Essays on Machine Learning in Risk Management, Option Pricing, and Insurance Economics

A DISSERTATION

Approved by the Faculty of Economics and Management Science,

Leipzig University,

for Obtaining the Academic Degree

Doctor rerum politicarum

Dr. rer. pol.

Presented

by Dipl.-Wirtsch.-Math. Simon Fritzsch

born on September 3, 1992 in Zwickau

Reviewers: Prof. Dr. Gregor Weiß

Prof. Dr. Roland Schuhr

Date of conferral: June 15, 2022

Bibliographic description

Fritzsch, Simon

Essays on Machine Learning in Risk Management, Option Pricing, and Insurance Economics

Leipzig University, dissertation

265 pages, 294 references, 46 figures, 42 tables, 4 annexes

Abstract:

Dealing with uncertainty is at the heart of financial risk management and asset pricing. This cumulative dissertation consists of four independent research papers that study various aspects of uncertainty, from estimation and model risk over the volatility risk premium to the measurement of unobservable variables.

In the first paper, a non-parametric estimator of conditional quantiles is proposed that builds on methods from the machine learning literature. The so-called leveraging estimator is discussed in detail and analyzed in an extensive simulation study. Subsequently, the estimator is used to quantify the estimation risk of Value-at-Risk and Expected Shortfall models. The results suggest that there are significant differences in the estimation risk of various GARCH-type models while in general estimation risk for the Expected Shortfall is higher than for the Value-at-Risk.

In the second paper, the leveraging estimator is applied to realized and implied volatility estimates of US stock options to empirically test if the volatility risk premium is priced in the cross-section of option returns. A trading strategy that is long (short) in a portfolio with low (high) implied volatility conditional on the realized volatility yields average monthly returns that are economically and statistically significant.

The third paper investigates the model risk of multivariate Value-at-Risk and Expected Shortfall models in a comprehensive empirical study on copula GARCH models. The paper finds that model risk is economically significant, especially high during periods of financial turmoil, and mainly due to the choice of the copula.

In the fourth paper, the relation between digitalization and the market value of US insurers is analyzed. Therefore, a text-based measure of digitalization building on the Latent Dirichlet Allocation is proposed. It is shown that a rise in digitalization efforts is associated with an increase in market valuations.

Acknowledgment

During the course of the last years, I have received strong support, encouragement, and inspiration from many sides. At this point, I would like to take the opportunity to thank all those people who have contributed greatly to the success of this dissertation.

First of all, I want to express my deep gratitude to my supervisor. Thank you very much, Gregor, for your constant advice and guidance as well as for continuously providing new ideas and perspectives. I also wish to thank the LBBW Asset Management and in particular the team Portfolio Analytics for providing the funding for a PhD-scholarship and for valuable insights into practical asset management. Furthermore, I am very grateful for the contributions of and fruitful discussions with my further co-authors Maike, Philipp, and Felix. I also thank my other colleagues Jana, Sebastian, and David for a great time at work, stimulating conversations, and not least for welcome distraction.

My greatest thanks goes to my family. I am immensely grateful to my mother and my siblings for always having an open ear and encouraging me. By far the deepest thanks goes to my girlfriend. Without you, Elli, and your unwavering support, patience, and understanding, this dissertation would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Dealing with uncertainty is at the heart of risk management and asset pricing. This concerns not only uncertainty with regard to future price movements but also with regard to latent factors like financial risk (e.g., measured by the Value-at-Risk or the Expected Shortfall) and volatility. As these latent factors are not directly observable they have to be captured via statistical models. Estimates from different models, however, can vary widely (Danielsson et al., 2016) as they are subject to two types of uncertainty. First, most models require parameters that have to be estimated from data. This leads to estimation risk. Secondly, the true data generating process is unknown and might not be adequately reflected by a particular model, giving rise to model risk (Lönnbark, 2013).

Regarding the estimation risk of popular risk measures, surprisingly little is known about the uncertainty of Value-at-Risk and Expected Shortfall predictors despite a lot of research on various other modeling issues. Early work is due to Jorion (1996) laying out a statistical methodology for analyzing the estimation error in Value-at-Risk models. Further studies in this vein include, amongst others, Christoffersen and Gonçalves (2005), Chan et al. (2007), Lan et al. (2010), and Kabaila and Mainzer (2018). However, all these studies rely on either Monte Carlo simulations, distribu-

tional assumptions on the Value-at-Risk or Expected Shortfall, or some other kind of parametric method for measuring estimation risk. This dissertation proposes a different non-parametric approach towards quantifying estimation risk that is based on the cross-section of risk estimates at a given point in time.

The dissertation also investigates the related issue of model risk for Value-at-Risk and Expected Shortfall models. Existing research focuses mainly on the factors that control model risk within models. This includes misspecification of the underlying theoretical models (Green and Figlewski, 1999) and assumptions made about distributions, parameters, or other model specifications (see, e.g., Hull and Suo, 2002, Alexander and Sarabia, 2012, Glasserman and Xu, 2014, Boucher et al., 2014). This dissertation studies a more general problem. Given a large variety of standard Value-at-Risk and Expected Shortfall models within the financial industry, uncertainty about the choice of a particular model creates model risk per se. This notion of model risk as uncertainty on the model choice itself in the presence of a large set of valid candidate models is most closely related to Cont (2006) and Danielsson et al. (2016). The theorem by Sklar (1959) enables the separate modeling of the marginals and the dependence structure of multivariate return data by means of copula functions. The dissertation focuses in particular on the contribution of each of these modeling steps to the overall model risk of multivariate risk models.

While the Value-at-Risk and Expected Shortfall are commonly used throughout the financial industry, the most basic measure of uncertainty is probably volatility. Given the prices of equity options, there are two main ways for deriving volatility estimates: via some model from the returns of the underlying stocks (realized volatility) or from option prices (implied volatility). Deviations between realized and implied volatility are then commonly referred to as the volatility risk premium. However, although volatility is an important driver of option prices and options themselves are ubiquitous in financial risk management, there is yet no conclusive answer on the role volatility and volatility risk play in the cross-section of option returns (see, e.g., Driessen et al., 2009, Carr and Wu, 2009 and Bakshi and Kapadia, 2003, Goyal and Saretto, 2009, Cao

and Han, 2013, Cao et al., 2019, Hu and Jacobs, 2020). This dissertation proposes a new approach for building long-short portfolios to exploit the volatility risk premium in the cross-section of option returns. Therefore, options with "extreme" deviations between implied and realized volatility are identified non-parametrically.

A key not only for determining these long-short portfolios but also more generally for dealing with uncertainty in the statistical relationship between a dependent and explaining variables is modeling their joint probability distribution. In the empirical literature, conditional means are predominantly used for capturing the dependence between a variable of interest and its covariates. This approach, however, neglects many information about the underlying joint probability distribution as the conditional mean only allows to assess the average relationship between the variables. There are many problems where more information about the distribution of an independent variable given that the covariates assume a particular value are needed. This is the application domain of quantile regression. Conditional quantiles have many favorable properties. For example, they can capture conditional asymmetry as well as heteroskedasticity and are more robust to outliers and censored data. Beginning with the introduction of quantile regression in the seminal paper by Koenker and Bassett (1978), there has been a lot of research in this field of study. Several extensions to the simple linear quantile estimator have been proposed (e.g., Koenker et al., 1994, Yu and Jones, 1998, Koenker and Mizera, 2004, Koenker, 2011). Other approaches include, among others, semiparametric quantile regression, kernel estimators, and methods for time series (e.g., Heagerty and Pepe, 1999, Li and Racine, 2008, Chen et al., 2009). Furthermore, machine learning based estimators (e.g., Bhattacharya and Gangopadhyay, 1990, Hwang and Shim, 2005, Meinshausen, 2006, Zheng, 2012, Charlier et al., 2015b, Rothfuss et al., 2019) have shown promising results. This dissertation introduces a new non-parametric estimator of conditional quantiles that relies on two methods from the field of machine learning: optimal quantization via the Competitive Learning Vector Quantization algorithm by Kohonen (1982, 1989) for grouping similar observations and leveraging for aggregating an ensemble of estimators to a stronger one.

Machine learning techniques have proven useful in various other fields. Over the last decades, advances in models and architectures, the increased availability of large datasets, and most importantly the rise in computing power have led to the advent of machine learning methods in countless areas of daily life and scientific research. In particular, machine learning methods are nowadays frequently used in finance for tasks such as portfolio construction and asset pricing (e.g., Moritz and Zimmermann, 2016, López de Prado and Lewis, 2019, Kelly et al., 2019, Goyenko and Zhang, 2020, Gu et al., 2020, Ivașcu, 2021, Bali et al., 2021, Bianchi et al., 2021) and time series forecasting (e.g., Rapach et al., 2013, Vedavathi et al., 2014, Rossi, 2018, Sirignano and Cont, 2019, Chen et al., 2020a, Freyberger et al., 2020, Bali et al., 2020, Kozak et al., 2020). These studies rely on both supervised and unsupervised learning algorithms. In supervised learning, algorithms are trained on data that are labeled while in unsupervised learning the goal is to find patterns or draw inference from data that are not. This dissertation relies on methods from the realm of unsupervised machine learning where the above mentioned Competitive Learning Vector Quantization algorithm is used for clustering and a probabilistic model from the field of textual analysis is employed for making inference about unobserved factors.

While in many industries machine learning has gained popularity only in recent years, digitalization in general has already radically transformed many economic sectors. The insurance industry, however, has yet to realize the full potential of digitalization over the whole insurance value chain (cf. Eling and Lehmann, 2018, Cappiello, 2020). This becomes all the more important considering the competitive pressure in the insurance industry that is increased by the zero interest rate policy, the aftereffects of the great financial crisis, and rising customer expectations. At the same time, there is only little empirical evidence on the effect of digitalization on firm outcomes in the insurance industry (see, e.g., Scott et al., 2017, Bohnert et al., 2019, Hanelt et al., 2020). One reason for this is that digitalization is a rather general concept making it difficult to quantify the degree to which insurers digitalize. A possible way of approaching this problem is to analyze the annual reports of insurance companies using

algorithms from the field of textual analysis. Over the last years, there has emerged a growing body of literature on textual analysis in finance (e.g., Hanley and Hoberg, 2010, Jegadeesh and Wu, 2013, Hoberg et al., 2014, Hoberg and Maksimovic, 2015, Jegadeesh and Wu, 2017, Ke et al., 2019). Machine learning provides possibilities that go far beyond mere word list approaches. Specifically, the Latent Dirichlet Allocation by Blei et al. (2003) allows inference to be drawn on the thematic structure within a collection of documents. This makes this probabilistic model a popular choice within the literature on machine learning and textual analysis in finance (see, e.g., Goldsmith-Pinkham et al., 2016, Ganglmair and Wardlaw, 2017, Hoberg and Lewis, 2017, Huang et al., 2018, Lopez-Lira, 2019, Lowry et al., 2020, Bellstam et al., 2020). In this dissertation, the Latent Dirichlet Allocation is employed to obtain a vector of topic loadings for each annual report that is subsequently used to construct a text-based measure of digitalization.

The dissertation consists of four self-contained research papers (Chapters 2-5), which can be read independently from each other. In Chapter 2, a new estimator of conditional quantiles is introduced that is based on ideas from machine learning, non-parametric, and can be applied to multivariate covariates. The estimator is then used to derive a measure of estimation risk for Value-at-Risk and Expected Shortfall models. In Chapter 3, this new estimator is applied to equity option data to analyze if the volatility risk premium is priced in the cross-section of option returns. Chapter 4 complements the empirical analysis from Chapter 2 by studying the model risk of multivariate Value-at-Risk and Expected Shortfall models (without relying on machine learning methods). Finally, in an additional paper (Chapter 5), a measure of digitalization based on the annual reports of insurance companies is proposed and its relation to firm outcomes is analyzed. In the following sections named after the corresponding chapters, more detailed information on the papers along with their main findings are provided.

1.2 Conditional quantile estimation via leveraging optimal quantization

In Chapter 2, a new non-parametric estimator of conditional quantile curves called the leveraging estimator is proposed. The estimator is based on two concepts from the machine learning literature: optimal quantization via the Competitive Learning Vector Quantization algorithm by Kohonen (1982, 1989) and leveraging.

The goal of optimal quantization is to replace a continuous random variable by another random variable that assumes only finitely many values and minimizes some kind of approximation error. A stochastic algorithm for performing optimal quantization on a finite sample of data is the Competitive Learning Vector Quantization algorithm. For the task of conditional quantile estimation, the algorithm is applied to the covariates yielding groups of similar observations. In each of these groups, the empirical (unconditional) quantiles of the response variable are computed. This is essentially the estimator introduced by Charlier et al. (2015b). Chapter 2 proposes the construction of an ensemble of these estimators where the single ensemble members are iteratively combined such that the performance of the aggregated estimator is improved stepwise. This concept is known as leveraging. The proposed estimator involves several hyperparameters that govern the extent to which it adapts to the data. Therefore, the paper also introduces a data-driven hyperparameter selection procedure. The theoretical description and discussion of the estimator is concluded by providing convergence results as the number of observations and the number of clusters in the Competitive Learning Vector Quantization algorithm go to infinity.

In an extensive simulation study, the performance of the leveraging estimator is analyzed in detail and compared to competing algorithms. For univariate covariates, the leveraging estimator produces conditional quantile curves that are both smooth and adapt well to the true curves of the underlying data generating model, even in the edges of the covariates' support. Additionally, the integrated squared errors of the leveraging

algorithm are very competitive among the considered estimators. The simulation study is then extended to account for multivariate covariates of up to four dimensions. Again, the estimator yields competitive integrated squared errors.

The analysis of the leveraging estimator is complemented by an empirical study where the estimator is used to derive a measure of estimation risk for Value-at-Risk and Expected Shortfall models in the US equity market. More , the estimator is applied to one day ahead Value-at-Risk and Expected Shortfall forecasts for the constituents of the S&P Composite 1500 Index derived from GARCH(1,1)-type models at a fixed date. For a given model, the quantiles of these parametric risk estimates conditional on their non-parametric counterparts obtained via historical simulation are then estimated. Based on the conditional 25 % and 75 % quantile curves the interquartile range averaged over all S&P Composite 1500 Index constituents is calculated and proposed as a measure of estimation risk. This approach of determining estimation risk non-parametrically from the cross-section of risk estimates without relying on Monte Carlo methods is new to the literature.

The paper finds that estimation risk varies substantially over the sample period January 2000 until March 2021 and is especially pronounced in the aftermath of the dot-com bubble, during the great financial crisis, and during the 2020 stock market crash due to the COVID-19 pandemic. When comparing various GARCH-type models, the EGARCH model by Nelson (1991) is associated with the highest estimation risk for both the Value-at-Risk and the Expected Shortfall while the GARCH model by Bollerslev (1986) emerges as the model with the lowest estimation risk. Overall, the results suggest that the estimation risk for the Expected Shortfall is in general higher than for the Value-at-Risk, regardless of the employed GARCH-type model. When conditioning on the realized volatility instead of the risk measures obtained via historical simulation, the results are similar for the Expected Shortfall but somewhat lower for the Value-at-Risk. This highlights the main weakness of the employed approach for measuring estimation risk, namely the reliance on a non-parametric benchmark.

Instead of computing a measure of estimation risk, the estimated conditional quan-

tile curves can be used directly to derive confidence intervals for the risk forecasts of a particular stock at a particular date. Aggregating these intervals over time yields confidence bands. The paper illustrates and discusses both average confidence bands and confidence bands for single stocks. As the interpretation of both the introduced measure of estimation risk and the confidence bands is somewhat delicate, the empirical study should be primarily seen as an illustration of the applicability of the proposed estimator in a risk management context. Furthermore, the non-parametric nature of the leveraging estimator and its applicability to multiple dimensions make it an interesting choice for many other applications. More generally, the paper argues that conditional quantiles can provide valuable insights beyond the commonly used conditional average and therefore should become a standard tool in empirical research.

