the conjecture that the researcher tries many different combinations and only reports the best, without ensuring that there is a broad superiority and to some extent robustness with respect to the architecture. We alleviate this problem by reporting positive Spearman's ρ, indicating that a good model in-sample is also a good model out-of-sample. Furthermore, we show that a monotonicity constraint can easily be implemented and standard linear quantile regression does not ensure monotonously increasing distribution functions. This also allows banks to use the QRNN on a loan level basis. To the best of our knowledge, this paper is the first in credit risk to disentangle the impact of variables in neural networks in a highly interpretable fashion. The first order feature importance measure in Horel et al. (2018) and Nagl (2021) allows us to quantify the relative importance of all features and calculate the direction of their impact. We find that macroeconomic variables account for up to one third in the US American sample, underlining the dependency of LGDs on the economic surrounding. On the contrary, the largest first order feature importance in the European dataset were collaterals and hence, loan characteristics. The macroeconomy accounts for only 10-15% of the overall importance. Therefore, we document highly diverging impacts with respect to the macroeconomy in these two regions. This may give further evidence that systematic behaviour, expressed by macroeconomic variables, is clearly different in the US and Europe. By using the second order and joint impact feature importance measure, we can see why the QRNN outperforms its counterpart. We quantify strong joint effects of the macroeconomy with other variables as the main driver of the superiority. The

contributions of this paper may have important implications for credit risk management, as in Europe and the United States the QRNN approach provides a higher precision in terms of quantile forecasts especially for higher quantiles. This suggests non-linear behaviour in quantiles of high LGD realizations. Furthermore, the empirical findings of high dependency of US American LGDs on the macroeconomy may have serious implications for banks and regulators to carefully account for this large impact. This points towards a highly cyclical behaviour of LGDs, which may result in higher losses of the US American banking system, especially in crisis periods.

Machine learning methods are often seen as problematic due to their black-box character, particularly from a regulatory perspective. The introduced feature importance measures are easy to implement and interpret and may enhance the adoption of machine learning approaches

Acknowledgements

We thank participants of the 14th International Conference on Computational and Financial Econometrics 2020 in London and the Global Credit Data European Conference 2021 for fruitful discussions and helpful comments. Furthermore, we thank two anonyms referees for their comments which have substantially improved the paper.

2.A Macroeconomic variables for the principal component analysis

Table 2.A.1: Macroeconomic variables for US American loans

(a) USA

Variable	Source
Economic Political Uncertainty | Three Component Model	https://www.policyuncertainty.com
Economic Political Uncertainty	https://www.policyuncertainty.com
Financial Stress Indicator	https://www.policyuncertainty.com
US Equity Market Volatility Index	https://www.policyuncertainty.com
Geopolitical Risk Index	https://www.policyuncertainty.com
Economic Uncertainty Related Queries	https://www.policyuncertainty.com
Financial Uncertainty Index	https://www.sydneyludvigson.com
Macroeconomic Uncertainty Index	https://www.sydneyludvigson.com
Real Uncertainty Index	https://www.sydneyludvigson.com
Unemployment Rate	https://fred.stlouisfed.org
Real Gross Domestic Product (yoy growth)	https://fred.stlouisfed.org
S&P/Case-Shiller U.S. National Home Price Index (yoy growth)	https://fred.stlouisfed.org
Industrial Production: Total Index US	https://fred.stlouisfed.org
Consumer Price Index for All Urban Consumers: All Items in U.S. City Average	https://fred.stlouisfed.org
CBOE Volatility Index: VIX	https://fred.stlouisfed.org
SP500	EIKON
M2 (yoy growth)	EIKON
TED Spread	https://fred.stlouisfed.org
Term Spread 10y-3m	https://fred.stlouisfed.org
Commercial and Industrial Loans, All Commercial Banks (yoy growth)	https://fred.stlouisfed.org

(b) Europe

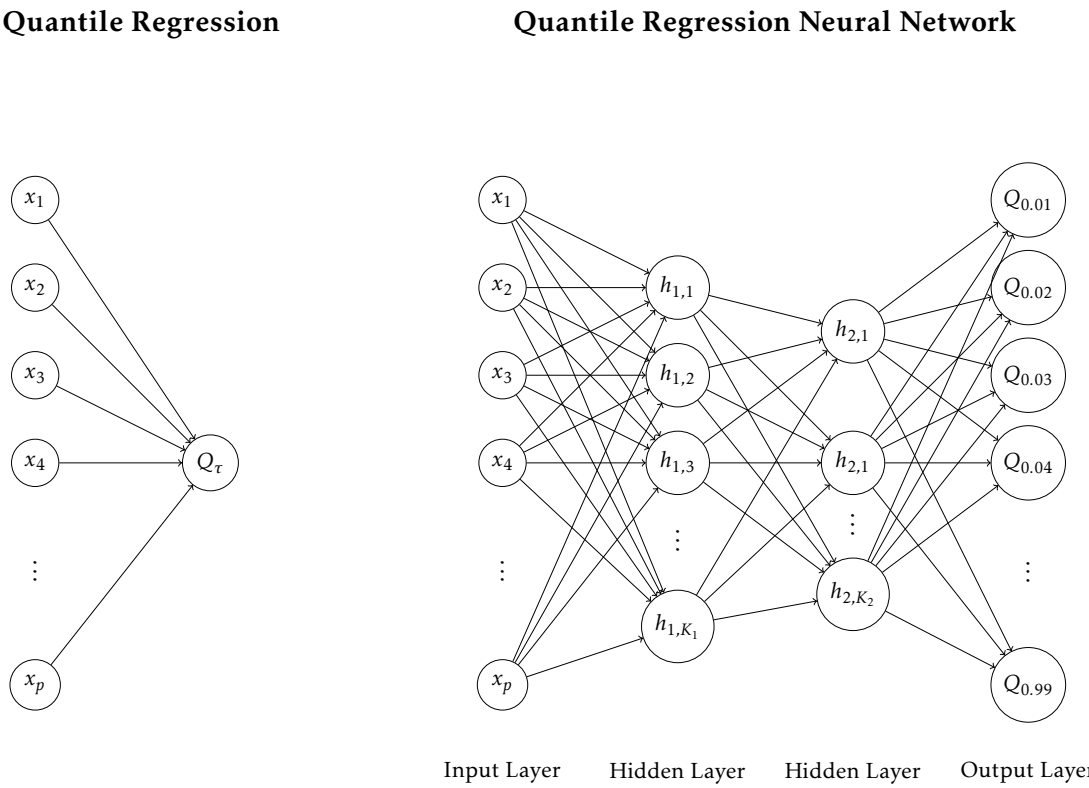
Variable	Source
Economic Political Uncertainty Europe	https://www.policyuncertainty.com
Harmonized Unemployment Rate: Total: All Persons for the Euro Area	https://fred.stlouisfed.org
Real Gross Domestic Product for Euro area (yoy growth)	https://fred.stlouisfed.org
Real Residential Property Prices for Euro area	https://fred.stlouisfed.org
Total Industry Production Excluding Construction for the Euro Area	https://fred.stlouisfed.org
Consumer Price Index: Harmonized Prices: Total All Items for the Euro Area	https://fred.stlouisfed.org
VSTOXX Europe	EIKON
EUROSTOXX 50	EIKON
M2 Europe (yoy growth)	EIKON
Total Loans to Corporate Euro Area (yoy growth)	https://www.euro-area-statistics.org
Business Survey Industry	https://ec.europa.eu/eurostat
Business Survey Construction	https://ec.europa.eu/eurostat
Economic Sentiment Indicator	https://ec.europa.eu/eurostat
Business Climate Indicator	https://ec.europa.eu/eurostat
International Trade (yoy growth)	https://ec.europa.eu/eurostat
Labor cost nominal value	https://ec.europa.eu/eurostat
Turnover in industry, total - quarterly data	https://ec.europa.eu/eurostat
Building Permits	https://ec.europa.eu/eurostat

Note: The table shows the employed macroeconomic variables for the principal component analysis. They are in line with papers in the literature concerning the estimation of workout LGDs of corporate loans and some variables which also may be suitable to account for variations of workout LGDs over the business cycle.

2.B Hyperparameter Optimization

This section elaborates in more detail how the QRNN relates to the linear quantile regression and the way we optimized the hyperparameters. Figure 2.B.1 shows the difference between the standard linear quantile regression (QR) on the left-hand side and the QRNN approach on the right-hand side.

Figure 2.B.1: Graphical overview | QR vs. QRNN



Note: This figure shows the setup of the standard linear quantile regression and the QRNN. There are three important enhancements compared to the QR. First, we model all quantiles of interest at once. Second, we can easily allow for non-linearities in each of the quantiles. Third, we provide more monotonic quantile estimates compared to the QR approach.

The standard quantile regression relates all input features directly and linearly to only one quantile of interest. To describe a full set of quantiles, one has to fit one separate linear quantile regression to each of them, resulting in 99 models overall for our empirical analysis. On the right-hand side, the QRNN approach is illustrated. We use the same set of input variables and estimate the same set of quantiles, but there are three important differences compared to the standard quantile regression. First, we model the full discrete set of quantiles at once, reducing the required models to only one single model. Second, the relation between input features and quantiles is no longer direct, but described by non-linear transformations in the hidden layers of the QRNN. This allows for all kinds of non-linearity to be present in all quantiles simultaneously.

Third, as the QRNN approach models the full set of quantiles, we can penalize non-monotonic quantile estimates, i.e. the estimated value must increase from the top ($Q_{0.01}$) to the bottom ($Q_{0.99}$). This is not possible in the standard QR approach as the models are fitted independently.

The QRNN network is somewhat special as the architecture requires that we have more output neurons (quantiles) than input neurons (features). Therefore, we choose to assume a so-called "baseline" structure, which ensures that the number of features in the hidden layers increases from one to the other. In classical applications, where only one output neuron is used, e.g., predicting probability of default or market returns, it has turned out that reducing the number of neurons in the hidden layer by half for each additional hidden layer seems to be a robust and suitable baseline for most applications, see e.g Gu et al. (2020). As we have the opposite starting point, we double the number of neurons in each hidden layer. Another positive side effect is that we now only have to validate the multiple of this baseline structure, which makes the validation task much more efficient[19]. We use eight base neurons in the first hidden layer and 16 neurons in the second hidden layer. Furthermore, due to the vanishing gradient problem of deep neural networks, we evaluate no more than two hidden layers, but rather use a large number of neurons in these layers.

Table 2.B.1: 5-fold Time Validation over time setup

Parameter	Possible Values
Learning Rate	0.00001 / 0.0001/ 0.001/ 0.01
Dropout	0.10 / 0.20 / 0.30 /0.40
Multiple	1 / 2 / 4 / 8
L1 Loss	0.01 / 0.005 / 0.0005
Hidden Layer	1 / 2
Activation	sigmoid / tanh
Epochs	100/ 150/ 200/ 250/ 300

Note: The table shows different values for the hyperparameter search. As avoiding overfitting is of major concern, we put much emphasis on regularization parameters (L1) and different designs of Dropout Layers. With respect to the difficulties of training very deep neural networks, we do not use more than two hidden layers, but rather increase the number of neurons in each layer.

Table 2.B.1 illustrates all used parameters during the „5-fold Time Validation". We estimate all possible combinations of the parameters and also apply a 3- and 10-fold approach, but the results remain the same. Various different learning rates and the adaptive moment-based (*Adam*) optimizer of Kingma and Ba (2014) are used. We further evaluate the levels of Dropout, ranging from very low to rather high. As an activation function, we use sigmoid and tanh, which are

[19] Please note that we also used other baseline models and the classical way, e.g. using several different numbers of neurons without assuming a baseline structure. We find this increasing fashion to be the most robust and computationally efficient one.

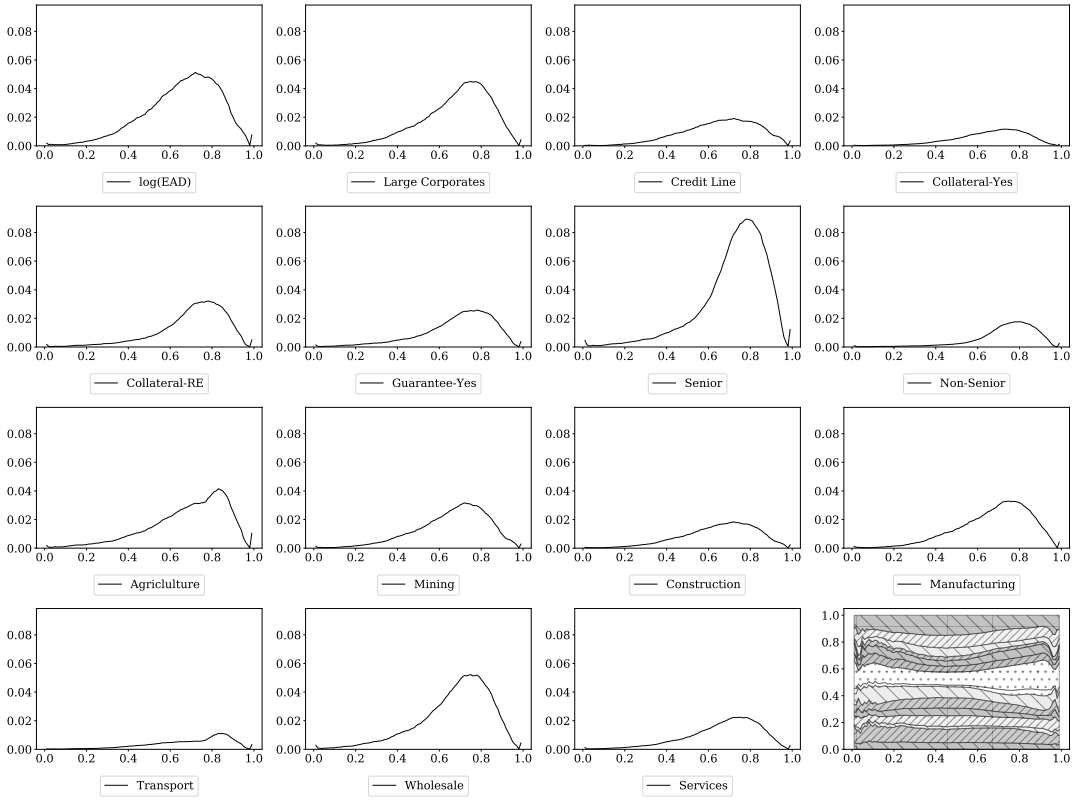
very common in neural network architectures and can be differentiated twice.

The special part of our hyperparameter set are the multiples. As we assume a baseline structure with eight, respectively 16 neurons in the first and second hidden layer, these multiples are an efficient way to validate the shallowness of our network. The most narrow network with multiple equals one, which coincides with the baseline structure with only one hidden layer. The most shallow network with multiple equals eight, has 64 neurons in the first and 128 neurons in the second hidden layer. In general, more complex models, e.g. a larger multiple and more layers, are prone to overfitting. Hence, we try to find a balance with respect to complexity and support this task with dropout layers and weight regularization.

2.C Joint effects with macro variables

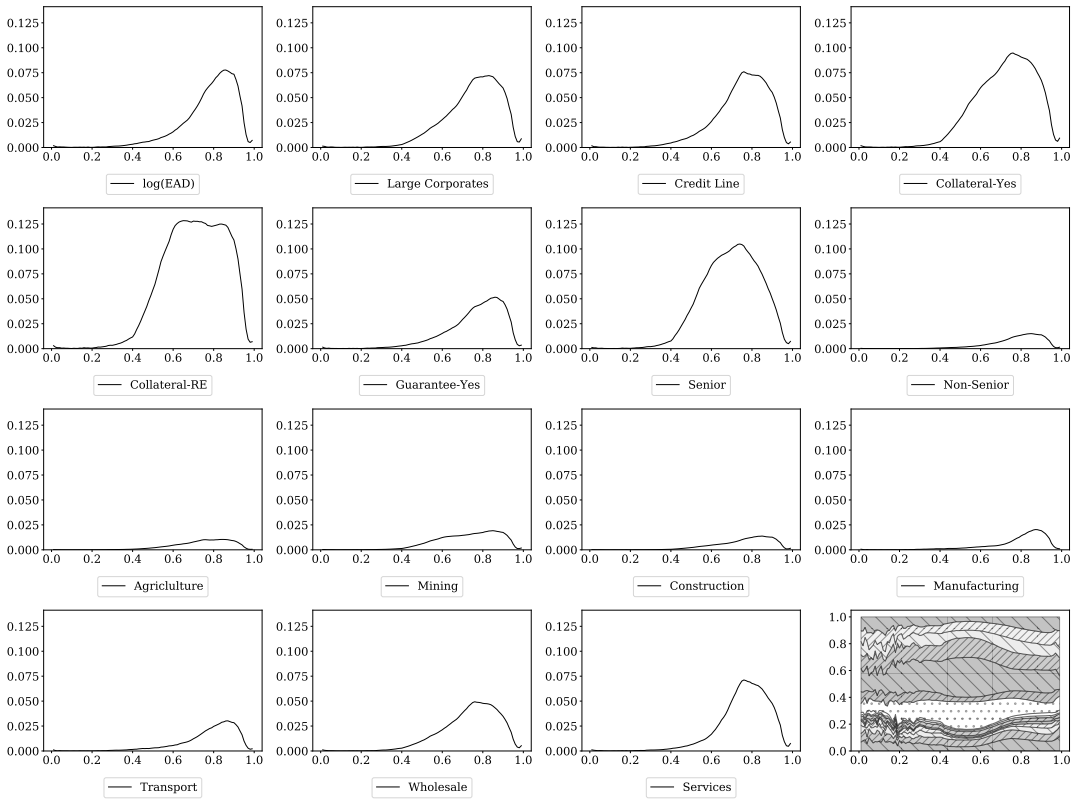
This Appendix shows the joint effects of all variables with the macroeconomic state. We can clearly observe differences among the two regions. In the United States the Seniority has the largest joint effects with the macroeconomic state, whereas in the European sample we observe large joint effects of the macroeconomic state and the collateralization.

Figure 2.C.1: $FI_{\tau}^{Joint}(x_{jl})$ of the macroeconomic state | United States



Note: These plots show the estimated values of $FI_{\tau}^{Joint}(x_{jl})$ for every variable relating to the macroeconomic state and for every quantile. The last plot illustrates the importance of all variables in a stacked fashion. This allows us to evaluate which variables especially have a joint effect with the macroeconomic environment.

Figure 2.C.2: $FI_{\tau}^{Joint}(x_{jl})$ of the macroeconomic state | Europe



Note: These plots show the estimated values of $FI_{\tau}^{Joint}(x_{jl})$ for every variable relating to the macroeconomic state and for every quantile. The last plot illustrates the importance of all variables in a stacked fashion. This allows us to evaluate which variables especially have a joint effect with the macroeconomic environment.

Chapter 3

Deep calibration of financial models: turning theory into practice

This chapter is joint work with Patrick Büchel[1], Michael Kratwochwil[2] and Daniel Rösch[3] published as:

Büchel, P., Kratochwil, M., Nagl, M., Rösch, D (2021). Deep calibration of financial models: turning theory into practice. ***Review of Derivatives Research*****, forthcoming** (available online since August, 2021)

https://doi.org/10.1007/s11147-021-09183-7

The calibration of financial models is laborious, time-consuming and expensive, and needs to be performed frequently by financial institutions. Recently, the application of artificial neural networks (ANNs) for model calibration has gained interest. This paper provides the first comprehensive empirical study on the application of ANNs for calibration based on observed market data. We benchmark the performance of the ANN approach against a real-life calibration framework that is in action at a large financial institution. The ANN based calibration framework shows competitive calibration results, roughly four times faster with less computational efforts. Besides speed and efficiency, the resulting model parameters are found to be more stable over time, enabling more reliable risk reports and business decisions. Furthermore, the calibration framework involves multiple validation steps to counteract regulatory concerns regarding its practical application.

Keywords:Deep learning ; derivatives ; model calibration ; interest rate term structure ; global optimizer

[1] Commerzbank AG, Mainzer Landstraße 157, 60327 Frankfurt am Main, Germany
email: patrick.buechel@commerzbank.com.

[2] Dr. Nagler & Company GmbH, Maximilianstraße 47, 80538 München, Germany
email: michael.kratochwil@nagler-company.com.

[3] University Regensburg, Chair of Statistics and Risk Management, 93040 Regensburg, Germany,
email: daniel.roesch@ur.de.

3.1 Introduction

The calibration of financial models is a laborious, time-consuming and expensive task performed by financial institutions on a regular basis (e.g., daily). Asset pricing models are used to determine the value of derivatives or to generate scenarios for Monte Carlo calculations in risk management. Hence, the outcomes of these models are crucial information required for investment and business decisions. The calibration of these models needs to be performed frequently to ensure the validity of their outcomes. In particular, the calibration of complex and multi-dimensional models is burdensome and requires significant computational efforts and time. The choice of an asset pricing model for a specific product involves balancing the accuracy of the model and the time required for its calibration.

Calibration of a financial model can be described as a reverse optimization task, where the inputs of a pricing function (model parameters) are determined to fit observable outputs (e.g., market prices). The solution of this problem usually requires calling a specific pricing function a large number of times with different parameter settings. Hence, the required time and computational resources have always been limiting factors when choosing a pricing model and models with fast (semi-)analytical solutions are generally preferred. Furthermore, these limitations have led to the broad application of local optimization algorithms for calibration, see Liu et al. (2019). The application of more advanced optimization algorithms is rarely considered. Particularly models with multiple parameters give rise to multiple minima for calibration. Hence, local optimization algorithms tend to struggle finding a robust solution.

Given the aforementioned issues and limitations, the application of machine learning for the calibration of asset pricing models has recently gained interest. In particular, the application of artificial neural networks (ANNs) for accelerating the pricing of derivatives is a topic of interest. As one of the first, Hutchinson et al. (1994) analyzed the applications of ANNs to estimate the pricing function for derivatives in a non-parametric, model-free way. This idea was resumed amongst others by Quek et al. (2008) and Culkin and Das (2017).[4] Recently various papers emerged dealing with a model-based approximation of derivative pricing functions under advanced asset pricing models. For example, Ferguson and Green (2018) apply a forward feed network to estimate the valuation function for equity basket options. Hirsa et al. (2019) analyse the performance of ANN pricing methods for European, Barrier and American options

[4] Ruf and Wang (2020) provide a comprehensive review of literature on the application of neural networks for option pricing and hedging.

under different mathematical regimes. Liu et al. (2019) use ANNs for the approximation of option values under the Black & Scholes and Heston model. With respect to interest rate models, Kienitz et al. (2020) analyze the application of ANNs for the approximation of swaption prices under the Hull-White and Trolle-Schwartz model.

Based on the application of ANNs for the pricing of derivatives, there are several papers on utilizing these trained ANNs for calibration. Hernandez (2017) firstly presented this idea by applying a feed forward ANN for the calibration of a single-factor Hull-White model based on real market data (Sterling ATM swaptions). Dimitroff et al. (2018) use convolutional neural networks for the calibration of stochastic volatility models. As the application of ANNs is expected to accelerate the pricing process, the application of more complex models is an intensively discussed issue. In particular, the calibration of rough volatility models is extensively analyzed by Bayer and Stemper (2018), Bayer et al. (2019), Horvath et al. (2021) and Stone (2020) . The general idea is the acceleration of the instrument valuation via the application of a neural network. The optimization itself is in most cases still based on a local optimization algorithm. Furthermore, most of the existing papers do not use real market data to assess the performance of the ANN, but use only simulated data. Correspondingly, there is no study which compares the ANN results to a real-life implementation at a financial institution to shed light on practical benefits.

We employ the calibration framework proposed by Liu et al. (2019). It involves a two-step procedure for the calibration of financial models. First, a feed forward ANN is trained based on simulated training data to approximate the valuation function under a given asset pricing model.[5] Second, the trained ANN is utilized in a backward manner for the calibration of model parameters. We apply the calibration framework to an interest rate (IR) term structure model based on Trolle and Schwartz (2009), as this setup is applied in the benchmark implementation.

While Liu et al. (2019) show the effectiveness of their approach on simulated data for the training of the ANN as well as the calibration of the model parameters, we empirically analyze the performance of this framework based on a comprehensive set of historic market data for a consecutive series of trading days (21 months). Hernandez (2017) uses historic market data for

[5] Horvath et al. (2021) train a neural network on a financial model in a first step and use this for the calibration in an consecutive step. The main difference between Liu et al. (2019) and Horvath et al. (2021) is the type of neural network employed. The latter authors use convolutional neural networks (CNN) as they focus on a 2-dimensional volatility surface, which can be interpreted as a picture. This enables Horvath et al. (2021) to lift all potentials of the CNN proved for pattern recognition and processing of pictures. In our empirical application, we use a financial model, where the volatility surface/prices are represented by 3-dimensions. As the transfer of the output layer of a CNN to higher dimensions is not trivial, we employ a feed-forward neural network similar to Liu et al. (2019). Hence, we follow the calibration framework of Liu et al. (2019) more closely than Horvath et al. (2021).

the calibration of the Hull-White model, but the data is limited to ATM swaptions. Furthermore, the adjustments to the Hull-White model, such as keeping the parameters constant across swaption maturities are considered as being too simplistic for practical application (Kienitz et al. (2020)). Hence, we consider our study as the first comprehensive empirical assessment that deeply examines the application of ANNs for calibration of financial models based on real market data. The purpose of the paper is to answer the question if current calibration frameworks of financial institutions can be accelerated, maintaining similar calibration accuracy. This would make it possible to use more advanced financial models or/and optimizers for the calibration tasks frequently performed by risk managers.

We extend the literature regarding the calibration of IR term structure models in three important ways. We are the first to establish an ANN for the valuation of swaptions under the Trolle-Schwartz (TS) model and validate the results based on historical market data, evaluating their performance in real-life situations. Second, we calibrate the Trolle-Schwartz model parameters for a consecutive series of trading days based on historic market data for EUR swaptions using a global optimization algorithm. We find that the resulting model parameters using a global optimizer are more stable compared to the benchmark implementation which is in action at a large financial institution. This has important managerial implications as more stable parameters might contribute to less volatile P&L figures over time, which is a desirable outcome for financial institutions. Furthermore, several more simplistic but widely used IR term structure models can be recovered from the Trolle-Schwartz model by using assumptions for certain parameters (Trolle and Schwartz (2009)). Therefore, we consider our results interesting not only for institutions using the TS model, but for a wide range of market participants applying less complex IR term structure models. Third, we outline lessons learned for the practical application of ANNs for financial model calibration and decision making in risk management.

The rest of the paper is structured as follows. In section 3.2, we briefly introduce the Trolle-Schwartz model and show the procedure for calibrating the model. Section 3.3 provides a detailed explanation of the ANN calibration approach and its subsequent components. The data, methodology and results of our comprehensive empirical study are presented in section 3.4. This includes the validation and benchmarking of our results. Section 3.5 concludes this paper.

3.2 Calibration of interest rate term structure models

3.2.1 The benchmark implementation

The calibration of interest rate term structure models is a widely faced task in the financial industry. In general, more complex models are accompanied by higher computational burden and an increase of time required for calibration. Therefore, financial institutions usually set up a costly infrastructure for the calibration of these financial models. However, they have to find a trade-off between the complexity of a financial model, the optimization algorithms and the available time in their daily calibration task. Hence, the computational resources are a limiting factor, when choosing pricing models and optimization algorithms. We set out to validate the ANN approach on empirical data and benchmark against a traditional calibration framework which is in action at a large financial institution. The traditional framework uses a semi-analytical solution of the Trolle-Schwartz model for the pricing of European swaptions, when performing the calibration task. The daily calibration at the financial institution is processed on a large computing cluster utilizing 72 CPU cores simultaneously. Due to time constraints in the productive setting, a local optimizer is used. This is called the "benchmark implementation" henceforth. To make a fair comparison, we use the exact same set of instruments and the same calibration loss function. The aim of the following sections is to show if an ANN can accelerate and increase the robustness of calibration frameworks at financial institutions, while maintaining similar calibration results.

3.2.2 Model calibration

The calibration of financial models is a reverse optimization problem. We assume that we can use a given model to calculate prices of certain financial instruments. The calculation of the price estimate ($\hat{p}_j^{(model)}$) under a specific model for a given instrument (j) requires a series of inputs. This includes the properties of the instrument (τ_j), the parameters of the model ($\Omega_t = (\omega_{t1}, ..., \omega_{tp})$), where p is the number of parameters to calibrate, and a set of market data (Λ_t) at a specific point in time (t). By applying a calibration procedure, the model parameters are set such that the difference between the resulting model prices and the observable market

prices is minimized given a specific loss function (L):

$$\underset{\Omega_t}{\arg\min} \sum_{j \in \mathscr{F}_t} L\left(p_j^{(market)}, \hat{p}_j^{(model)}(\Omega_t \mid \tau_j, \Lambda_t)\right), \tag{3.1}$$

where $\mathscr{F}_t$ represents a set of financial instruments, which have observable market prices ($p_j^{(market)}$). The calibration requires a reasonable and thoughtful choice of calibration instruments. Instruments used for calibration should be liquid, frequently traded and inherit all relevant risk drivers of the instruments it will be applied to. Furthermore, the quality of the calibration is limited by the ability of the model to capture all relevant risk drivers and dependencies of the observable market prices. Nevertheless, the calibration of a complex and high-dimensional model might be quite burdensome from a methodological and computational point of view. Hence, the choice of an appropriate model requires balancing accuracy and computational performance. Especially, if these models are used for pricing financial instruments the ability to perform the calibration in a reasonable amount of time is a crucial prerequisite for their practical application, e.g., for investment or hedging decisions. In addition, the traceability and interpretability of the model is an important feature and considered a key aspect in supervisory oversight and validation.