1.3 Cross-section of option returns and the volatility risk premium

Chapter 3 provides another application of the leveraging estimator from Chapter 2. The estimator is applied to realized and implied volatility estimates of US stock options between January 1996 and June 2019 to empirically test if the volatility risk premium is priced in the cross-section of option returns. Therefore, a long-short strategy that is based on the conditional 10 % and 90 % quantile curves of implied volatility conditional on realized volatility is implemented. Using conditional quantiles helps in capturing the non-linear relationship between implied and realized volatilities thereby avoiding biases stemming from systematic differences in realized volatility that are known to affect the cross-section of option returns (cf. Cao and Han, 2013, Hu and Jacobs, 2020). The obtained results provide strong and robust evidence for the existence of such a premium in the cross-section of option returns.

A trading strategy that is long (short) in high (low) deviations between realized and implied volatilities delivers positive returns that are both statistically and economically

significant. This applies to call and put delta-hedged and raw option strategies for at the money options. For example, average monthly delta-hedged returns of 1-month at the money options are 2.0 % for call and 1.7 % for put options with monthly Sharpe ratios of 0.842 and 0.796, respectively. When additionally conditioning on option moneyness, the results can be extended to options of arbitrary moneyness. For example, a long-short delta-hedged trading strategy for options of arbitrary moneyness yields average monthly returns of 2.4 % for call and 2.5 % for put contracts with monthly Sharpe ratios of 0.816 and 0.844, respectively. The results are robust to the inclusion of dividend-paying stocks, alternative estimators of conditional quantiles, reasonable transaction costs (delta-hedged returns), the expansion of the long-short portfolios to less extreme options, different levels of trading volume, and controlling for skewness and kurtosis of the underlyings' return distribution.

The key to these findings, distinguishing the paper from previous work, is the use of conditional quantiles instead of conditional portfolios sorts or regression techniques. While conditional portfolio sorts are frequently used throughout the literature, they can usually not control for more than two characteristics at the same time due to the curse of dimensionality. The use of machine learning techniques by the leveraging estimator allows for a more data-efficient modeling of the relevant characteristics and makes the inclusion of more covariates feasible. At the same time, the approach can be seen as a possible solution to the empty portfolio problem that arises in standard portfolio sorts when sorting on too many variables (see, e.g., Goyal, 2012). Nevertheless, the number of covariates one can control for with the leveraging estimator is still limited as the curse of dimensionality, albeit later, comes into effect in higher dimensions.[1] This is typically the field of application of cross-sectional regressions. Unfortunately, such regressions only provide information on long-short trading strategies that involve trading in all securities with portfolio weights that may vary widely. However, in empirical asset pricing one is typically interested in trading strategies that only involve a relatively small number of assets and are easy to interpret. While the proposed non-parametric

[1] In the corresponding paper, up to four covariates are included at the same time.

approach cannot account for as many covariates as cross-sectional regressions, it yields a trading strategy that is easy to implement and interpret and makes no assumption on the functional form of the relationship between implied and realized volatility. Most notably, the approach is not limited to the study of options but is sufficiently general to be applied to other assets as well.

1.4 Marginals versus copulas: Which account for more model risk in multivariate risk forecasting?

Chapter 4 complements the empirical analysis of estimation risk in Chapter 2 by analyzing the model risk of multivariate Value-at-Risk and Expected Shortfall models. The focus of the chapter is on copula GARCH models that provide a very flexible way for modeling multivariate time series and risk forecasts (see, e.g., Jondeau and Rockinger, 2006, Patton, 2006, Aas and Berg, 2009, Fischer et al., 2009, Brechmann and Czado, 2013, Jiang et al., 2018, Chabi-Yo et al., 2018). This class of models is based on the theorem by Sklar (1959) stating that the modeling of the marginals can be separated from the modeling of the dependence structure of a multivariate probability distribution. GARCH-type models are employed to filter the univariate time series while copulas are subsequently used to model the dependence between the different assets in a portfolio. Combining various copula functions and GARCH-type models yields a large number of model specifications that produce differing forecasts for the same risk measure. Model risk now arises from the uncertainty on the model choice in the presence of many possible alternative models (cf. Danielsson et al., 2016).[2]

The paper proposes the usage of the mean absolute deviation of the various risk forecasts (for a given risk measure at a given day) as a measure of model risk. As banks are required by the Basel III regulation to backtest their (internal) market risk models, for them uncertainty on the model choice is essentially uncertainty on the choice of

[2]Note that the goal of the paper is not to identify an optimal model or rank models by their forecasting accuracy like, e.g., in Santos et al. (2013) but to quantify the extent of non-conformity of risk forecasts for the same risk measure.

models that have not been rejected in backtests. Therefore, at a given day only those risk forecasts enter in the calculation of the model risk measure that were not rejected by a standard backtest. Risk forecasts by copula GARCH models are produced in a two step procedure by first modeling the marginals and subsequently the multivariate dependence structure. Consequently, the question which of these two steps contributes more to the overall model risk arises. To answer this question, different groups of models are analyzed where either the marginals, the copula, or neither are fixed.

The empirical analysis is based on returns from a well diversified portfolio consisting of equity, bond, commodity, and real estate indices of developed and emerging markets from January 2001 until December 2018. Overall, 180 copula GARCH models are employed for producing one day ahead risk forecasts. The focus is on the 99 % Value-at-Risk and the 97.5 % Expected Shortfall in line with the Basel II and Basel III market risk regulations. Over the entire sample period, model risk is on average 0.165 % (of the portfolio value) for Value-at-Risk and 0.092 % for Expected Shortfall forecasts, while daily model risk is quite volatile with a standard deviation of more than half the average model risk. Model risk is especially high during times of financial turmoil. For example, during the years 2008 and 2009 the average model risk is 0.286 % and 0.145 % for the Value-at-Risk and the Expected Shortfall, respectively, which is more than double as high as the average over the period before 2008. This finding, though not surprising, can only partly be explained by an increase in volatility as the average increase in model risk is higher than the average increase in the absolute level of the risk forecasts. A possible explanation for this disproportionately high increase of model risk in periods of financial crisis is that models treat history and shocks quite differently such that a changing statistical regime can lead to higher disagreements between risk forecasts (Danielsson et al., 2016).

Turning to the question concerning the contribution of the modeling of the marginals and copulas to the overall model risk, the results suggest that model risk is mainly driven by the modeling of the copulas. When fixing the marginal distribution, the average model risk is 0.157 % for Value-at-Risk and 0.068 % for Expected Shortfall

forecasts. Model risk is significantly lower when fixing the copula instead (0.052 % for Value-at-Risk and 0.058 % for Expected Shortfall forecasts). This main finding is robust to other choices of the model risk measure, different confidence levels for the risk measures (except for the Expected Shortfall on the 95 % confidence level), and randomly generated portfolio weights. While the results also hold when not considering any backtest before determining the model risk at a given day, in case of the Value-at-Risk the results are not robust to considering an alternative backtest (the dynamic quantile backtest by Engle and Manganelli (2004) instead of the duration based backtest by Christoffersen (2004)). This highlights the importance of the choice of a particular backtest.

As a possible means to reduce model risk the usage of the model confidence set procedure by Hansen et al. (2011) is proposed. This iterative procedure yields a set of models that contains the best model with a given confidence. Narrowing down the set of candidate models by applying the procedure to all models that have not been rejected by the backtests leads to significant reductions in model risk, in particular for the Value-at-Risk. Over the entire period, the average model risk is reduced to 0.127 % for the Value-at-Risk and 0.089 % for the Expected Shortfall.

Summing up, the paper provides new empirical insights on the contribution of the modeling of the marginals and the multivariate dependence structure on the model risk of copula GARCH models. The findings are of particular importance for practitioners and highlight the need to especially pay attention to the modeling of the multivariate dependence structure.

1.5 Estimating the relation between digitalization and the market value of insurers

Chapter 5 studies the relation between digitalization and the market value of US insurance companies.[3] Therefore, a new measure of digitalization is proposed. The measure

[3]The corresponding paper was published in Fritzsch et al. (2021).

is based on the distribution of (latent) topics in the annual reports of US insurers that is extracted via the Latent Dirichlet Allocation.

The Latent Dirichlet Allocation is a probabilistic model that has only recently been introduced to the finance literature with Huang et al. (2018) being one of the first applications in this field. The Latent Dirichlet Allocation yields a finite set of common topics that best reflect a collection of documents. In contrast to commonly used word list approaches where the lists have to be provided by the researcher, these topics and corresponding word distributions arise endogenously from the data. That is, the underlying machine learning algorithm determines the words that are most suitable to discriminate between documents and topics in an unsupervised fashion. Applying the model to a specific document yields a vector of topic loadings that represent how intensively each topic is discussed in the report and serve as a low-dimensional representation of the document (cf. Blei et al., 2003). The topic distribution inside each report is then compared to the topic distribution within a reference document on digitalization (Bohnert et al., 2019) via the Kullback Leibler divergence to proxy for the extent to which insurers digitalize.

This digitalization measure and its relation to the market value of US insurers is subsequently analyzed in an empirical study on 86 publicly-listed US insurance companies between 2006 and 2015 available via Thomson Reuters Datastream. The results from a multivariate ordinary least squares model including firm and time fixed effects suggest that an increase in the digitalization measure is strongly related to an increasing market value and market-to-book value of US insurance companies. That is, market participants expect more digitalized insurance companies to exhibit a higher future profitability and consequently firm value. More detailed, an increase in the digitalization measure by one standard deviation is associated with an increase of the market value by about 7.48 % and of the market-to-book value by about 8.04 % in the subsequent year. The estimation results are robust to different specifications of the Latent Dirichlet Allocation (the number of topics) and the sentiment in which the annual reports are written. Furthermore, the results neither depend on the choice of the reference

document nor the particular calculation of the digitalization measure.

There are some limitations to the approach outlined in the paper. First, the Latent Dirichlet Allocation model requires some assumptions, most notably the bag of words assumption, that might be considered problematic. However, while there exists a variety of extensions to the original model addressing some of its shortcomings, the question which topic model to use when being confronted with a new set of texts and a new task is still an open question in topic modeling (Blei, 2012). Second, the proposed approach has limitations in discriminating between digitalization and mere innovation because these two concepts are closely related to each other. Finally, the regression results might be subject to endogeneity due to reverse causality. Consequently, a causal link cannot be established unequivocally.

The rise of machine learning methods in the field of textual analysis in combination with massive increases in computational power provides researchers and practitioners with powerful tools for gaining new insights from large amounts of textual data. In this sense, the proposed approach can be seen as a first step towards a new empirical analysis of the impact of digitalization in the insurance sector and beyond.

Chapter 2

Conditional Quantile Estimation via Leveraging Optimal Quantization

2.1 Introduction

It is standard to analyze the statistical relationship between dependent and explaining variables via the conditional mean function. For this purpose, many methods have been introduced with the (linear) ordinary least squares estimator still being the most frequently used one. Machine learning algorithms are, in this regression context, also typically concerned about estimating conditional means or closely related quantities. However, the conditional mean only models the average relationship between variables and there are many problems where more information about the conditional distribution of a (scalar-valued) random variable Y given that a random vector of covariates X assumes a particular value are needed. This is the application domain of quantile regression, which provides a more complete picture of conditional distributions. Indeed, conditional quantiles (e.g., median or quartiles) can capture heteroskedasticity, conditional asymmetry, and can also yield conditional prediction intervals and reference curves (hypersurfaces for multivariate covariates). Furthermore, conditional quantiles are more robust to outliers and censored data. These favorable features have led to a widespread use of quantile regression in many different areas, in particular in finance.

Quantile regression was introduced in the seminal paper by Koenker and Bassett (1978) expanding the ordinary least squares estimator to linear quantile regression. In this paper, we introduce a novel non-parametric estimator called *leveraging estimator*. Therefore, we rely on two methods from the field of machine learning: *optimal quantization* via the Competitive Learning Vector Quantization (CLVQ) algorithm by Kohonen (1982, 1989) for grouping similar observations and *leveraging* for combining an ensemble of estimators to a stronger estimator. Optimal quantization is about replacing a continuous random variable by another random variable that assumes only finitely many values such that some quantization error is minimized. The quantization of the covariates yields groups of observations for which the empirical quantiles of the response variable can be calculated. This is essentially the estimator introduced by Charlier et al. (2015b).

We propose to construct an ensemble of these estimators by iteratively combining the ensemble members such that in each step the performance of the aggregated estimator is improved. This concept is called leveraging.[4] Furthermore, we propose a data-driven procedure for selecting the hyperparameters of the algorithm and discuss it in detail. We provide convergence results for the proposed estimator and compare it to the base estimator and other competitors in an extensive simulation study. In the case of univariate covariates, we find that the proposed estimator produces smooth quantile curves that adapt well to the true conditional quantile curves of the underlying data generating model and yields integrated squared errors (ISEs) that are very competitive among the considered algorithms. The estimator generalizes naturally to multivariate covariates. We extend the simulation study to consider up to four dimensional covariates and again yield competitive ISEs.

In an empirical study, we apply the leveraging estimator to one day ahead Value-at-Risk (VaR) and Expected Shortfall (ES) forecasts to study the associated estimation risk in the broad US equity market across time. Therefore, we compute VaR

[4]The concept of leveraging is very similar to boosting and the terms are often used interchangeably, see Section 2.3 for more details.

and ES estimates for each constituent of the S&P Composite 1500 Index via various GARCH(1,1)-type models based on log-returns from January 2000 until March 2021. For a given point in time and a given VaR or ES model, we then derive conditional quantile curves of the (parametric) risk estimates conditional on their non-parametric counterparts obtained via historical simulation. Subsequently, estimation risk is quantified as the average interquartile range (iqr) of the conditional quantiles. Additionally, the procedure yields non-parametric confidence bands for VaR and ES forecasts at the stock level. There is substantial variation in estimation risk over time. Estimation risk is especially high in the aftermath of the bursting dotcom bubble, during the great financial crisis, and during the 2020 stock market crash. Additionally, we find that among the considered GARCH-type models the GARCH model exhibits the lowest estimation risk for both VaR and ES, while the EGARCH model is associated with the highest estimation risk. Furthermore, the results suggest that estimation risk is in general higher for the ES than for the VaR. When determining the estimation risk of the GARCH model by conditioning on realized volatility (RV) instead of historical simulation ES/VaR we obtain similar values for the ES and somewhat lower values for the VaR.

The paper is related to a growing body of literature on the estimation of conditional quantiles. Starting with the introduction of the linear quantile estimator by Koenker and Bassett (1978) there has been a lot of research on quantile regression. Several extensions to the simple linear model have been proposed, e.g., via quantile smoothing splines (Koenker et al., 1994, Koenker and Mizera, 2004), additive models (Koenker, 2005, 2011), and local linear quantile regression (Fan et al., 1994, Yu and Jones, 1998). Further approaches (among many others) include semiparametric quantile regression (Heagerty and Pepe, 1999), quadratic programming based estimators (Takeuchi et al., 2006), kernel estimators (Li and Racine, 2008, Li et al., 2013), non-crossing estimators (Dette and Volgushev, 2008), single-index quantile regression (Wu et al., 2010), local quantile regression (Spokoiny et al., 2013), copula based estimators (Noh et al., 2015, Kraus and Czado, 2017), and methods for time series (e.g., Chen et al., 2009, Xiao

and Koenker, 2009). For a more complete coverage of (earlier) methods see also the monograph by Koenker (2005).

In the machine learning context, Bhattacharya and Gangopadhyay (1990) introduce nearest-neighbor estimators while Hwang and Shim (2005), Meinshausen (2006), and Zheng (2012) propose estimators based on support vector machines, random forests, and gradient boosting, respectively. More recently, Charlier et al. (2015b,a) derive an estimator based on the concept of optimal quantization via the CLVQ algorithm and Rothfuss et al. (2019) propose neural network based estimators.

Predicting conditional quantiles of a random variable given that some covariates assume a particular value has found numerous applications, in particular in finance. For example, Bouyé and Salmon (2009) study dependencies in the foreign exchange market, Spokoiny et al. (2013) analyze tail dependence in the Hong Kong stock market, and Adrian and Brunnermeier (2016) introduce a measure of systemic risk, the CoVaR, that is calculated from the conditional loss distribution of one financial institution conditional on other institutions being under distress.

As in the empirical application in this paper we analyze estimation risk of GARCH-type models, this paper is also related to the literature on estimation risk in risk models. Interestingly, although there is a lot of research on different modeling issues, only little is known about the uncertainty of VaR and ES predictors. There are some papers that focus on how uncertainty in risk estimates influences the accuracy of VaR and ES backtests (e.g., Escanciano and Olmo, 2010, Lönnbark, 2013), coverage levels of VaR (Figlewski, 2003), or how VaR estimators should be corrected for estimation errors to avoid underestimation of the portfolio risk (Lönnbark, 2010). However, this paper aims to directly quantify estimation risk in a key figure and to provide confidence bands. Early research in this direction includes work by Jorion (1996) laying out the statistical methodology for analyzing estimation errors in VaR models, and Christoffersen and Gonçalves (2005) and Chan et al. (2007) providing confidence bands around point VaR and ES forecasts. Further studies in this vein include Gao and Song (2008), Lan et al. (2010), and Kabaila and Mainzer (2018). However, all these studies hinge

on either Monte Carlo simulations, distributional assumptions on VaR or ES, or some other kind of parametric method towards measuring estimation risk. This paper proposes a different approach that is based on the cross-section of risk estimates at a given point in time.

The paper makes two main contributions. First, we introduce a novel estimator of conditional quantiles that yields competitive quantile estimates and generalizes naturally to multiple dimensions. In an extensive simulation study, we illustrate the added value of leveraging an ensemble of quantization-based estimators. Second, in an empirical application we study the estimation risk of various GARCH-type models in the US equity market via a non-parametric approach that does not rely on Monte Carlo simulations and provide confidence bands for single stocks.

The remainder of the paper is organized as follows. Section 2.2 introduces the concept of optimal quantization along with the CLVQ algorithm. In Section 2.3, we derive the proposed estimator and discuss the effect and the data-driven choice of the employed hyperparameters in Section 2.4. In Section 2.5, an extensive simulation study for both univariate and multivariate covariates is performed. Subsequently, the usefulness of the estimator is studied in an empirical application in Section 2.6. Section 2.7 concludes.

2.2 Optimal quantization

In this section we introduce the concept of L_p-optimal quantization of random variables and the related concept of Voronoi tessellations along with basic existence and convergence results. Furthermore, we present a stochastic gradient algorithm for performing optimal quantization in the finite sample case. Optimal quantization addresses the problem of finding the (in some way) best *discrete* approximation of a continuous random variable X. That is, for $N \in \mathbb{N}$ one aims to replace X by another random variable $\tilde{X}^N$ that assumes at most N pairwise distinct values and minimizes some error functional. The concept of L_p-optimal quantization is not new but has been barely

used in statistics (Charlier et al., 2015b). A first extensive treatment of the topic can be found in Zador (1964).

We start by introducing some notation. Let X denote a d-dimensional random variable with distribution P_X on the probability space $(\Omega, \mathfrak{A}, P)$. Let further $p \geq 1$ be such that the Euclidean norm of X is in L_p. That is, we require $\|X\|_p < \infty$, where

$$\|X\|_p := \Big(\int_\Omega |X|^p dP \Big)^{1/p} \tag{2.1}$$

with $|\cdot|$ denoting the Euclidean norm on $\mathbb{R}^d$. For $N \in \mathbb{N}$ the aim of optimal quantization is to find a random variable $\tilde{X}^N$ assuming at most N pairwise distinct values such that the L_p-norm quantization error $\|X - \tilde{X}^N\|_p$ is minimized. This problem is equivalent to finding a set of points $\Gamma_N = \{\xi_1, \xi_2, \ldots, \xi_N\} \subset \mathbb{R}^d$ which we refer to as quantizers[5] such that the L_p-norm quantization error $\|X - Proj_{\Gamma_N}(X)\|_p$ is minimized, where $Proj_{\Gamma_N}(X)$ is defined as the projection of X on the nearest point (in the Euclidean norm) of the N-grid Γ_N. In the sequel we will denote such a grid simply as *optimal N-grid*.

Each grid Γ_N gives rise to a Voronoi tessellation $C = C(\Gamma_N)$ of $\mathbb{R}^d$. As the concept of a Voronoi tessellation is crucial for an intuitive understanding of the quantization method employed in this paper we provide a formal definition. We say $C = (C_j)_{j=1}^N$ is a Voronoi tessellation of the N-grid $\Gamma_N = \{\xi_1, \ldots, \xi_N\} \subset \mathbb{R}^d$ if and only if $(C_j)_{j=1}^N$ is a Borel partition satisfying

$$C_j \subset \{x \in \mathbb{R}^d \Big| \, |\xi_j - x| = \min_{1 \leq i \leq N} |\xi_i - x|\} \text{ for all } 1 \leq j \leq N.$$

The projection $Proj_{\Gamma_N}(X)$ of X on the N-grid Γ_N can than be expressed in terms of the Voronoi tessellation $C = (C_j)_{j=1}^N$ as

$$Proj_{\Gamma_N}(X) = \sum_{j=1}^N \xi_j \mathbb{I}_{C_j}(X), \tag{2.2}$$

[5]In the literature the term quantizer is also sometimes used to denote the map of X to its quantized version $\tilde{X}$. However, in this paper we restrict the usage of the term to the points in the optimal N-grid.

where $\mathbb{I}_{C_j}$ denotes the indicator function of the set C_j (Bally et al., 2005).

The existence of an optimal N-grid is ensured by Abaya and Wise (1984) under the assumption that the distribution P_X is continuous. Even though an optimal N-grid Γ_N might not be unique,[6] the convergence of the sequence $(\|Proj_{\Gamma_N}(X) - X\|_p)_{N\in\mathbb{N}}$ towards zero can easily be established as a consequence of Lebesgue's dominated convergence theorem, see the Appendix for details. For results on the rate of convergence we refer to Graf and Luschgy (2002).

Although existence of an optimal N-grid is guaranteed for continuous distributions, we are still left with the problem of actually finding an optimal N-grid. Closed form solutions only exist in very specific situations, therefore we rely on a stochastic gradient algorithm for determining a grid based on a finite sample of observations. The algorithm can be specified for arbitrary $p \geq 1$. However, in the sequel we focus on the most common case $p = 2$ for which convergence results are much more satisfactory than for arbitrary $p \geq 1$, see Pagès (1998) for more details. Restricting to $p = 2$ leads to the CLVQ algorithm, also known as Kohonen algorithm with 0 neighbors, which has emerged in the mid-1980s as the degenerate version of self-organizing algorithms (Kohonen, 1982, 1989, Bouton and Pagès, 1997, Pagès and Printems, 2003). The algorithm can be understood as a special case of an artificial neural network where weight vectors are adjusted based on competitive learning (Grossberg, 1976, 1987). CLVQ tries to represent the feature space by a set of N so-called prototypes. These prototypes are determined iteratively according to the following intuition. An input vector is presented to all N neurons to determine the neuron with the minimum distortion between its weight and the input vector. For the winning neuron the weight is then adjusted towards the input vector (Ahalt et al., 1990). More formally, the algorithm proceeds as follows: Let $(x_j)_{j\in\mathbb{N}} \subset \mathbb{R}^d$ be observations generated independently from the distribution P_X. The initial N-grid is defined as the first N input vectors with pairwise distinct

[6] Let for example Γ_N be an optimal N-grid for a two-dimensional random variable X following a multivariate normal distribution with expectation 0 and Σ being equal to the identical matrix. Rotating Γ_N around $(0, 0)$ by an arbitrary angle again yields an optimal N-grid for X.

entries. At iteration $t \in \mathbb{N}$ the algorithm is presented a new input vector x_t[7] and the current grid is updated according to

$$\xi_j^{(t+1)} := \begin{cases} \xi_j^{(t)} - \delta_t(\xi_j^{(t)} - x_t), & \text{if } |\xi_j^{(t)} - x_t| = \min_{1\leq i\leq N} |\xi_i^{(t)} - x_t| \\ \xi_j^{(t)}, & \text{otherwise,} \end{cases} \qquad \forall 1 \leq j \leq N, \quad (2.3)$$

where $|\cdot|$ denotes the Euclidean norm and $(\delta_t)_{t\in\mathbb{N}}$ the learning rate (Bouton and Pagès, 1997). At each point in time, only one prototype of the current grid is updated to better represent the observed data where the degree of adaptation is governed by the learning rate. Typically, the learning rate is reduced monotonously to zero during the learning process. The grids provided by the CLVQ algorithm are based on a finite sample of observations and are therefore most likely not optimal in the L_p sense. However, Pagès (1998) provides results for the convergence of the grids provided by the CLVQ algorithm to optimal grids with the number of iterations going to infinity.[8] We further refer to Charlier et al. (2015b) for a discussion of these results.

2.3 Conditional quantiles through leveraging optimal quantization

In this section we propose a new non-parametric estimator for conditional quantiles called *leveraging estimator*. Therefore we combine the concept of optimal quantization from the previous section with a machine learning approach called leveraging. Leveraging is an ensemble method very similar to boosting that "combine[s] simple 'rules' to form an ensemble such that the performance of the single ensemble member is improved" (Meir and Rätsch, 2003, p. 118). Following this approach, we update the proposed estimator in an iterative manner. Therefore, at each iteration step data

[7] For notational convenience, after making use of N input vectors for defining the initial grid we discard the remaining indexes and start counting at 1 again.

[8] To make use of convergence results for the CLVQ algorithm we assume that the learning rate δ_t is a $(0, 1)$-valued sequence satisfying $\sum_{t=1}^{\infty} \delta_t = \infty$ and $\sum_{t=1}^{\infty} \delta_t^2 < \infty$. A natural choice is $\delta_t = \frac{1}{t}$ (Bally et al., 2005).

weights are readjusted to give more weight to observations for which the current estimator produces a high estimation error, and less weight to observations associated with a low estimation error.[9]

In Section 2.3.2 we introduce an estimator based on the whole distribution of X. This is, however, infeasible in practice. Therefore, based on this infeasible estimator we derive an estimator for the finite sample case in Section 2.3.3. We start with some notation.

2.3.1 Notation

In order to introduce the leveraging estimator we first define the conditional quantile function $q_\alpha(\cdot)$. Let therefore (X, Y) denote a $(d+1)$-dimensional random variable on the probability space $(\Omega, \mathfrak{A}, P)$, where Y is a scalar response and X is a d-dimensional random vector of covariates. For $\alpha \in (0,1)$ and $x \in S_X$ with S_x denoting the support of X we set

$$q_\alpha(x) := \inf\{y \in \mathbb{R} |\ F(y|x) \geq \alpha\}, \tag{2.4}$$

where $F(\cdot|x)$ denotes the conditional cumulative distribution function of Y given $X = x$.

Throughout the section, we fix $\lambda \in (0,1), \gamma \in [\lambda, 1), p \geq 1$, $\alpha \in (0,1)$ and $N \in \mathbb{N}$ with $\lceil N \cdot \gamma \rceil < N$.[10] To ensure that all integrals are finite we assume $Y \in L_1$ and to guarantee the existence of an L_p-optimal N-grid we further assume that the Euclidean norm of X is in L_p (see Equation (2.1)) and that the distribution P_X of X is continuous (cf. Abaya and Wise, 1984). We denote the proposed estimator of the conditional quantile function with $\tilde{q}_\alpha$.

[9] Although leveraging and boosting are very similar concepts, we follow Duffy and Helmbold (1999) and restrict the usage of the term *boosting* to algorithms for which a so-called PAC-property can be proved to hold and use the term *leveraging* for all other related ensemble learning approaches. The concept of *probably approximately correct* (PAC) learning has been introduced by Valiant (1984) and guarantees, loosely speaking, that weak learners, each only performing slightly better than random, can indeed be aggregated to an arbitrarily good ensemble estimator (Kearns et al., 1994). However, as the concepts of boosting and leveraging are very similar to each other in the literature both terms are often used interchangeably.

[10] The effect of the hyperparameters N, λ, and γ as well as the data-driven selection of these parameters is discussed in detail in Section 2.4.

2.3.2 Infeasible estimator in the full sample case

Initialization

We start by defining the first estimator $\tilde{q}_\alpha^{(0)}$. Let therefore $\Gamma_N^{(0)} = \{\xi_1^{(0)}, \xi_2^{(0)}, \ldots, \xi_N^{(0)}\} \subset \mathbb{R}^d$ denote an optimal N-grid for X.[11] Let $\tilde{X}^{(0)}$ be defined as

$$\tilde{X}^{(0)} := Proj_{\Gamma_N^{(0)}}(X^{(0)}),$$

where $Proj_{\Gamma_N^{(0)}}$ denotes the projection of X on the N-grid $\Gamma_N^{(0)}$. That is, for $\omega \in \Omega$ we derive $\tilde{X}^{(0)}(\omega) \in \mathbb{R}^d$ from $X^{(0)}(\omega)$ by replacing the latter value with the $\xi_j \in \Gamma_N^{(0)}$ nearest to it in the Euclidean norm, see Equation (2.2). Consequently $\tilde{X}^{(0)}(\Omega) \subseteq \Gamma_N^{(0)}$ is fulfilled.

It is well known (see, e.g., the seminal paper by Koenker and Bassett, 1978) that for $x \in S_X$ the following identity holds:

$$q_\alpha(x) = \arg\min_{a \in \mathbb{R}} E(\rho_\alpha(Y - a) \Big| X = x), \tag{2.5}$$

where the check-function $\rho_\alpha : \mathbb{R} \to [0, \infty), z \mapsto \rho_\alpha(z)$ is given by

$$\rho_\alpha(z) := \begin{cases} -(1-\alpha)z & \text{for } z \leq 0, \\ \alpha z & \text{for } z > 0. \end{cases} \tag{2.6}$$

Building on this identity, we define the first base estimator $\tilde{q}_\alpha^{(0)}$ as

$$\tilde{q}_\alpha^{(0)}(x) = \tilde{q}_{\alpha,N}^{(0)}(x) := \arg\min_{a \in \mathbb{R}} E(\rho_\alpha(Y - a) \Big| \tilde{X}^{(0)} = \tilde{x}), \tag{2.7}$$

where $\tilde{x} \in \Gamma_N^{(0)}$ denotes the projection of $x \in \mathbb{R}^d$ on the N-grid $\Gamma_N^{(0)}$. Note that the base learner $\tilde{q}_\alpha^{(0)}$ is piecewise constant and the same as the (non bootstrapped) estimator by Charlier et al. (2015b).

[11] For keeping track of the current iteration step, we add a superscript in brackets to most of the variables and begin with the initialization as step 0. However, to make notation less heavy, we will sometimes drop this superscript as well as other possible sub- and superscripts when it is clear from context which variable is referenced.