3.2.3 The Trolle-Schwartz model

In this paper, we perform an empirical study for the application of an ANN based framework to calibrate an interest rate term structure model. We use a term structure model based on Trolle and Schwartz (2009), the so called Trolle-Schwartz model (TS henceforth), used by the real-life benchmark implementation. The TS model is an advanced stochastic volatility model based on the Heath-Jarrow-Morton framework (Heath et al. (1992)). We use the TS model in its risk-neutral setting. The TS model consists of two stochastic processes for the instantaneous forward rate and the variance of the rate process. The dynamics of the forward rate are modelled as follows (see Trolle and Schwartz (2009)):[6]

$$df(t,T) = \mu_f(t,T)dt + \sum_{i=1}^{N} \sigma_{f,i}(t,T)\sqrt{v_i(t)}dW_i^{\mathbb{Q}}(t) \tag{3.2}$$

[6] Within this paper we provide an overview of the Trolle-Schwartz model based on Trolle and Schwartz (2009). Hence, we do not provide mathematical derivation, proofs and background of the model. For additional information on the model and its methodological foundations, please refer to Trolle and Schwartz (2009) and Kienitz et al. (2020).

$$dv_i(t) = \kappa_i(\theta_i - v_i(t))\,dt + \sigma_i\sqrt{v_i(t)}\left(\rho_i dW_i^{\mathbb{Q}}(t) + \sqrt{(1-\rho_i^2)}dZ_i^{\mathbb{Q}}(t)\right) \tag{3.3}$$

Given these differential equations, the evolution of the forward rate is defined based on $2N$ standard Wiener processes ($W_i^{\mathbb{Q}}(t)$,$Z_i^{\mathbb{Q}}(t)$). N defines the number of dimensions of the model. In equation (3.2), $\mu_f(t,T)$ equals the forward drift. Under the assumption of no-arbitrage, Heath et al. (1992) have shown that this term is defined as:

$$\mu_f(t,T) = \sum_{i=1}^{N} v_i(t)\sigma_{f,i}(t,T)\int_t^T \sigma_{f,i}(t,u)du \tag{3.4}$$

Based on this property, the evolution of the forward rate under the risk-neutral measure is solely driven by the initial forward rate curve, the volatility state variables ($v_i(t)$) and the volatility function ($\sigma_{f,i}$). Within the TS model, the volatility function is set to a specific form (see equation(3.5)) to ensure that the forward rate can be represented by a finite-dimensional Markov process and a time-homogeneous volatility structure as:

$$\sigma_{f,i}(t,T) = \left(\alpha_{0,i} + \alpha_{1,i}(T-t)\right)\cdot e^{-\gamma_i(T-t)} \tag{3.5}$$

The TS model offers semi-analytical pricing for the most common interest rate products. In this paper, we use swaptions prices as input for the calibration of the TS model, in line with the benchmark implementation. Hence, we need to calculate the prices of swaptions under the TS model. The TS model provides a semi-analytical solution for an option on a zero-coupon bond. We perform the pricing of swaptions by utilizing these pricing functions and mapping the swaptions based on the stochastic duration method (Munk (1999)).[7]

The TS model is applied in the given benchmark implementation and considered to be suitable to assess the performance of the calibration framework. Furthermore, the TS model offers a semi-analytical solution for pricing European Swaptions, which will be used as calibration instruments for our empirical study. Hence, we are able to generate train and test data in a fast and efficient way. Nevertheless, the model is complex enough to capture the structure and properties of the market-implied volatility / price cube. The TS model can be transformed into more simplistic IR term structure models by simply using specific settings for the parameters of the volatility function (see Trolle and Schwartz (2009)). Hence, our results are also relevant for the application of ANNs to calibrate more simplistic IR term structure models, which are also common in practical implementations.

[7] For additional details and background on the pricing of swaptions under the TS model, please refer to Trolle and Schwartz (2009) and Kienitz et al. (2020).

Table 3.1: Parameters of the Trolle-Schwartz model

Parameter	Interpretation
κ	Mean reversion speed of the variance process
θ	Long-term variance
σ	Volatility of the variance
ρ	Correlation between forward rate and volatility state variables
α_0	Free parameter of the volatility function $\sigma_f(t,T)$
α_1	Free parameter of the volatility function $\sigma_f(t,T)$
γ	Free parameter of the volatility function $\sigma_f(t,T)$

Notes: This table provides an overview of the model parameters in the TS model and their interpretation.

As discussed above, the calibration of a model requires the setting of model parameters such that the model prices fit the observable market prices. The calibration of the TS model requires the determination of Nx7 parameters (see Table 3.1). We consider these parameters as elements of N parameter vectors Ω_i. In line with the setup of the benchmark implementation, we set $N = 1$ which reduces the calibration problem to the determination of seven parameters.[8] In our empirical study, we perform a daily calibration of these parameters by using the sum of squared errors over a set of observable swaption prices as loss function. Hence, the specific calibration procedure for the TS model can be written as:

$$\underset{\Omega_t}{\arg\min} \sum_{j \in \mathscr{F}_t} \left(p_j^{(market)} - \hat{p}_j^{(model)} \left(\Omega_t \mid \tau_j, \Lambda_t \right) \right)^2, \tag{3.6}$$

where Ω_t equals the parameter vector ($\Omega_t = (\kappa_t, \theta_t, \sigma_t, \rho_t, \alpha_{t0}, \alpha_{t1}, \gamma_t)$) for a specific trading day (t). In case of IR swaptions, τ_j equals a vector of properties describing the instrument, such as expiry date of the swaption, tenor and swap rate of the underlying swap. Λ_t represents the yield curve (and discount factors) in the respective currency. Based on these inputs a model price is calculated. The calibration procedure optimizes Ω_t such that the loss function is minimized. The number of available instruments in the empirical application is much higher than the number of parameters to calibrate in the TS Model ($\mathscr{F}_t > \Omega_t$). Therefore, we do not add an additional penalty term in Equation (3.6) to counteract overfitting, in contrast to the original CaNN framework of Liu et al. (2019).[9] The swaptions used in the empirical section are

[8] We are aware that Trolle and Schwartz (2009) propose to use more dimensions. However, our focus is not TS model and its practical implementation. Our paper tries to provide evidence whether an implementation at a financial institution can be substituted or accelerated by an ANN calibration framework. Hence, we follow exactly the setup of the given benchmark to get a reliable and adequate comparison. Therefore, we have to choose $N = 1$ dimensions. We thank participants of the 9^{th} International Conference on Futures and Other Derivatives (ICFOD) 2020 and the 33^{rd} Australasian Finance and Banking Conference (AFBC) 2020 for putting emphasis on that point.

[9] In the original paper of the CaNN framework by Liu et al. (2019) a penalty term of $10 \cdot 10^{-6}$ is added to the calibration loss to avoid overfitting. They used 35 instruments per calibration task and determined five parameters in the Heston model and eight parameters in the Bates model. Hence, the number of instruments is higher than

consistent with the price observations entering the calibration in the benchmark implementation. This means, that we only use swaptions that are sufficiently liquid. Furthermore, we do not introduce a weighting function in Equation (3.6) to focus on the calibration of ATM swaptions, which is in line with the calibration setting at the financial institution. We refer to Ω_t^{BM} for the values calibrated by the benchmark implementation at the financial institution and to Ω_t^{ANN} for the calibrated values of our approach. The observable market prices are structured along three dimensions (expiry tenor, swap tenor, strike). Hence, the observable swaption data can be thought of as a cube of swaption prices.

3.3 ANN calibration approach

3.3.1 Methodological overview

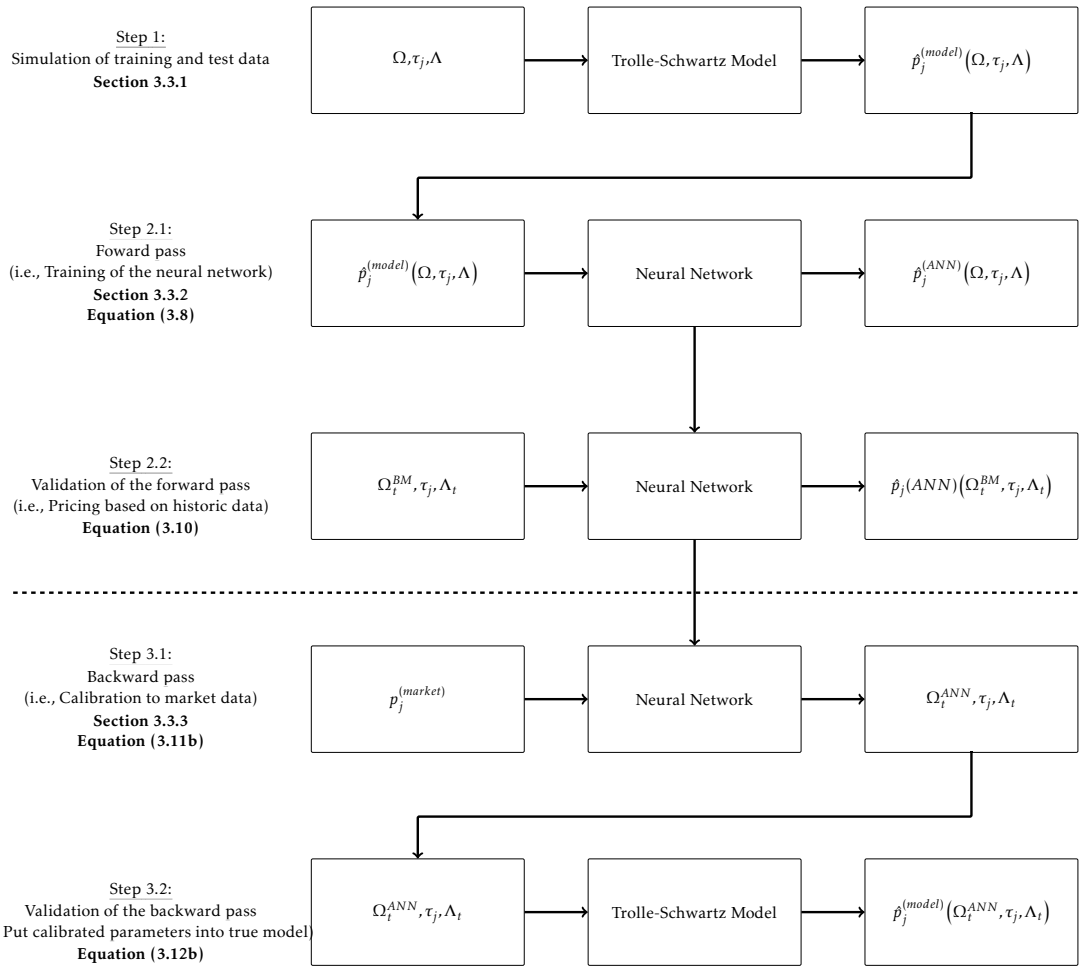
In general, a calibration framework should be flexible, robust, fast and accurate. All these properties are combined in ANNs. They became widespread in the financial domain due to their flexibility and approximation properties. We use the calibration framework (CaNN) proposed by Liu et al. (2019), which involves two consecutive components (two-step or indirect approach). First, we train an ANN to learn the pricing functions for swaptions under the TS model (forward pass). Second, the resulting ANN is applied within a calibration procedure, to fit the model parameters (Ω) to a set of observable market prices. There are other publications that suggest a one-step (direct) approach, where model parameters are learned from market prices directly (e.g. Gambara and Teichmann (2020) or Hernandez (2017)).[10] The indirect approach has a series of advantages compared to the direct approach when it comes to the practical application of ANNs for calibration of financial models (see Horvath et al. (2021) and Bayer et al. (2019) for a comprehensive discussion of reasons for preferring the two-step approach). Most importantly, the two-step approach leverages on existing knowledge and experiences with respect to traditional pricing models and leads to a deterministic calibration framework (Horvath et al. (2021)). These aspects could ease the discussion with regulators, when introducing the prevailing calibration framework in practice. Furthermore, the separation

the parameters to calibrate, but the ratio is lower than in our application. In the empirical section, we calibrate the seven parameters of the TS model on approxmately 800 swaptions per calibration task.

[10] In addition, there are discussions to use ANNs trained on market data for pricing and calibration without using a traditional pricing model at all. While this method could theoretically provide a better fit to market data, it imposes several issues with respect to its lack of traceability and the arbitrary choice of the ANN's properties (i.e., number of parameter, feature selection). Furthermore, evidence for the stability and robustness in practical applications of these approaches still need to be provided.

of the pricing and calibration procedure makes it easier to explain results and identify sources of deviations from market prices. Based on this discussion, we prefer an indirect (two-step) approach for the practical application of the ANN calibration framework. Figure 3.1 illustrates the subsequent steps of the calibration framework, which are outlined in the rest of this section.

Figure 3.1: Workflow of the CaNN framework



Notes: This figure is a detailed description of the calibration framework (CaNN). In the first step, we simulate millions of swaptions based on the Trolle-Schwartz model. In step 2.1, we train the neural network such that the sum of squared differences between the model prices and the ANN prices is as small as possible. Step 2.2 is an important validation step. We put the real historic values of Ω_t^{BM}, which are calibrated by the benchmark implementation of the financial institution, into the trained neural network and compare the squared difference between the $\hat{p}_j^{(model)}\left(\Omega_t^{BM}, \tau_j, \Lambda_t\right)$ and $\hat{p}_j^{(ANN)}\left(\Omega_t^{BM}, \tau_j, \Lambda_t\right)$. The smaller the value, the better our ANN approximates the semi-analytical pricing function used in the benchmark implementation.
In step 3.1 we put the observed market prices of each trading day into the neural network and try to find the values of Ω_t^{ANN} which produces the smallest deviations of $\hat{p}_j^{(ANN)}\left(\Omega_t^{ANN}, \tau_j, \Lambda_t\right)$ and $p_j^{(market)}$ for all observable swaptions for a given trading day. To ensure that the parameter combination Ω_t^{ANN} is also a valid solution in the true model, we put the values Ω_t^{ANN} into the Trolle-Schwartz model in step 3.2 and compare the differences between $\hat{p}_j^{(model)}\left(\Omega_t^{ANN}, \tau_j, \Lambda_t\right)$ and $p_j^{(market)}$.
In each of these steps we want to achieve a similar level of accuracy compared to the given benchmark implementation of the financial institution, as the advantages of the CaNN framework are speed and less computational resources providing similar calibration errors.

ANNs are capable of approximating any continuous function that maps input variables to

outputs, see Cybenko (1989) and Hornik (1991). Our approach utilizes this principle to map input features on swaption prices in a highly non-linear and complex fashion. For each swaption, the neural network starts with covariates $(\Omega, \tau_j, \Lambda) \in \mathbb{R}^p$ as inputs which are called input neurons. The network consists of stacked layers $l = 1,...,L$ whereby each layer consists of $k_l = 1,...,K_l$ neurons $\boldsymbol{h}_{l_{K_l}} \in \mathbb{R}^{K_l}$ that are determined by an affine combination of neurons in the previous layer which is composed with an arbitrary (non-linear) activation function δ. Formally, the ANN is defined by:[11]

$$\boldsymbol{h}_{l_{K_l}} = \delta\left(\boldsymbol{W}_l \boldsymbol{h}_{(l-1)_{K_{l-1}}} + \boldsymbol{b}_l\right)$$

with $\boldsymbol{W}_l \in \mathbb{R}^{K_l \times K_{l-1}}, \boldsymbol{b} \in \mathbb{R}^{K_l}$ as parameters which are usually called weights and biases. Estimates are derived from the last layer, the so called output layer and are given by choosing the identity function for δ, resulting in:

$$F(y|\boldsymbol{x}) = \boldsymbol{W}_{L+1} \boldsymbol{h}_{K_L} + \boldsymbol{b}_{L+1}$$

3.3.2 The forward pass: Learning the pricing function

Step 1 is a prerequisite for the approximation of the TS model using the ANN. In this step, we generate millions of different swaptions to train the ANN. However, this is the most computationally intense part of the whole setup. A detailed description of the swaption characteristics and the range of parameters can be found in Section 3.4.1. Step 2.1 of the calibration framework consists of learning the mapping function, i.e. the Trolle-Schwarz Model, via an Artificial Neural Network (ANN). Finding a suitable architecture which holds the balance between computational time, complexity and accuracy is the main task in this subsection. As our goal is a highly accurate approximation, we use a rather large and complex neural network, as it ensures a high approximation accuracy. As ANNs are sensitive to diverging dimensions of input parameters, we normalize all features $\xi \in (\Omega, \tau_j, \Lambda)$ to a predefined range, i.e $\xi \in [\xi_{min}, \xi_{max}]$, closely following Horvath et al. (2021). This makes it also easier in the backward pass to set optimization bounds. The features are normalized by:

$$\frac{2\xi - (\xi_{max} + \xi_{min})}{\xi_{max} - \xi_{min}} \in [-3, 3]. \tag{3.7}$$

Usually, ANNs are prone to the problem of overfitting, meaning, that the network is able to approximate the training data very well, but fails to approximate unseen test data. This is

[11] Within this paper we provide a short overview on the mathematical foundations of ANNs only. For a comprehensive summary of the most common mathematical concepts of deep learning, please refer to Kraus et al. (2020).

usually the case in out-of-time prediction in the financial context. Our approach is not designed to provide a prediction in an out-of-time fashion, as we want to approximate a specific mapping function as accurate as possible. In our case, the mapping function of training and test data is equal, as both datasets are generated via the (highly complex) pricing function for swaptions under the TS model. As stated by Srivastava et al. (2014), some of these relationships will occur only due to sample noise, resulting in overfitting complex relations in the training set. This could be averted by increasing the number of observations. As we use simulated data for the training of the ANN, we can ensure a large sample size. Furthermore, the data generating process we want to approximate has no inherit noise, as the relation between input parameters and the resulting prices in the TS model is deterministic. Therefore, the ANN is not prone to the problem of overfitting the noise of the data.

Furthermore, in the empirical section, the number of simulated swaptions is larger than the parameters to be estimated by the neural network. Hence, this optimization is overdetermined, which also reduces the chance of overfitting, see Bishop (2006).[12] Therefore, we are confident that approximating the training data ensures that the test data is approximated similarly well. Hence, the issue of overfitting can be neglected in the prevailing use case, as shown by our empirical results in Section 3.4.2. Furthermore, this is supported by findings of previous papers, such as Liu et al. (2019) and Liu et al. (2019). These authors conduct hyper parameter searches, including techniques to reduce overfitting. In none of their final models, an overfitting reducing technique is found to be beneficial for the quality of the ANN's approximation. Hence, these findings underline the above mention indications that the problem of overfitting can be neglected when learning the mapping function within an ANN based calibration framework. Of course, this only holds if we generate a vast amount of training data, which can easily be ensured here. For a detailed description of the generation of the training data, we refer to Section 3.4.1.

The ANN is trained to minimize the following loss function[13] with respect to weights $\boldsymbol{W}$ and biases $\boldsymbol{b}$:

$$\underset{\boldsymbol{W},\boldsymbol{b}}{\arg\min} \sum \left(p_j^{(model)}(\Omega, \tau_j, \Lambda) - \hat{p}_j^{(ANN)}(\boldsymbol{W}, \boldsymbol{b} \mid \Omega, \tau_j, \Lambda) \right)^2 \tag{3.8}$$

As a precaution, we also generated test samples to calculate the loss of equation (3.8) in an out-of-sample task. In general, the ANN is trained over 5,000 epochs to ensure the weights and

[12] We use 7,68 million swaptions as training data to estimate roughly 2,9 million parameters of the ANN.

[13] We also investigated different variants of the loss functions, such as the mean absolute error (MAE), the mean absolute precentage error (MAPE) and an inverse weighting scheme, where we multiply the squared differences by a scaling factor of $\frac{1}{p_j^{(model)}}$ to put more weight on small prices, but find no superior performance in the calibration task.

biases are estimated as accurate as possible. In the additional validation step 2.2, we test the approximation properties of the ANN on real market data. We use the historical parameter values calibrated by the financial institution, put them into the ANN and compare the resulting prices with the observed market prices. We do this for a time period not included in the training of the ANN, i.e. parameter values and yield curves are unseen to the ANN. This step should give a first indication of robustness to unseen market periods.

3.3.3 The backward pass: Calibration of model parameters

Step 3.1 of the framework is to calibrate the input parameters Ω_t given the observed market prices at a specific trading day (t). After the forward pass is successfully accomplished, the weights and biases describing the relation of the input parameters $(\Omega_t, \tau_j, \Lambda_t)$ to the prices of a swaption p_j are known. This means that the mapping function is now deterministic in the sense that simple and fast matrix multiplications map the input to the corresponding swaption prices $(\hat{p}_j^{(ANN)})$. Hence, we have now a very fast way to price a swaption given $(\Omega_t, \tau_j, \Lambda_t)$. For calibration purposes, we are interested in Ω_t which expresses the observed market prices $p_j^{(market)}$ based on the TS model as good as possible. Hence, we basically invert the trained neural network by setting the values of Ω_t as degrees of freedom in a optimization problem:

$$\underset{\Omega_t}{\arg\min} \sum_{j \in \mathscr{T}_t} \left(p_j^{(market)} - \hat{p}_j^{(ANN)}(\Omega_t \mid \tau_j, \Lambda_t, \boldsymbol{W}, \boldsymbol{b}) \right)^2 \tag{3.9}$$

The optimization problem in equation (3.9) is essentially the calibration problem widely faced in the financial industry. To solve this problem, usually local optimizers are widely used due to their speed (see Liu et al. (2019) or Gambara and Teichmann (2020)). In our analysis, several local minima exist, see e.g., Gilli and Schumann (2012). This may be a bottleneck for local optimizers. As we gain a high amount of speed by using the neural network approach, we are able to use slower, but in terms of minimization more robust optimizers. In the calibration framework, we apply a global optimizer called differential evolution (see Storn and Price (1997) for more details).[14] This stochastic optimization scheme is probably able to find a global minimum even if the optimization problem is non-convex. We speed up the calibration framework by using the (transformed) values of Ω_{t-1} as initial values for the optimization (this is also done by the benchmark implementation).

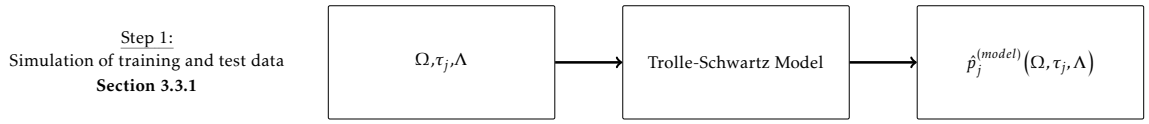
[14] Please note that we use the default values in the implementation of the Python package *SciPy*, except for the population size which we set to 49.

3.4 Empirical study

3.4.1 Data

The first step of the calibration framework is to simulate millions of different swaptions to train the ANN. This is a computationally intensive step, but has to be done only once. Figure 3.2 illustrates this initial step of the calibration framework.

Figure 3.2: The CaNN framework | Simulation of training and test data



Notes: In the first step, we simulate millions of swaptions based on the Trolle-Schwartz model.

Our empirical study is based on a comprehensive set of daily prices for EUR swaptions. These prices are used as input for the calibration procedure. The available market data covers 439 consecutive trading days from January 2019 to September 2020. Hence, our dataset includes the stressed market period in the context of the COVID-19 pandemic in spring 2020. The daily swaption data is available for different expiry tenor, swap tenor and strike values:

- Option Tenor: 1M, 3M, 6M, 9M, 1Y, 2Y, 5Y, 10Y, 15Y, 20Y
- Swap Tenor: 1Y, 2Y, 5Y, 10Y, 15Y, 20Y, 30Y
- Strike (ATM ± bp): 0, 12.5, 25, 50, 100, 150, 200

On each trading day, we observe valid prices for about 800 swaptions. This amounts to a total number of more than 350,000 price observations. In practical applications, financial institutions tend to use a reduced set of swaptions for the calibration of IR term structure models to reduce the calibration time. For our empirical study, we do not further reduce the amount of swaptions entering the calibration procedure to be in line with the benchmark implementation. In addition to swaption data, we obtain the yield curve (6m EURIBOR) for each trading day as well as the relevant forward rate for each swaption. The yield curve is transformed into discount factors for 53 tenors. We compare our calibration performance against the benchmark implementation, which is using a Levenberg-Marquardt optimization algorithm (see Levenberg (1944), Marquardt (1963)) by iterating the traditional pricing formula using a large computing

cluster using 72 CPU cores simultaneously. In contrast, the ANN calibration procedure is based on a standard office computer with 8 CPU cores used at the same time.[15]

The data for each trading day includes the model parameters and model prices estimated by the benchmark implementation. Table 3.2 provides an overview of the observed values for each TS parameter and the associated model prices.

Table 3.2: Training and market data

Parameter	observed (Benchmark)	Sampling (CaNN)
Kappa (κ)	[0.0031,2.80]	[0.005,3]
Theta (θ)	[0.037,3.89]	[0.01,4.0]
Sigma (σ)	[0.24,1.73]	[0.1,2.0]
Rho (ρ)	[-0.047,0.60]	[-0.50,0.80]
Alpha0	[0.00001,0.006]	[0.00001,0.008]
Alpha1	[0.0007,0.005]	[0.0005,0.005]
Gamma (γ)	[0.048,0.089]	[0.01,0.1]
Prices $\left(\hat{p}_i^{(model)}\right)$	[0.0,0.64]	[0.0,1.06]

Notes: This table provides observed values for Trolle-Schwartz parameters as well as the value ranges used for sampling of training data.

As discussed in section 3.3.1, we do not perform the training with real swaption market data. While our swaption dataset includes 350,000 observations, it only provides 439 combinations of TS model parameters. Hence, the number of observations is not sufficient to ensure a satisfying performance of the ANN.

For Step 1 in Figure 3.1, i.e. to train the ANN, we need to generate a large amount of artificial (synthetic) swaption data. We get the required dataset by sampling swaption data for 12,000 synthetic trading days. By using synthetic swaption data for training and testing, we are able to set aside the swaption prices obtained from real market data for the validation of the ANN. The properties of the synthetic swaptions are set to the discrete values shown above. The values for the TS model parameters are randomly sampled from predefined ranges (see Table 3.2) using a uniform distribution. Please note that in general the value ranges used for sampling of parameter values exceed the observed parameter values of the benchmark implementation. Thereby, we ensure that the calibration procedure is able to provide prices for parameter values outside of observed ranges. Furthermore, the CaNN framework is able to find optimal parameter values outside the observed ranges in the calibration procedure.

[15] We use Intel Core i7-9700 CPU cores with 3.00 GHz.

The yield curve for each synthetic trading day is randomly sampled from a collection of yield curve data. The yield curve dataset is constructed by a blended approach, where we combine historically observed market data with synthetic yield curve data. First, we collect yield curves for eight different currencies[16] for a historic two-year time period (Apr 2018 - Apr 2020). This includes about 3,700 different yield curves. We do not include the yield curves observed from May until September 2020 to obtain a real out-of-time validation of the CaNN calibration results within our empirical analysis. Second, we enrich the dataset by adding 20,000 synthetic yield curves. These yield curves are generated by using an algorithm based on the Nelson-Siegel-Svensson methodology (see Nelson and Siegel (1987), Svensson (1994)). Our blended approach provides a comprehensive and representative yield curve dataset. On the one hand, we consider recent historic market environment in the training process. On the other hand, we ensure that the resulting ANN is flexible enough to cope with new unseen market data. Furthermore, this approach offers the possibility for recurring generation of training data and re-training of the CaNN framework based on newly observed yield curves.

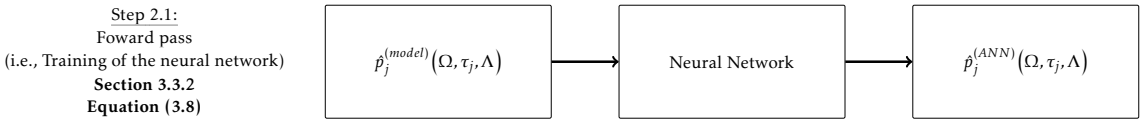
By following the generation procedure outline above, we obtain a total number of 9,6 million synthetic swaptions. The prices of these swaptions are calculated by applying the pricing procedure outlined in section 3.2.3. The resulting dataset is used for training and testing the ANN in Step 2.1, see Figure 3.1, of the calibration framework. In general, we consider the generation of training and test data as a crucial and probably the most laborious task within the calibration framework. The composition of the dataset and its granularity are important drivers of the CaNN's estimation power. Please note that the initial training of the ANN is time consuming and requires significant computational capacities. Nevertheless, this step has to be performed only once. The application of the CaNN framework can be accompanied by frequent re-training, which is significantly less time consuming.

3.4.2 ANN architecture and forward pass (pricing)

After simulating millions of swaptions, the training of the ANN is the subsequent step. Hereafter, we optimize the network architecture and determine the weights and biases to approximate the TS model as close as possible. Figure 3.3 provides a graphical representation for this step of the calibration framework.