The first iteration step

Following the leveraging approach, we proceed by iteratively learning new estimators and aggregating them to a stronger estimator while stage-wise reducing the value of an appropriate cost function. Typically, a leveraging algorithm readjusts the data weights at each stage in such a way that data examples associated with a higher approximation error in the previous step are given more weight while at the same time weights for data points associated with a lower approximation error are decreased. This forces future base learners to focus more on previously misclassified examples. As not all weak learners can be adapted to directly include weights, Freund and Shapire (1996) rely on resampling the data based on the weights for correctly and incorrectly classified examples. Inspired by this strategy, our approach translates the regression error associated with each Voronoi cell into new N-grids as described in the following.

For each Voronoi cell of the Voronoi tessellation associated with the optimal N-grid $\Gamma_N^{(0)} = \{\xi_1^{(0)}, \xi_2^{(0)}, \ldots, \xi_N^{(0)}\}$ we want to quantify the regression error of $\tilde{q}_\alpha^{(0)}$ inside that cell. Because we do not know the true conditional quantile function q_α, we cannot directly compare it with $\tilde{q}_\alpha^{(0)}$. Instead, we rely on the previously introduced check-function ρ_α to quantify the regression error inside each Voronoi cell. Therefore, we define

$$M_j = M_{j,\alpha}^{(1)} := E\big[\rho_\alpha(Y - \tilde{q}_\alpha^{(0)} \circ X)\big|\tilde{X}^{(0)} = \xi_j^{(0)}\big], \quad 1 \leq j \leq N. \tag{2.8}$$

We continue by choosing the λ-proportion of Voronoi cells with the highest regression error according to Equation (2.8) and define the index set $\mathcal{I}^{(1)} \subseteq \{1, \ldots, N\}$ by requiring that $\#\mathcal{I}^{(1)} = N_\lambda := \lceil N \cdot \lambda \rceil$ and $M_j \geq M_k$ for all $j \in \mathcal{I}^{(1)}$ and $k \in \{1, \ldots, N\} \setminus \mathcal{I}^{(1)}$. We now want the algorithm to focus more on the data in the Voronoi cells corresponding to $\mathcal{I}^{(1)}$ and reduce "weight" for the data corresponding to the remaining Voronoi cells. Instead of actually changing any weights, our "re-weighting scheme" follows a strategy similar to that in Freund and Shapire (1996). However, we do not resample more data from the Voronoi cells corresponding to $\mathcal{I}^{(1)}$. Instead, we introduce a new grid for approximating these data examples more accurately. Therefore, we introduce a new

grid based on more than the previously used $\#\mathcal{I}^{(1)}$ quantizers. The actual number of quantizers employed depends on the value of γ.

Formally, we define the set $A^{(1)}$ of data examples belonging to the Voronoi cells corresponding to $\mathcal{I}^{(1)}$ as $A^{(1)} := \bigcup_{j\in\mathcal{I}^{(1)}}\{\tilde{X}^{(0)} = \xi_j\} \in \mathfrak{A}$. Making use of the increased number of quantizers $N_\gamma := \lceil N \cdot \gamma \rceil \geq N_\lambda$, we introduce $\Gamma^{(1)}_{N_\gamma}$ as the optimal N_γ-grid for the restriction $X_{\big|A^{(1)}}$ of X to $A^{(1)}$. Analogously, based on the number of remaining quantizers $N - N_\gamma$, we define $\Gamma^{(1)}_{N-N_\gamma}$ as the optimal $(N - N_\gamma)$-grid for $X_{\big|\Omega\setminus A^{(1)}}$. That is, instead of calculating an optimal N-grid for all data examples, we calculate two separate grids: one for approximating the data from the Voronoi cells exhibiting high approximation errors and one for the remaining Voronoi cells associated with lower approximation errors. By using more quantizers for the first grid, we increase weight on the data with higher approximation errors and reduce weight for the remaining data. Combining these two grids yields $\Gamma^{(1)}_N := \Gamma^{(1)}_{N_\gamma} \cup \Gamma^{(1)}_{N-N_\gamma}$. The number of quantizers in the combined grid $\Gamma^{(1)}_N$ has not increased which is important to avoid overfitting. Note that the new grid is not necessarily L_p-optimal for X any longer.

Based on the N-grid $\Gamma^{(1)}_N$ we introduce the new base learner $\bar{q}^{(1)}_\alpha$ analogously to Equation (2.7) as

$$\bar{q}^{(1)}_\alpha(x) := \arg\min_{a\in\mathbb{R}} E[\rho_\alpha(Y - a)|\tilde{X}^{(1)} = \tilde{x}^{(1)}], \tag{2.9}$$

where $\tilde{X}^{(1)}$ respectively $\tilde{x}^{(1)}$ denotes the projection of X respectively $x \in \mathbb{R}^d$ on the current N-grid $\Gamma^{(1)}_N$. Following the leveraging approach, we aggregate the new base learner $\bar{q}^{(1)}_\alpha$ and the preceding estimator $\tilde{q}^{(0)}_\alpha$ to yield a stronger estimator $\tilde{q}^{(1)}_\alpha$. This is done by setting

$$\tilde{q}^{(1)}_\alpha := \beta^{(1)}_{\text{opt}} \cdot \bar{q}^{(1)}_\alpha + (1 - \beta^{(1)}_{\text{opt}}) \cdot \tilde{q}^{(0)}_\alpha$$

with $\beta^{(1)}_{\text{opt}} \in [0, 1]$, that is, $\tilde{q}^{(1)}_\alpha$ is defined as a convex combination of $\bar{q}^{(1)}_\alpha$ and $\tilde{q}^{(0)}_\alpha$. We choose $\beta^{(1)}_{\text{opt}} \in [0, 1]$ such that the check-loss $\int_\Omega \rho_\alpha(Y - (\beta \cdot \bar{q}^{(1)}_\alpha + (1 - \beta) \cdot \tilde{q}^{(0)}_\alpha) \circ X)dP$ is

minimized.[12] Consequently, we have

$$\int_\Omega \rho_\alpha(Y - \tilde{q}_\alpha^{(1)} \circ X)dP \leq \int_\Omega \rho_\alpha(Y - \tilde{q}_\alpha^{(0)} \circ X)dP. \tag{2.10}$$

That is, the overall error of the conditional quantile estimator $\tilde{q}_\alpha^{(0)}$ as measured by the right hand side of Inequality (2.10) is reduced and a stronger estimator $\tilde{q}_\alpha^{(1)}$ is formed. This finishes the first iteration step.

The following iteration steps

The next iteration step is performed analogously to the first iteration step. By replacing $\Gamma_N^{(1)}$ with $\Gamma_N^{(0)}$ and $\tilde{q}_\alpha^{(1)}$ with $\tilde{q}_\alpha^{(0)}$ we obtain a new N-grid $\Gamma_N^{(2)}$, base learner $\bar{q}_\alpha^{(2)}$, and $\beta_{\text{opt}}^{(2)} := \arg\min_{\beta\in[0,1]} \int_\Omega \rho_\alpha(Y - (\beta \cdot \bar{q}_\alpha^{(2)} + (1-\beta) \cdot \tilde{q}_\alpha^{(1)}) \circ X)dP$. This yields the further improved estimator

$$\tilde{q}_\alpha^{(2)} := \beta_{\text{opt}}^{(2)} \cdot \bar{q}_\alpha^{(2)} + (1 - \beta_{\text{opt}}^{(2)}) \cdot \tilde{q}_\alpha^{(1)}.$$

Proceeding in this manner, in the k'th iteration step we have a sequence of estimators $(\tilde{q}_\alpha^{(j)})_{j=0}^k$ as well as a sequence of check-losses

$$(I^{(j)})_{j=0}^k := \Big(\int_\Omega \rho_\alpha(Y - \tilde{q}_\alpha^{(j)} \circ X)dP \Big)_{j=0}^k \tag{2.11}$$

for which $I^{(0)} \geq I^{(1)} \geq I^{(2)} \cdots \geq I^{(k)} \geq 0$ holds by construction. That is, each iteration step reduces the overall error (measured by the check-function) associated with the conditional quantile estimator $\tilde{q}_\alpha^{(j)}$ of the previous step and provides a stronger estimator $\tilde{q}_\alpha^{(j+1)}$. We stop iterating when

$$\frac{I^{(j)}}{I^{(j+1)}} < 1 + tol$$

[12] As by construction, $\bar{q}_\alpha^{(1)}$ and $\tilde{q}_\alpha^{(0)}$ only assume finitely many values, it holds that $\sup_{x_1,x_2\in\mathbb{R}^d} \left|\bar{q}_\alpha^{(1)}(x_1) - \tilde{q}_\alpha^{(0)}(x_2)\right| < \infty$. Therefore, the function $[0,1] \to [0,\infty), \beta \mapsto \int_\Omega \rho_\alpha(Y - (\beta \cdot \bar{q}_\alpha^{(1)} + (1-\beta) \cdot \tilde{q}_\alpha^{(0)}) \circ X)dP$ is continuous in β and attains a minimum.

holds for a previously specified number s of successive j's and a tolerance $tol > 0$.[13] Let κ denote the final number of iteration steps.[14] We set

$$\tilde{q}_\alpha = \tilde{q}_{\alpha,N} := \tilde{q}_\alpha^{(\kappa)}. \tag{2.12}$$

Based on results obtained in Charlier et al. (2015b) one can show the following for arbitrary $p \geq 1$:

Theorem 1 *Fix $\alpha, \lambda, \gamma \in (0,1)$. Then under Assumptions A.3 and A.4 we have*

$$\sup_{x \in S_X} \left| \tilde{q}_{\alpha,N}(x) - q_\alpha(x) \right| \to 0 \text{ as } N \to \infty.$$

For the assumptions and a proof of the theorem see the Appendix.

Usually, we do not know the true distribution of X and Y which is why the present procedure is not feasible in practice. In the next section we therefore adapt the algorithm to the finite sample case.

2.3.3 The proposed estimator in the finite sample case

Let S be a sample of $n \in \mathbb{N}$ data points $(x_j, y_j)_{j=1}^n \subset \mathbb{R}^{d+1}$, where the (x_j, y_j) are generated independently at random from the joint distribution of (X, Y). As previously discussed, we focus on the CLVQ algorithm, that is $p = 2$, for determining empirical N-grids as this is the most common case and convergence results are much more satisfactory. However, generalizing our procedure to arbitrary $p \geq 1$ is straightforward.[15] By replacing some of the previous expressions with their sample counterparts,

[13] Note that e.g. for $\beta_{\text{opt}}^{(j+1)} = 0$ the ratio $\frac{I^{(j)}}{I^{(j+1)}}$ is equal to $1 < 1 + tol$ although in the next iteration step $\frac{I^{(j+1)}}{I^{(j+2)}} > 1 + tol$ is still possible. The algorithm should therefore only stop after the ratio has fallen below $1 + tol$ for several successive iteration steps. This motivates a choice of $s > 1$. In case $I^{(j)} = 0$ for some $j \in \mathbb{N}$, the iteration process is stopped, too.

[14] Note that the algorithm always comes to a halt after a finite number of iterations. This is due to the fact that because of Identity (2.5) we have $I \leq I^{(k)}$ with $I := \int_\Omega \rho_\alpha(Y - q_\alpha \circ X) dP$ and q_α denoting the true conditional quantile function. One can now show by complete induction that $I \leq I^{(k)} \leq \frac{I^{(0)}}{(1+tol)^{\lfloor k/s \rfloor}}$ is fulfilled as long as the stopping criterion is not met. As $Y \in L_1$ we have $I^{(0)} < \infty$. The case $I = 0$ is not relevant in our context because this would imply that Y is a deterministic function of X P_X-a.e.

[15] For more details on the stochastic gradient algorithm for determining empirical N-grids for arbitrary $p \geq 1$ see, e.g., Pagès (1998) and Charlier et al. (2015b).

we obtain the finite sample quantile estimator $\hat{q}_\alpha(x)$ of the conditional quantile function $q_\alpha(x)$.

Initialization

By applying the CLVQ algorithm from Section 2.2 to the sample $x_1, \ldots, x_n$, we obtain the initial N-grid $\hat{\Gamma}_N^{(0)} = \hat{\Gamma}_{N,n}^{(0)}$.[16] For $x \in \mathbb{R}^d$ we define

$$\hat{q}_\alpha^{(0)}(x) = \hat{q}_{\alpha,N,n}^{(0)}(x) := \arg\min_{a \in \mathbb{R}} \sum_{j=1}^{n} \rho_\alpha(y_j - a) \mathbb{I}_{\{\hat{x}_j\}}(\hat{x}),$$

where $\hat{x}_j := Proj_{\hat{\Gamma}_N^{(0)}}(x_j)$, $\hat{x} := Proj_{\hat{\Gamma}_N^{(0)}}(x)$, ρ_α denotes the previously introduced check-function, and $\mathbb{I}_{\{\hat{x}_j\}}(\cdot)$ is the indicator function of the set $\{\hat{x}_j\}$. In practice, we do not determine $\hat{q}_\alpha^{(0)}(x)$ as $\arg\min_{a \in \mathbb{R}}$ but instead we simply compute it as the sample quantile of the y_j's for which the $\hat{x}_j$'s equal $\hat{x}$. That is, the initial estimator $\hat{q}_\alpha^{(0)}$ is constant in each Voronoi cell from the Voronoi tessellation corresponding to $\hat{\Gamma}_N^{(0)}$.

Iteration

Let $\hat{\Gamma}_N^{(0)}$ be given by $\hat{\Gamma}_N^{(0)} = \{\hat{\xi}_1^{(0)}, \hat{\xi}_2^{(0)}, \ldots, \hat{\xi}_N^{(0)}\}$. Analogously to Equation (2.8) we quantify the regression error in each Voronoi cell via

$$\hat{M}_j = \hat{M}_{j,\alpha}^{(1)} := \frac{1}{\sum_{i=1}^{n} \mathbb{I}_{\{\hat{\xi}_j^{(0)}\}}(\hat{x}_i)} \sum_{i=1}^{n} \rho_\alpha(y_i - \hat{q}_\alpha^{(0)}(x_i)) \mathbb{I}_{\{\hat{\xi}_j^{(0)}\}}(\hat{x}_i),$$

where the notation is as above. As $\mathbb{I}_{\{\hat{\xi}_j^{(0)}\}}(\hat{x}_i) = 1$ if $\hat{x}_i = \hat{\xi}_j^{(0)}$ and 0 otherwise, $\hat{M}_j$ is simply the average regression error measured by the check-function for the observations belonging to the Voronoi cell corresponding to the j'th quantizer $\hat{\xi}_j^{(0)}$. Based on the $\hat{M}_j$'s we identify the Voronoi cells with the highest regression error and choose $\hat{\mathcal{I}}^{(1)} \subseteq \{1, \ldots, N\}$ such that $\#\hat{\mathcal{I}}^{(1)} = N_\lambda := \lceil N \cdot \lambda \rceil$ and $\hat{M}_j \geq \hat{M}_k$ for all $j \in \hat{\mathcal{I}}^{(1)}$ and $k \in \{1, \ldots, N\} \setminus \hat{\mathcal{I}}^{(1)}$.