[16] We use the historically observed yield curves for the following currencies: EUR, USD, GBP, JPY, CHF, DKK, NOK, SEK

Figure 3.3: The CaNN framework | The forward pass



Notes: In step 2.1, we train the neural network such that the sum of squared differences between the model prices and the ANN prices is as small as possible.

Finding a suitable ANN architecture is a major cornerstone of the successful approximation of the pricing function. As usual, one has to find the balance between approximation accuracy and computational burden, hence a so called random search of the hyper parameters with a subset of the training data is employed. Resulting from this, four hidden layers with 2048, 1024, 512 and 256 neurons are used. To optimally train the ANN, we use the *Adam* optimizer and Relu activation function. As described above, we do not use any dropout layer or early stopping criterion. To ensure convergence with the TS model, we train the ANN with 5000 epochs. An overview of the hyper parameters is illustrated in Table 3.3.

Table 3.3: Hyper parameter of the CaNN

Parameter	Value
Number Features (X)	66
Hidden Layers	4
Neurons per Layer	[66, 2048, 1024, 512, 256, 1]
Number of parameters	2,891,777
Loss function	Sum of squared errors
Activation function	ReLu
Optimizer	Adam
Initialization	Glorot-Uniform
Batch Size	16,384

Notes: This table provides the applied hyper parameters of the final CaNN. In total, a neural network with four hidden layers and 2,891,777 parameters is trained to approximate swaption prices under the TS model.

For illustration, we also employed and validated the hyper parameter setting proposed by Liu et al. (2019) with 200 neurons in each of the four hidden layers. The accuracy in terms of mean squared error is 10 times worse than with our architecture. This gives rise to the conjecture that any calibration framework needs a tailored set of hyper parameters to provide the a sufficiently accurate estimation of model prices. This also suggests, that the model complexity of the ANN should increase with the complexity of the IR dynamics.[17] To train the ANN, we randomly split the 12,000 synthetic trading days into a training set (7,68 million swaptions) and a test set (1,92

[17] In unreported results, we employed a hyper parameter search only for ATM options, and found the same tendency towards more complex and deeper neural networks. This gives rise to the conjecture that the complexity of the ANN is by a large part determined by the complexity of the IR dynamics, and the number of instruments plays only a minor role. This is plausible, as all instruments share the same yield curve and TS model parameters at a specific trading day, which accounts for a large number of the input parameters.

million swaptions). Table 3.4 shows key evaluation metrics in the train and test sample.

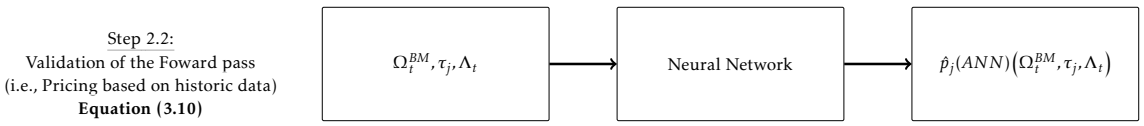
Table 3.4: Results of ANN training

CaNN	MSE	MAE	RMSE
Training	1.47e-07	2.38e-04	3.52e-04
Testing	1.80e-07	2.45e-04	4.24e-04

Notes: This table provides the performance measures for the ANN training. As all six measures are quite low, we are confident that the ANN approximates the TS model very well.

We observe only small differences, when comparing the results for the train and test set. This may imply that the ANN generalizes well and we do not encounter overfitting. Furthermore, the metrices are well in line with results of previous studies, see e.g. Liu et al. (2019) or Horvath et al. (2021). The very similar performance for the train and test data may also be attributed to the comparatively large training sample, which is imminent to approximate the mapping function accurately. After training the neural network, we validate our results against an implementation of a large financial institution in step 2.2, see Figure 3.4:

Figure 3.4: The CaNN framework | Validation of the forward pass



Notes: Step 2.2 is an important validation step. We put the historic values of Ω_t^{BM}, which are calibrated by the benchmark implementation into the trained neural network and compare the squared difference between the $\hat{p}_j^{(model)}\left(\Omega_t^{BM}, \tau_j, \Lambda_t\right)$ and $\hat{p}_j^{(ANN)}\left(\Omega_t^{BM}, \tau_j, \Lambda_t\right)$. The smaller the value, the better our ANN approximates the semi-analytical pricing function used in the benchmark implementation.

In contrast to most other papers on the application of ANNs for pricing and calibration, we perform an additional validation of the forward pass based on historic pricing data obtained from a benchmark implementation (BM). We call this step the "out-of-simulation validation", as the data used to assess the ANN's pricing performance has not been generated with the same process as the train and test sample, but historically based on real-life market data. Thereby, we ensure that the ANN has learned the TS pricing function correctly and performs well in a true out-of-sample evaluation. From our point of view, the validation based on results from a benchmark model is a prerequisite for the practical application of an ANN based calibration framework. To perform the out-of-simulation validation, we pass the observed parameters estimated by the benchmark implementation ($\Omega^{(BM)}$) together with the historic market data for the respective trading day through the ANN for all swaptions across available trading days. Afterwards, we compare the predicted prices of the trained ANN with the model prices

generated by the benchmark implementation (see equation (3.10) for mathematical illustration).

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\sum_{j\in\mathscr{T}_t}\left(\hat{p}_j^{(model)}\left(\Omega_t^{(BM)} \mid \tau_j, \Lambda_t\right) - \hat{p}_j^{(ANN)}\left(\Omega_t^{(BM)} \mid \tau_j, \Lambda_t, \boldsymbol{W}, \boldsymbol{b}\right)\right)^2 \tag{3.10}$$

The results of this validation step are displayed in Table 3.5. First, we check the performance for the time period from January 2019 to April 2020. The swaption data from this period was used for setting the parameter ranges and yield curves for the simulation of synthetic swaptions. As the evaluation metrics are close to the results obtained in the training and testing, we may conclude that the ANN is robust in real-life market situations. As a next step, we use the benchmark parameters from the out-of-time period (May 2020 – September 2020). Data and information from this period, such as parameter values and yield curves, has not been used in the previous steps and is therefore completely new to the framework. The results for this period of time indicate that we achieved generalization even in an out-of-time perspective with unseen circumstances. These results may serve as a first proof of concept for a practical implementation.

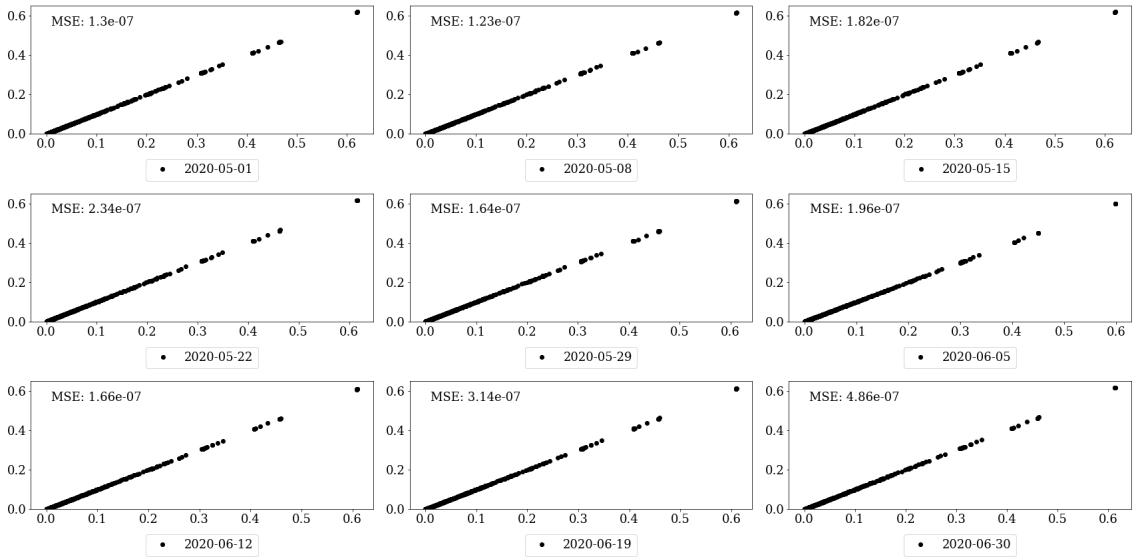
Table 3.5: Results of ANN training

CaNN	MSE	MAE	RMSE
Out-of-simulation (Jan 2019 – Apr 2020)	5.47e-07	2.98e-04	7.24e-04
Out-of-simulation (May 2020 – Sept 2020)	2.48e-07	2.65e-04	4.98e-04

Notes: This table show key evaluation metrics in the out-of-simulation validation. We divide the samples into data building the basis of our training (January 2019 to April 2020) and true out-of-time data (May to September 2020)

Figure 3.5 provides real fit plots for selected trading days taken from the out-of-time period. The plots compare the prices estimated by the ANN (x-axis) with model prices from the benchmark implementation (y-axis). As we can see, the points are on the bisecting line which implies a very good convergence of the ANN prices to BM model prices. To each real fit plot, the MSE for the respective trading day is added. For some days, we obtain much better results than in training, whereas for other days we are slightly worse. In summary, we find sufficient evidence that the trained ANN generalizes very well even if confronted with unseen data. Hence, the ANN provides a very good approximation of the TS pricing function for swaptions.

Figure 3.5: Real fit plots for selected trading days

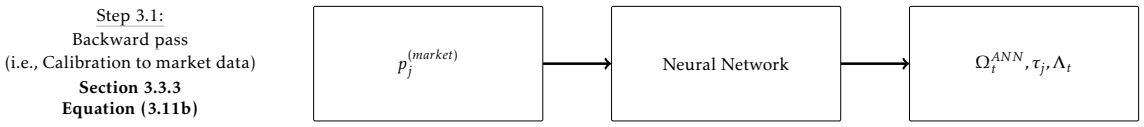


Notes: These figures show the real fit plots of selected historic trading days. Furthermore, the day specific MSE is displayed. The price estimations of the ANN are displayed on the x-axis, whereas the model prices of the benchmark implementation is shown on the y-axis.

3.4.3 The backward pass (calibration)

For the rest of the section, we are now concerned with the calibration task frequently performed by the given benchmark implementation. For step 3.1 of the calibration framework, we utilize the trained ANN to calibrate the TS model parameters to a daily set of observable swaption prices, see Figure 3.6. Furthermore, we validate our results against the real-life benchmark implementation of a large financial institution.

Figure 3.6: The CaNN framework | The backward pass



Notes: In step 3.1 we put the observed market prices of each trading day into the neural network and try to find the values of Ω_t^{ANN} which produces the smallest deviations of $\hat{p}_j^{(ANN)}\left(\Omega_t^{ANN}, \tau_j, \Lambda_t\right)$ and $p_j^{(market)}$ for all observable swaptions for a given trading day.

For each of the 439 trading days, we obtain two calibrated parameter sets. One parameter set is returned from the benchmark implementation ($\Omega_t^{(BM)}$), while the other parameter set results from the ANN based calibration framework ($\Omega_t^{(ANN)}$).

For clarification, we restate and concretize the general formulation of the calibration problem in equation (3.9) and provide a specific notation for both calibration processes:

$$\underset{\Omega_t^{(BM)}}{\arg\min} \sum_{j \in \mathscr{F}_t} \left(p_j^{(market)} - \hat{p}_j^{(model)}(\Omega_t^{(BM)} \mid \tau_j, \Lambda_t) \right)^2 \tag{3.11a}$$

$$\underset{\Omega_t^{(ANN)}}{\arg\min} \sum_{j \in \mathscr{F}_t} \left(p_j^{(market)} - \hat{p}_j^{(ANN)}(\Omega_t^{(ANN)} \mid \tau_j, \Lambda_t, \boldsymbol{W}, \boldsymbol{b}) \right)^2 \tag{3.11b}$$

Both calibration approaches aim to minimize the sum of squared errors for each trading day. By minimizing the loss function, an optimal set of TS model parameters is selected. The benchmark implementation performs the calibration by applying a local optimization algorithm (Levenberg-Marquardt) and repeatedly calls the traditional implementation of the semi-analytic pricing formula (see equation (3.11a)) and sets parameter restrictions for the TS parameters to ensure that the optimizer returns a result. For this empirical analysis, the benchmark model parameters ($\Omega_t^{(BM)}$) are obtained from the historical calibration results of the benchmark implementation. The CaNN framework utilizes the forward pass by frequently estimating swaption prices based on the trained neural network for different parameter settings (see equation (3.11b)). Please note that the weights and biases of the ANN have already been set in the training phase (forward pass) and are not altered during the calibration procedure.

With respect to the substantial acceleration using the ANN, a global optimization algorithm (Differential Evolution) can be used to minimize the loss function given by equation (3.11b). Due to time constraints in the productive workflow of the financial institution, only a local optimizer is used in the benchmark setup. The application of the differential evolution (DE) algorithm shall avoid the problem of stopping at local minima and offers the advantage that no starting values are required (see Liu et al. (2019)). However, we use the parameter values of the previous trading day as starting values for the DE algorithm. We observe that using starting values leads to a faster convergence and significantly accelerates the calibration process. In practical applications, such as the referred benchmark implementation, the parameter values of the previous trading day are commonly used as starting point for the optimization process. This could potentially lead to a deterioration of the minimization, when applying local optimizers, but should not be an issue for global optimization algorithms. Hence, we are confident that there is no downside in setting starting values for the DE algorithm in the CaNN framework. On the contrary, we observed that setting starting values speeds up the ANN calibration by roughly 50 times. Thereby, the calibration for each trading day can be performed in about 30 seconds. This is four times faster than the benchmark implementation, although it uses a

local optimizer and 72 CPU cores. This means, that our approach, i.e. using a global optimizer and only 8 CPU cores, is faster than the benchmark implementation. Summarizing, we can achieve a very similar calibration error, see Table 3.6, but are faster, require less computational resources and are able to use a global optimizer. Even more benefits could be realized if the financial institutions use financial models without analytical solutions, i.e. the prices can only be determined via Monte Carlo simulations. However, this would increase the computational burden of the first step greatly, as the generation of enough training data could take extremely long.

Table 3.6: Calibration results

Period	daily MSE (BM)	daily MSE (ANN)	daily SSE (BM)	daily SSE (ANN)
Jan 2019 – Apr 2020	1.36e-06	1.29e-06	1.11e-03	1.10e-03
May 2020 – Sept 2020	1.61e-06	1.63e-06	1.13e-03	1.13e-03

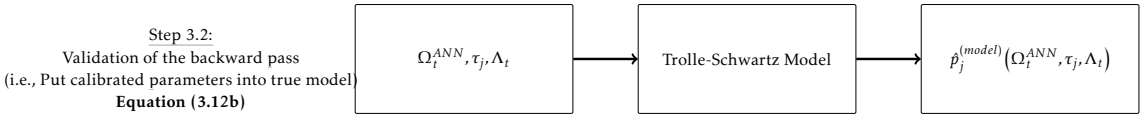
Notes: This table show key evaluation metrics of the ANN and benchmark calibration result. We divide the samples into data building the basis of our training (January 2019 to April 2020) and true out-of-time data (May to September 2020)

Table 3.6 provides an overview of the calibration results equal to the average daily values of the loss function calculated by equations (3.11a) and (3.11b) as well as the daily mean squared error (MSE) for both calibration approaches. The results show that the CaNN framework provides calibration results that are very close to the benchmark implementation for both time periods.

Nevertheless, there might be a concern that these results do not provide sufficient evidence for the practical applicability of the CaNN framework. We expect that supervisory authorities will have a critical view on the application of ANNs for pricing and calibration as the ANN pricing function constructed in the forward pass is not considered traceable given the high amount of parameters in the neural network.

To prove that the CaNN provides reliable parameter values, the calibration framework involves an additional validation step 3.2. Hence, the CaNN parameter set ($\Omega_t^{(ANN)}$) is used as input for the semi-analytical pricing formula for swaptions under the TS model. By comparing the resulting prices with observable market prices, we are able to prove that the CaNN calibration results hold true in the Trolle-Schwartz model framework, see Figure 3.7:

Figure 3.7: The CaNN framework | Validation of the backward pass



Notes: To ensure that the parameter combination Ω_t^{ANN} is also a valid solution in the true Trolle-Schwartz model, we put in the values Ω_t^{ANN} into the Trolle-Schwartz model in step 3.2 and compare the differences between $\hat{p}_j^{(model)}\left(\Omega_t^{ANN}, \tau_j, \Lambda_t\right)$ and $p_j^{(market)}$.

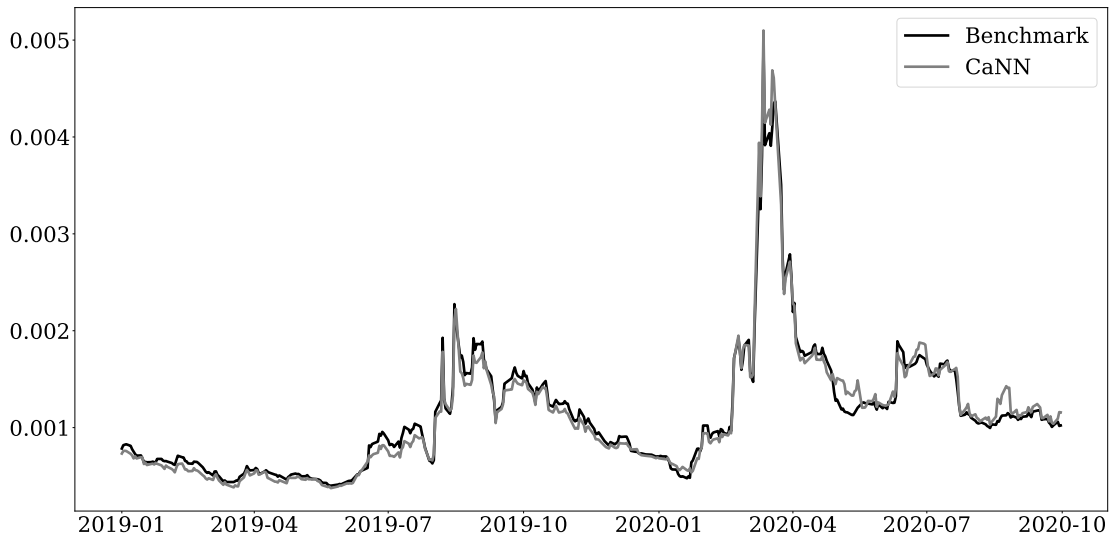
Hence, we apply equation (3.12b) to validate the ANN solution for each trading day. The result will provide insights with respect to the true quality of the CaNN calibration results.

$$SSE^{(BM)}(t) = \sum_{j \in \mathscr{F}_t} \left(p_j^{(market)} - \hat{p}_j^{(model)}(\Omega_t^{(BM)} \mid \tau_j, \Lambda_t) \right)^2 \tag{3.12a}$$

$$SSE^{(ANN)}(t) = \sum_{j \in \mathscr{F}_t} \left(p_j^{(market)} - \hat{p}_j^{(model)}(\Omega_t^{(ANN)} \mid \tau_j, \Lambda_t) \right)^2 \tag{3.12b}$$

Figure 3.8 illustrates the daily performance measure (SSE) for both calibration approaches over time. The black line represents the benchmark result (equation (3.12a)), while the grey line represents the performance measure for the CaNN framework (equation (3.12b)). In general, we find that the performance of both calibration approaches significantly varies over time. In the early months of 2019 the losses are comparatively low whereas in the fourth quarter of 2019, we observe a considerable increase. A remarkable spike can be observed after the break-out of the COVID-19 pandemic, meaning that the calibrated TS model prices strongly deviates from market prices. These results clearly indicate that a thorough assessment of ANN calibration approaches should be done in different market environments to ensure their practical applicability.

Figure 3.8: Sum of squared errors over trading days

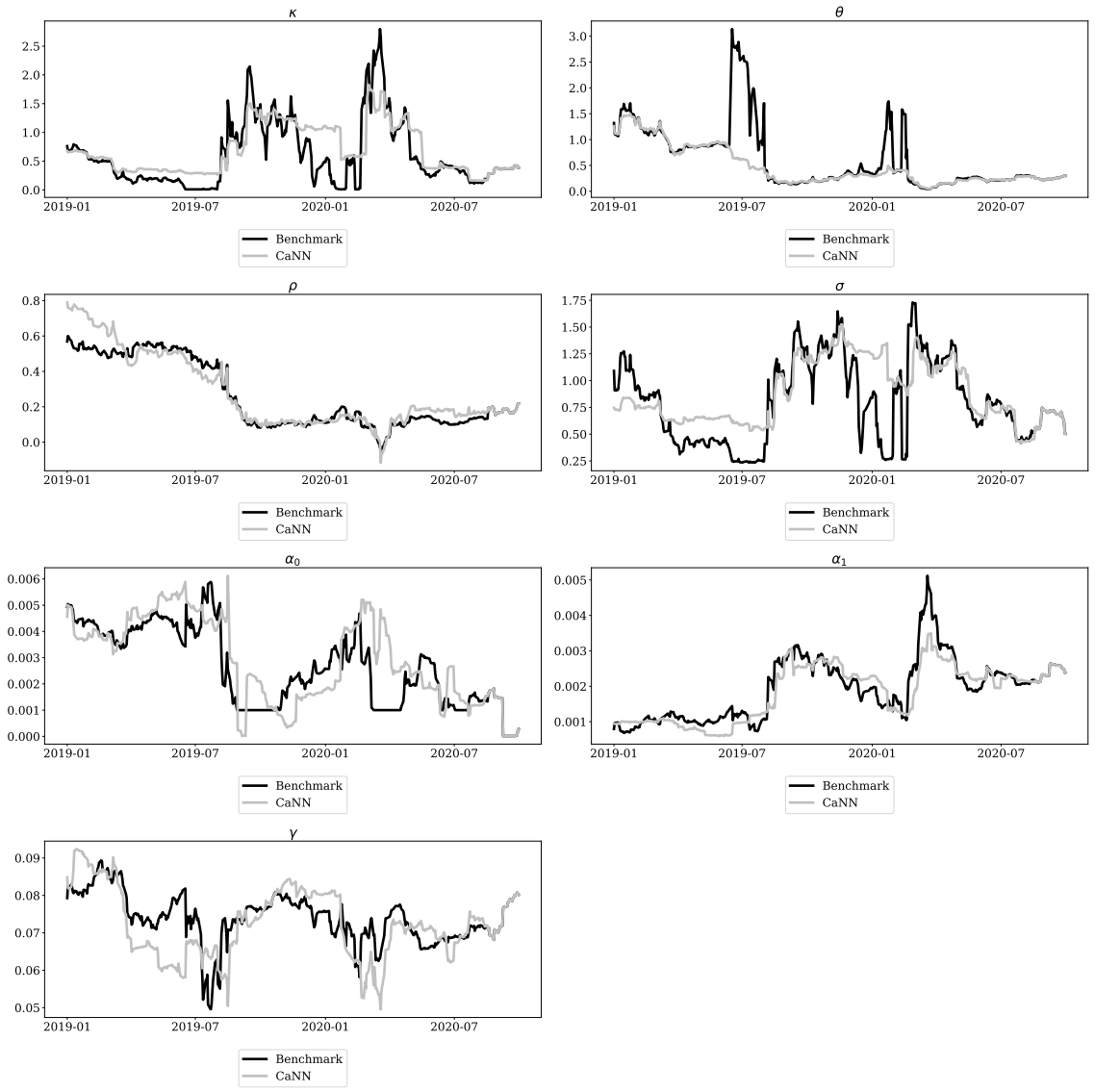


Notes: This figure shows the sum of squared errors of trading days for the whole time span. The grey line corresponds to the SSE using the CaNN approach, whereas the black line coincides with the SSE of the benchmark implementation.

The results presented in Figure 3.8 show that the CaNN framework produces competitive results compared to the benchmark implementation in terms of daily performance. For some market periods we can even find better solution for the parameters, see e.g. the period from June 2019 to August 2019 or the early months of 2019. The largest deviation between the CaNN and the benchmark implementation can be observed during the COVID-19 period in the March 2020. Nevertheless, the daily performance of both approaches does not differ significantly even in this stressed market environment. Hence, the CaNN framework does provide comparable calibration results even in extreme and unusual market situations in a faster and computationally more efficient manner. Furthermore, the very good results for the out-of-time period (May to September 2020) indicate that the performance of the CaNN framework does not depend on including current market data during training.

In addition to analyzing the performance of the CaNN framework, we are interested in a comparison of the parameter estimates for both calibration approaches. Figure 3.9 illustrates the different estimates for all elements of Ω_t over time. The black line represents the parameter estimated by the benchmark implementation, while the grey line represents the respective element of $\Omega_t^{(ANN)}$.

Figure 3.9: Calibrated parameters over trading days



Notes: These figures show the calibrated values of $\Omega_t^{(BM)}$ and $\Omega_t^{(ANN)}$. The black line represents the values gathered form the benchmark implementation, whereas the grey line illustrates $\Omega_t^{(ANN)}$. For details on the parameters, please refer to section 3.2.

Overall, the analysis reveals that parameter estimates from both calibration procedures are quite close to each other and have a similar evolution over time. However, the results indicate that the CaNN parameters are more stable over time and therefore more robust against taking extreme values.[18] For example the BM estimates for θ show four considerable peaks in the analyzed period, while the CaNN estimates show a relatively smooth evolution over time. On some days, the benchmark implementation obtains extreme values for certain parameters, which are equal to a boundary of the parameter restrictions. This may imply that the local optimizer used by the

[18] In both calibration frameworks the respective calibrated parameter values of the previous day are used as starting values for the next day. Hence, the more stable results of the CaNN approach may not be attributed to the way the staring values are set.

benchmark implementation ended up in a different local minimum on the respective trading days, leading to a compensation of the high θ value by extreme settings for other parameters. As the parameters in the TS model are not completely "independent", in the sense that different combinations of parameter values may result in the more or less same calibration loss, we achieve much less fluctuating parameters while maintaining a similar calibration result. This can be seen for example in the period around July 2019, where we observe simultaneous peaks respectively lows in θ and γ values, whereas our parameter values are more or less stable trough this period.

A similar issue can be observed for the parameter κ. In the period from September 2019 to mid January 2020, the estimated parameter of the benchmark implementation starts with values from 0.777 to 2.11 in early September, decrease to 0.56 mid September and then plumbs to 0.04 in mid January 2020 and increases sharply afterwards to 2.5 in spring 2020. In contrast, the CaNN parameter fluctuates from September 2019 with values around 1.5 to end of January 2020 with values of 1.07 with considerably less fluctuations within this period. The same behavior can be observed for σ in the aforementioned time period. The evolution of CaNN estimates for different parameters show significantly lower fluctuation and that the parameters are less likely to take extreme values.

Based on these observations, we conclude that the CaNN framework generally provides more stable parameter estimates over time. From our point of view, the stability of parameter estimates over time is a desirable property of a calibration procedure. The estimated model parameters are not only required as inputs for the pricing function, but also to specify stochastic processes in Monte-Carlo simulations for the purpose of calculating P&L components, such as Credit Valuation Adjustments (CVA), and risk measures. Hence, more stable parameters might significantly contribute to a reduction of day-to-day P&L volatility and costs of hedging in the trading business. Furthermore, more stable calibration results will lead to less volatile and more reliable risk measures, which enables managers to take more profound business decisions. This makes the CaNN approach highly relevant for risk managers of financial institutions.

3.4.4 Discussion and additional results

In summary, the results of our empirical study give rise to the conjecture that an ANN based calibration framework does not only provide competitive results compared to traditional approaches, but also offers further benefits and advantages with respect to the stability and reliability of resulting parameter values. Hence, we conclude that there is indeed a practical applicability for ANN based calibration frameworks. However, we recognize that the practical application of a CaNN framework might involve challenges with respect to the fulfilment of regulatory requirements. Especially, with respect to risk management there are extensive regulatory requirements for the application of internal models (e.g. ECB (2019), OCC, et. al (2011)). Amongst others, the European Central Bank's guide on internal models (ECB (2019)) introduces regulatory requirements and expectations for the validation of pricing functions and calibration procedures. As an example, ECB (2019) defines a pricing function in the context of an internal Counterparty Credit Risk (CCR) model as the dedicated implementation of a pricing model also taking into account its method for calibration. Furthermore, it requires the inclusion of pricing functions used for calculating or calibrating exposure methods into the model's framework and governance. Based on this definition, institutions are required to implement a framework that allows for a granular identification of pricing deficiencies (on transaction level). According to ECB (2019) the validation framework needs to include all pricing functions used in the internal model. Hence, we argue that methods and pricing functions used for calibration are subject to the same requirements as pricing functions applied for valuation of derivatives within the exposure simulation.

The proposed calibration framework is a two-step approach, where pricing and calibration are separated. The pricing function is approximated explicitly via an ANN before the actual calibration step. In contrast, a one-step approach calibrates parameters of a dedicated pricing model from market prices directly. Nevertheless, the one-step approach involves an implicit approximation of the model's pricing function that should be validated according to regulatory requirements. This might be challenging as no explicit pricing function is available in the calibration process and the parameters of the ANN are hard to interpret. In a two-step approach, the validation of the pricing function used within the calibration procedure is straightforward. We are able to identify deviations of ANN prices to the traditional pricing function and market prices on transaction level. Furthermore ,the validation of the ANN's approximation of the pricing function as well as the results of the calibration process can easily be integrated in the validation framework including various materiality thresholds for deviations. Hence, a two-

step approach might allow for a straightforward fulfilment of the aforementioned regulatory requirements. In our opinion, the framework proposed in this paper is generally compliant with supervisory expectations as we offer a staggered approach involving additional and separate validation steps 2.2 and 3.2 for the ANN based pricing as well as calibration procedure.

Neural networks are often considered black boxes as it is somewhat difficult to explain and track the mapping function due to the high complexity and high amount of parameters. Hence, regulators may not be fully convinced of a full replacement of traditional calibration frameworks with ANN based calibration procedures. But in contrast to other use cases of machine learning algorithms, such as prediction of future stock returns or risk figures, we know the ground truth of the mapping function we want to approximate, i.e. the TS model. Hence, it is possible to validate our pricing results, i.e. step 2.2, and our calibration results as outlined in step 3.2. These, to some extend unique validation steps of this framework, are strong arguments in the discussion with regulators.

Moreover, we argue that this framework can be utilized to generate initial values for the currently implemented calibration procedures, which should lead to a faster and more robust calibration process. As the initial calibration is performed by calling the ANN, financial institutions are able to reduce dependencies between pricing and calibration procedures in daily production, especially if the solution of the financial model can only be determined by Monte Carlo simulations. Hence, financial institutions could be able to monetize the benefits of ANN based calibration without replacing traditional approaches for now. Based on our results this could increase the stability of results over time and reduce the probability of a local optimizer getting stuck in a local minimum. Additionally, we find that the number of function evaluations required for the local optimizer can be reduced by more than one third using the start values obtained from the CaNN calibration instead of values of the previous day. We are able to provide empirical evidence for the latter aspect in the following case study, where we repeat the calibration process of section 3.4.3 in two different settings.
In the first setting we only use the local Levenberg-Marquardt (LM) optimization algorithm (see Levenberg (1944), Marquardt (1963)) to calibrate the parameters. In the second setting, we first use the differential evolution algorithm and *afterwards* pass these values to the LM optimization as initial values. We measure the performance over the out-of-time period based on the function evaluations required by the LM algorithm to arrive at the optimum on each day. Both optimizations are performed in the CaNN framework and on the same hardware to ensure comparability. On average the stand-alone LM algorithm (with previous day start

values) requires 253 evaluations per trading day, while the combined optimization only requires 161 evaluations. Hence, we were able to decrease the number of function evaluations by about 36%, while keeping the level of accuracy. This is a considerable reduction leading to a faster calibration process and reduces the computational capacities required and additionally lead to more robust parameter values over time. Furthermore, it is a cheap and efficient way for financial institutions to use a global optimizer, without altering their actual calibration framework. The generation of the daily start values with the DE algorithm does not take longer than 30 seconds, which probably is considerably less than the potential speed up due to less function evaluations. These results support our conclusion that the implementation of a CaNN framework provides added value, even if traditional calibration procedures are not fully replaced yet.

3.5 Conclusion

This paper provides the first comprehensive proof of concept regarding the practical application of artificial neural networks (ANNs) for the calibration of asset pricing models. We propose additional steps for the CaNN framework based on Liu et al. (2019) to accelerate practical applicability and counteract regulatory concerns for the practical implementation. First, we provide a blended concept for the generation of train and test data. Second, we introduce additional validation procedures based on real-life historic market data to ensure that results of the CaNN are conform with observed pricing and calibration results. Third, we perform a real out-of-time validation to provide evidence that the CaNN framework can cope with unseen data.

Based on a comprehensive time series of historic market data, we are able to show that the calibration framework produces competitive calibration results for a complex IR term structure model compared to a benchmark implementation of a large financial institution. Our empirical analysis covers 1.75 years of swaption data, including the stressed market environment following the break-out of the COVID-19 pandemic. Hence, the calibration approach is suitable for real-life calibration problems and the CaNN framework performs well in different market environments. Given the substantial acceleration of the calibration process by using the CaNN framework, the efficient application of a global optimizer is feasible. As shown in the empirical analysis, the global optimizer is less likely to adopt boundary solutions, leading to more stable parameter results over time compared to the benchmark implementation. At the same time the CaNN framework is able to cope with changing market environments, while maintaining a comparable

level of calibration error. The more stable parameter estimates from the CaNN framework might help to reduce the P&L volatility over time, while still ensuring that the model is consistent with the risk-neutral expectations of market participants. Hence, a CaNN framework will provided added value, beyond a potential acceleration of the calibration process. The assessment of the potential benefit with respect to P&L volatility is complex and subject to further analysis. Further conclusions for the practical implementation of an ANN based calibration framework are as follows. First, the composition and quality of train and test data is a major driver of the CaNN's performance. Historic swaption data should not be used for training and testing as the data is more valuable for validation. Hence, we propose a blended approach, which produces synthetic data by combining information from historic market data with an algorithm that simulates synthetic datasets. Second, we recommend to set start values for the global optimizer based on the previous day's results as this significantly accelerates the CaNN calibration process.

We are aware that our empirical analysis is limited to one IR term structure model for a single currency (EUR). The decision to use the Trolle-Schwartz model was based on the aspiration to analyze the performance of the calibration framework for a rather complex, but practically implemented model. Hence, this is the first study to investigate whether ANNs are faster and more robust compared to an implementation of a large financial institution. Furthermore, the TS model can be easily reduced to more simplistic term structure models. However, we believe that the application of this framework to further currencies, models and asset classes will provide further findings regarding the performance of ANN based calibration frameworks. Future work may also focus on obtaining additional insights with respect to the calibration procedure from the CaNN framework, such as information on parameter sensitivity or importance of different inputs.

Although we believe that the framework generally adheres to regulatory requirements, its practical application might be viewed critical by supervisory authorities as the training process and resulting ANN pricing function may seen as not fully traceable. To counteract this, we offer a staggered approach involving additional and separate validation steps for the ANN based pricing as well as calibration procedure. However, regulators might still have concerns about the replacement of traditional implementations with the CaNN framework. Nevertheless, the implementation of this framework and the subsequent integration of its results could significantly improve traditional calibration procedures in terms of accuracy, robustness, speed and provide additional insights for validation processes. These aspects give rise to the conjecture that the CaNN framework is of high practical relevance and has the potential to improve model calibration, risk assessment and business decisions.

Acknowledgements: We would like to thank participants of the 9^{th} International Conference on Futures and Other Derivatives (ICFOD) 2020 and the 33^{rd} Australasian Finance and Banking Conference (AFBC) 2020 for fruitful discussions and review comments. We also send special thanks to participants of the Frankfurter Institut für Risikomanagement und Regulierung e.V. (FRIM e.V.) Round Table Artificial Intelligence for providing valuable insights and thoughts with respect to the practical application of the calibration framework and possible regulatory concerns. Further, we would like to thank the two anonymous referees for their comments which greatly improved the paper.

Chapter 4

Does non-linearity in risk premiums vary over time?

This chapter is corresponds to a working paper with the same name (has been reviewed by *Management Science*).

This paper proposes a model agnostic measure of non-linearity to study the hidden dynamics in the cross-section of expected returns. Thereby, a significant inverse relationship of linearity in return predictions and uncertainty expressed by the VIX is documented. Linear asset pricing models work quite well in normal times as the share of non-linearity is on average 15%, but it more than doubles in crisis periods. This indicates that the relation of firm-characteristics on the risk premium changes in times of high uncertainty. Especially in crisis periods, non-linearity plays a crucial role. With extensions to state-of-the art explainable machine learning techniques, we can identify past return volatility and the expected market volatility as main driver of the inverse relationship.

Keywords: Machine Learning, Explainable Machine Learning, Risk Premiums, Non-Linearity, Uncertainty

JEL Classification: C52, C55, C58, G0, G1, G17

4.1 Introduction

Capturing non-linearity with machine learning methods is an increasing strand of literature not only in financial research but also in many other fields. They are used to optimize the targeting of promotions for new customers, see, e.g., Simester et al. (2020) or for search personalization, see, e.g., Yoganarasimhan (2020). Furthermore, the productivity and selection of human capital in social policy applications is modelled using machine learning algorithms, see Chalfin et al. (2016). Additionally, in the increasingly important field of personal health and climate change, these models gain interest. Deryugina et al. (2019) use them to predict the life-years lost due to air pollution exposure of US elderly. Gibson et al. (2021) trains machine learning methods on large climate models to increase the forecasting accuracy.

This paper focusses on the prediction of excess stock returns, also labelled as risk premiums in the academic literature. In many applications important drivers are regressed on subsequent returns using linear models, see, e.g., Fama and French (2008) and Lewellen et al. (2015). However, there is a growing body of literature showing that some drivers have a non-linear relationship to risk premiums. Gu et al. (2020) compare a variety of statistical and machine learning models and find neural networks and regression trees to be best, statistically and economically. Especially their ability to include interactions between variables is named as an important advantage. This is also in line with findings of Bryzgalova et al. (2020), who use decision trees to group similar stocks together and put a special emphasis on their interactions. Their portfolio sorts show up to three times higher Sharpe ratios in the cross-section compared to traditional sorts and machine learning prediction-based portfolios. The importance of interactions between firm-specific characteristics is also confirmed by Chen et al. (2020) who apply a sequence of advanced machine learning algorithms to estimate an asset pricing model for individual stock returns. Furthermore, they include a no-arbitrage condition as criterion, which increases the performance. Freyberger et al. (2020) use adaptive group LASSO to select the most important characteristics for expected return predictions. They find that only a small number of predictors have an (time-varying) incremental explanatory power and non-linear relationships matter. Feng et al. (2020) use hidden states of a neural network to reduce the dimension of the input, which automatically allow non-linearities and interactions. Rossi (2018) use boosted regression trees to forecast stork returns and volatility. He finds a stronger predictive performance compared to linear models. Furthermore, he use the machine learning algorithm to construct mean-variance efficient portfolios and document a superior performance

compared to the linear framework.[1] Recently, also risk premiums of bonds and hedge funds are targeted with machine learning. Bianchi et al. (2020) apply a battery of machine learning algorithms to forecast bond returns and find neural networks and extreme trees to perform best. Wu et al. (2021) applies machine learning methods to forecast hedge fund returns and use them for selection. Again, neural networks are the best choice.

Summarizing, there is a broad evidence that the dependence of predictors and risk premiums is non-linear, which results in a superior performance of machine learning methods. Furthermore, Freyberger et al. (2020) document a time-varying impact of some predictors. The paper contributes to the literature by addressing two important questions left unanswered so far.

How much non-linearity is actually modelled by machine learning methods?

And, with respect to the time-varying nature of the stock market and the different phases of the business cycle, a second question emerges:

Does the amount of non-linear dependencies vary over time?

These questions are the next step to Freyberger et al. (2020), as the focus is on the shape of the relationship in addition to the impact. This may shed light at the economic mechanisms buried in the hidden dynamics of risk premiums. Furthermore, this can extend the understanding how firm characteristics drive risk premiums in different phases of the business cycle. To answer this important questions this paper extends two well known approaches of the explainable machine learning literature. First, a novel model agnostic measure of non-linearity is proposed, which builds on the work by Apley and Zhu (2020). This new measure quantifies the amount of non-linearity in predictions in one single number. Therefore, it is an easy, widely applicable and intuitive way to answer the question how much non-linearity is actually modelled by the machine learning algorithm. To identify the drivers of this non-linearity, the work by Sadhwani et al. (2021) is extended to quantify the non-linear relationship of every variable in one single number. Furthermore, we illustrate a way to quantify the direction of impact, additionally to the overall importance of a predictor variable calculated so far.

[1] Another strand of literature focusses on the application of machine learning methods to factor models. For example, Kelly et al. (2019) focus on the extension of traditional principal component analysis (PCA) and show that their instrumented PCA explains the cross-section of average returns significantly more than existing factor models. To capture the time variation of factor models non-parametrically, Pelger (2020) applies PCA to high-frequency data. Pelger and Xiong (2021) document the importance of macroeconomic states in capturing the time-variation in PCA-based factors. Lettau and Pelger (2020) refine the PCA to include no-arbitrage restrictions. This penalty helps to overcome the imminent signal-to-noise ratio problem in financial data.

The empirical results show that the share of non-linearity in risk premium predictions is around 15% in non-crisis periods, but more than doubles in crisis periods. Thereby, this paper documents an inverse relationship of linearity in risk premium predictions and uncertainty measured by the VIX. That is, in crisis periods and periods of high uncertainty the non-linearity increases considerably but is rather low in certain periods. This accompanies findings of Adrian et al. (2019) who find a non-linear dependence between VIX and stock as well as bond returns. Furthermore, Jackson et al. (2020) documents a large non-linear relationship between the VIX and the real economy. The novel non-linearity measure is more general in the sense, that it quantifies the non-linearity of all predictor variables. The empirical findings suggest that especially stock-level volatility measures show large non-linear behaviours in periods with high uncertainty. Furthermore, the study documents joint effects of variables with large impacts in crisis periods. The novel non-linearity measure is not only of relevance for the finance literature, but for all fields of science which use machine learning models and want to unveil the deep hidden relationships modelled.

The remainder of this paper is structured as follows. Section 4.2 reviews methods to detect non-linear relationships in machine learning methods. Section 4.3 describes the data. Section 4.4 introduces neural networks and the novel non-linearity measure. Subsequently, Section 4.5 provides the empirical results and Section 4.6 concludes.

4.2 Literature Review

Opening the black-box of machine learning models to identify the important drivers or to investigate the modelled non-linear relationships gained interest in recent years. This strand of literature is commonly labelled as explainable artificial intelligence (XAI). As the methodical contributions of this paper build on existing methods, the following section gives a short overview of other ways to find important drivers in machine learning methods. The section starts with model agnostic methods, which can be used for any machine learning algorithm and followed by methods tailored to neural networks and finally put emphasis on the detection of joint and higher order effects, which is new to the literature.

One of the earliest and most intuitive interpretation method are graphical explanations, starting with the partial dependence plot (PDP) introduced by Friedman (2001). This plot shows the average marginal effect of a feature by varying over its marginal distribution. By taking the average, positive and negative values can cancel out and, thus, the plots can be misleading. A solution to this problem are so called individual conditional expectation (ICE) plots of Goldstein et al. (2015). An ICE plot visualizes the dependence of the prediction on a feature for each observation separately, resulting in one line per observation. An extension to PDP can be found in Accumulated Local Effects (ALE) plots by Apley and Zhu (2020). The feature space is divided in several sub intervals to compute the difference in prediction, focusing on the conditional distribution of the features. Especially this method solves many problems of PDP and ICE and may therefore seen as the best choice to graphically open the black-box of machine learning methods. It is also closely related to gradient based methods, which are introduced later.

Accessing the importance of variables by permutation was introduced for random forests by Breiman (2001). The basic idea is, that if one permutes (ignores) the values of a given (important) feature, the loss function must increase. The higher the increase of loss, the more important is the feature for the machine learning model. However, the model usually has to be refitted a several hundred times for each feature. Hence, the computational burden sharply rises with the number of data, features and model complexity. An alternative permutation method are shapely additive explanations (SHAP) introduced by Lundberg and Lee (2016). SHAP relies on the theory of coalition games and is the only explainability method with an economic foundation. Instead of answering the question what *increased* the loss of the prediction, SHAP asks how did the feature *contribute* to the prediction. However, this approach is computationally expensive and a full representation of the data is in most empirical applications infeasible. Similar to

SHAP, local interpretable model-agnostic explanations (LIME) method introduced by Ribeiro et al. (2016) tries to approximate black-box predictions locally. The more interpretable models can be linear regressions, but also tree based methods. With all the methods in this paragraph, the underlying relationship between predictor and prediction cannot be lifted, as their focus is on identifying the most important variable and not identifying their shape of relationship.

Next, we focus on explanation methods tailored to neural networks. They are in these terms special, as their estimation procedure, i.e., backpropagation using gradient information, offers straightforward and intuitive ways to explain their predictions. The first order gradient is commonly known as marginal effect. Hence, one tries to answer the question "How much does the prediction change if the value of x_{1i} changes?". In the machine learning literature this approach is also called sensitivity analysis. The gradient can be calculated for every observation i and then aggregated to a global explanation by taking the mean of the actual values or of the squared values to avoid that positive and negative importance cancel each other out. Sadhwani et al. (2021) shows that squared gradients can easily represent important features and Horel and Giesecke (2020) derive a test statistic for single layer neural networks. The latter authors are the first to provide a sound statistical test statistic for feature importance of neural networks.

The main interest of this paper is the quantification of higher order and joint effects to illustrate the potential non-linear relationship. In most cases, the superior performance of machine learning methods is attributed to capture these effects. However, there are only a little number of studies which try to quantify these effects and show how much of the performance can be attributed to these effects.

Friedman and Popescu (2008) derive Friedman's H-statistic based on partial dependence plots. The interaction effect is defined as the share of variance that is explained by the interaction. ALE Plots by Apley and Zhu (2020) can also be used to visualize second-order importances, i.e. pair-wise interactions. Nevertheless, as the number of variables p in the data rises, the number of plots for the second order effects rises with $\frac{p^2-p}{2}$, i.e, with $p = 5$ in total $\frac{5^2-5}{2} = 10$ plots and with $p = 25$ one would have to interpret $\frac{25^2-25}{2} = 300$ plots. However, Apley and Zhu (2020) offer a R^2-like measure to quantify the explanation power of main effects, i.e. single variable effects, and higher order effects, i.e. pairwise or higher order interaction effects. Furthermore, Sadhwani et al. (2021) extend the first order gradient methods to cross derivatives for calculating the importance of joint effects. This gradient information can easily be aggregated to give a clear overview of most important variables.

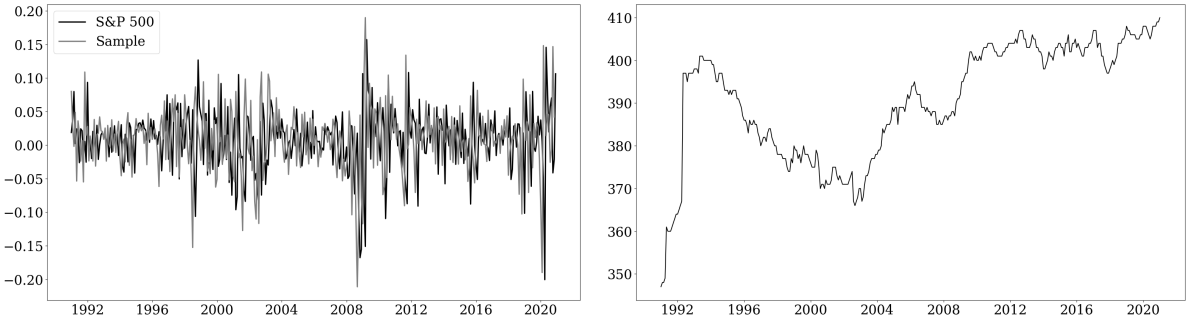
This paper builds on Apley and Zhu (2020) and Sadhwani et al. (2021) to derive a novel measure of non-linearity for machine learning predictions and trace this non-linearity back to specific variables. Both extensions aggregate the amount of non-linearity in the overall model respectively for every single variable into one singe number. Combining these two extensions gives a deep dive into the relationships modelled by machine learning methods. This may help economists to extract hidden patterns and derive new insights on complex relations.

4.3 Data

The data for this paper consist of monthly observations for companies listed in the S&P500 from 01.01.1991 until 31.12.2020. We gather a comprehensive collection of stock-level characteristics following Lewellen et al. (2015). They are retrieved from Thomson Reuters Datastream and Worldscope. We check at the beginning of every month which stocks are included in the index and use their history for training, validation and testing. In total, we have information about 1,135 companies over 30 years using 26 stock level characteristics and the VIX to forecast the next month return.

Figure 4.1: Sample comparison to the S&P 500

(a) Mean return **(b)** Number of constitutens



The plot on the left hand side shows the monthly returns of the Standard & Poors 500 Composite index illustrated by the black solid line and the monthly returns of the constituents for which we have all available variables illustrated by the grey solid line. Both lines show a very comparable evolution over time. The plot on the right hand side shows the number of constituents for which we have all variable information. The number is low in earlier years, but increases over time. On average we have information for roughly 390 out of 505 constituents. Therefore, we may label our sample as representative for the majority of the index constituents.

Figure 4.1 shows on the left a comparison of the monthly mean returns of the S&P 500 and the constituents for which we have all firm characteristics. Over this long history of data, we are not able to retrieve the firm characteristics for all roughly 500 constituents in every month.

However, following Panel (a) in Figure 4.1, we can conclude that the monthly returns are similar and, thus, our sample roughly follows the S&P 500. Panel (b) on the right shows the number of

constituents for which we have all firm characteristics value. In the early years, the number of complete information sets is rather low but increases until the end of our sample period. Over the whole timeline, we have roughly 390 constituents per month. We opt against replacing non-available firm characteristics with their monthly sample mean, as e.g., Green et al. (2017), as this may distort our feature importance calculation. We winsorize the data at the 1th and 99th percent quantile every month. This is done on a monthly basis to ensure that we do not incorporate information of future data points or constituents not included in the index any more. Furthermore, we standardize the variables to lie within the range of $[-3, 3]$, similar to Gu et al. (2020). The standardization is based on the training data set values and the standardization scheme is applied to validation and test sample.

Table 4.1: Overview of firm characteristics

Acronym	Firm characteristic	Original paper	Frequency
Accruals	Change in working capital from $t-13$ to $t-1$ divided by book value of common equity $t-1$	Sloan (1996)	annual
Beta	CAPM beta using excess market returns on excess stock returns over previous 60 months	Fama and MacBeth (1973)	monthly
Beta_{daily}	CAPM beta using daily excess market returns on excess stock returns over previous 12 months	Fama and MacBeth (1973)	monthly
BM	log book-to-market ratio	Rosenberg et al. (1985)	annual
CashRatio	Log of cash and equivalents divided by total debt	Ou and Penman (1989)	annual
CF/Price	Funds from operations divided by market capitalization	Asness et al. (2000)	annual
Debt/Price	Log of total debt divided by market capitalization	Bhandari (1988)	annual
DY	Dividends over previous 12 months to end of month share price	Litzenberger and Ramaswamy (1982)	annual
EarningsGrowth	Relative change in net income after preferred dividends from $t-13$ to $t-1$	Basu (1977)	annual
Earnings/Price	Net income after preferred dividends divided by market capitalization	Basu (1977)	annual
GrossProfit	Net sales or revenues minus cost of goods sold divided by total assets	Novy-Marx (2013)	annual
Investment	Relative change in total assets from $t-13$ to $t-1$	Fairfield et al. (2003)	annual
Issues	Log change of split-adjusted number of shares outstanding from $t-36$ to $t-1$	Pontiff and Woodgate (2008)	annual
Ivol	Log idiosyncratic volatility from beta and excess market return volatility on previous 60 months	Fama and MacBeth (1973)	monthly
Ivol_{daily}	Idiosyn. CAPM volatility of daily prices over the last 12 months	Fama and MacBeth (1973)	monthly
MAX	Maximum daily return in previous month	Bali et al. (2011)	monthly
$\text{Mom}_{1,0}$	Short term reversal based on last month's return	Jegadeesh and Titman (1993)	monthly
$\text{Mom}_{12,2}$	Momentum based on return of previous month $t-12$ to $t-2$	Jegadeesh (1990)	monthly
$\text{Mom}_{36,13}$	Momentum based on return of previous month $t-13$ to $t-36$	Jegadeesh and Titman (1993)	monthly
$\text{Mom}_{60,13}$	Momentum based on return of previous month $t-60$ to $t-13$	Jegadeesh and Titman (1993)	monthly
SalesGrowth	Log change of net sales or revenues from $t-13$ to $t-1$	Lakonishok et al. (1994)	annual
Sales/Price	Log of net sales or revenues divided by market capitalization	Barbee Jr et al. (1996)	annual
Size	Log end of month market capitalization	Banz (1981)	monthly
Turnover	Log stock turnover by volume divided by number of shares outstanding	Datar et al. (1998)	monthly
Vol	Log excess stock return volatility based on previous 60 months	Ang et al. (2006)	monthly
Vol_{daily}	Log excess stock return volatility based on previous 12 months of daily data	Ang et al. (2006)	monthly

This table provides an overview of every employed variable in this study. The first column shows the acronym used in subsequent sections. A description of every employed variable can be found in the second column. Furthermore, the original paper which introduced the variable as an important factor for return prediction is provided in the third column. The last column shows the available frequency of the variable. In total half of the variables are available on monthly frequency and half can be retrieved annually. The market index returns and the risk-free return are from Kenneth French's website http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Table 4.1 describes the retrieved firm characteristics, the original paper which introduced the firm characteristics and the accompanying frequency. The selected variables are a subset of the 94 company specific variables used in Gu et al. (2020). They have information of about roughly 6,000 firms every month, whereas the sample of this paper consists of roughly 390 firms per months. Therefore, the number of variables are reduced to avoid the change of overfitting. To do so, we used the most prominent ones in the literature, following Lewellen et al. (2015). Furthermore, the empirical analysis in Section 4.5 identifies the same important drivers as in Gu et al. (2020). Therefore, we might argue that the overall findings of this paper also hold in larger samples, as the main drivers and conclusions of Gu et al. (2020) can be recovered. We estimate stock volatility measures in two frequencies. For long-term relationships we use the previous 60 months and for short-term relationships we use daily data of the previous 12 month. The rationale behind this is to have a vivid and fast reacting variable on the one hand, but also long-term information on the other hand. Figure 4.A.1 and 4.A.2 in Appendix 4.A show their evolution over time for the interested reader.

4.4 Methods

The great popularity of neural networks can be traced baked to their theoretical foundation. The universal approximation theorem states that they can represent any smooth connection between predictors and predictions (Cybenko, 1989; Hornik, 1991). This is probably the reason, why neural networks gain large interest in a variety of scientific fields. Their flexibility stems from their information processing using subsequent non-linear transformations.

A neural network consists basically of three types of layers. The first one is the input layer, which entails the predictor information. Subsequently, we find hidden layers with non-linear activation functions. The final layer is the output layer which contains the final prediction. More formally, the neural network starts with covariate matrix $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ as inputs in the input neurons. The network subsequently shows stacked hidden layers $l = 1, ..., L$ whereby each layer entails K_l neurons $\boldsymbol{h}_l \in \mathbb{R}^{K_l}$ that are determined by an affine combination of neurons in the previous layer. These are processed by an arbitrary (non-linear) activation function σ.

$$\boldsymbol{h}_l = \sigma\left(\boldsymbol{W}_l \boldsymbol{h}_{(l-1)} + \boldsymbol{b}_l\right)$$

with $\boldsymbol{W}_l \in \mathbb{R}^{K_l \times K_{l-1}}, \boldsymbol{b}_l \in \mathbb{R}^{K_l}$ as parameters which are usually called weights and biases. The final

prediction $f(\boldsymbol{X})$ is derived from the last layer, the so-called output layer $\boldsymbol{h}_{L+1}$ and is given by choosing the identity function for σ, resulting in:

$$f(\boldsymbol{X}) = \boldsymbol{h}_{L+1} = \boldsymbol{W}_{L+1}\boldsymbol{h}_L + \boldsymbol{b}_{L+1}.$$

The weights and biases are estimated via a backpropagation algorithm based on Rumelhart et al. (1986). This paper benchmarks the neural network to a linear model using the Huber loss instead of the mean squared error (MSE) loss, following Huber (1964). The Huber loss is commonly used if the data exhibit extreme observation to make the inference and prediction more robust. The MSE weighs large errors very much, which can reduce the stability of all predictions. Especially heavy tails, i.e., very low or high returns, are very common in the context of individual stock returns and, thus, robustness is of major concern. For an excellent and more detailed introduction to neural networks and the discussion whether to use the Huber loss instead of MSE, we refer to Gu et al. (2020).

Quantifying important variables and non-linearity

This paper uses deep neural networks and, thus, follows Sadhwani et al. (2021) to quantify important drivers. The focus is on the "learned" relations of the neural network and therefore all importance measures are estimated on the training data.

The first order feature importance $\theta_\tau^{First}(x_r)$ quantifies the overall importance of an input variable $r = 1,\dots,p$. It is defined as:

$$\theta^{First}(x_r) = \frac{1}{C}\operatorname{sgn}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial f(\boldsymbol{x}_i)}{\partial x_{ir}}\right)\right)\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial f(\boldsymbol{x}_i)}{\partial x_{ir}}\right)^2}, \tag{4.1}$$

with $\boldsymbol{x}_i \in \mathbb{R}^p$ as a vector of p covariates for any observation $i = 1,...,N$.

$\theta^{First}(x_r)$ is the feature importance of covariate x_r and C is a normalizing constant that ensures $\sum_{r=1}^{p}|\theta^{First}(x_r)| = 1$. The $\operatorname{sgn}(\cdot)$ operator defines the direction in which the feature drives the prediction. This feature importance employs the gradient for every covariate x_r in relation to the individual prediction $f(\boldsymbol{x}_i)$. The gradients are squared to avoid cancellations of positive and negative values. Furthermore, the normalization allows a quick interpretation of relative importance.