We now compute a new grid $\hat{\Gamma}_{N_\gamma}^{(1)}$ by applying the CLVQ algorithm to the data points

[16] Again, for notational convenience we will sometimes drop the superscript determining the current iteration step when it is clear from the context which variable is referenced.

belonging to the set $\{x \in \{x_1, \ldots, x_n\} \big| Proj_{\hat{\Gamma}_N^{(0)}}(x) \in \{\hat{\xi}_j |\ j \in \hat{I}^{(1)}\}\}$ containing all observations inside the Voronoi cells from the index set $\hat{I}^{(1)}$. As we have pointed out previously, by approximating data examples associated with higher regression errors with more quantizers ($N_\gamma \geq N_\lambda$) we increase weight on these data examples. Inversely, we reduce weight for the remaining data points by computing the $(N - N_\gamma)$-grid $\hat{\Gamma}_{N-N_\gamma}^{(1)}$ for the data points corresponding to the Voronoi cells associated with a lower regression error. We set

$$\hat{\Gamma}_N^{(1)} := \hat{\Gamma}_{N_\gamma}^{(1)} \cup \hat{\Gamma}_{N-N_\gamma}^{(1)}.$$

This leads to the estimator

$$\bar{q}_\alpha^{(1)}(x) = \bar{q}_{\alpha,N,n}^{(1)}(x) := \arg\min_{a \in \mathbb{R}} \sum_{j=1}^{n} \rho_\alpha(y_j - a) \mathbb{I}_{\{\hat{x}_j\}}(\hat{x}),$$

where $\hat{x}_j := Proj_{\hat{\Gamma}_N^{(1)}}(x_j)$, $\hat{x} := Proj_{\hat{\Gamma}_N^{(1)}}(x)$, and the rest of the notation is as above. We aggregate the new base learner $\bar{q}_\alpha^{(1)}$ and the preceding estimator $\hat{q}_\alpha^{(0)}$ to form a stronger estimator $\hat{q}_\alpha^{(1)}$. We do so by setting

$$\hat{q}_\alpha^{(1)} := \beta_{\text{opt}}^{(1)} \cdot \bar{q}_\alpha^{(1)} + (1 - \beta_{\text{opt}}^{(1)}) \cdot \hat{q}_\alpha^{(0)}$$

with $\beta_{\text{opt}}^{(1)} := \arg\min_{\beta \in [0,1]} \sum_{j=1}^{n} \rho_\alpha(y_j - (\beta \cdot \bar{q}_\alpha^{(1)}(x_j) + (1 - \beta) \cdot \hat{q}_\alpha^{(0)}(x_j)))$. That is, we introduce $\hat{q}_\alpha^{(1)}$ as a convex combination of $\bar{q}_\alpha^{(1)}$ and $\hat{q}_\alpha^{(0)}$ that minimizes the empirical check-loss.[17] We define the average overall error associated with $\hat{q}_\alpha^{(1)}$ as $\hat{I}^{(1)} := \frac{1}{n} \sum_{j=1}^{n} \rho_\alpha(y_j - \hat{q}_\alpha^{(1)}(x_j))$.

The following iteration steps

The next iteration step is performed analogously to the first one. By using $\hat{\Gamma}_N^{(1)}$ instead of $\hat{\Gamma}_N^{(0)}$ and $\hat{q}_\alpha^{(1)}$ instead of $\hat{q}_\alpha^{(0)}$ we obtain a new N-grid $\hat{\Gamma}_N^{(2)}$, base learner $\bar{q}_\alpha^{(2)}$, parameter $\beta_{\text{opt}}^{(2)}$, and finally the new estimator $\hat{q}_\alpha^{(2)} := \beta_{\text{opt}}^{(2)} \cdot \bar{q}_\alpha^{(2)} + (1 - \beta_{\text{opt}}^{(2)}) \cdot \hat{q}_\alpha^{(1)}$. Proceeding in this

[17] In practice we approximate β_{opt} via grid search. That is, we search exhaustively through a finite, manually specified set of reasonable values for β. Note that the exact choice of β does not have an influence on the consistency results.

manner, in the k'th iteration step we obtain a sequence of estimators $(\hat{q}_\alpha^{(j)})_{j=0}^k$ as well as a sequence of empirical check-losses $(\hat{I}^{(j)})_{j=0}^k := \left(\frac{1}{n}\sum_{i=1}^n \rho_\alpha(y_i - \hat{q}_\alpha^{(j)}(x_i))\right)_{j=0}^k$ for which $\hat{I}^{(0)} \geq \hat{I}^{(1)} \geq \hat{I}^{(2)} \cdots \geq \hat{I}^{(k)} \geq 0$ holds by construction. That is, in each iteration step we further reduce the check-loss of the conditional quantile estimator $\hat{q}_\alpha^{(j)}$ of the previous stage and form a stronger estimator $\hat{q}_\alpha^{(j+1)}$. We stop iterating when

$$\frac{I^{(j)}}{I^{(j+1)}} < 1 + tol$$

is fulfilled for a number s of successive j's and a tolerance $tol > 0$, or when a previously specified maximum number of iteration steps is reached.[18] For κ denoting the final number of iteration steps, we set

$$\hat{q}_\alpha = \hat{q}_{\alpha,N,n} := \hat{q}_\alpha^{(\kappa)}. \tag{2.13}$$

One can now show the following:

Theorem 2 *Fix $\alpha, \lambda, \gamma \in (0,1)$, and $x \in S_X$. Given that the grids are obtained in the quadratic case ($p = 2$), we have under Assumptions A.3, A.4, A.8, and A.9:*

$$p - \lim_{N\to\infty} \; p - \lim_{n\to\infty} |\hat{q}_{\alpha,N,n}(x) - q_\alpha(x)| = 0.$$

For the assumptions and a proof of the theorem see the Appendix. For empirical results concerning the convergence of the estimator on simulated datasets we refer to Section 2.5.2.

[18] Throughout the paper, we choose 30 as the maximum number of iterations and $s = 5$. This offers a good compromise between the accuracy and smoothness of the quantile estimates and the runtime of the algorithm. Note that this choice has no influence on the convergence results. For completeness, we further note that the estimation procedure is also stopped in case $I^{(j)} = 0$ for some $j \in \mathbb{N}$.

2.4 The hyperparameters N, λ, and γ

2.4.1 Effect of the hyperparameters on the estimates

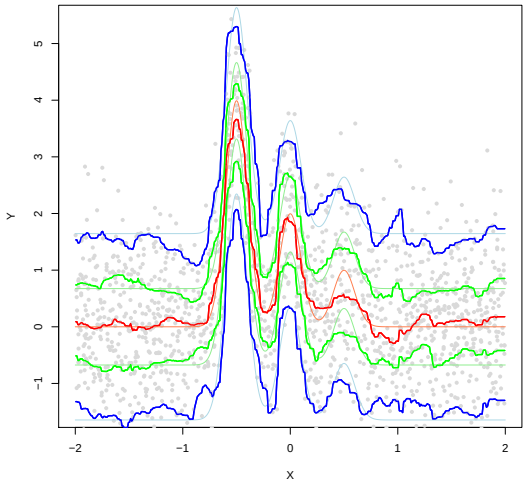
The quantile estimates rely crucially on the choice of the hyperparameters N, λ, and γ. These parameters are set before training and control the learning process. The parameter N has an influence on how well the estimator adapts to the distribution of X, whereas λ and γ determine how well the estimator adapts to changes in the conditional distribution of Y.

In the three panels of Figure 2.1 we provide quantile estimates for $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$ based on 1500 data points sampled randomly according to model $\mathcal{M}_3$ (see Section 2.5.2). All differences between the three panels are due to the choice of hyperparameters. In Panels A and B we fix $\lambda = \gamma = 0.5$ and vary the number of quantizers ($N = 30$ and $N = 15$, respectively). There is a bias-variance trade-off in the choice of the parameter N. While the third bell-shaped curve from model $\mathcal{M}_3$ is captured in Panel A, overfitting is evident especially in the edges of the interval $[-2, 2]$. While overfitting issues can be mitigated by reducing the number of quantizers to 15 in Panel B, in particular the third bell-shaped curve from model $\mathcal{M}_3$ is no longer captured as variability is reduced. Thus, N mainly behaves as a smoothing parameter.

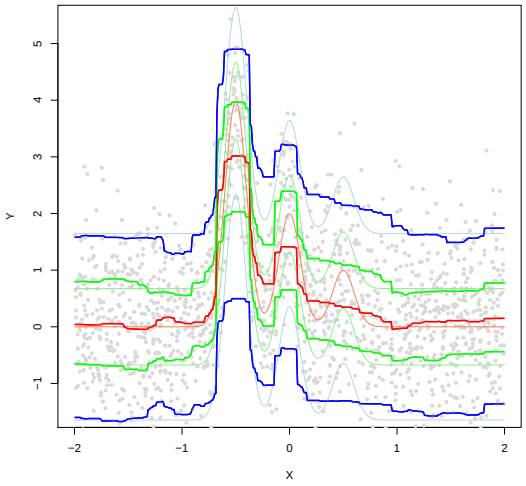
Because the leveraging estimator is based on a quantization of X, the estimator yields a good adaptation to the distribution of X. However, quantizing with respect to X does not take into account the conditional distribution of Y and especially not the variability of the associated conditional quantile curves. As one can clearly see in Figure 2.1, variability of the quantile curves is high in the middle of the interval $[-2, 2]$ and low in the edges. Capturing these effects requires more quantizers in the middle of $[-2, 2]$ and fewer in the edges. As X is uniformly distributed in $[-2, 2]$, so are the quantizers (approximately). In Panels 2.1(a) and 2.1(b) we set $\lambda = \gamma$ and consequently did not increase weight for data examples associated with a higher regression error. Thus,

Figure 2.1: Estimated conditional quantile curves for different values of the hyperparameters N, λ, and γ

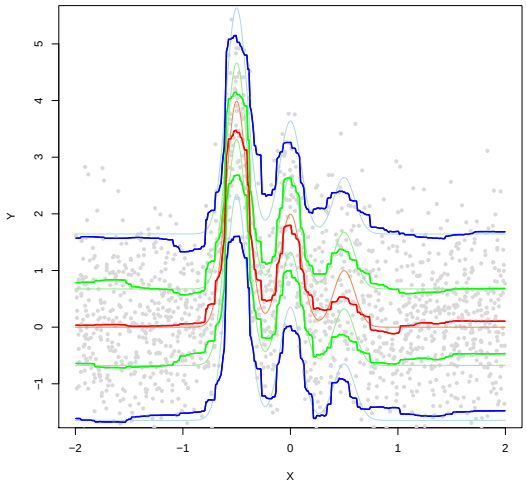
This figure shows quantile curves estimated by the proposed leveraging estimator for quantile levels $\alpha = 0.05$ (blue), 0.25 (green), 0.5 (red), 0.75 (green), and 0.95 (blue). The lighter quantile curves correspond to the population ones. The data examples ($n = 1500$) are generated according to model $\mathcal{M}_3$ (see Section 2.5.2). In Panel 2.1(a) estimation is performed with the number of quantizers N being equal to 30 and $\lambda = \gamma = 0.5$, in Panel 2.1(b) estimation is realized with $N = 15$ and again $\lambda = \gamma = 0.5$. In the last Panel 2.1(c) however, learning is performed based on 15 quantizers and $\lambda = 0.3, \gamma = 0.5$. That is, in each iteration step data examples associated with a higher estimation error gain more weight in the next iteration step.



(a) $N = 30,\ \lambda = \gamma = 0.5$



(b) $N = 15,\ \lambda = \gamma = 0.5$



(c) $N = 15,\ \lambda = 0.3, \gamma = 0.5$

there is no adaptation to the conditional distribution of Y and especially not to areas of higher or lower variability of the quantile curves. In this case, our iterative procedure works less like a leveraging and more like a bagging approach (Breiman, 1996).

In contrast to the two previous panels, in Panel 2.1(c) we allow λ to be different from γ. Specifically, we set $\lambda = 0.3$ and $\lambda = 0.5$ (and $N = 15$). That is, in each iteration step we identify the $5 = \lceil N \cdot \lambda \rceil$ Voronoi cells with the highest regression error[19] and approximate the data points associated with these Voronoi cells with $8 = \lceil N \cdot \gamma \rceil$ quantizers in the next iteration step. Analogously, the data examples corresponding to the remaining 10 quantizers associated with a lower approximation error are approximated with only 7 quantizers in the next iteration step. As desired, data examples in areas with a higher variability of the conditional quantile function are approximated with more quantizers while we use less quantizers in the edges of the interval $[-2, 2]$. This leads to a better approximation of the true quantile curves. Especially, we are able to capture the third bell-shaped curve from model $\mathcal{M}_3$ while at the same time reducing variability of our estimator in areas where the true quantile curves are constant. The next section is concerned with the data-driven choice of the hyperparameters.

2.4.2 Data-driven hyperparameter selection

The choice of the hyperparameters is critical for the estimation process. One usually determines the parameters in such a way that some error function is minimized. Naturally, we would like to minimize $\int_{S_X} e(q_\alpha(x) - \hat{q}_\alpha(x))dP_X(x)$, where $e(\cdot)$ denotes an appropriate error function. As we normally do not know the true quantile curve $q_\alpha(\cdot)$, this approach is not feasible in practice. Instead, we make use of Identity (2.5), stating that for all $x \in S_x$ we have $q_\alpha(x) = \arg\min_{a \in \mathbb{R}} E(\rho_\alpha(Y - a)|X = x))$ and consequently

$$\int_\Omega \rho_\alpha(Y - \hat{q}_\alpha \circ X)dP \geq \int_\Omega \rho_\alpha(Y - q_\alpha \circ X)dP. \tag{2.14}$$

[19] See Section 2.3.3 for how the regression error is calculated.

In the case of risk measures like the quantile based Value-at-Risk this translates into the more general concept of elicitability.[20] It is therefore a natural choice to try to minimize the left handside of Inequality (2.14), or more specifically the corresponding sample expression

$$\frac{1}{n}\sum_{j=1}^{n}\rho_\alpha(y_j - \hat{q}_\alpha(x_j)), \tag{2.15}$$

where $\rho_\alpha(\cdot)$ denotes the check-function and $(x_j, y_j)_{j=1}^n$ is as before.

Simply selecting hyperparameters that minimize the quantity in (2.15) would lead to serious overfitting issues as the minimum would be reached when the number of quantizers N is equal to the size of the sample n. In this case we would simply obtain $\hat{q}_\alpha(x_j) = y_j,\ j = 1, \ldots, n$. To mitigate the overfitting problem we rely on cross-validation, a very popular method for parameter selection and assessment of the generalization performance for a great variety algorithms (Picard and Cook, 1984, Zhang, 1993, Yang, 2007) which is by now the standard in the literature (Hastie et al., 2017). The key idea is to partition the data into complementary subsets and perform training in one (the training set) and calculation of the estimation error in another subset (the validation set). In our case, this implies computing the estimator $\hat{q}_\alpha$ based on data from the training set and then calculating the check-loss as defined in (2.15) by using only the data points from the validation set. To consider more than one possible partition of the dataset and to use each data point exactly once in the validation step, we rely on k-fold cross-validation, where k is usually set to 5 or 10 (Breiman and Spector, 1992, Hastie et al., 2017). To perform k-fold cross-validation, one randomly partitions the original data into k equal sized complementary subsamples. Each of the k subsamples then serves once as the validation set whereas the remaining $k-1$ subsamples are used as the training set.