The novelty of calculating the importance of any predictor variable with Equation (4.1) lies in the sign operator, which is an extension to Sadhwani et al. (2021). Usually, in the financial

context, but also in many other fields, we don't just want to know what is important but also in what direction drives the variable our prediction. For example, does the risk premium increasing or decreasing with a larger market capitalization of the company? The extension of this paper also quantifies the direction of the feature importance by taking the mean values of the gradients for variable x_r. This is a simple, but efficient way to leverage the direction of impact as well.

The gradients can be also used to quantify joint impacts of features using cross-derivatives, see Sadhwani et al. (2021). As a further extension, this paper calculates the second partial derivative with respect to the same input to quantify the (single) non-linear impact. Taking the second partial derivative with respect to the same variable gives a simple and cheap quantification of the non-linearity modelled by the neural network. If this value is zero, the predictor relates linearly to the predictions. Otherwise, there is non-linearity. The second order feature importance $\theta^{Second}(x_r)$ measures the extent of non-linear relationships of an input variable r and $\theta^{Joint}(x_{rs})$ quantifies the strength of joint effects of two variables r and $s = 1, \ldots, p$. [2]

$$\theta^{Second}(x_r) = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\frac{\partial^2 f(\boldsymbol{x}_i)}{\partial x_{ir} \partial x_{ir}} \right)^2}, \tag{4.2}$$

$$\theta^{Joint}(x_{rs}) = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\frac{\partial^2 f(\boldsymbol{x}_i)}{\partial x_{ir} \partial x_{is}} \right)^2}. \tag{4.3}$$

If $\theta^{Joint}(x_{rs})$ and $\theta^{Second}(x_r)$ are close to zero there are no single non-linear and joint impacts of the input variables.

The gradient based methods summarize the importance, non-linearity and interaction in one single metric. However, for some neural networks a more in-depth investigation of the modelled relationship between input and final prediction could be interesting. The ALE Plots by Apley and Zhu (2020) rely on gradients as well and, thus, can be seen as an graphical illustration of $\theta^{First}(x_r)$, $\theta^{Second}(x_r)$ and $\theta^{Joint}(x_{rs})$. They introduce $\theta^{Main}_{ALE}(x_r)$ as the marginal relationship between the input variable x_r and the prediction. $\theta^{Joint}_{ALE}(x_{rs})$ calculates the impact of joint effects over the conditional distribution of x_r and x_s. For the sake of clarity, we refer for mathematical expressions and derivations to the original paper. Apley and Zhu (2020) proposed also an

[2] We do not calculate the direction of joint impacts, as the interpretation is tedious. We would have to take into account the sign of the first order importance as well. Rather, we are interested in how much non-linearity is modelled.

R^2 like measure which quantifies how well each effect ($\theta_{ALE}^{Main}(x_r)$ and $\theta_{ALE}^{Joint}(x_{rs})$) explains the prediction of the neural network $f(\boldsymbol{X})$. They define their R^2 measure as:

$$R_{Main}^{2,ALE} = \frac{\text{var}\left\{\sum_{r=1}^{p} \theta_{ALE}^{Main}(x_r)\right\}}{\text{var}\left\{f(\boldsymbol{X})\right\}}, \tag{4.4}$$

$$R_{Joint}^{2,ALE} = \frac{\text{var}\left\{\sum_{r=1}^{p} \theta_{ALE}^{Main}(x_r) + \sum_{r=1}^{p}\sum_{s=1}^{p} \theta_{ALE}^{Joint}(x_{rs})\right\}}{\text{var}\left\{f(\boldsymbol{X})\right\}}. \tag{4.5}$$

The R^2 measure can be extended up to order p, which would imply $R_p^{2ALE} = 1$. This paper extents their R^2 measure one step further to quantify how much of the prediction can be explained by using only linear relationships. That can be achieved by fitting an OLS regression on the values of $\theta_{ALE}^{Main}(x_r)$, labelled as $\theta_{ALE}^{Linear}(x_r)$.

If the relation between the input variable and the prediction is linear, the regression line coincides with the graph of $\theta_{ALE}^{Main}(x_r)$. To quantify the amount of non-linearity, we reformulate the R^2 measure as:

$$R_{Linear}^{2,ALE} = \frac{\text{var}\left\{\sum_{r=1}^{p} \theta_{ALE}^{Linear}(x_r)\right\}}{\text{var}\left\{f(\boldsymbol{X})\right\}}. \tag{4.6}$$

Therefore 1- $R_{Linear}^{2,ALE}$ can be seen as a novel measure of how much non-linearity is modelled. This is a simple extension to Apley and Zhu (2020) but a clear and intuitive measure of non-linearity. Furthermore, the additional calculation comes at almost no cost, as OLS regressions are estimated very quickly. Therefore, if one assesses the important drivers via the work by Apley and Zhu (2020), the additional quantification of how much non-linearity is modelled is an insightful but inexpensive information. Furthermore, this approach can be applied to any machine learning methods, as the approach by Apley and Zhu (2020) is model agnostic. Therefore, this novel non-linearity measure may be helpful for a variety of scientific fields, where machine learning methods are used.

Table 4.2 shows three examples to illustrate these R^2 measures, starting with a very simple linear model. The second example models a non-linear relationship between the variable and the target. The last example models marginal non-linearity and a pairwise interaction to allow for joint effects. The independent variables follow a uniform distribution $\mathcal{U} \sim [-3;3]$ and ϵ follows a normal distribution $\mathcal{N} \sim (0, 0.01)$. The Data Generating Processes (DGP) are simulated

with 25,000 observations and fitted using a neural network with two hidden layers with 256 neurons each. We use the Adam algorithm (Kingma and Ba, 2014) with a learning rate of 0.0001 and 500 epochs.

Table 4.2: Illustrative examples

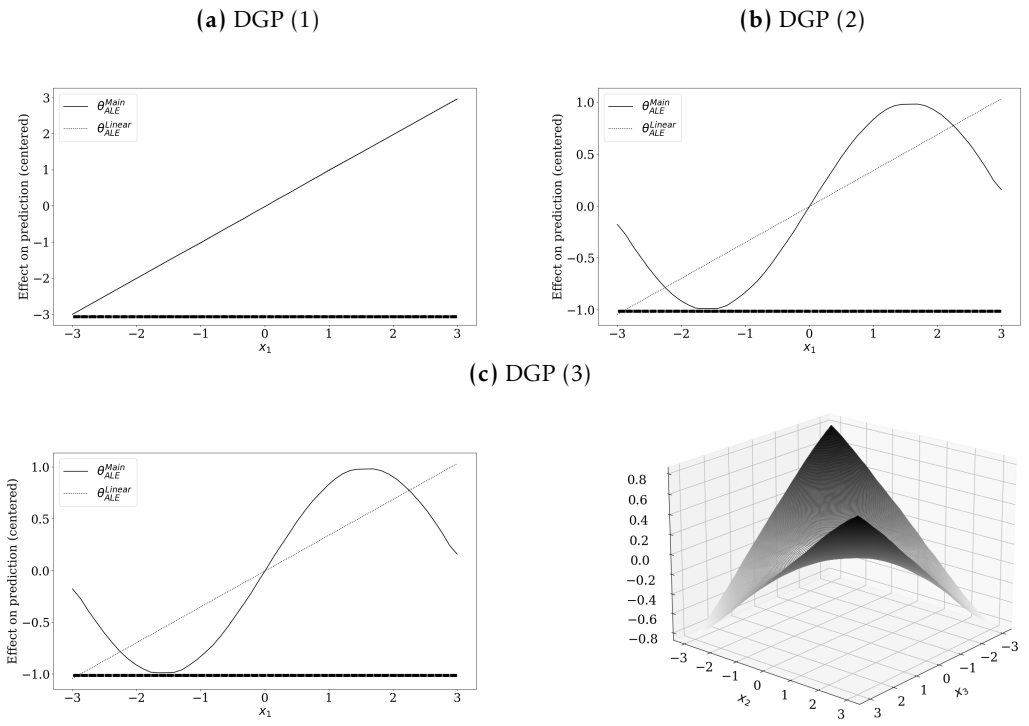
	DGP	$R^{2,ALE}_{Linear}$	$R^{2,ALE}_{Main}$	$R^{2,ALE}_{Joint}$
(1)	$y \sim x_1 + \epsilon$	1.00	1.00	1.00
(2)	$y \sim \sin(x_1) + \epsilon$	0.69	1.00	1.00
(3)	$y \sim \sin(x_1) + 0.1 \cdot x_2 x_3 + \epsilon$	0.60	0.86	1.00

Note: This table shows three simple examples to show how the R^2 measures behave and how they can be interpreted. The first example consists of a simple linear model with only one (linearly dependent) variable and, thus, all three R^2 measures are equal to one. Example two and three add marginal non-linearity via the sinus function or joint effects via a pairwise interaction. We can see that the $R^{2,ALE}_{Linear}$ decreases as other effects are added. Please note that the $R^{2,ALE}_{Linear}$ in example two coincides with the standard R^2 measures of the OLS regression, as we use only one explanatory variable.

Starting with the first row, we can see that all three R^2 measures are equal to one. This is expected as there are no non-linear or joint effects and, thus, the prediction can be replicated using only linear relationships. In the second example non-linearity via the sinus function is introduced, but no joint effects. We can see that the $R^{2,ALE}_{Linear}$ drops to 0.69 as the sinus function cannot be approximated properly with a linear relationship. $R^{2,ALE}_{Main}$ and $R^{2,ALE}_{Joint}$ are equal to 1 as $\theta^{Main}_{ALE}(x_1)$ approximates the non-linear relationship well and no joint effects are present. The last example models additionally a pairwise interaction to allow for joint effects. The $R^{2,ALE}_{Linear}$ drops further to 0.60 as the overall non-linearity increased. $R^{2,ALE}_{Main}$ shows a value 0.86, as the joints effects are neglected. The $R^{2,ALE}_{Joint}$ has a value of 1 as now the joint effect of x_2 and x_3 is now incorporated. These three simple examples show that, using the strategy of Apley and Zhu (2020) with the extension of this paper, we can easily calculate how much non-linearity is modelled. Figure 4.2 shows the calculated values for $\theta^{Linear}_{ALE}(x_1), \theta^{Main}_{ALE}(x_1)$ and $\theta^{Joint}_{ALE}(x_{23})$.

In Panel a) we can see that $\theta^{Linear}_{ALE}(x_1)$ and $\theta^{Main}_{ALE}(x_1)$ coincide as the true DGP assumes a linear relationship between y and x_1. Panel b) shows the approximation error of $\theta^{Linear}_{ALE}(x_1)$. The linear approximation deviates strongly from the modelled non-linearity and, thus, explain the low value of 0.69 for $R^{2,ALE}_{Linear}$. Hence, only roughly 70% of the neural network prediction can be explained if we approximate the relationship linearly. Panel c) shows on the left hand side a very similar plot to Panel b), as in both examples a sinus function is assumed. This means that the $\theta^{Main}_{ALE}(x_1)$ recovers the true relation of x_1 with the prediction, irrespective of the additional interaction term. The right hand side shows the modelled joint effect. We see positive values if x_2 and x_3 move in the same direction and negative values if they move in the opposite direction.

Figure 4.2: ALE Plots of the illustrative examples



Note: These figures show $\theta_{ALE}^{Linear}(x_1), \theta_{ALE}^{Main}(x_1)$ and $\theta_{ALE}^{Joint}(x_{23})$ for the three illustrative examples. The latter two ALE Plots clearly approximate the true Data Generating Process.

Summarizing the three illustrative examples, we can calculate the amount of non-linearity modelled in the neural network by calculating $R^{2,ALE}_{Linear}$. The lower this value, the higher the amount of non-linearity. Furthermore, by calculating $R^{2,ALE}_{Main}$ and $R^{2,ALE}_{Joint}$, we can illustrate how much of the prediction we can explain by incorporating marginal non-linearities and pairwise interactions. This procedure could be extended to higher order interactions, but this would be infeasible to visualize and we will see in Section 4.5 that a very large amount of the prediction can be explained by $R^{2,ALE}_{Main}$ and $R^{2,ALE}_{Joint}$.

4.5 Empirical Results

As this paper focusses on the non-linearity over time, the hyperparameters of the neural network are evaluated along the timeline, instead on a cross-sectional basis. Two approaches are common in the literature, see Gu et al. (2020). First, the so-called ”rolling” scheme, where training and validation splits are shifted monthly holding the length of training and validation split constant, i.e., using the last five years of observations for training. The advantage of this approach is that only the ”most recent” information is used for training. An alternative is the so-called

"recursive" approach, where the training data is extended in each iteration, employed by Gu et al. (2020). The subsequent sections present the "rolling" window approach, as we want to quantify the non-linearities in the most recent time period. We use five years for the training data, three years for the validation data and one year for the final out-of-time prediction. Every constellation of the hyperparameters is fitted on the training data and subsequently applied to the validation data. The hyperparameter search is conducted on an annual basis to reduce the computational burden. Hence, a neural network is fitted at the end of every year and applied to the subsequent 12 months rolling window. The hyperparameters are sampled by a random search algorithm, which tests 1,000 constellation every year sampled from a predefined range. Overall, we controlled for overfitting and the dependence of the neural network on its weight initialization. Furthermore, we use advanced activation functions to reduce well-known problems, such as vanishing gradient and dying ReLU. For a detailed overview of the hyperparameter search and the activation functions, we refer to Appendix 4.B. As the composition of the S&P 500 index changes over time, also our sample changes, potentially inducing a survivorship bias. This can occur if one uses only stocks which are listed at the end of the sample periods, see, e.g., Brown et al. (1992) or Elton et al. (1996). To counteract this, we check which companies are listed in the S&P 500 at the end of every year and use their history for training and their future returns for validation and testing[3]. Overall, the aim of this study is not the replication of the S&P 500 or the application of trading strategies. Our aim is to show the determinants of the majority of the market capitalization in the U.S. stock market. Hence, the S&P 500 serves only as a guidance which companies to choose.

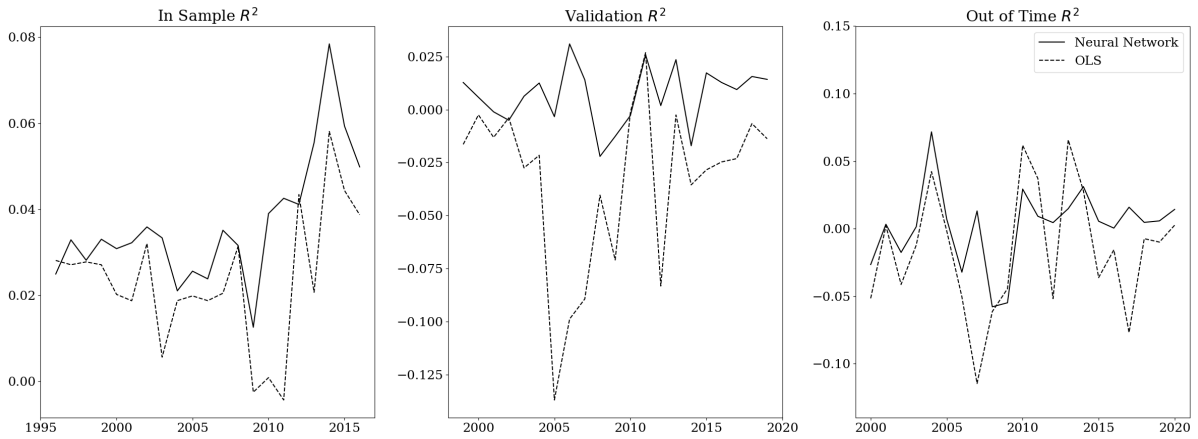
To compare the predictive performance for individual excess stock return forecasts of the neural network and the linear model, we follow Gu et al. (2020) and compute the following metric likewise for the training, validation and test sample:

$$R^2 = \frac{\sum_{i=1}^{\omega}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{\omega} y_i^2}. \tag{4.7}$$

The value of ω represents the number of observations in the training, validation or test sample. This metric benchmarks the predictions against a forecast value of 0 which is more suitable to assess the performance of individual stock return predictions (Gu et al., 2020). Figure 4.3 shows the annual mean of the R^2 measure for training, validation and testing dataset.

[3] This is a similar approach as Fischer and Krauss (2018), who use this procedure in the context of statistical arbitrage strategies.

Figure 4.3: Performance metrics and their evolution over time



This panel shows the R^2 measure of Gu et al. (2020) for the training, validation and test sample. This measure is calculated monthly using rolling window approach. For illustration purposes the annual mean is plotted. The dashed black line coincides to the values of the OLS, whereas the black solid line refers to the performance of the Neural Network. The higher the R^2 measure the better is the performance and, thus, we see some evidence that the Neural Network outperforms the OLS in most years and samples.

The first training sample ranges from January 1991 to December 1995. Hence, the calculated value coincides to the performance over these five years. The average R^2 for the linear model is negative in the validation and test sample, contrary to the neural network which results in positive values in all three samples. It outperforms the linear model in almost every year in the training and validation sample. In the testing sample, the performance is mixed, but in 21 of 26 years it outperforms the linear model. Furthermore, the estimated values of the neural networks vary considerably less than the values of the linear model.[4] The setting for the test sample is quite challenging as we use models fitted on data three years ago (length of the validation sample) to forecast the next 12 months of individual returns.[5]

In most papers using machine learning methods, the superior findings of these models are explained by the fact that they automatically model non-linear relations, including interactions. So there arises one obvious but unanswered question:

How much non-linearity is actually modelled?

With respect to the asset pricing literature and the aim to predict subsequent individual stock returns a second affiliating question follows.

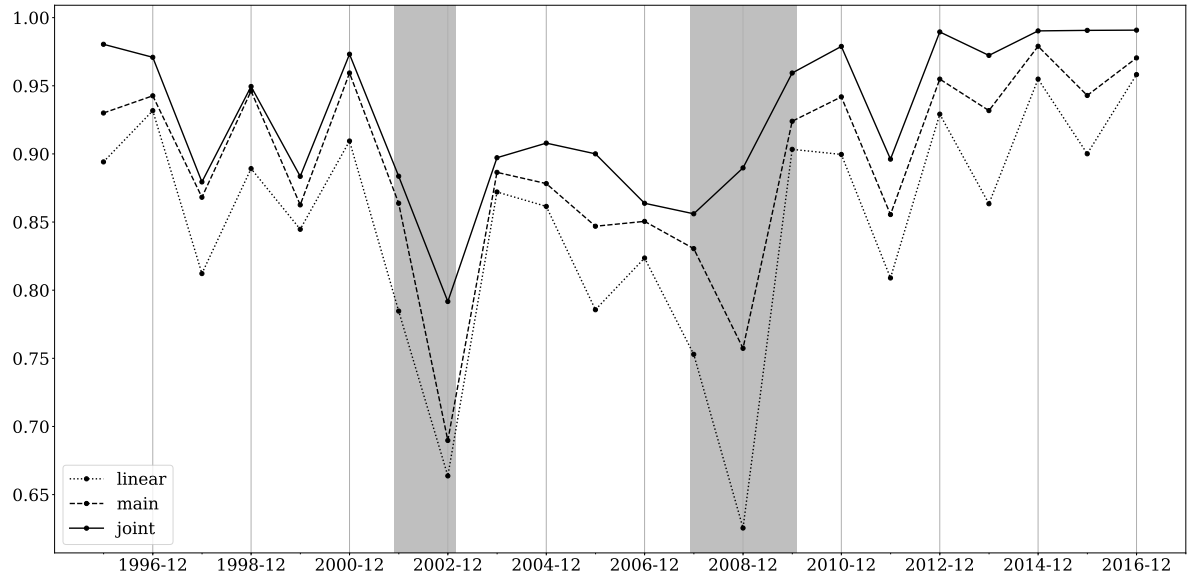
[4] This result is somewhat contrary to Gu et al. (2020) as they report large negative R^2 values for the linear model. The difference originates in the different number of variables used. Gu et al. (2020) used more than 900 variables, which increases the chance of overfitting and the occurrence of unstable estimates due to multicollinearity. However, in this paper, we use in total 26 variables and, thus, this problem is smaller which results in a smaller difference between linear model and neural network.

[5] One might argue that we could have use a larger training sample, e.g., 10 years or a shorter validation sample to boost the performance on the test sample. However, the main aim of this paper is not to find the most predictive model, as Gu et al. (2020) do, but to evaluate how much non-linearity is modelled over the business cycle. Hence, we argue for a shorter training sample to evaluate the drivers of the most recent time.

Is this non-linearity is stable over the business cycle?

These two central questions can be answered by using the R^2_{linear} measure. The difference between 1 and the R^2_{linear} can be interpreted as the amount of non-linearity modelled by the neural network.

Figure 4.4: R^2 by Apley and Zhu (2020) over time



The plot shows the estimated values for R^2_{linear},R^2_{main} and R^2_{joint} over time. The maximum values is 1, so the difference to this value can be interpreted as how much (higher-order) non-linearities or interactions are modelled by the Neural Network, but not included. For example, if the value of R^2_{joint} is 1, we can completely recover the predictions of the Neural Network for (non-linear) main effects and pairwise interactions. These values are estimated annually based on the five years used for training. The dot-com bubble and the Global Financial Crisis are specified according to the OECD recession indicator (available at https://fred.stlouisfed.org/release?rid=242) and shaded in grey.

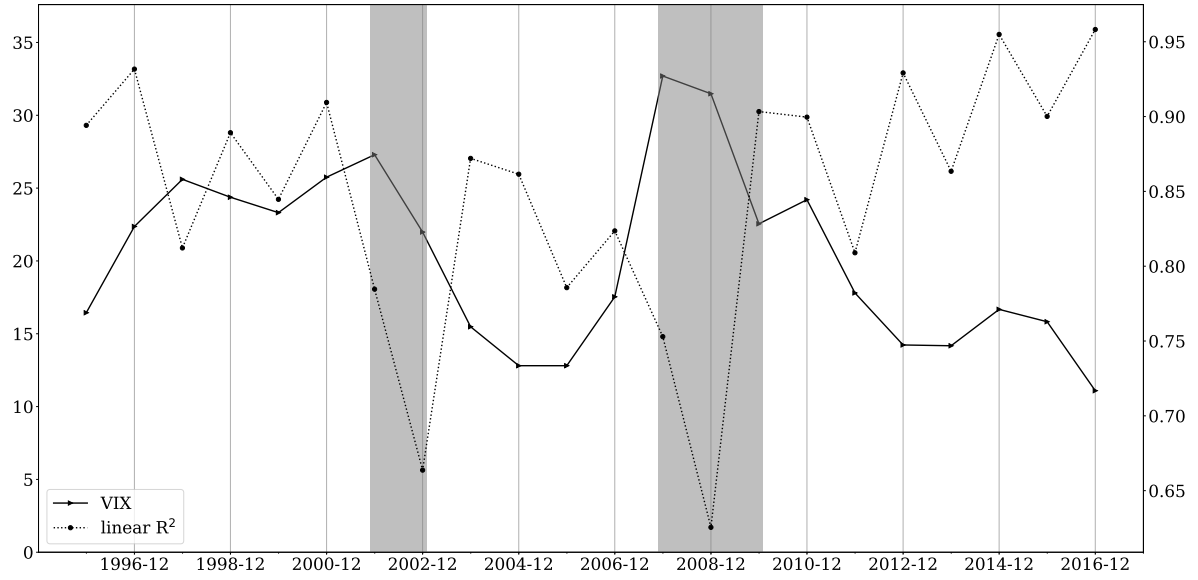
Figure 4.4 shows the estimated values for R^2_{linear},R^2_{main} and R^2_{joint} over time. The values are calculated on an annual basis using the last five years of data in the training sample. The values for all three measures vary over time and show different behaviours in crisis and non-crisis periods. The empirical analysis shows evidence that the non-linearity increases if crisis periods are included in the training sample. As a first crisis one can name the dot-com bubble in the early 2000s. The R^2_{linear} shows a first larger drop to 0.78 as the year 2001 is included in the training data and a subsequent one as the year 2002 is included. Until the end of 2000, the values of R^2_{linear} vary around 0.86, but drop to 0.67 at the end of 2002. that is, in crisis periods non-linear relationships are present and the neural networks approximate these. Hence, the neural network models a considerable amount of non-linearity in these crisis periods.

Furthermore, the levels of R^2_{joint} drop also considerably, which means that the higher order non-linearities and interactions are modelled. Moving to subsequent years, the values for all

R^2 measures increase until the Global Financial Crisis (GFC). A first drop to 0.76 can be seen as the year 2007 is included, marking the beginning of the upcoming turbulences. The lowest value of R^2_{linear} with 0.64 can be observed as the crisis year 2008 is included. After the crisis, the modelled non-linearity decreases. The predictions of the neural network in the years 2014 to 2016 can almost fully recovered from non-linear relationships and pairwise interactions, indicated by R^2_{joint} close to 1. The R^2_{linear} varies in non-crisis periods around 0.85. This means that a large part of the prediction can be recovered by approximating the $\theta^{ALE}_{main}(x_j)$ linearly and only a small part of the prediction requires non-linearity or interactions. The R^2_{joint} varies in non-crisis periods around 0.95, showing that the very large share of predictions can be recovered by non-linear relationships between variables and pairwise interactions.

Another interesting fact is the inverse relationship of the R^2_{linear} and the VIX, a prominent proxy for uncertainty in the financial market. Figure 4.5 shows the evolution of both over time. In times of high uncertainty, i.e. in crisis periods, the R^2_{linear} is considerably lower than in more certain periods.

Figure 4.5: R^2_{linear} and VIX over time



The plot shows the estimated values for R^2_{linear} and the VIX over time. The dotted black line indicates the R^2_{linear}, whereas the solid black line refers to the VIX. The left vertical axis refers to the VIX, whereas the right vertical axis refers to the R^2_{linear}. This plot indicates that in times of high uncertainty, indicated by high values for the VIX, the R^2_{linear} decreases. Hence, in times of high uncertainty more non-linearity is present. The dot-com bubble and the Global Financial Crisis are specified according to the OECD recession indicator (available at `https://fred.stlouisfed.org/release?rid=242`) and shaded in grey.

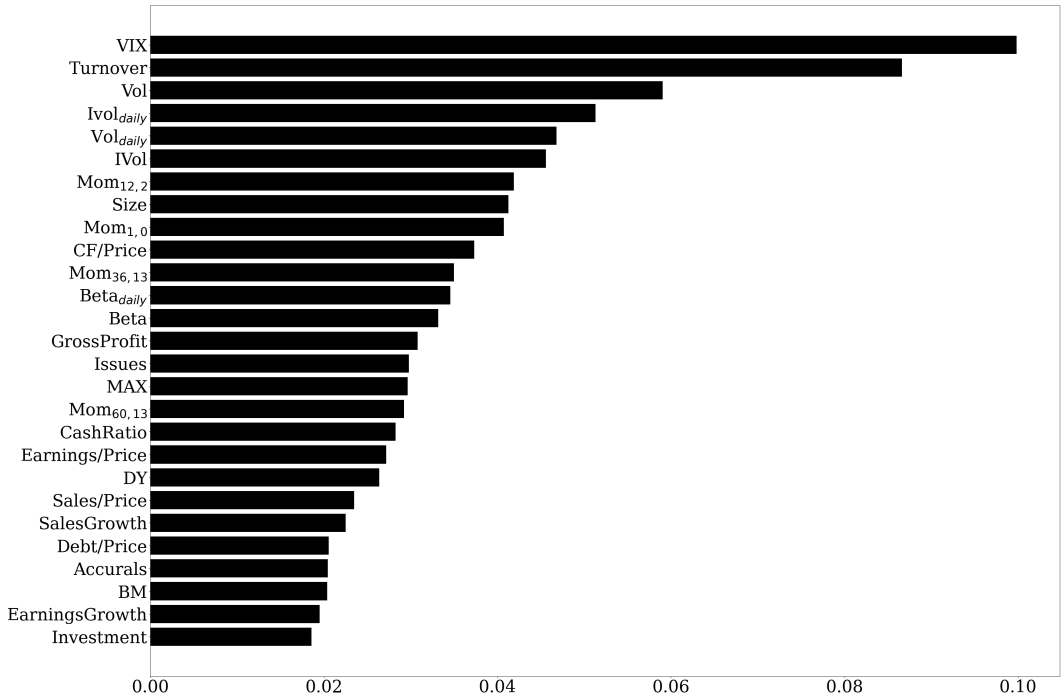
Figure 4.4 and 4.5 show that non-linearity increases in crisis and periods with high uncertainty. This accompanies evidence by Adrian et al. (2019), who document a non-linear dependence of the VIX on future returns. We go in the same direction by showing that the relationship of our input variables in general is non-linear and this non-linearity increases in periods of

high uncertainty. Our metric can be seen as a broader definition of non-linearity in terms of the overall model and is not restricted to single input variables. We document an inverse relationship between uncertainty and linear dependence in individual stock returns. The Pearson correlation coefficient ρ of R^2_{linear} and the VIX is -0.499 with a p-value of 0.017.[6] Hence, the empirical findings show that there is a statistically significant relationship between R^2_{linear} and the VIX. This is in line with Jackson et al. (2020), who find higher non-linearity of real economic variables, e.g., industrial production, real GDP growth or inflation, in times of higher uncertainty.

Most important drivers

The following paragraph aims at evaluating the most important drivers of the neural network predictions and shed light on their behaviour over time. Therefore, the gradient based methods and their extensions are used in the first place. Followed by an in-depth evaluation of the neural network showing the highest amount of non-linearity. Figure 4.6 starts with a global overview of the ten most important variables over the whole training period. It is calculated by taking the mean of the absolute values of $\theta^{First}(x_r)$ and normalize the result such that all sum up to 1.

Figure 4.6: Global values of $\theta^{First}(x_r)$



The plot shows the top 10 of most important drivers over the whole training sample. The values on the horizontal axis are calculated by taking the absolute mean of each variable of the timespan and normalised such that all means sum to 1. This figure reveals that the VIX is clearly the most dominating driver of future individual stock returns.

[6] The p-value corresponds to a two-sided test with $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$.

As may suspected from the previous findings, the VIX is clearly the most dominating variable over the whole training period.[7] The second most important variable is Turnover, which reflects how actively a share is traded. The following important variables all reflect the volatility of previous excess stock returns, calculated as the realized return volatility and based on the CAPM similar to Gu et al. (2020), who report the return volatility as the third most important variable.

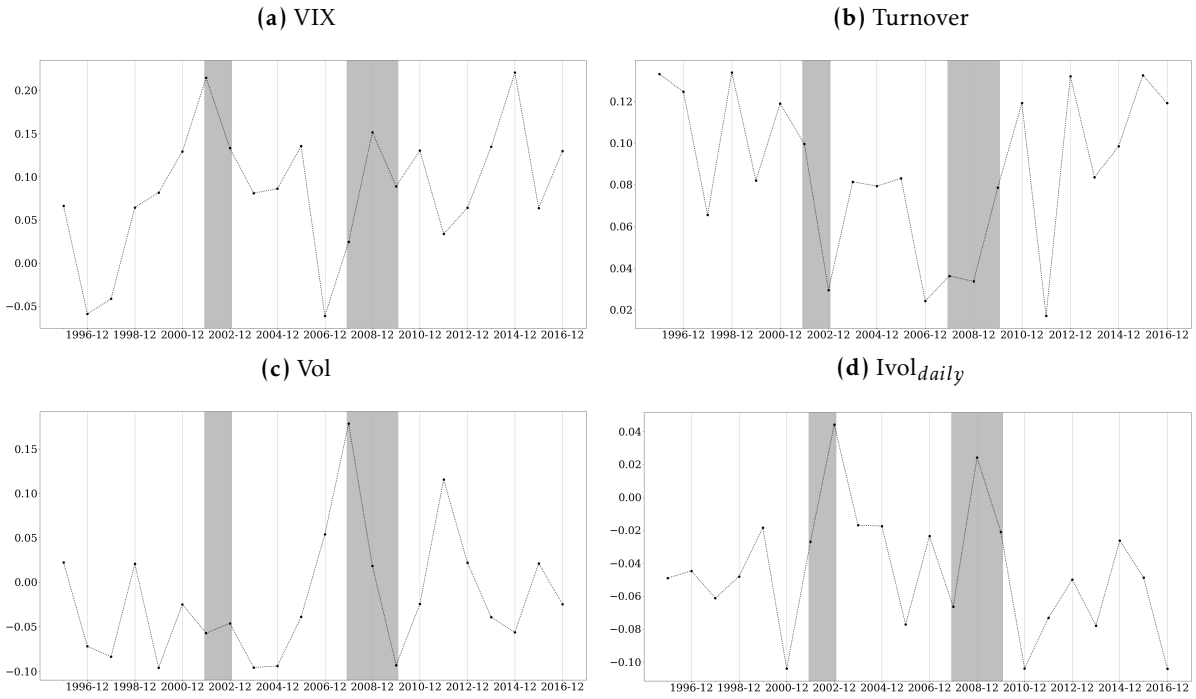
The next important variables are Momentum and the Size. This importances are in line with the findings of Gu et al. (2020). However, there are some differences in the raking of the variables. Gu et al. (2020) find their best performing neural network, labelled as NN3 in their paper, variables reflecting the Momentum and Size of the company as the most important ones. This may attributed to the fact that Gu et al. (2020) use a more broader universe of the stock universe and include much more smaller stocks. As this paper uses the largest stocks in the United States, the Size of the company may play a minor role. The same applies to momentum related variables, which are also more important for small stocks than for very large ones, see for example Fama and French (2008) or Novy-Marx (2012). Overall, the empirical analysis shows up with the same important drivers.

[7] We re-estimate the whole empirical section without the VIX as predictor to rule out the possibility that the inverse relationship between VIX and R^2_{linear} comes indirectly from the VIX. The values of the R^2 measures are slightly different, but the main conclusion holds. We can identify a negative relationship between R^2_{linear} and the VIX. The results are available upon request.

Time variation of important drivers

We focus now on the time variation of these importances. Figure 4.7 shows the evolution of $\theta^{First}(x_j)$ over time for the four most important variables.

Figure 4.7: Time variation of $\theta^{First}(x_j)$



Note: These figures show the time variation of $\theta^{First}(x_j)$ over the whole trainings period for the four most important variable. We can clearly see, that the calculated importances vary considerably over time and we frequently observe peaks around periods of high uncertainty and crisis. The dot-com bubble and the Global Financial Crisis are specified according to the OECD recession indicator (available at https://fred.stlouisfed.org/release?rid=242) and shaded in grey.

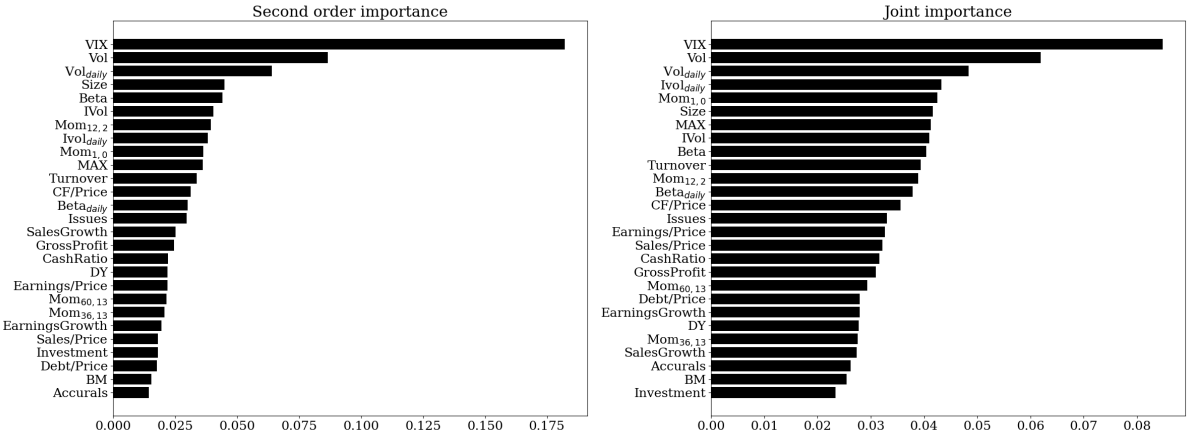
Throughout the variables we observe a considerable variation in time and most peaks of importance around the crisis and periods with high uncertainty. The values of $\theta^{First}(x_r)$ are normalized such that the absolute some in every year is equal to 1. The VIX shows for the most estimated values a positive sign, which is in line with Adrian et al. (2019) who shows that in times of high uncertainty the expected return is higher as well. The highest value corresponds the sample until the end of 2014 and accounts for more than 20% of the overall importance in this year. The remaining peaks lie around crisis periods. The second highest value can be observed in the sample covering the dot-com bubble end of 2001 covering roughly 20% and the third peak corresponds to the sample until the end of 2008, including the Global Financial Crisis. Hence, we may conclude that the importance of the VIX increases towards crisis periods and accounts for roughly one fifth of the total importance. The Turnover has a throughout positive sign, which is plausible as it measures the trading activity of the previous month. This

means, that a higher trading activity should result in higher expected returns.[8] The Vol peaks clearly in the sample of 2007, which marks the start of the Global Financial Crisis and a time of high uncertainty. This variable accounts for roughly 15 % of the overall importance in this year. Furthermore, the Ivol_{daily} peaks in both crisis periods with positive values. Usually, one would expect a negative sign as high volatility should lead to lower expected returns. This is true for all samples except these two peaks. This may indicate that there is a difference in crisis and non-crisis periods.

Most important non-linearities and joint effects

Figure 4.8 shows on the left hand side the variables with the most non-linear relationship over time and on the right hand side the most interacting variables over time. Both figures are calculated by taking the mean of the respective variable over time and normalize them such that the absolute sum is equal to 1. The following paragraph is an extension to McLean and Pontiff (2016) and Freyberger et al. (2020) who document a time variation of the predictor's impact on the return prediction. To the best of our knowledge, this is the first paper to evaluate the shape of relationship and joint effects of predictors over time. The values of $\theta^{Joint}(x_{rs})$ in every year are calculated taking sum of all pairwise interactions with all other variables for every variable x_r.

Figure 4.8: Higher order importance



The plots show the second order and joint effects over the whole training sample. The values on the horizontal axis are calculated by taking the absolute mean for each variable of the timespan and normalised such that all means sum to 1. This figure reveals that the VIX is clearly the most dominating source of non-linearity and joint effects.

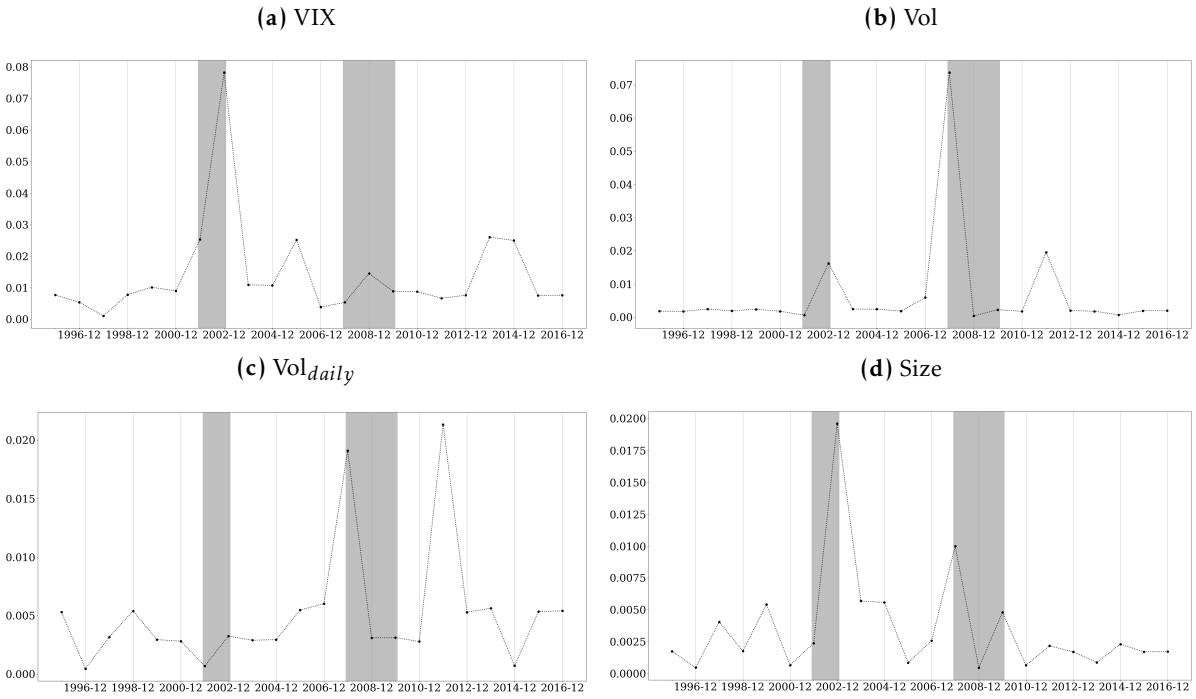
[8] The positive sign is contrary to the proposed relationship by Datar et al. (1998). This is due to two reasons. First, Datar et al. (1998) interpret their turnover variable as a proxy for liquidity and refer to the illiquidity premium, i.e., stocks with lower liquidity need to provide higher gross returns in comparison to more liquid assets. However, this paper uses the largest stocks in the United States and, thus, the liquidity is of minor concern. Second, we use a different definition of the turnover value. Datar et al. (1998) use the previous 12 months turnover and we opt for the last month turnover to get more temporary and vivid information about the current trading activity as the aim of this paper is to predict the next month return and not to explain premiums assets have to pay.

Abstracting from this figure we can see that the VIX shows the largest non-linear relationship over time, followed by the two measures of stock volatility Vol and Vol_{daily}. Overall, the ranking is quite similar to Figure 4.6. The top ranks of pairwise interactions, indicated by the joint importance, are filled by the VIX and three measures of individual stock volatility. This means that, over the whole timespan, the volatility measures show the largest interactions with other variables and, thus, have large joint effects. In both figures the momentum variables are on mid ranks.

Non-linearities and joint effects over time

Subsequently, this paper ask how the shape of the relationship between input variables and the prediction varies across time by looking at the values of $\theta^{Second}(x_r)$ along the timeline. If this value is zero, one would expect a linear dependence between input and prediction. Figure 4.9 shows the variation over time for the four most non-linear variables, namely the VIX, Vol, Vol_{daily} and Size.

Figure 4.9: Time variation of $\theta^{Second}(x_j)$

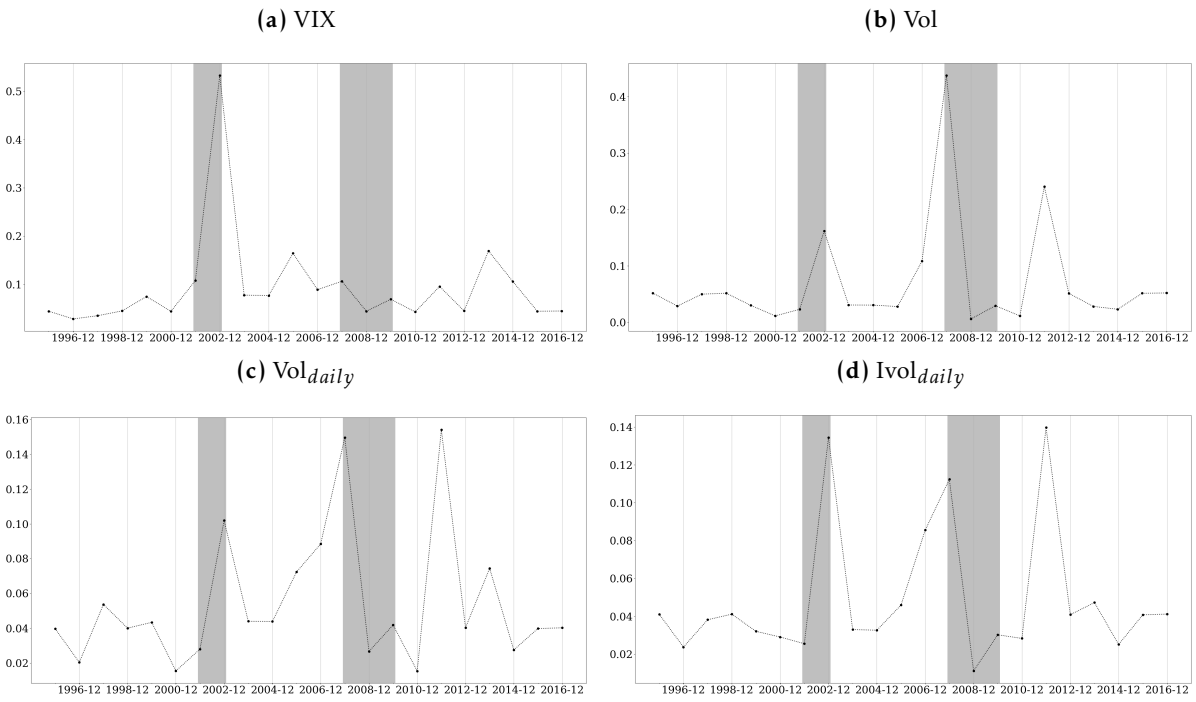


Note: These figures show the time variation of $\theta^{Second}(x_j)$ over the whole trainings period for the four most important variable. We can clearly see, that the calculated importances varies considerably over time and we frequently observe peaks around periods of high uncertainty and crisis. The dot-com bubble and the Global Financial Crisis are specified according to the OECD recession indicator (available at `https://fred.stlouisfed.org/release?rid=242`) and shaded in grey.

Overall, Figure 4.9 indicates that non-linearity is not always present and more specifically, in most time periods show up only very little. Remarkable exceptions are times of high uncertainty

and crisis periods. For example, we observe a large peak of the VIX non-linearity in the sample of 2001 and 2002. Hence, the relation of the VIX on the prediction is highly non-linear in this crisis period, compared to all other periods. Interestingly, in the Global Financial Crisis the VIX shows up a quite linear relationship. A possible reason could be that the dot-com bubble directly originated in the stock market and, thus, the VIX has a stronger non-linear relationship. The GFC mainly originated in the housing market and the pressure spilled over to the stock market subsequently. The Size variable shows high non-linearity in the 2002 sample, but also a considerable peak in 2007. The remaining volatility variables show small peaks in the 2002 samples, but major peaks in 2007. A considerable increase of non-linearity can also be observed in the 2011 sample may referring to the European Debt Crisis in 2011. A similar picture can be observed in Figure 4.10, where the sum of $\theta^{Joint}(x_j)$ for every variable with all other variables is illustrated over time.

Figure 4.10: Time variation of $\theta^{Joint}(x_j)$



Note: These figures show the time variation of $\theta^{Second}(x_j)$ over the whole trainings period for the four most important variable. We can clearly see, that the calculated importances varies considerably over time and we frequently observe peaks around periods of high uncertainty and crisis. The dot-com bubble and the Global Financial Crisis are specified according to the OECD recession indicator (available at `https://fred.stlouisfed.org/release?rid=242`) and shaded in grey.

The VIX shows the largest joint effects of every variable at the end of 2002, indicating that this uncertainty measure has had large joint effects in the dot-com crisis. The variable Vol, which covers the long term variation of the stock prices shows its highest peak at the Global Financial

Crisis, followed by the European Debt Crisis and the dot-com bubble. The remaining volatility variables show the peaks in the same periods.[9]

ALE Plots of selected variables

The rest of this section focuses on an in depth analysis of the neural network covering the sample until 2007. We choose this one, as its prediction entails the largest amount of non-linearity. Figure 4.11 shows the $\theta^{ALE}_{main}(x_r)$ for the four most important variables. If we observe an affine line in the ALE Plots, the neural network intrinsically models a linear relationship between the variable of interest and the prediction. Any deviation of this affine line can be interpreted as a non-linear dependence. The ALE plots can be interpreted as an visual exploitation of the gradient information used to calculate the shape of non-linearity in $\theta^{Second}(x_r)$ as both approaches are based on the gradient information. Hence, if the value of $\theta^{Second}(x_r)$ is high for a specific variable, the ALE Plot can be used to visualize the exact shape of the modelled non-linear relationship. The vertical axis of the ALE plot shows the impact on the prediction, whereas the horizontal axis shows the scaled values of the variable of interest. The marks on the vertical axis are a rug plot to illustrate the distribution of the data.[10] This allows us to interpret and compare their absolute impact on the final prediction.

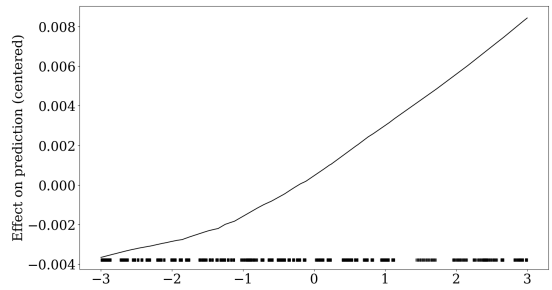
Comparing all vertical axes, we can see that the variable Vol shows the largest range of impact compared to all the others. This coincides with the very large value in Figure 4.7. It is plausible as both, the ALE Plot and $\theta^{First}(x_j)$ use the gradient information to assess the importance of the variable of interest. The graph of the variable VIX shows the expected positive slope indicating a higher expected return in times of higher uncertainty. The relation is slightly non-linear with a positive slope. This coincides with the small value of $\theta^{Second}(x_j)$ in Figure 4.9. More interestingly is the u-shaped relationship of Vol. The effect for low and high values is negative on the prediction, whereas for the mid range we observe positive values. This means that the well documented negative relationship is especially present for low and high values, but not for mid values. This amplifies that especially extreme values of the Vol have a large impact on the return prediction. Another interesting shape offers the variable Vol_{daily}. It is similar to shape of the Cumulative Distribution Function (CDF) of a normal or logit distribution. This means that the neural network models large relative changes in the middle of the observed values, but the

[9] We also calculated the joint effects over time excluding the interactions between Vol, Vol_{daily}, Ivol and Ivol_{daily}. The results are similar in the sense that we observe the peaks at the same time, although the actual value is 10-15% lower on average. This indicates that there are also joint effects between these volatility measures.

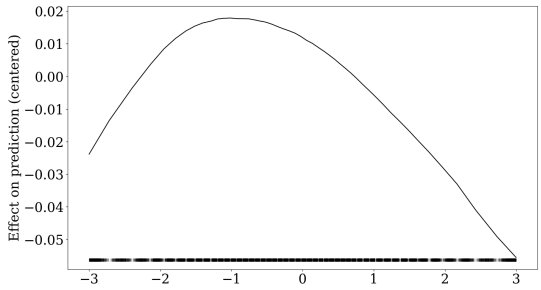
[10] We opt for showing the scaled values on the horizontal axis to allow a easier comparison with respect to low and high values for each variable.

Figure 4.11: $\theta^{ALE}_{main}(x_r)$ of the most important variables

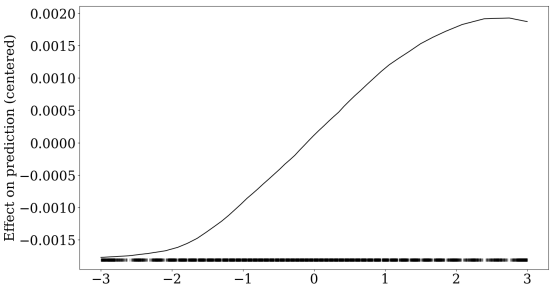
(a) VIX



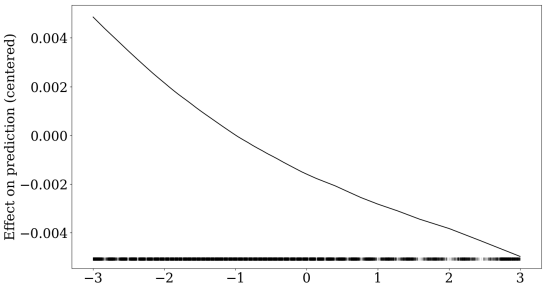
(b) Vol



(c) Vol_{daily}



(d) Size

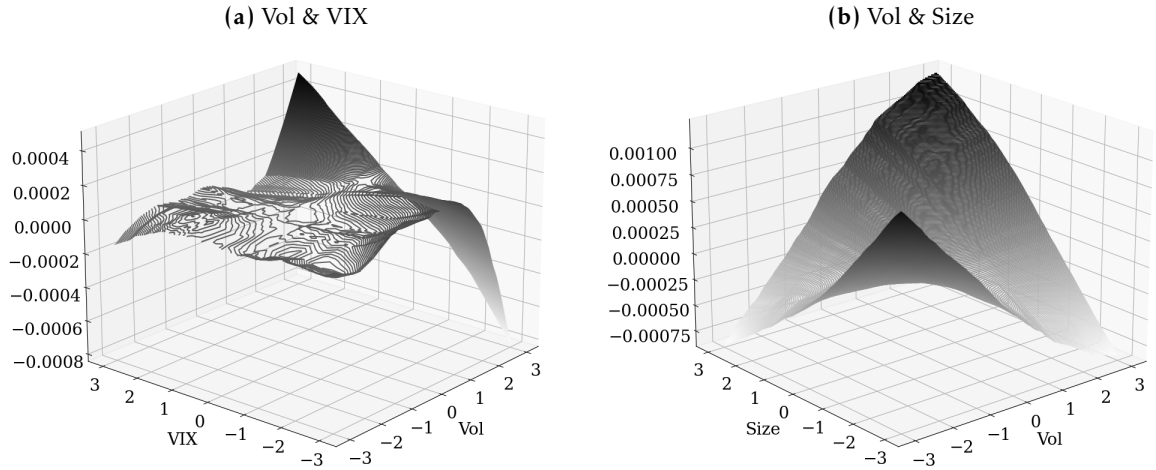


Note: These figures show the time variation of $\theta^{ALE}_{main}(x_r)$ for selected variables. An affine line in the plots indicates a linear relationship, whereas any deviation form that can be labelled as non-linear. The largest non-linearity can be observed with the variable Vol.

relative impact decreasing in the tails. Hence, an increasing Vol_{daily} leads to higher expected return moving from the middle to the right tail, but the relative impact is decreasing. The relationship between Size and the final prediction is negative as expected and shows a slight non-linear behaviour.

The final Figure 4.12 shows two exemplary modelled ALE interaction plots, which give some interesting new insights. Panel a) shows the interaction between Vol and VIX, whereas Panel b) shows the Vol and Size.

Figure 4.12: ALE Plot of end of 2007 sample



Note: These figures show the joint effects of two selected variable pairs. We can see that Size and Vol have a reinforcing joint impact of both variables go in the same direction. The joint effect of VIX and Vol are more asymmetric, as the effects are only observable if the Vol reaches high levels.

The joint effect of Vol and VIX shows positive values for very low levels of Vol and high values of VIX. As future returns are expected to increase with the VIX and decrease with Vol, we can interpret this as an amplifying effect that if both variables show extreme values that would increase future stock returns. Hence, the marginal effects are reinforced by their joint effects. Panel b) shows the interaction between Vol and Size. Both variables are expected to have a negative relationship with future stock returns. Similar to the former plot, we can see an amplifying effect if both variables have large negative values. However, this relationship is u-shaped, i.e., if both variables have large positive values, the returns also increase. Comparing the effect of the interactions with the main effects in Figure 4.11, we see a lower order of magnitude. Nevertheless, these plots are a first indication that there are some joint effects between variables, at least in crisis periods and periods with high uncertainty.

The empirical section showed that the relation between drivers of individual stock returns and their prediction is non-linear in crisis periods and times of high uncertainty. In less exceptional times the approximated relations in the neural network are largely linear, indicated by comparatively high values of R^2_{linear}. This paper illustrates a new way to quantify this non-linearity using neural networks and novel state-of-the-art methods. The agenda of evaluating the non-linearity in predictions by first looking at it globally, using the R^2 measures of Apley and Zhu (2020), then asking which variables drive this non-linearity, using the gradient based feature importance measures $\theta^{First}(x_r)$, $\theta^{Second}(x_j)$ and $\theta^{Joint}(x_r)$ over time, and finally analyse

interesting neural networks in great depth using $\theta^{ALE}_{main}(x_r)$ and $\theta^{ALE}_{Second}(x_{rs})$, can be used for variety of business research applications.

4.6 Conclusion

Time series and cross-sectional patterns of risk premiums are critical for many tasks in finance, including determining a firm's cost of capital, constructing trading strategies and testing asset pricing models. Moreover, they help us to understand how firm returns are affected by their firm characteristics and the business cycle. These determinants can affect strategic decisions of firms and, thus, it is imminent to estimate the underlying patterns as good as possible. Approximating highly non-linear and hidden patters is certainly a task machine learning methods are made for. In recent years the body of literature applying machine learning methods on risk premium forecasting has grown fast, see, e.g., Gu et al. (2020); Bryzgalova et al. (2020) and Chen et al. (2020). This paper proposes a novel measure of non-linearity of predictions which is model agnostic. Therefore, it can be applied to any machine learning model applied in any field of research.