The check-loss is calculated for each of the validation sets, leading to k error estimates which are subsequently averaged to a single error estimate, say $\hat{E}(N, \lambda, \gamma)$. This

[20] Generally speaking, a risk measure is elicitable if it minimizes the expected value of a so-called scoring function, see Gneiting (2011) for a comprehensive literature review on elicitability as well as Frongillo and Ian A. Kash (2015), Fissler et al. (2016), Ziegel (2016), Nolde and Ziegel (2017) for more recent advances in the field.

procedure ensures that we make use of each data point for both training and validation and that training and validation sets are independent of each other. A special case is leave-one-out cross-validation where k is set to n. However, this approach is not feasible in our setting because of the computational burden related to each training and validation step. We therefore choose $k = 5$ throughout the paper.[21]

At the core of our data-driven hyperparameter selection procedure, we try to find a tuple $(N_{\text{opt}}, \lambda_{\text{opt}}, \gamma_{\text{opt}})$ such that the empirical error $\hat{E}(N_{\text{opt}}, \lambda_{\text{opt}}, \gamma_{\text{opt}})$ is minimized. However, we do not minimize the error over all possible combinations of N, λ, and γ, as this would be computationally infeasible and the regression error is not very sensitive to small variations in the parameters. Instead, we perform a grid search, that is, we only consider a finite subset of reasonable parameter combinations.[22] Extensive pre-tests suggest that the selection of the parameter N can be performed separately from (λ, γ), increasing the computational efficiency. This is illustrated in Figure 2.2. Throughout the paper, we therefore first calculate N based on $\lambda = \gamma = 0.5$ and subsequently determine the optimal values for λ and γ.

Furthermore, we slightly adapt the estimation procedure from Section 2.3.3 to avoid the quantile crossing problem (cf. Bassett and Koenker, 1982). One possibility to deal with this problem is the rearrangement procedure introduced by Chernozhukov et al. (2010). However, we propose a different solution to this problem, which arises naturally from our estimation procedure. Instead of performing the iterative procedure separately for each confidence level α, at each iteration step we jointly calculate the quantile estimates based on the same N-grid. As the conditional quantiles at each iter-

[21] In each iteration step, the quantile estimates inside a particular Voronoi cell are based on the empirical quantiles of the y_j's corresponding to the cell. Therefore, the average number of observations per quantizer, that is the ratio n/N has an influence on the range of the quantile estimates, especially for small values of this ratio. This is because when calculating sample quantiles the largest (smallest) value is usually considered to correspond to the 100 % (0 %) quantile. We therefore increase the training sample consisting of only 80 % of the original data points ($k = 5$) to 100 % by resampling from the training sample. Note that training and validation set stay independent of each other.

[22] Based on pre-tests, we consider the values $\lambda, \gamma \in \{0.3, 0.35, 0.4, 0.45, 0.5\}$ throughout all calculations in the paper. Because our leveraging approach is based on giving data examples that have performed poorly in the previous iteration more weight in the next one, we only consider tuples (λ, γ) with $\lambda \leq \gamma$. Reasonable values for N are chosen depending on the specific model and the size of the random sample. For more details we refer to Section 2.5.

ation step are essentially calculated as empirical quantiles inside a particular Voronoi cell, this ensures monotonicity of the quantile estimates in α. Furthermore, we choose β_{opt} such that the average check-loss over all values of α is minimized. Analogously, we calculate the errors $\hat{M}_j$ associated with each Voronoi cell of the current grid as an average over the various confidence levels. This procedure guarantees that the quantile crossing problem is avoided. Additionally, this approach provides the possibility to calculate quantile curves for multiple values of α with nearly no increase in the computational effort. To stay consistent with this approach, the hyperparameters N, λ, and γ have to be chosen the same for all confidence levels α. This is achieved by determining the parameters based on the average of the cross-validation errors over the confidence levels. Figure 2.3 illustrates that one obtains very similar estimates for N when considering each α level separately and when considering all levels jointly. The figure further demonstrates that $\alpha = 0.5$ contributes stronger to the average check error than, e.g., $\alpha = 0.05$, that is, the conditional median estimate has a stronger influence on the parameter selection than the 5 % conditional quantile.[23]

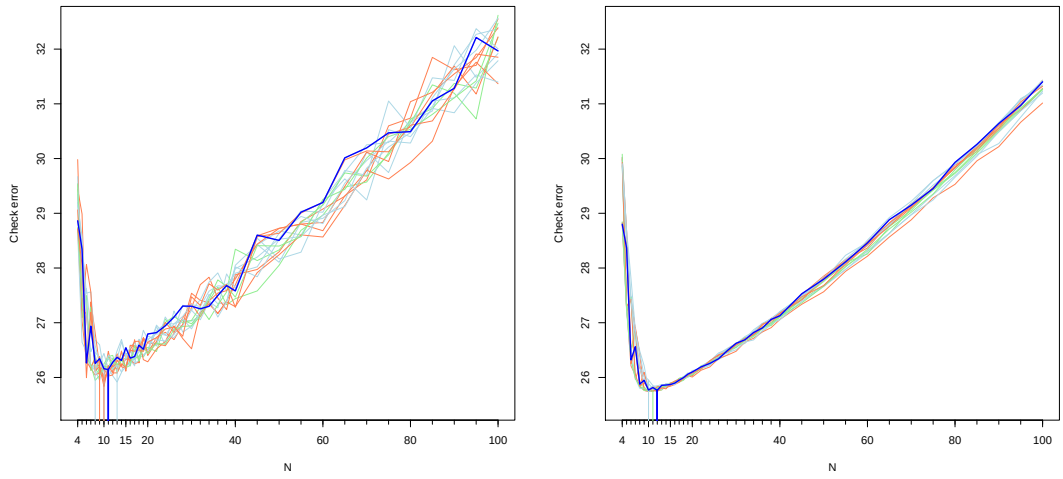
2.5 Simulation study

In this section we compare the proposed leveraging estimator with some competitors in the field of conditional quantile estimation. Thereby we aim to give a more complete picture on the strengths and weaknesses of the leveraging estimator and to provide more detailed information on how to employ the algorithm. Furthermore, we investigate the behavior of the algorithm both when the dimension of the covariates and when the number of observations increases. In one dimension, we compare the shapes of the estimated quantile curves to each other as in practice it is often favorable to obtain smooth curves. The analysis of the regression accuracy is based on the integrated

[23] In various pre-tests we observed that the median has a smoothing effect on the parameter selection and helps to mitigate overfitting issues. Of course, one could re-weight every quantile level to guarantee an equal contribution to the average check-error. However, a further analysis of this aspect is beyond the scope of this paper.

Figure 2.2: Selection of the optimal number of quantizers N depending on λ and γ

These panels illustrate the data-driven choice of the hyperparameters for the proposed leveraging estimator, see Section 2.4.2. In Panel 2.2(a), the optimal number of quantizers is determined separately for each tuple (λ, γ) with $\lambda, \gamma \in \{0.3, 0.35, 0.4, 0.45, 0.5\}$ and $\lambda \leq \gamma$. The optimal value of N for each of the tuples is indicated by a vertical line. The dark blue line corresponds to the error associated to $\lambda = \gamma = 0.5$ as proposed in Section 2.4.2. The calculations are performed based on a random sample of size $n = 500$ generated according to model $\mathcal{M}_1$. In the two panels, the check error denotes the sum of check-losses on a grid of uniformly distributed points in the interval $[-2, 2]$ averaged over all values of $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$. We consider $N = 4, 5, 6, ..., 19, 20, 22, 24, ..., 38, 40, 45, 50, ..., 95, 100$ as possible values for the number of quantizers. Panel 2.2(b) presents the same results as Panel 2.2(a) but as an average over 50 random samples.



(a) one random sample (b) average over 50 random samples

squared error (ISE) defined as

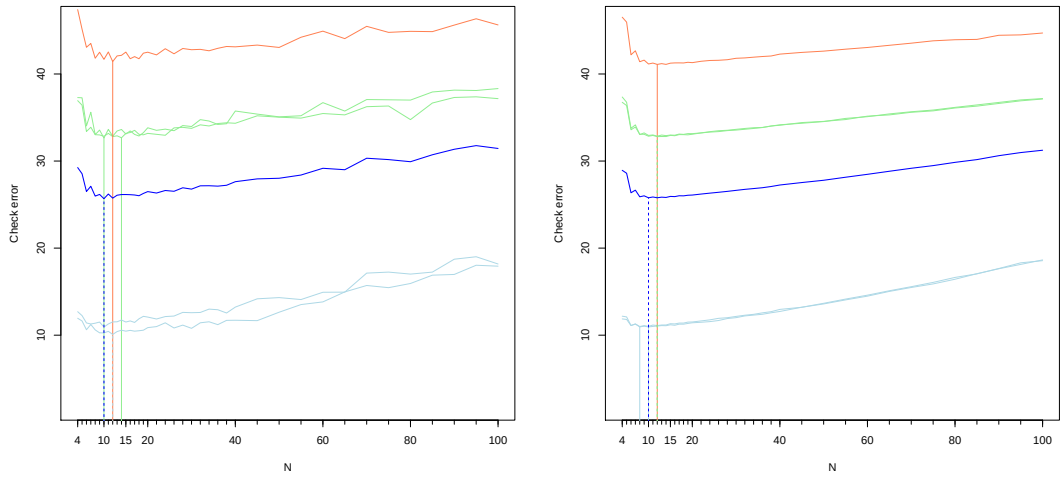
$$ISE := \frac{1}{M} \sum_{m=1}^{M} \left(\hat{q}_\alpha(\xi_m) - q_\alpha(\xi_m)\right)^2, \tag{2.16}$$

where $\hat{q}_\alpha$ denotes an estimator for the conditional α-quantile function, q_α is the true conditional α-quantile function, and $\xi_1, \ldots, \xi_M$ are equi-spaced points in the interval $[-2, 2]$ and $[-3, 3]$, respectively (in the one-dimensional case). In the multi-dimensional case of the simulation study, $\xi_1, \ldots, \xi_M$ are given as equi-spaced points in the hyper-cube $[-2, 2]^d$, where $d = 2, 3, 4$ denotes the dimension.[24] Note that this approach of measuring the regression error is not feasible in practice as the true conditional quantile functions are usually unknown.

[24] In the one-dimensional case we choose $M = 400$ for models $\mathcal{M}_1$ and $\mathcal{M}_3$ and $M = 600$ for model $\mathcal{M}_2$. In the multi-dimensional case we set $M = 20^d$.

Figure 2.3: Selection of the optimal number of quantizers N depending on α

The two panels illustrate the data-driven choice of the hyperparameters for the proposed leveraging estimator for different values of α, see Section 2.4.2. In Panel 2.3(a), the optimal number of quantizers is determined separately for each level of $\alpha = 0.05$ (blue), 0.25 (green), 0.5 (red), 0.75 (green), and 0.95 (blue). The optimal value for each of these confidence levels is indicated by a vertical line. The dark blue line corresponds to the average error over all considered values of α. The calculations are performed based on a random sample of size $n = 500$ generated according to model $\mathcal{M}_1$. In the two panels, the check error denotes the sum of check-losses on a grid of uniformly distributed points in the interval $[-2, 2]$. All calculations are performed based on $\lambda = \gamma = 0.5$. We consider $N = 4, 5, 6, ..., 19, 20, 22, 24, ..., 38, 40, 45, 50, ..., 95, 100$ as possible values for the number of quantizers. Panel 2.3(b) shows the same results as Panel 2.3(a) but as an average over 50 random samples.



(a) one random sample

(b) average over 50 random samples

We employ the olvq1 method from the popular *class* R-package by Venables and Ripley (2002) for an implementation of the CLVQ algorithm.[25] For an efficient assignment of observations to the quantizers of a respective grid we rely on k-d trees implemented in the get.knnx function from the *FNN* R-package by Beygelzimer et al. (2019).

2.5.1 The competitors considered

The proposed leveraging estimator relies on the *quantization estimator* by Charlier et al. (2015b) as base learner. In Charlier et al. (2015a), the authors compare their quantization estimator to some classical competitors. To demonstrate the usefulness

[25]The optimized-learning-rate LVQ1 algorithm (OLVQ1) by Kohonen (1992) is actually a supervised classification algorithm. However, performing this algorithm for only one class yields the CLVQ algorithm from Section 2.2. We also refer to Kohonen (2001) for further information on the algorithm.

of our leveraging approach we therefore include the quantization estimator along with some estimators from this study, namely the *smoothing splines estimator* by Koenker et al. (1994) and the *kNN estimator* by Bhattacharya and Gangopadhyay (1990).[26] We further consider another estimator that is based on ensemble learning, the *xgboost estimator*. The xgboost estimator is constructed based on the very popular and versatile Extreme Gradient Boosting (XGBoost) algorithm due to Chen and Guestrin (2016). As boosting and leveraging are very similar techniques, including the xgboost estimator allows for a more comprehensive comparison. In the following we shortly introduce the estimators. Let therefore $\alpha \in (0,1)$ denote the quantile level of interest and $(x_1, y_1), \ldots, (x_n, y_n)$ realizations chosen independently from the joint distribution of (X, Y).

The quantization estimator $\hat{q}_\alpha^{quant}(x)$ is defined as

$$\hat{q}_\alpha^{quant}(x) := \arg\min_{a \in \mathbb{R}} \sum_{j=1}^{n} \rho_\alpha(y_j - a) \cdot \mathbb{I}_{\{\hat{x}\}}(\hat{x}_j),$$

where $\hat{x}$ and $\hat{x}_j$ denote the projection of x and x_j on an (approximately) optimal L^2-grid obtained by the CLVQ algorithm, respectively, $\rho_\alpha(\cdot)$ is the check-function from Equation (2.6), and $\mathbb{I}_{\{\hat{x}\}}(\cdot)$ denotes the indicator function of the set $\{\hat{x}\}$. Note that this is the same estimator as the base learner $\tilde{q}_\alpha^{(0)}$ obtained in the initialization step, see Section 2.3.3. To further improve the estimator, Charlier et al. (2015b) suggest to use a bootstrapped version of the quantization estimator obtained by sampling B times with replacement from the original sample and calculating B different optimal grids. The bootstrapped estimator is then given as the average over the estimators obtained for each of the B grids. For determining the optimal number N of quantizers Charlier et al. (2015a) propose a method that is based on minimizing the sample equivalent of the ISE. In the sequel, $\hat{q}_\alpha^{quant}$ will denote the bootstrapped version of the estimator ($B = 50$) along with this parameter selection procedure. As the authors explicitly

[26] We do not consider the kernel estimators (local linear, local constant) by Yu and Jones (1998) in our simulation study as the results in Charlier et al. (2015a) suggest that these estimators are inferior to the before mentioned ones.

suggest the estimator to be used for multivariate covariates, we consider this estimator in the multivariate part of our simulation study, too.

The smoothing splines estimator $\hat{q}_\alpha^{spline}$ by Koenker et al. (1994) is defined as

$$\hat{q}_\alpha^{spline} := \underset{g \in \mathcal{G}}{\arg\min} \sum_{j=1}^{n} \rho_\alpha(y_j - g(x_j)) + \lambda \Big(\int_0^1 |g''(x)|^p dx \Big)^{1/p} \tag{2.17}$$

with $\lambda \geq 0$, $p \geq 1$, and $\mathcal{G}$ denoting an appropriate functional space. There is a considerable scope for the form of the roughness penalty. However, as in the original paper the authors focus on the total variation of the first derivative of g, we also choose $p = 1$ in the penalty term. For this penalty and an appropriate choice $\mathcal{G}$ the authors conclude that the estimator is piecewise linear with breakpoints $x_1, \ldots, x_n$. As λ weights the roughness penalty term, it works as a smoothing parameter. In the literature, there exist several methods for choosing an optimal value. In the sequel, we will determine λ based on the AIC criterion, that is minimizing

$$AIC(\lambda) = \ln\Big(\frac{1}{n} \sum_{j=1}^{n} \rho_\alpha(y_j - \hat{g}_\lambda(x_j))\Big) + \frac{p_\lambda}{n},$$

where $\hat{g}_\lambda$ is the arg min from Equation (2.17) for a given λ and p_λ denotes the effective dimension of $\hat{g}_\lambda$, see Koenker and Mizera (2004) for details.[27] Koenker and Mizera (2004) generalize the estimator to the bivariate case by introducing triogram-based splines. However, this estimator does not easily extend to dimensions greater than 2. Therefore, we do not consider the spline based estimator in the multivariate part of the simulation study.

The next competitor is the kNN estimator $\hat{q}_\alpha^{kNN}$, where $\hat{q}_\alpha^{kNN}(x) = \hat{q}_{\alpha,k}^{kNN}(x)$ is defined as the empirical α-quantile of the y_j's corresponding to the k x_j's that are closest to x in Euclidean distance. For more information on the estimator as well as for convergence results we refer to Bhattacharya and Gangopadhyay (1990). The parameter k works as a smoothing parameter with higher values of k leading to smoother estimates and

[27] We implement this estimator based on the *rqss* function from the *quantreg* R-package by Koenker (2020). The optimal value of λ is determined separately for each considered confidence level.

lower k values being prone to overfitting. Like for the leveraging estimator, we rely on 5-fold cross-validation to determine the optimal value of k in a data-driven manner. More exactly, for fixed k and each data point (x_j, y_j) in the validation set, we calculate $\hat{q}_\alpha^{kNN}(x_j)$ based on data from the test set only. We then calculate the regression error based on the check-function and choose the k, for which the average error over all points from the validation set and rounds of the 5-fold cross-validation is minimal.[28]

Finally, we introduce the *xgboost estimator*. The inclusion of the estimator in the simulation study is motivated by the fact that the proposed estimator is based on the ensemble technique leveraging, an approach very similar to the concept of boosting. Both techniques share the same idea of combining simple estimators to an ensemble that improves the performance of each ensemble member (Meir and Rätsch, 2003). A very popular type of boosting is *gradient boosting* which is based on the view of boosting as an optimization algorithm that minimizes a suitable cost function (cf. Breiman, 1998). In this manner, Zheng (2012) estimates conditional quantile functions by using the check-function as the cost function to be minimized. More detailed, the author performs gradient descent over the space of linear functions thereby obtaining a linear estimator. For this reason, we cannot directly build on this approach as in our simulation study the quantile curves are highly non linear.

Instead, we rely on regression trees as weak learners in the gradient boosting framework (Friedman et al., 2000, Friedman, 2001) in the form of the XGBoost algorithm, a very powerful algorithm that has proved successful in many applied machine learning and kaggle competitions. For the implementation of the $\hat{q}_\alpha^{xgboost}$ estimator we rely on the *xgboost* R-package by Chen et al. (2020b). The package offers the possibility to hand over user-defined cost functions. As the check-function is not differentiable in 0, we replace it by a smoothed version.[29] Apart from that, we use the default parameter

[28] In our simulation study, we consider multiple confidence levels α. To avoid quantile crossing, we therefore choose the optimal k such that the average cross-validation error over all confidence levels is minimized. This has the additional advantage that for each x we only have to determine the k nearest neighbors once and can calculate the conditional quantiles simultaneously for all confidence levels. This raises the computational efficiency of the algorithm and makes it one of the fastest among the considered competitors.

[29] The check-function from Equation (2.6) can be rewritten as $\rho_\alpha(z) = z\mathbb{I}_{(0,\infty)}(z) - (1-\alpha)z$. We follow

values of the package except for the parameter γ that we determine based on the 5-fold cross-validation check-error value. The parameter γ works as a smoothing parameter aiming to prevent overfitting. More exactly, γ is the minimum reduction in the loss-function required to make a further partition on a leaf node of the tree.[30] Although for applying this algorithm to multivariate covariates only small changes have to be made, we do not report results of this algorithm in the multivariate setting because of the poor performance we observed in pre-tests.

Table 2.1: Parameters considered for the various estimators

This table summarizes the considered values in the parameter selection procedures of the algorithms introduced in Sections 2.4.2 and 2.5.1. For the leveraging estimator we determine the number of quantizers N as well as the parameters λ and γ. For the quantization estimator one has to choose the number of quantizers, for the kNN estimator the number of neighbors, for the smoothing splines estimator the smoothing parameter λ, and for the xgboost estimator the parameter γ, see Sections 2.4.2 and 2.5.1 for details. We determined the sets of parameter values for the different algorithms and models based on extensive pre-tests. Information on the models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ in the one-dimensional case (dim=1) can be found in Section 2.5.2 while details on the model $\mathcal{M}_1'$ in the multi-dimensional case (dim=2, 3, 4) are provided in Section 2.5.3.

		dim = 1		**dim = 2, 3, 4**
		$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ $n = 500$	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ $n = 1500$	$\mathcal{M}_1'$ $n = 5000$
leveraging	N	$4, 6, 8, \ldots, 38, 40$	$4, 6, 8, \ldots, 48, 50$	$10, 20, 30, 50, 100, 150, \ldots, 650, 700$
	λ, γ	$0.3, 0.35, 0.4, 0.45, 0.5$	$0.3, 0.35, 0.4, 0.45, 0.5$	$0.3, 0.35, 0.4, 0.45, 0.5$
quantization		$4, 6, 8, \ldots, 38, 40$	$4, 6, 8, \ldots, 48, 50$	$10, 20, 30, 50, 100, 150, \ldots, 650, 700$
kNN		$4, 6, 8, \ldots, 38, 40$	$4, 6, 8, \ldots, 48, 50$	$10, 20, 30, 50, 100, 150, \ldots, 650, 700$
smoothing splines		$0, 0.025, \ldots, 1.975, 2$	$0, 0.025, \ldots, 1.975, 2$	
xgboost		$1, 2, 3, \ldots, 9, 10$	$1, 2, 3, \ldots, 9, 10$	

Details on the considered parameters for each of the various estimators in the simulation study can be found in Table 2.1. The actual parameters for the leveraging estimator chosen by the proposed hyperparameter selection procedure are provided in Table A.1 in the Appendix.

Zheng (2012) and replace the indicator function by the cumulative distribution function of the standard normal distribution.

[30] We also considered randomization of the second derivative of the smoothed check-function to force additional partitions on leaf nodes of the tree. However, this approach did not improve the results in pre-tests and is therefore not considered in the simulation study. Of course, more efforts can be made in tuning other parameters of the algorithm. This is, however, beyond the scope of this paper.

2.5.2 Analysis of the one-dimensional case

The models considered

In the one-dimensional setting, samples of sizes 500 and 1500 are generated according to the three models

$$\begin{aligned} (\mathcal{M}_1) \quad & Y = X_1^2 + \epsilon_1, \\ (\mathcal{M}_2) \quad & Y = \sin(X_2) + m(X_2) \cdot \epsilon_2, \\ (\mathcal{M}_3) \quad & Y = \psi(X_3) + \epsilon_3, \end{aligned} \tag{2.18}$$

where ϵ_1 and ϵ_2 follow a standard normal distribution, and ϵ_3 a χ^2-distribution with one degree of freedom. The function $m(\cdot)$ in model $\mathcal{M}_2$ is defined as $m : [-3, 3] \to \mathbb{R}, x \mapsto 0.5 + 1.5 \sin^2(\frac{\pi}{2}x)$ and the function ψ in model $\mathcal{M}_3$ is given as $\psi : [-2, 2] \to \mathbb{R}, x \mapsto 20\Big(0.5 \cdot \phi(10(x + 0.5)) + 0.5^2 \cdot \phi(10x) + 0.5^3 \cdot \phi(10(x - 0.5))\Big)$, where ϕ denotes the standard normal density function. Consequently, we include both symmetric and asymmetric as well as homoscedastic and heteroskedastic conditional distributions in the simulation study.

We also consider different distributions for the covariate X. For the models $\mathcal{M}_1$ and $\mathcal{M}_3$ we assume X_1 and X_3 to follow a continuous uniform distribution on the interval $[-2, 2]$ and for model $\mathcal{M}_2$ we set $X_2 = 6Z - 3$, where Z follows a beta distribution with shape parameters $(2, 2)$. For $1 \leq j \leq 3$, X_j and ϵ_j are assumed to be stochastically independent.[31] A visualization of the three models including quantile curves for $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$ can be found in Figure 2.4.

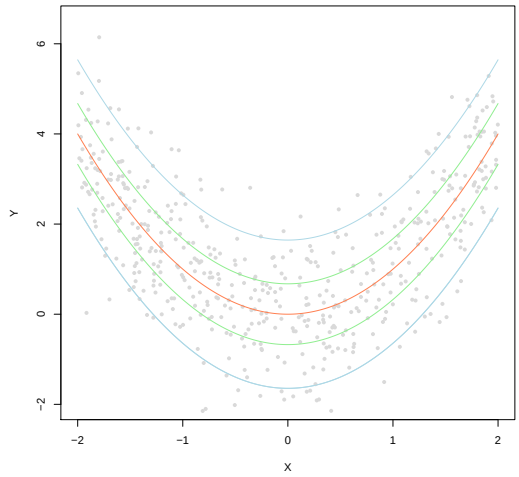
Analysis of the quantile curves and error statistics

The conditional quantile curves for the models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ and random samples of sizes $n = 500$ and $n = 1500$ produced by the leveraging estimator as well as by the

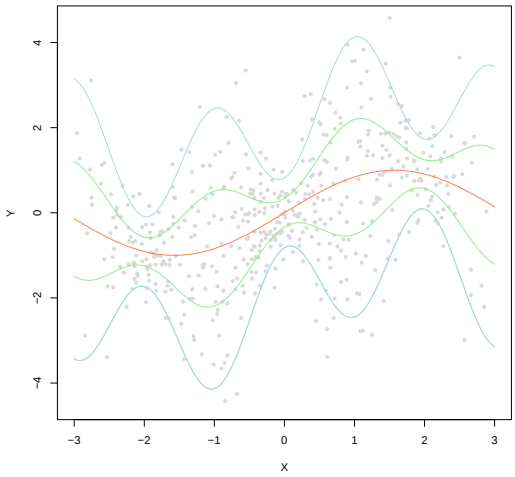
[31] We adopted some of the models from Charlier et al. (2015a) with only slight changes. For further simulation results, especially for other distributions of X_1, X_2, and X_3, we refer the interested reader to this paper.

Figure 2.4: True quantile curves for the models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$

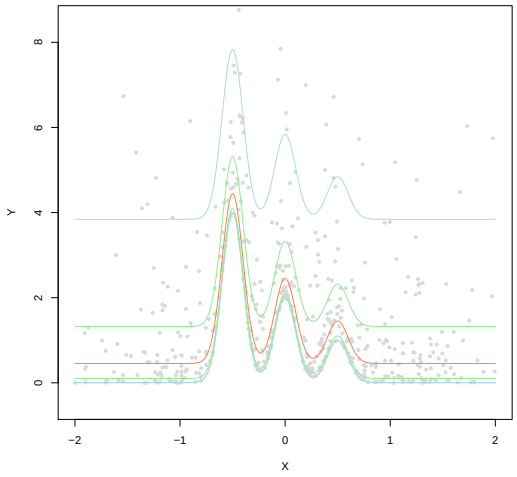
The panels show random samples of size $n = 500$ generated according to the models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, see Section 2.5.2. The curves are the true conditional quantile functions for $\alpha = 0.05$ (blue), 0.25 (green), 0.5 (red), 0.75 (green), and 0.95 (blue).



(a) model $\mathcal{M}_1$



(b) model $\mathcal{M}_2$



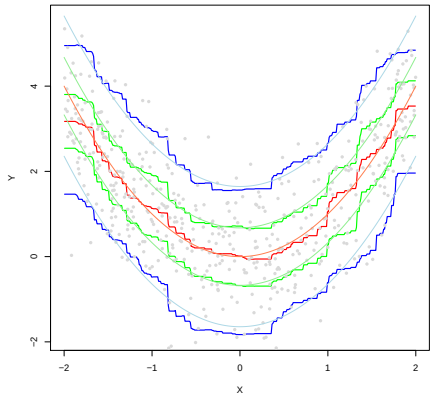
(c) model $\mathcal{M}_3$

four competing estimators are presented in Figures 2.5, 2.6, 2.7, 2.8, A.1, and A.2.

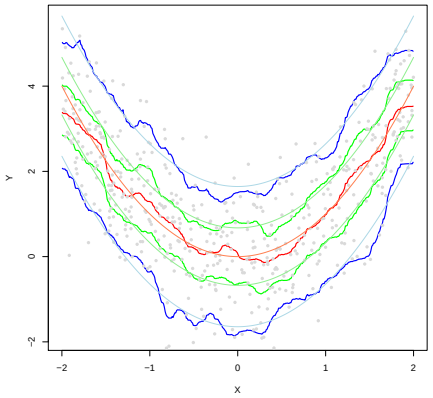
The proposed estimator produces smooth curves that capture the conditional distributions very well. For model $\mathcal{M}_3$ and a sample size of $n = 500$ (Panel 2.7(a)) we observe some overfitting issues, especially for $\alpha = 0.95$, due to the conditional χ^2 distribution. However, these difficulties for $\alpha = 0.95$ are shared by the competing algorithms. When increasing the sample size to $n = 1500$ (Panel A.2(a)) the estimator adapts quite well to the complex link function ψ (Equation (2.18)). It further detects

Figure 2.5: Estimated conditional quantile curves for model $\mathcal{M}_1$ and $n = 500$

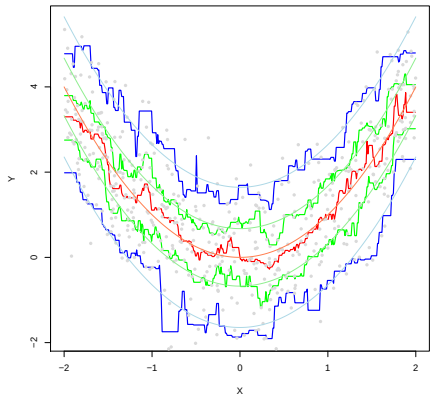
The conditional quantile curves are estimated by the leveraging and four competing algorithms (quantization, kNN, smoothing splines, and xgboost estimator), see Section 2.5.1 for details on the estimators and Table A.2 for details on the employed parameters. The quantile curves are estimated based on a random sample of size $n = 500$ generated according to model $\mathcal{M}_1$. In all panels, the quantile levels considered are $\alpha = 0.05$ (blue), 0.25 (green), 0.5 (red), 0.75 (green), and 0.95 (blue). The more transparent quantile curves are the population ones.