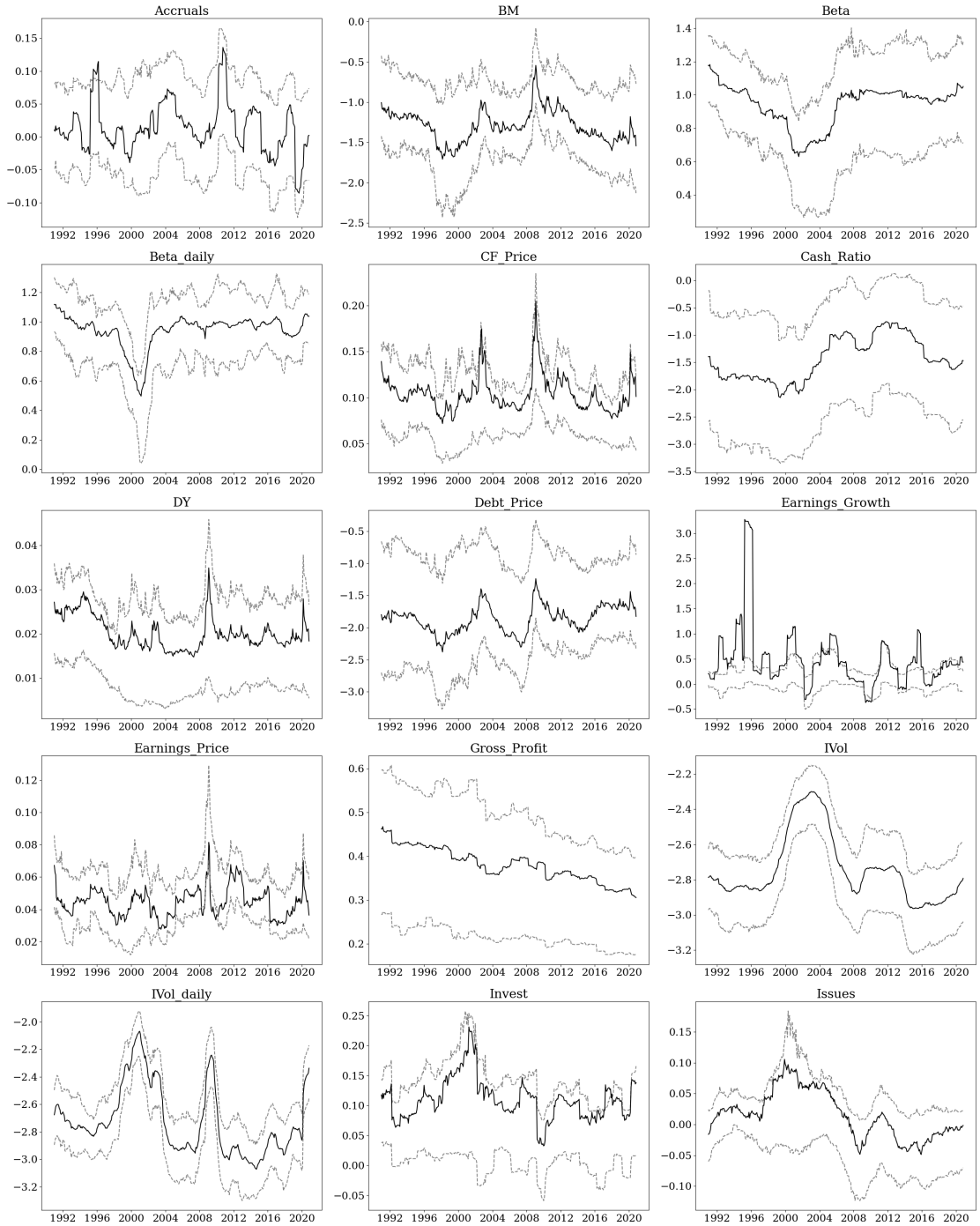
The main empirical contribution of this paper is the strong inverse relationship between the linearity of risk premium predictions and uncertainty measured by the VIX. We can abstract from the empirical findings, that non-linearity plays a less important role in normal times. This indicates that classical linear asset pricing models, as for example employed in Lewellen et al. (2015), are comparatively good to approximate the relationship between predictors and predictions in times of low uncertainty. However, in periods with high uncertainty, non-linearity plays a much greater role. Hence, the underling patters of predictors and risk premiums change. In these times, the actual set of firm characteristics is much more important as a small change in the firm characteristics can have a disproportionately high or low impact on the risk premium. Therefore, one single (linear) coefficient does not necessarily characterize the change in risk premiums from a given change in their drivers during crisis periods. Furthermore, the large amount of interactions implies that we cannot determine the effect of a single firm characteristics on the risk premium separately. We observe larger joint effects for extreme realizations of firm characteristics in crisis periods. Hence, firms with abnormal realizations of their firm value faces reinforced effects on their risk premium. Non-linearity and joint effects on risk premiums in crisis periods challenges linear assumptions which are, however, well suited for normal times.

This paper gives a first attempt at identifying marginal and joint effects over time using a neural network. The proposed non-linearity measure and the extension of state-of-the art explainable machine learning techniques may ease the identification of (non-linear) economic mechanisms behind asset pricing phenomena. Furthermore, these extensions can be applied in various fields of research and are not restricted to applications in finance.

4.A Evolution of firm characteristics over time

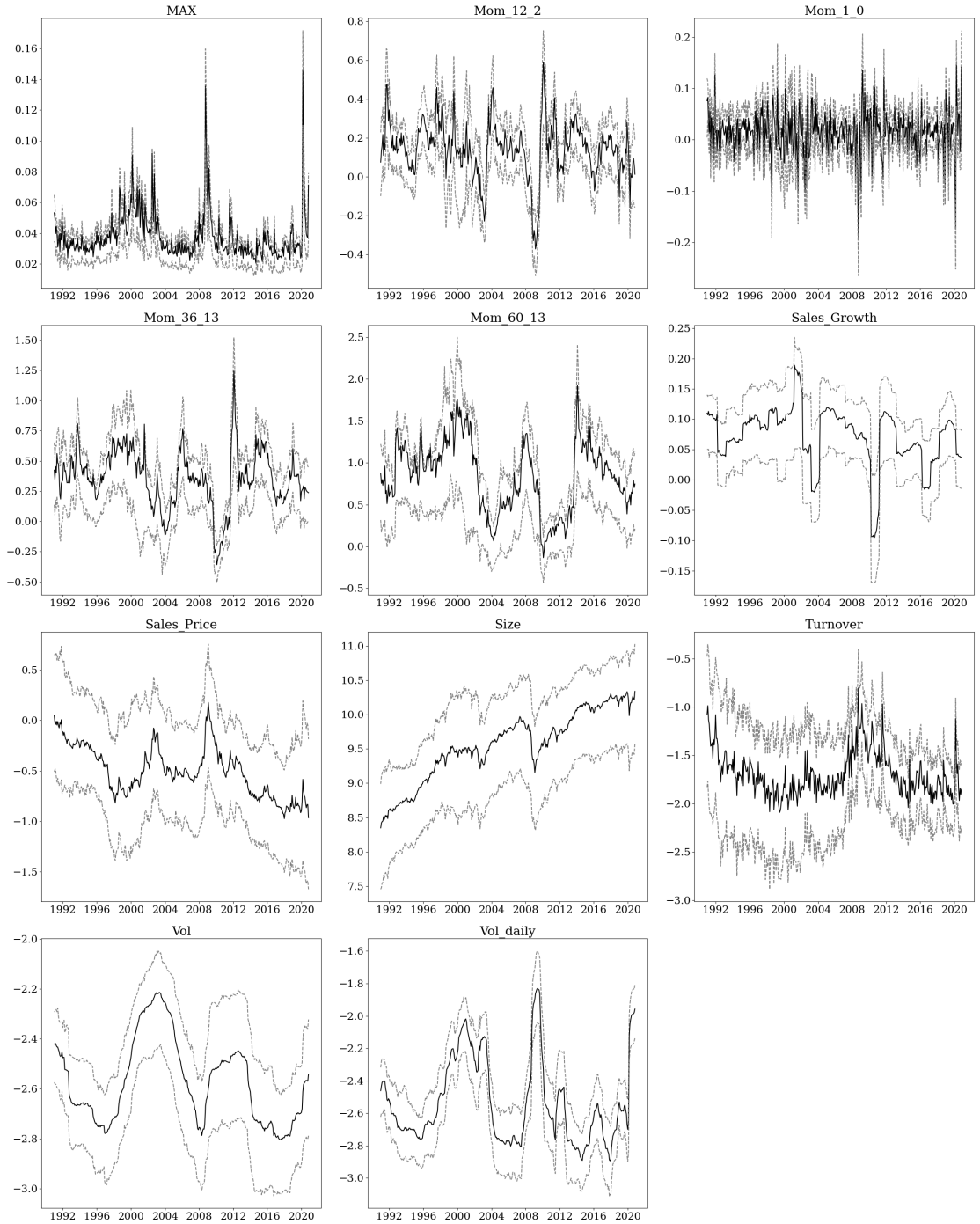
This Appendix illustrates the firm characteristics over time. Figure 4.A.1 and 4.A.2 show the monthly winzorized values. The solid black line illustrates the monthly mean, whereas the grey dashed lines show the 25% respectively 75% quantile every month.

Figure 4.A.1: Firm characteristics and their evolution over time



The plot shows the evolution of the winzorized firm characteristics over time. The black solid line corresponds to the monthly means using all available constituencies. The lower grey dashed line corresponds to the 25% quantile, whereas the upper grey dashed line corresponds to the 75% quantile of the monthly available data. The variable *Earnings_Growth* shows a rather unusual behaviour as the mean is larger than the 75% quantile, which is a result of the strongly positive skewed distribution.

Figure 4.A.2: Firm characteristics and their evolution over time



The plot shows the evolution of the winzorized firm characteristics over time. The black solid line corresponds to the monthly means using all available constituencies. The lower grey dashed line corresponds to the 25% quantile, whereas the upper grey dashed line corresponds to the 75% quantile.

4.B Hyperparameter search

The setup of the hyperparameter is inspired by Gu et al. (2020). We use a random search algorithm to draw 100 combinations of hyperparameter sets every year. However, we extend the procedure to a large number of hyperparameters of the neural network, instead of using only the regularization rate as hyperparameter.

Advanced activation functions

Activation functions are a sensible choice when training especially deep neural networks. It is well known that standard activation functions such as sigmoid and tanh frequently suffer from the so-called vanishing gradient problem, i.e. earlier layers learn much slower that later layers, due to their small gradients. The update of weights is based on the gradient information using the chain rule. This has the effect of multiplying many small values to compute gradients for early layers in a deep neural network. It can be shown that the gradient (error signal) decreases exponentially with the number of layers and, thus, early layers learn much slower or even stop learning in deep neural networks (Hochreiter, 1991). A common solution is the Rectified Linear Unit (ReLU) activation function, which reduces the vanishing gradient problem considerably (Hochreiter, 1998). However, they suffer from the dying ReLU problem, i.e. the activation becomes inactive and only output 0 for any input, see Lu et al. (2019) for an overview. There are several alternative activation functions proposed to avoid the vanishing gradient problem and avoid the dying ReLU problem as well. There is evidence in recent literature that the well-known ReLU function can be easily approximated by a continuous activation function, maintaining the performance. Table 4.B.1 shows an overview of advanced activation functions used in this paper.

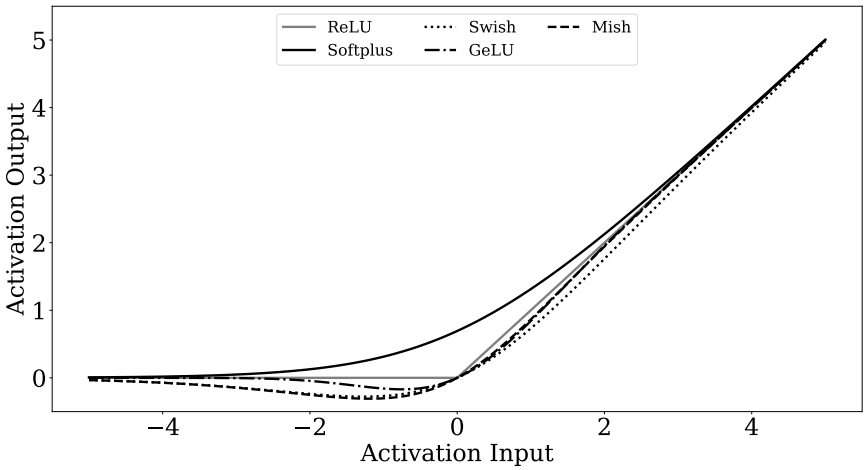
Table 4.B.1: Advanced activation functions

Activation	Formula	Orignal Paper
Softplus	$\log(\exp(x)+1)$	Dugas et al. (2001)
Swish	$x \cdot \text{sigmoid}(\beta \cdot x)$	Ramachandran et al. (2017)
GeLU	$x \cdot \Phi(x)$	Hendrycks and Gimpel (2016)
Mish	$x \cdot \tanh(\text{softplus}(x))$	Misra (2019)

Note: This table shows the advanced activation functions used in this paper. The first column describes the name, the second the mathematical expression and the last column shows the original paper which introduced the activation function. This papers uses extensions to common activations such as tanh and sigmoid to avoid the vanishing gradient problem. Furthermore, all four activation functions avoid the dying ReLU problem and higher performance than the original ReLU activation function. We opt for $\beta = 1$ using the standard Swish activation formulation.

All of these methods are continuous approximations of the ReLU activation functions using combinations of well known activations such as sigmoid and tanh. Figure 4.B.1 illustrates the different activation outputs for the employed activation functions. Comparing the ReLU with Swish, Mish and GeLU, we can clearly see that they follow closely the output for positive input values, but allow for negative outputs for negative inputs. Especially the latter counteracts the dying ReLU problem. Overall, these three activations functions differ more for negative input values. On the contrary the Softplus activation has a positive support over the illustrated range. In the empirical section all four advanced activation functions are used in the hyperparameter search.

Figure 4.B.1: Advanced activation functions



This figure shows the advanced activation functions over [-5,5]. It is imminent that especially Swish, Mish and GeLU follow more closely the ReLU activation but allow for small negative values. On the contrary, the Softplus activation is positive over the illustrated range.

The hyperparameter search

Table 4.B.2: Setup of the hyperparamter search

Parameter	Distribution
Learning Rate	$U^c \sim [0.00001, 0.02]$
Lambda L1	$U^c \sim [0.00001, 0.05]$
Dropout	$U^c \sim [0.10, 0.50]$
Hidden Layer	$U^d \sim [1, 4]$
Multiple	$U^d \sim [1, 5]$
Activation Function	Softplus, Swish, Mish, GeLU

Note: The table shows different values for the hyperparameter search. U^c labels the continuous uniform distribution, whereas U^d labels the discrete uniform distribution. As avoiding overfitting is of major concern, we put much emphasis on regularization parameters (L1) and different designs of Dropout Layers. The network architecutre employs a baseline structure of halving the number of neurons over the hidden layers, following Gu et al. (2020). The minimum number of neurons in the first hidden layer is 32.

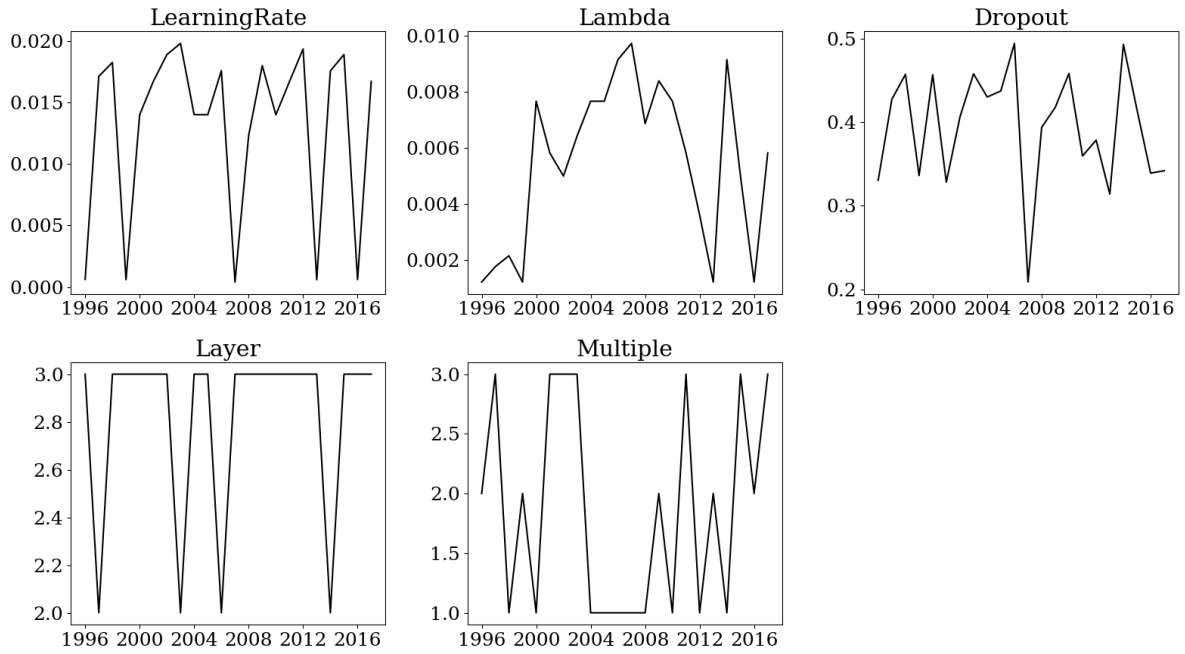
Following Gu et al. (2020), we assume that the number of neurons halves over the hidden layer, i.e. 32 neurons in the fist hidden layer and 16 in the second hidden layer and so on. Hence, instead of validating the actual number of neurons, we validate a multiple of a baseline structure. We assume 32 neurons as minimum for the fist hidden layer. Hence, for a multiple of 1 and four hidden layers, we have 32-16-8-4 neurons. Using a multiple of 2 we get 64-32-16-4. If only two hidden layers and a multiple of 2 is selected, we have 64-32 as number of neurons. This approach gives us a great flexibility, but ensures an efficient way to validate the shallowness of the neural network. The loss function is the common mean-squared-error (MSE), which is commonly used in regression tasks. To avoid overfitting we also use a *Early Stopping*, which stops the training if the validation loss increases a selected number of iterations (so-called *patience*). In this paper, we use a *patience* of 100, a maximum number of 5000 epochs and a batch size of 1,024.

As the the training of neural network also depends on the initialization of the weights we do not use the best neural network in terms of the lowest mean squared error on the validation sample but choose the architecture with the lowest variability over 10 fits. This means we rerun the network architecture 10 times for each setting and calculate the coefficient of variation.:

$$CV^{adj.} = \overline{MSE} \cdot var(MSE) \tag{4.8}$$

As the MSE is usually a very low number and always lower than one, we multiply the mean and variance instead of dividing them. Holding the variance constant, a lower MSE results in a lower value of $CV^{adj.}$. Conversely holding the mean constant, a lower variance of the MSE values leads to a lower $CV^{adj.}$. Consequently, we balance the expected value of the MSE with its variation, and, thus, use the most stable network architecture. The best model according to this metric is the one with the lowest $CV^{adj.}$. However, the chosen architecture is within the top 5 lowest MSE in each year. Hence, we chose one of the best fitting model, but maintaining stability of the MSE estimates. Furthermore, we exclude architecture which collapse in the sense that they predict one single number for all observations in the sample. This is a sort of dying ReLU problem and may be tracked back to the low signal-to-noise ratio in financial data. To skip these collapsed models automatically, we drop all models where the variance of the prediction is more than 100 times lower than the variance of the risk premiums in the training set. This occurs only for roughly 14% to 21% of the sampled architectures.

Figure 4.B.2: Hyperparamters over time



The plot shows the evolution of the hyper parameters over time. Interestingly, the number of layers is either two or three. This might be due to the effect, that we put emphasis on the independence of the weight initialization which usually results in less complex neural networks. However, we can see that the multiple is higher for neural networks trained in 2001 and 2002 onwards. This may indicated that in crisis periods more flexibility is needed.

Figure 4.B.2 shows the validated hyperparameters over time. The interpretation of the values is not straightforward as they depend on each other. However, it is interesting that the number of hidden layer is either two or three. This indicates that the hyperparameter search favours less deep neural networks. This may due to the fact that we also account for the stability w.r.t the dependence on the weight initialization, which usually favours less parameters in the neural networks. For the years 2001 and 2002 we observe a considerable deep neural network with three hidden layers and a multiple of 3. The multiple drops in 2007 and 2008 to a value of 1. Interestingly, throughout any hyperparamter search the swish activation function was selected. Hence, we omit this hyperparamter in Figure 4.B.2.

Conclusion

Summary

This thesis focuses on the application of statistical and machine learning methods for credit and market risk. It is composed of a profound empirical analysis of various central aspects. In the first research paper *Credit line exposure at default modelling using Bayesian mixed effect quantile regression* (see Chapter 1), a Bayesian mixed effect quantile regression is used to address the challenging distribution of conversion factors to determining the EAD of credit lines. The empirical analysis documents strong varying covariate effects over the conditional distribution and differences between Europe and the United States. Furthermore, the extension with random effects is needed in Europe to generate appropriate downturn estimates. The empirical analysis of LGDs is subject to the second research paper *Opening the Black Box – Quantile Neural Networks for Loss Given Default Prediction* (see Chapter 2). A combination of linear quantile regression and neural networks is proposed to allow for any kind of non-linearity in every quantile of the conditional distribution. Non-linearity is especially important in higher quantiles. Furthermore, the macroeconomic environment shows large non-linearity and joint effects. The third research paper *Deep calibration of financial models: turning theory into practice* (see Chapter 3) takes another view on neural networks. The aim of this paper is not to achieve better predictions, but to accelerate a main objective in market risk management. This study is the first to benchmark calibration frameworks using a neural network with a real-life implementation at a large financial institutions. The empirical analysis shows that the application of neural networks for calibration can lift several benefits for financial institutions, such as the speed and robustness of calibration results over time. The fourth and last research paper *Does non-linearity in risk premiums vary over time?* (see Chapter 4) focus on the amount of non-linearity modelled in stock price predictions. By proposing a new model agnostic approach, an inverse relationship between linearity in risk premium predictions and uncertainty measured by the VIX is documented. The empirical results show that linear asset pricing models work quite well in normal times, but in crisis periods non-linearity gets more important.

Outlook

The topics of this thesis are of high relevance for financial institutions as well as regulators. The first research paper deals with the downturn behaviour of defaulted credit lines. At the time of writing, the COVID-19 pandemic is still prevailing and the impact on credit risk in general, but on the EAD of credit lines in particular is not foreseeable yet. The paper uses access to one of the world's largest databases of defaulted loans provided by GCD. Future research is able to investigate the effects of the pandemic on credit lines using a very broad perspective on the banking universe. Furthermore, the pandemic has not originated in the financial system, contrary to the Global Financial Crisis. Therefore, investigating whether the structural relationship between the macroeconomy and credit lines has changed in the recent crisis periods may also be a promising path for future research. With respect to the topic of LGD modelling in Chapter 2, there are several areas of subsequent research. First, it may be fruitful to look at other possible combinations of classical statistical and machine learning models. Finite Mixture Models have been used by various authors to tackle the distribution of workout LGDs, see, e.g., Altman and Kalotay (2014); Calabrese (2014) or Betz et al. (2018). Hence, it is maybe possible to lift the potential of both approaches as well. Second, alternative data- e.g., Moody's ultimate recovery database- might be used. Recent studies show, that there is non-linearity between predictors and marked-based LGDs, see, e.g., Qi and Zhao (2011); Loterman et al. (2012) and Sopitpongstorn et al. (2021). Therefore, applying the quantile regression neural network to these data may gain new insights into the non-linear behaviour in different parts of the market-based LGD's conditional distribution. The calibration of financial models with neural networks is relatively new to the literature as the first approach was introduced by Hernandez (2017). Banks usually use a battery of asset pricing models to calibrate various instruments. Hence, the application of the deep calibration framework on more complex models in financial institutions can lift large potentials in practice as well for academics. Furthermore, this may enable financial institutions to use more complex financial models as well as optimizers. Answering the question of how much non-linearity is modelled by complex models is a cornerstone of understanding machine learning algorithms and a first step is taken in Chapter 4. XAI approaches will become more and more important and, thus, are a fruitful path for future research. This thesis deals with classical feed-forward neural networks and determines their drivers. However, there are other architectures, such as Long Short Term Memory (LSTM) neural networks which deal with time series data. The application of XAI methods on them is rather sparse, see, e.g., Arras et al. (2017); Murdoch and Szlam (2017); Murdoch et al. (2018) or Guo et al. (2019). Time series are faced quite frequently in the financial context, for example when dealing with stock market

data. Therefore, it would be interesting to unveil the hidden dynamics in time series models to get a deeper understanding and maybe derive novel economic mechanisms.

Bibliography

Acharya, V., H. Almeida, F. Ippolito, and A. P. Orive (2020). Bank lines of credit as contingent liquidity: Covenant violations and their implications. *Journal of Financial Intermediation 44*, 100817.

Acharya, V., H. Almeida, F. Ippolito, and A. Perez (2014). Credit lines as monitored liquidity insurance: Theory and evidence. *Journal of Financial Economics 112*(3), 287–319.

Acharya, V. V., H. Almeida, and M. Campello (2013). Aggregate risk and the choice between cash and lines of credit. *The Journal of Finance 68*(5), 2059–2116.

Acharya, V. V. and N. Mora (2015). A crisis of banks as liquidity providers. *The Journal of Finance 70*(1), 1–43.

Adrian, T., R. K. Crump, and E. Vogt (2019). Nonlinearity and Flight-to-Safety in the Risk-Return Trade-Off for Stocks and Bonds. *The Journal of Finance 74*(4), 1931–1973.

Agarwal, S., B. W. Ambrose, and C. Liu (2006). Credit Lines and Credit Utilization. *Journal of Money, Credit and Banking 38*(1), 1–22.

Alhamzawi, R. (2016). Bayesian Analysis of Composite Quantile Regression. *Statistics in Biosciences 8*(2), 358–373.

Altman, E. I. and E. A. Kalotay (2014). Ultimate recovery mixtures. *Journal of Banking & Finance 40*, 116–129.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *The Journal of Finance 61*(1), 259–299.

Antoniou, A., Y. Guney, and K. Paudyal (2008). The determinants of capital structure: capital market-oriented versus bank-oriented institutions. *Journal of Financial and Quantitative Analysis 43*(1), 59–92.

Apley, D. W. and J. Zhu (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(4), 1059–1086.

Araten, M. and M. Jacobs Jr (2001). Loan equivalents for revolving credits and advised lines. *The RMA Journal 83*(8), 34–39.

Arellano, M. and S. Bonhomme (2017). Quantile Selection Models With an Application to Understanding Changes in Wage Inequality. *Econometrica 85*(1), 1–28.

Arras, L., G. Montavon, K.-R. Müller, and W. Samek (2017). Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 159–168.

Asness, C. S., R. B. Porter, and R. L. Stevens (2000). Predicting stock returns using industry-relative firm characteristics. *Working Paper*. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=213872.

Bakoben, M., T. Bellotti, and N. Adams (2020). Identification of credit risk based on cluster analysis of account behaviours. *Journal of the Operational Research Society 71*(5), 775–783.

Bali, T. G., N. Cakici, and R. F. Whitelaw (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics 99*(2), 427–446.

Bank of Canada (2018). Financial system survey. Technical report, Bank of Canada. Available at https://www.bankofcanada.ca/2018/11/financial-system-survey-highlights/.

Bank of England (2019). Machine learning in uk financial services. Technical report, Bank of England and Financial Conduct Authority. Available at https://www.bankofengland.co.uk/report/2019/machine-learning-in-uk-financial-services.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics 9*(1), 3–18.

Barakova, I. and H. Parthasarathy (2013). Modeling corporate exposure at default. *Working Paper*. Available at http://dx.doi.org/10.2139/ssrn.2235218.

Barbee Jr, W. C., S. Mukherji, and G. A. Raines (1996). Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal 52*(2), 56–60.

Barraza, S. and A. Civelli (2020). Economic Policy Uncertainty and the Supply of Business Loans. *Journal of Banking & Finance*, 105983.

Basel Committee on Banking Supervision (2000). Principles for the management of credit risk. Technical report, Bank for International Settlements. Available at https://www.bis.org/bcbs/publ/d457.htm.

Basel Committee on Banking Supervision (2017). Basel iii: Finalising post-crisis reforms. Technical report, Bank for International Settlements. Available at https://www.bis.org/bcbs/publ/d424.pdf.

Basel Committee on Banking Supervision (2019a). High-level summary: BCBS SIG industry workshop on the governance and oversight of artificial intelligence and machine learning in financial services.

Basel Committee on Banking Supervision (2019b). Minimum capital requirements for market risk. Technical report, Bank for International Settlements. Available at https://www.bis.org/publ/bcbs75.htm.

Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance 34*(10), 2510–2517.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance 32*(3), 663–682.

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *The Review of Financial Studies 9*(1), 69–107.

Bayer, C., B. Horvath, A. Muguruza, B. Stemper, and M. Tomas (2019). On deep calibration of (rough) stochastic volatility models. *Working Paper*. Available at https://arxiv.org/pdf/1908.08806.pdf.

Bayer, C. and B. Stemper (2018). Deep calibration of rough stochastic volatility models. *Working Paper*. Available at https://arxiv.org/abs/1810.03399.

Bellotti, A., D. Brigo, P. Gambetti, and F. Vrins (2021). Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting 37*(1), 428–444.

Bellotti, T. and J. Crook (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting 28*(1), 171–182.

Bernardi, M., G. Gayraud, and L. Petrella (2015). Bayesian Tail Risk Interdependence Using Quantile Regression. *Bayesian Analysis 10*(3), 553–603.

Berrospide, J. M. and R. R. Meisenzahl (2015). The real effects of credit line drawdowns. *Finance and Economic Discussion Series 2015-007*. Available at https://ideas.repec.org/p/fip/fedgfe/2015-07.html.

Betz, J., R. Kellner, and D. Rösch (2021). Time matters: How default resolution times impact final loss rates. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 70*(3), 619–644.

Betz, J., R. Kellner, and D. Rösch (2018). Systematic Effects among Loss Given Defaults and their Implications on Downturn Estimation. *European Journal of Operational Research 271*(3), 1113–1144.

Betz, J., S. Krüger, R. Kellner, and D. Rösch (2020). Macroeconomic effects and frailties in the resolution of non-performing loans. *Journal of Banking & Finance 112*, 105212.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance 43*(2), 507–528.

Bianchi, D., M. Büchner, and A. Tamoni (2020). Bond Risk Premiums with Machine Learning. *The Review of Financial Studies 34*(2), 1046–1089.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer-Verlag.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1), 5–32.

Brown, S. J., W. Goetzmann, R. G. Ibbotson, and S. A. Ross (1992). Survivorship bias in performance studies. *The Review of Financial Studies 5*(4), 553–580.

Brumma, N., N. Rainone, and R. Crecel (2020a). Downturn lgd study 2020. *Report*. Available at https://www.globalcreditdata.org/library/downturn-lgd-study-2020.

Brumma, N., N. Rainone, and R. Crecel (2020b). Lgd report 2020-large corporate borrowers. *Report*. Available at https://www.globalcreditdata.org/news/how-can-banks-project-losses-in-the-current-covid-19-crisis-asks-global-credit-data-in-latest.

Bryzgalova, S., M. Pelger, and J. Zhu (2020). Forest through the trees: Building cross-sections of stock returns. *Working Paper*. Available at https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3493458.

Calabrese, R. (2014). Downturn Loss Given Default: Mixture distribution estimation. *European Journal of Operational Research 237*(1), 271–277.

Carriero, A., T. E. Clark, and M. G. Marcellino (2020). Nowcasting Tail Risks to Economic Activity with Many Indicators. SSRN Scholarly Paper ID 3599285, Social Science Research Network, Rochester, NY.

Chalfin, A., O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan (2016). Productivity and Selection of Human Capital with Machine Learning. *American Economic Review 106*(5), 124–127.

Chen, L., M. Pelger, and J. Zhu (2020). Deep learning in asset pricing. *Working Paper*. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3350138.

Chen, Z., W. Chen, and Y. Shi (2020). Ensemble learning with label proportions for bankruptcy prediction. *Expert Systems with Applications 146*, 113155.

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics 33*(2), 806–839.

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

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and Quantile Effects in Nonseparable Panel Models. *Econometrica 81*(2), 535–580.

Clements, A., C. Drovandi, and D. Li (2020). Reducing the Risk in Tail Risk Forecasting Models. SSRN Scholarly Paper ID 3750440, Social Science Research Network, Rochester, NY.

Cochrane, J. H. (2011). Presidential Address: Discount Rates. *The Journal of Finance 66*(4), 1047–1108.

Colla, P., F. Ippolito, and K. Li (2013). Debt specialization. *The Journal of Finance 68*(5), 2117–2141.

Cornett, M. M., J. J. McNutt, P. E. Strahan, and H. Tehranian (2011). Liquidity risk management and credit supply in the financial crisis. *Journal of Financial Economics 101*(2), 297–312.

Cowden, C., F. J. Fabozzi, and A. Nazemi (2019). Default prediction of commercial real estate properties using machine learning techniques. *The Journal of Portfolio Management 45*(7), 55–67.

Culkin, R. and S. R. Das (2017). Machine Learning in Finance: The Case of Deep Learning for Option Pricing. *Journal of Investment Management 15*(4), 92–100.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems 2*, 301–314.

Datar, V. T., N. Y. Naik, and R. Radcliffe (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets 1*(2), 203–219.

Deryugina, T., G. Heutel, N. H. Miller, D. Molitor, and J. Reif (2019). The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction. *American Economic Review 109*(12), 4178–4219.

Deutsche Bundesbank (2020). The use of artificial intelligence and machine learning in the financial sector. Available at https://www.bundesbank.de/resource/blob/598256/d7d26167bceb18ee7c0c296902e42162/mL/2020-11-policy-dp-aiml-data.pdf.

Dimitroff, G., D. Röder, and C. P. Fries (2018). Volatility Model Calibration With Convolutional Neural Networks. *Working Paper*. Available at https://www.ssrn.com/abstract=3252432.

Doshi, H., R. Elkamhi, and C. Ornthanalai (2018). The Term Structure of Expected Recovery Rates. *Journal of Financial and Quantitative Analysis 53*(6), 2619–2661.

Dugas, C., Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia (2001). Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 472–478.

Dumitrescu, E., S. Hue, C. Hurlin, and S. Tokpavi (2021). Machine Learning for Credit Scoring: Improving Logistic Regression with Non-Linear Decision-Tree Effects. *European Journal of Operational Research (forthcomming)*.

ECB (2019). Guide to internal models. *European Central Bank (ECB): Banking Supervision*.

Elton, E. J., M. J. Gruber, and C. R. Blake (1996). Survivor bias and mutual fund performance. *The Review of Financial Studies 9*(4), 1097–1120.

European Banking Authority (2016). Regulatory technical standards on materiality threshold of credit obligation past due. *Techincal Report*. Available at https://www.eba.europa.eu/regulation-and-policy/credit-risk/regulatory-technical-standards-on-materiality-threshold-of-credit-obligation-past-due.

European Banking Authority (2020). Risk assessment of the european banking system. Technical report. Available at https://www.eba.europa.eu/risk-analysis-and-data/risk-assessment-reports.

Fairfield, P. M., J. S. Whisenant, and T. L. Yohn (2003). Accrued earnings and growth: Implications for future profitability and market mispricing. *The Accounting Review 78*(1), 353–371.

Fama, E. F. and K. R. French (2008). Dissecting anomalies. *The Journal of Finance 63*(4), 1653–1678.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*(3), 607–636.

Feng, G., N. Polson, and J. Xu (2020). Deep learning in characteristics-sorted factor models. *Working Paper*. Available at https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3243683.

Ferguson, R. and A. Green (2018). Deeply Learning Derivatives. *Working Paper*. Available at https://arxiv.org/abs/1809.02233.

FERMA (2019). Artificial intelligence applied to risk management. Technical report, Federation of European Risk Management Associations. Available at https://www.ferma.eu/publication/artificial-intelligence-ai-applied-to-risk-management.

Ferrara, L., M. Mogliani, and J.-G. Sahuc (2021). High-frequency monitoring of growth at risk. *International Journal of Forecasting*.

Fischer, T. and C. Krauss (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research 270*(2), 654–669.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies 33*(5), 2326–2377.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics 29*(5), 1189–1232.

Friedman, J. H. and B. E. Popescu (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics 2*(3), 916–954.

Fritz-Morgenthal, S., B. Hein, and J. Papenbrock (2021). Financial Risk Management and Explainable Trustworthy Responsible AI. *Working Paper*. Available at https://papers.ssrn.com/abstract=3873768.

Galvao, A. F. and K. Kato (2017). Quantile Regression Methods for Longitudinal Data. In *Handbook of Quantile Regression*. Chapman and Hall/CRC.

Galvao, A. F., C. Lamarche, and L. R. Lima (2013). Estimation of Censored Quantile Regression for Panel Data With Fixed Effects. *Journal of the American Statistical Association 108*(503), 1075–1089.

Galvao, A. F. and A. Poirier (2019). Quantile Regression Random Effects. *Annals of Economics and Statistics* (134), 109–148.

Gambara, M. and J. Teichmann (2020). Consistent Recalibration Models and Deep Calibration. *Working Paper*. Available at https://arxiv.org/abs/2006.09455.

Gambetti, P., G. Gauthier, and F. Vrins (2019). Recovery rates: Uncertainty certainly matters. *Journal of Banking & Finance 106*, 371 – 383.

Gambetti, P., F. Roccazzella, and F. Vrins (2020). Meta-learning approaches for recovery rate prediction. *Working Paper*. Available at https://dial.uclouvain.be/pr/boreal/object/boreal:229301.

Gatev, E. and P. E. Strahan (2006). Banks' advantage in hedging liquidity risk: Theory and evidence from the commercial paper market. *The Journal of Finance 61*(2), 867–892.

Gelman, A., D. B. Rubin, et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Geraci, M. and M. Bottai (2006). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics 8*(1), 140–154.

Geraci, M. and M. Bottai (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics 8*(1), 140–154.

Gibson, P. B., W. E. Chapman, A. Altinok, L. Delle Monache, M. J. DeFlorio, and D. E. Waliser (2021). Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment 2*(1), 1–13.

Gilli, M. and E. Schumann (2012). Calibrating Option Pricing Models with Heuristics. In *Natural Computing in Computational Finance: Volume 4*, Studies in Computational Intelligence, pp. 9–37. Berlin, Heidelberg: Springer.

Glorot, X. and Y. Bengio (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics 24*(1), 44–65.

Graham, B. S., J. Hahn, A. Poirier, and J. L. Powell (2018). A quantile correlated random coefficients panel data model. *Journal of Econometrics 206*(2), 305–335.

Green, J., J. R. Hand, and X. F. Zhang (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies 30*(12), 4389–4436.

Grunert, J. and M. Weber (2009). Recovery rates of commercial lending: Empirical evidence for german companies. *Journal of Banking & Finance 33*(3), 505–513.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies 33*(5), 2223–2273.

Guo, T., T. Lin, and N. Antulov-Fantulin (2019). Exploring interpretable lstm neural networks over multi-variable data. In *International Conference on Machine Learning*, pp. 2494–2504. PMLR.

Gürtler, M., M. T. Hibbeln, and P. Usselmann (2018). Exposure at default modeling – a theoretical and empirical assessment of estimation approaches and parameter choice. *Journal of Banking & Finance 91*, 176 – 188.

Hartmann-Wendels, T., P. Miller, and E. Töws (2014). Loss given default for leasing: Parametric and nonparametric estimations. *Journal of Banking & Finance 40*, 364–375.

Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *The Review of Financial Studies 29*(1), 5–68.

Heath, D., R. Jarrow, and A. Morton (1992). Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation. *Econometrica 60*(1), 77–105.

Heidelberger, P. and P. D. Welch (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM 24*(4), 233–245.

Heidelberger, P. and P. D. Welch (1983). Simulation run length control in the presence of an initial transient. *Operations Research 31*(6), 1109–1144.

Hendrycks, D. and K. Gimpel (2016). Gaussian error linear units (gelus). *Working Paper*. Available at https://arxiv.org/abs/1606.08415.

Hernandez, A. (2017). Model calibration with neural networks. *Risk*.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies 6*(2), 327–343.

Hirsa, A., T. Karatas, and A. Oskoui (2019). Supervised Deep Neural Networks (DNNs) for Pricing/Calibration of Vanilla/Exotic Options Under Various Different Processes. *Working Paper*. Available at https://arxiv.org/abs/1902.05810.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München 91*(1).

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6*(02), 107–116.

Hon, P. S. and T. Bellotti (2016). Models and forecasts of credit card balance. *European Journal of Operational Research 249*(2), 498 – 505.

Horel, E. and K. Giesecke (2020). Significance tests for neural networks. *Journal of Machine Learning Research 21*(227), 1–29.

Horel, E., V. Mison, T. Xiong, K. Giesecke, and L. Mangu (2018). Sensitivity based Neural Networks Explanations. *Working Paper*. Available at http://arxiv.org/abs/1812.01029.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks 4*(2), 251–257.

Horowitz, J. L. and S. Lee (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association 100*(472), 1238–1249.

Horvath, B., A. Muguruza, and M. Tomas (2021). Deep learning volatility: A deep neural network perspective on pricing and calibration in (rough) volatility models. *Quantitative Finance 21*(1), 11–27.

Hoshino, T. (2014). Quantile regression estimation of partially linear additive models. *Journal of Nonparametric Statistics 26*(3), 509–536.

Hu, Y., H. J. Wang, X. He, and J. Guo (2021). Bayesian joint-quantile regression. *Computational Statistics 36*(3), 2033–2053.

Huang, H. and Z. Chen (2015). Bayesian composite quantile regression. *Journal of Statistical Computation and Simulation 85*(18), 3744–3754.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics 35*(1), 73–101.

Hull, J. and A. White (1990). Pricing interest-rate-derivative securities. *The Review of Financial Studies 3*(4), 573–592.

Hutchinson, J. M., A. W. Lo, and T. Poggio (1994). A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks. *The Journal of Finance 49*(3), 851–889.

Hwang, R.-C. and C.-K. Chu (2018). A logistic regression point of view toward loss given default distribution estimation. *Quantitative Finance 18*(3), 419–435.

Hwang, R.-C., C.-K. Chu, and K. Yu (2020). Predicting LGD distributions with mixed continuous and discrete ordinal outcomes. *International Journal of Forecasting*.

Ivashina, V. and D. Scharfstein (2010). Bank lending during the financial crisis of 2008. *Journal of Financial Economics 97*(3), 319–338.

Jackson, L. E., K. L. Kliesen, and M. T. Owyang (2020). The nonlinear effects of uncertainty shocks. *Studies in Nonlinear Dynamics & Econometrics 24*(4).

Jacobs Jr, M. (2010). An empirical study of exposure at default. *Journal of Advanced Studies in Finance 1*(1), 31–59.

Jacobs Jr, M. and P. Bag (2011). What do we know about exposure at default on contingent credit lines? - a survey of the literature, empirical analysis and models. *Journal of Advanced Studies in Finance 2*(1), 26–46.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance 45*(3), 881–898.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance 48*(1), 65–91.

Jiménez, G., J. A. Lopez, and J. Saurina (2009). Empirical analysis of corporate credit lines. *The Review of Financial Studies 22*(12), 5069–5098.

Jing, J., W. Yan, and X. Deng (2021). A hybrid model to estimate corporate default probabilities in china based on zero-price probability model and long short-term memory. *Applied Economics Letters 28*(5), 413–420.

Kalotay, E. A. and E. I. Altman (2017). Intertemporal Forecasts of Defaulted Bond Recoveries and Portfolio Losses. *Review of Finance 21*(1), 433–463.

Kaposty, F., J. Kriebel, and M. Löderbusch (2020). Predicting loss given default in leasing: A closer look at models and variable selection. *International Journal of Forecasting 36*(2), 248–266.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kellner, R., M. Nagl, and D. Rösch (2022). Opening the black box – Quantile neural networks for loss given default prediction. *Journal of Banking & Finance 134*, 106334.

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Khieu, H. D., D. J. Mullineaux, and H.-C. Yi (2012). The determinants of bank loan recovery rates. *Journal of Banking & Finance 36*(4), 923–933.

Kienitz, J., S. K. Acar, Q. Liang, and N. Nowaczyk (2020). Deep option pricing - term structure models. *Journal of Machine Learning in Finance 1*.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *Working Paper*. Available at https://arxiv.org/abs/1412.6980.

Koenker, R. (2005). *Quantile regression*. Camebridge University Press.

Koenker, R. and G. Bassett (1978). Regression Quantiles. *Econometrica 46*(1), 33–50.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017). *Handbook of quantile regression*. CRC press.

Koenker, R., P. Ng, and S. Portnoy (1994). Quantile smoothing splines. *Biometrika 81*(4), 673–680.

Kraus, M., S. Feuerriegel, and A. Oztekin (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research 281*(3), 628–641.

Krüger, S., T. Oehme, D. Rösch, and H. Scheule (2018). A copula sample selection model for predicting multi-year LGDs and Lifetime Expected Losses. *Journal of Empirical Finance 47*, 246–262.

Krüger, S. and D. Rösch (2017). Downturn LGD modeling using quantile regression. *Journal of Banking & Finance 79*, 42–56.

Kvamme, H., N. Sellereite, K. Aas, and S. Sjursen (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications 102*, 207–217.

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994). Contrarian investment, extrapolation, and risk. *The Journal of Finance 49*(5), 1541–1578.

Leow, M. and J. Crook (2016). A new mixture model for the estimation of credit card exposure at default. *European Journal of Operational Research 249*(2), 487 – 497.

Leow, M., C. Mues, and L. Thomas (2014). The economy and loss given default: evidence from two UK retail lending data sets. *Journal of the Operational Research Society 65*(3), 363–375.

Lettau, M. and M. Pelger (2020). Factors That Fit the Time Series and Cross-Section of Stock Returns. *The Review of Financial Studies 33*(5), 2274–2325.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *The Quarterly of Applied Mathematics 2*(2), 164–168.

Lewellen, J. et al. (2015). The cross-section of expected stock returns. *Critical Finance Review 4*(1), 1–44.

Li, Q., J. Lin, and J. S. Racine (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics 31*(1), 57–65.

Li, Q. and J. S. Racine (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics 26*(4), 423–434.

Li, Y. and W. Chen (2019). Entropy method of constructing a combined model for improving loan default prediction: A case study in China. *Journal of the Operational Research Society 72*(5), 1099–1109.

Lins, K. V., H. Servaes, and P. Tufano (2010). What drives corporate liquidity? an international survey of cash holdings and lines of credit. *Journal of Financial Economics 98*(1), 160–176.

Litzenberger, R. H. and K. Ramaswamy (1982). The effects of dividends on common stock prices tax effects or information effects? *The Journal of Finance 37*(2), 429–443.

Liu, S., A. Borovykh, L. A. Grzelak, and C. W. Oosterlee (2019). A neural network-based framework for financial model calibration. *Journal of Mathematics in Industry 9*(1), 9.

Liu, S., C. W. Oosterlee, and S. M. Bohte (2019). Pricing Options and Computing Implied Volatilities using Neural Networks. *Risks 7*(1), 16.

Loterman, G., I. Brown, D. Martens, C. Mues, and B. Baesens (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting 28*(1), 161–170.

Lu, L., Y. Shin, Y. Su, and G. E. Karniadakis (2019). Dying relu and initialization: Theory and numerical examples. *Working Paper*. Available at https://arxiv.org/abs/1903.06733.

Lundberg, S. and S.-I. Lee (2016). An unexpected unity among methods for interpreting model predictions. *Working Paper*. Available at http://arxiv.org/abs/1611.07478.

Lundberg, S. M. and S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc.

Luo, J., X. Yan, and Y. Tian (2020). Unsupervised quadratic surface support vector machine with application to credit risk assessment. *European Journal of Operational Research 280*(3), 1008–1017.

Luo, Y., H. Lian, and M. Tian (2012). Bayesian quantile regression for longitudinal data models. *Journal of Statistical Computation and Simulation 82*(11), 1635–1649.

Mai, F., S. Tian, C. Lee, and L. Ma (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research 274*(2), 743–758.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics 11*(2), 431–441.

Matuszyk, A., C. Mues, and L. C. Thomas (2010). Modelling LGD for unsecured personal loans: decision tree approach. *Journal of the Operational Research Society 61*(3), 393–398.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–32.

Michael, J. R. (1983). The stabilized probability plot. *Biometrika 70*(1), 11–17.

Misra, D. (2019). Mish: A self regularized non-monotonic neural activation function. *Working Paper*. Available at https://arxiv.org/abs/1908.08681.

Moral, G. (2006). Ead estimates for facilities with explicit limits. In *The Basel II risk parameters*, pp. 197–242. Springer.

Munk, C. (1999). Stochastic duration and fast coupon bond option pricing in multi-factor models. *Review of Derivatives Research 3*, 157–181.

Murdoch, W. J., P. J. Liu, and B. Yu (2018). Beyond word importance: Contextual decomposition to extract interactions from lstms. In *International Conference on Learning Representations*.

Murdoch, W. J. and A. Szlam (2017). Automatic rule extraction from long short term memory networks. *Working Paper*. Available at https://arxiv.org/abs/1702.02540.

Nagl, M. (2021). Does non-linearity in risk premiums vary over time? *Working Paper*, 1–44.

Nazemi, A., F. Baumann, and F. J. Fabozzi (2021). Intertemporal Defaulted Bond Recoveries Prediction via Machine Learning. *European Journal of Operational Research (forthcomming)*.

Nazemi, A. and F. J. Fabozzi (2018). Macroeconomic variable selection for creditor recovery rates. *Journal of Banking & Finance 89*, 14–25.

Nazemi, A., F. Fatemi Pour, K. Heidenreich, and F. J. Fabozzi (2017). Fuzzy decision fusion approach for loss-given-default modeling. *European Journal of Operational Research 262*(2), 780–791.

Nazemi, A., K. Heidenreich, and F. J. Fabozzi (2018). Improving corporate bond recovery rate prediction using multi-factor support vector regressions. *European Journal of Operational Research 271*(2), 664–675.

Nelson, C. R. and A. F. Siegel (1987). Parsimonious modeling of yield curves. *Journal of Business 60*(4), 473–489.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics 103*(3), 429–453.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics 108*(1), 1–28.

OCC, et. al (2011). Interagency Supervisory Guidance on Counterparty Credit Risk Management. *Office of the Comptroller of the Currency Federal Deposit Insurance Corporation Board of Governors of the Federal Reserve SystemOffice of Thrift Supervision*.

Odom, M. D. and R. Sharda (1990). A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on neural networks*, pp. 163–168. IEEE.

Ou, J. A. and S. H. Penman (1989). Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics 11*(4), 295–329.

Paulsen, B., A. Misback, J. Sheesley, D. Uejio, and M. Conyers-Ausbrooks (2021). Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning. Available at https://www.federalregister.gov/documents/2021/03/

31/2021-06607/request-for-information-and-comment-on-financial-institutions-use-of-artificial-intelligence.

Pelger, M. (2020). Understanding systematic risk: A high-frequency approach. *The Journal of Finance 75*(4), 2179–2220.

Pelger, M. and R. Xiong (2021). State-varying factor models of large dimensions. *Journal of Business & Economic Statistics 0*(0), 1–19.

Petropoulos, A., V. Siakoulis, E. Stavroulakis, and N. E. Vlachogiannakis (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting 36*(3), 1092–1113.

Pontiff, J. and A. Woodgate (2008). Share issuance and cross-sectional returns. *The Journal of Finance 63*(2), 921–945.

Qi, M. (2009). Exposure at default of unsecured credit cards. *Office of the Comptroller of the Currency Working Paper*. Available at https://occ.treas.gov/publications-and-resources/publications/economics/working-papers-archived/economic-working-paper-2009-2.html.

Qi, M. and X. Yang (2009). Loss given default of high loan-to-value residential mortgages. *Journal of Banking & Finance 33*(5), 788–799.

Qi, M. and X. Zhao (2011). Comparison of modeling methods for Loss Given Default. *Journal of Banking & Finance 35*(11), 2842–2855.

Quek, C., M. Pasquier, and N. Kumar (2008). A novel recurrent neural network-based prediction system for option trading and hedging. *Applied Intelligence 29*(2), 138–151.

Ramachandran, P., B. Zoph, and Q. V. Le (2017). Searching for activation functions. *Working Paper*. Available at https://arxiv.org/abs/1710.05941.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 1135–1144. Association for Computing Machinery.

Ribeiro, M. T., S. Singh, and C. Guestrin (2018). Anchors: High-precision model-agnostic explanations. *Conference on Artificial Intelligence (AAAI) 18*, 1527–1535.

Rolnick, D. and M. Tegmark (2018). The power of deeper networks for expressing natural functions. In *International Conference on Learning Representations*.

Rösch, D. and H. Scheule (2014). Forecasting probabilities of default and loss rates given default in the presence of selection. *Journal of the Operational Research Society 65*(3), 393–407.

Rosenberg, B., K. Reid, and R. Lanstein (1985). Persuasive evidence of market inefficiency. *The Journal of Portfolio Management 11*(3), 9–16.

Rossi, A. G. (2018). Predicting stock market returns with machine learning. *Working Paper*. Available at https://mendoza.nd.edu/wp-content/uploads/2019/07/2018-Alberto-Rossi-Fall-Seminar-Paper-1-Stock-Market-Returns.pdf.

Ruf, J. and W. Wang (2020). Neural networks for option pricing and hedging: a literature review. *Journal of Computational Finance*, *Forthcoming*.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning representations by back-propagating errors. *Nature 323*(6088), 533–536.

Sadhwani, A., K. Giesecke, and J. Sirignano (2021). Deep learning for mortgage risk. *Journal of Financial Econometrics 19*(2), 313–368.

Salinas, D., V. Flunkert, J. Gasthaus, and T. Januschowski (2019). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting 36*(3), 1181–1191.

Sariev, E. and G. Germano (2019). Bayesian regularized artificial neural networks for the estimation of the probability of default. *Quantitative Finance*, 1–18.

Segura, A. and J. Zeng (2020). Off-balance sheet funding, voluntary support and investment efficiency. *Journal of Financial Economics 137*(1), 90–107.

Sigrist, F. and C. Hirnschall (2019). Grabit: Gradient tree-boosted Tobit models for default prediction. *Journal of Banking & Finance 102*, 177–192.

Simester, D., A. Timoshenko, and S. I. Zoumpoulis (2020). Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science 66*(6), 2495–2522.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review*, 289–315.

Sopitpongstorn, N., P. Silvapulle, J. Gao, and J.-P. Fenech (2021). Local Logit Regression for Loan Recovery Rate. *Journal of Banking & Finance*, 106093.

Sriram, K., R. V. Ramamoorthi, and P. Ghosh (2013). Posterior Consistency of Bayesian Quantile Regression Based on the Misspecified Asymmetric Laplace Density. *Bayesian Analysis 8*(2), 479–504.

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

Stone, H. (2020). Calibrating rough volatility models: a convolutional neural network approach. *Quantitative Finance 20*(3), 379–392.

Storn, R. and K. Price (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization 11*(4), 341–359.

Sufi, A. (2009). Bank lines of credit in corporate finance: An empirical analysis. *The Review of Financial Studies 22*(3), 1057–1088.

Sun, H. S. and Z. Jin (2016). Estimating credit risk parameters using ensemble learning methods: an empirical study on loss given default. *Journal of Credit Risk 12*(3).

Svensson, L. E. (1994). Estimating and interpreting forward interest rates. *IMF Working Paper*. Available at `https://www.nber.org/papers/w4871`.

Takeuchi, I., Q. V. Le, T. D. Sears, and A. J. Smola (2006). Nonparametric Quantile Estimation. *Journal of Machine Learning Research 7*, 1231–1264.

Tam, K. Y. (1991). Neural network models and the prediction of bank bankruptcy. *Omega 19*(5), 429–445.

Tam, K. Y. and M. Y. Kiang (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management Science 38*(7), 926–947.

Tanoue, Y. and S. Yamashita (2019). Loss given default estimation: a two-stage model with classification tree-based boosting and support vector logistic regression. *Journal of Risk 21*(4).

Thackham, M. and J. Ma (2019). Exposure at default without conversion factors - evidence from global credit data for large corporate revolving facilities. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 182*(4), 1267–1286.

Tian, Y., H. Lian, and M. Tian (2017). Bayesian composite quantile regression for linear mixed-effects models. *Communications in Statistics - Theory and Methods 46*(15), 7717–7731.

Tobback, E., D. Martens, T. V. Gestel, and B. Baesens (2014). Forecasting Loss Given Default models: impact of account characteristics and the macroeconomic state. *Journal of the Operational Research Society 65*(3), 376–392.

Tomarchio, S. D. and A. Punzo (2019). Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 182*(4), 1247–1266.

Tong, E. N., C. Mues, I. Brown, and L. C. Thomas (2016). Exposure at default models with and without the credit conversion factor. *European Journal of Operational Research 252*(3), 910 – 920.

Trolle, A. B. and E. S. Schwartz (2009). A General Stochastic Volatility Model for the Pricing of Interest Rate Derivatives. *The Review of Financial Studies 22*(5), 2007–2057.

Valvonis, V. (2008). Estimating ead for retail exposures for basel ii purposes. *Journal of Credit Risk 4*(1), 79–110.

Wagenvoort, R. (2006). Comparing distributions: The harmonic mass index: Extension to m samples. Technical report, Economic and Financial Report.

Wu, Q. and X. Yan (2019). Capturing deep tail risk via sequential learning of quantile dynamics. *Journal of Economic Dynamics and Control 109*, 103771.

Wu, W., J. Chen, Z. Yang, and M. L. Tindall (2021). A cross-sectional machine learning approach for hedge fund return prediction and selection. *Management Science 67*(7), 4577–4601.

Xu, Q., X. Liu, C. Jiang, and K. Yu (2016). Quantile autoregression neural network model with applications to evaluating value at risk. *Applied Soft Computing 49*, 1–12.

Yang, B. H. and M. Tkachenko (2012). Modeling exposure at default and loss given default: empirical approaches and technical implementation. *The Journal of Credit Risk 8*(2), 81.

Yao, X., J. Crook, and G. Andreeva (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research 263*, 679–689.

Yashkir, O. and Y. Yashkir (2013). Loss given default modeling: a comparative analysis. *The Journal of Risk Model Validation 7*(1), 25–59.

Yoganarasimhan, H. (2020). Search Personalization Using Machine Learning. *Management Science 66*(3), 1045–1070.

Yu, K. and R. A. Moyeed (2001). Bayesian quantile regression. *Statistics & Probability Letters 54*(4), 437–447.

Yu, K. and J. Stander (2007). Bayesian analysis of a tobit quantile regression model. *Journal of Econometrics 137*(1), 260–276.

Yu, K., P. van Kerm, and J. Zhang (2005). Bayesian Quantile Regression: An Application to the Wage Distribution in 1990s Britain. *Sankhyā: The Indian Journal of Statistics (2003-2007) 67*(2), 359–377.

Yu, K. and J. Zhang (2005). A Three-Parameter Asymmetric Laplace Distribution and Its Extension. *Communications in Statistics - Theory and Methods 34*(9-10), 1867–1879.

Yue, Y. R. and H. Rue (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis 55*(1), 84–96.

Zhao, J. Y., D. W. Dwyer, and J. Zhang (2014). Usage and exposures at default of corporate credit lines: an empirical study. *The Journal of Credit Risk 10*(1), 65–86.