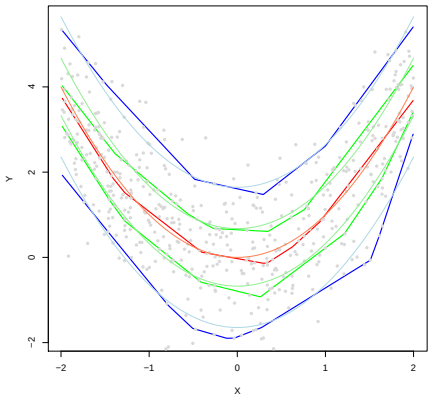
(a) leveraging estimator



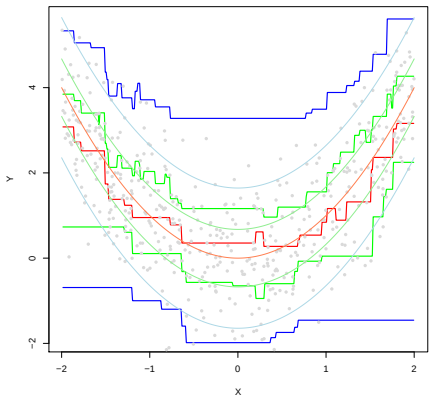
(b) quantization estimator



(c) kNN estimator



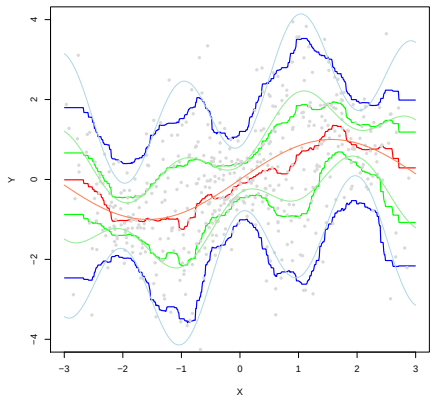
(d) smoothing splines estimator



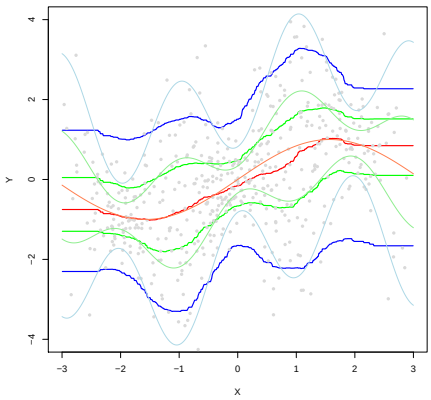
(e) xgboost estimator

Figure 2.6: Estimated conditional quantile curves for model $\mathcal{M}_2$ and $n = 500$

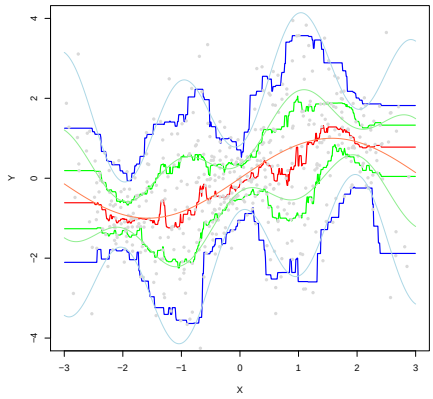
This figure presents the same plots as Figure 2.5 but for a random sample generated according to model $\mathcal{M}_2$.



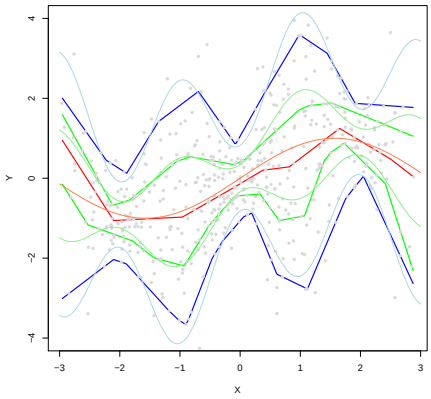
(a) leveraging estimator



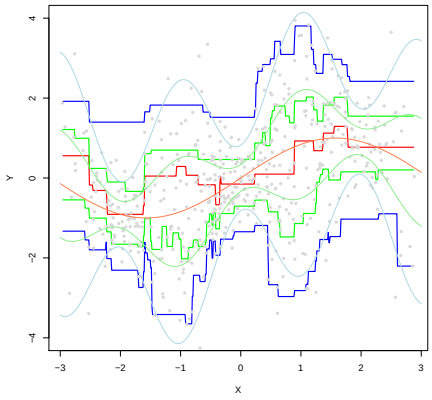
(b) quantization estimator



(c) kNN estimator



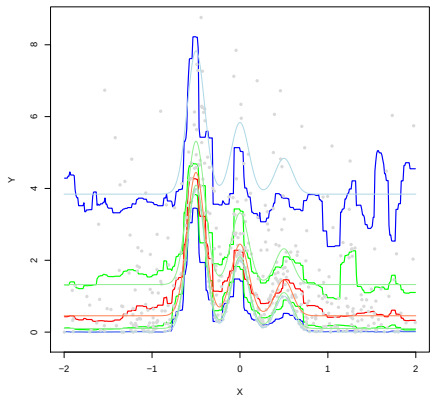
(d) smoothing splines estimator



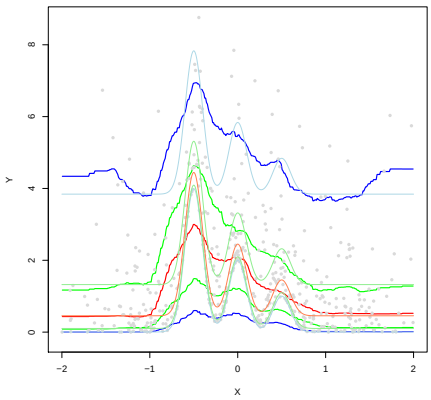
(e) xgboost estimator

Figure 2.7: Estimated conditional quantile curves for model $\mathcal{M}_3$ and $n = 500$

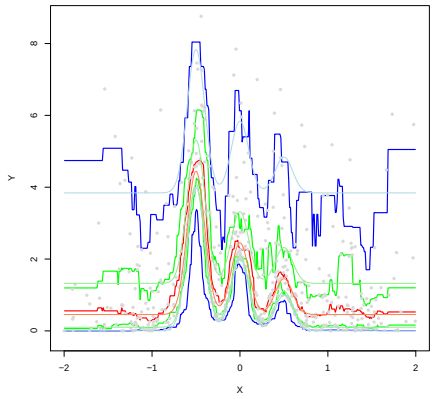
This figure presents the same plots as Figure 2.5 but for a random sample generated according to model $\mathcal{M}_3$.



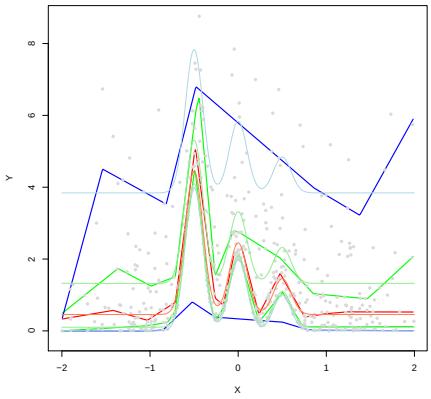
(a) leveraging estimator



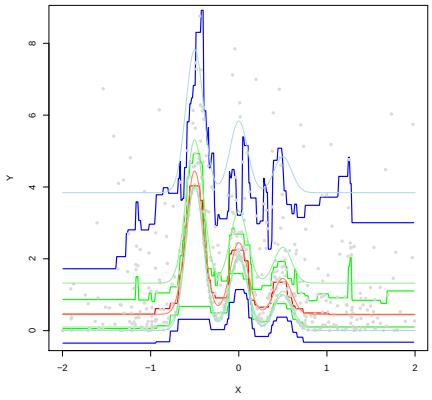
(b) quantization estimator



(c) kNN estimator



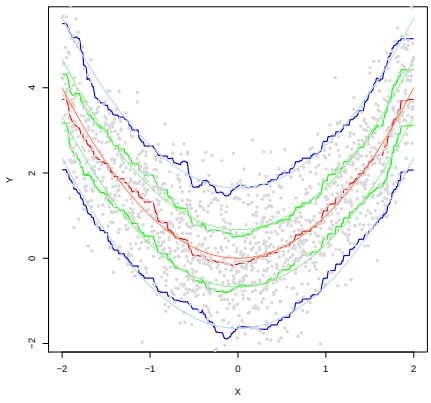
(d) smoothing splines estimator



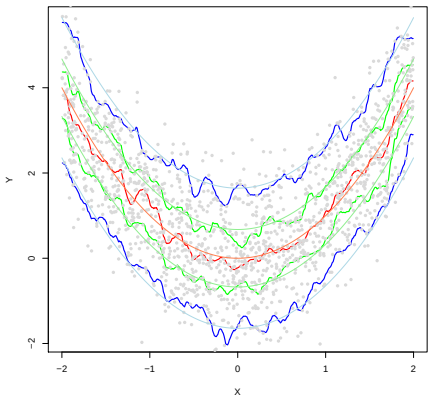
(e) xgboost estimator

Figure 2.8: Estimated conditional quantile curves for model $\mathcal{M}_1$ and $n = 1500$

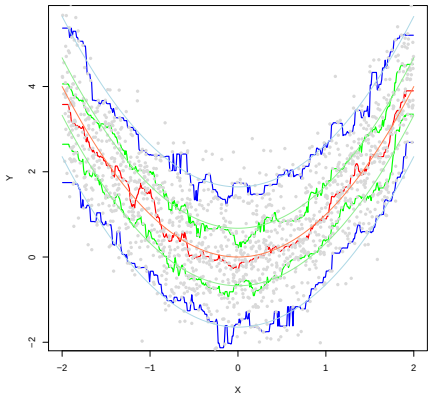
This figure presents the same plots as Figure 2.5 but for a random sample of size $n = 1500$.



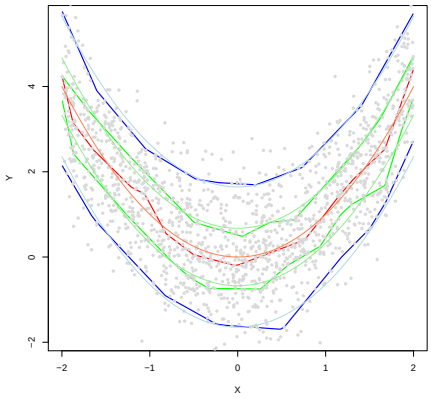
(a) leveraging estimator



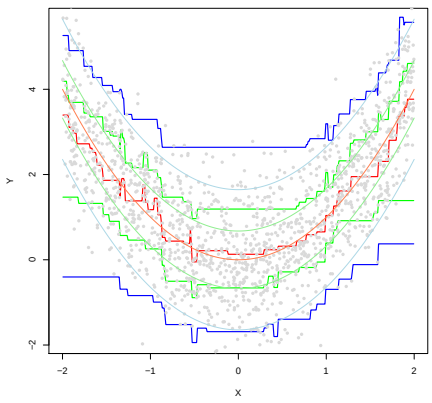
(b) quantization estimator



(c) kNN estimator



(d) smoothing splines estimator



(e) xgboost estimator

the domains of the interval $[-3, 3]$ where the link function is constant fairly well as opposed to the kNN, smoothing splines, and xgboost estimator. This is due to the leveraging approach that puts less weight (fewer quantizers) in the edges and more in the middle of the interval $[-3, 3]$.

The quantization estimator produces smooth quantile curves, too, but sometimes shows signs of underfitting, see, e.g., Panels 2.6(b) and 2.7(b). One reason is that the optimal number of quantizers is determined only based on the distribution of X, not taking into account the conditional distribution of Y. Furthermore, the parameter selection method by Charlier et al. (2015a) aims to minimize variation in the quantile estimates favoring smaller numbers of quantizers. This is especially pronounced in model $\mathcal{M}_3$, see Panels 2.7(b) and A.2(b). The kNN estimates are overall relatively close to the true conditional quantile curves but the estimator is prone to overfitting. Opposed to this, the smoothing splines estimator produces piecewise linear quantile curves that capture the underlying link function quite well for models $\mathcal{M}_1$ and $\mathcal{M}_2$. For model $\mathcal{M}_3$ and some values of α however, the estimator completely misses the link function, see Panel 2.7(d). The xgboost estimator produces piecewise constant estimates due to its tree based nature and is therefore prone to underfitting (see, e.g., Panel 2.5(e)). Especially for $\alpha = 0.05$ and $\alpha = 0.95$, the xgboost estimator faces serious difficulties, which is due to the fact that the gradient of the error function employed is near zero for y values above the 5 % and below the 95 % quantile, respectively.[32]

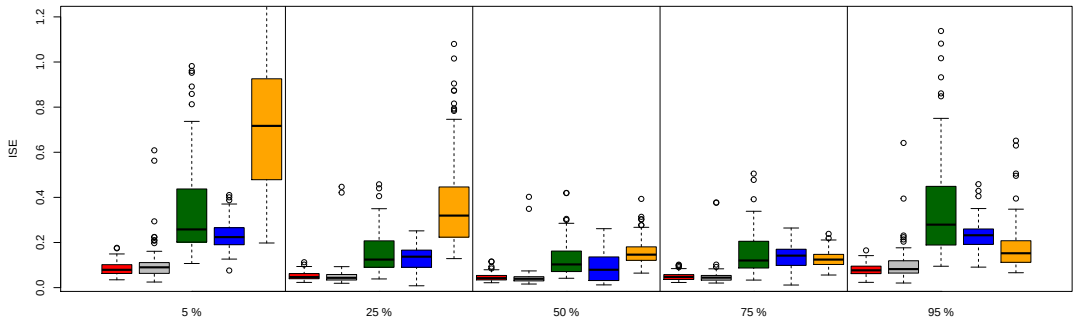
We now compare the accuracy of the estimators by sampling 100 times from the models $\mathcal{M}_1, \mathcal{M}_2$, and $\mathcal{M}_3$ with sample sizes $n = 500$ and $n = 1500$ and comparing the ISE values. The results are illustrated in the boxplots 2.9, 2.10, and 2.11. The boxplots confirm that the leveraging estimator performs well for all of the models, samples sizes, and quantile levels and often provides the best results of all considered algorithms. It is further evident from the plots that the interquartile range of the leveraging-based ISEs is low, indicating that the estimation quality of the estimator is relatively stable. This

[32] In further numerical experiments we added some randomization to the gradient and the second derivative of the error function to overcome this issue. However, this did not lead to substantial improvements.

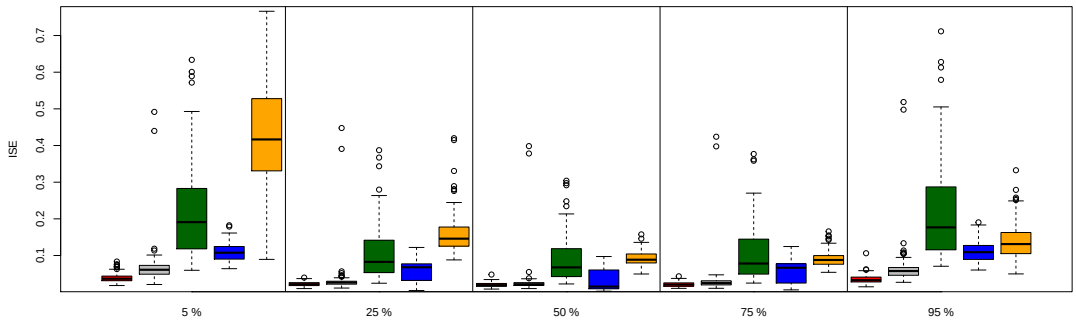
conclusion is supported by the fact that the leveraging estimator produces only few outliers. Of course, compared to $\alpha = 0.5$, ISEs increase for lower and higher values of α as in these cases the estimates are based on fewer observations. However, these increases are low compared to most of the competing algorithms. As expected, estimation errors decrease when the sample size rises. Summarizing, the results confirm that the leveraging estimator produces stable conditional quantile estimates of good quality.

Figure 2.9: ISEs of the proposed and competing estimators for model $\mathcal{M}_1$

The panels show boxplots of the ISEs for random samples of sizes $n = 500$ (top) and $n = 1500$ (bottom) generated according to model $\mathcal{M}_1$. For each sample size, 100 independent repetitions are performed yielding 100 ISE estimates for each estimator and $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$. We include the leveraging estimator (red), the quantization estimator (grey), the kNN estimator (green), the smoothing splines estimator (blue), and the xgboost estimator (orange).



(a) model $\mathcal{M}_1, n = 500$



(b) model $\mathcal{M}_1, n = 1500$

The quantization estimator often provides good results, too, but has difficulties in capturing the complex link function of model $\mathcal{M}_3$. The estimates provided by the kNN estimator produce a relatively large interquartile range between the ISE values with frequent outliers. However, especially for models $\mathcal{M}_2$ and $\mathcal{M}_3$ the estimation accuracy

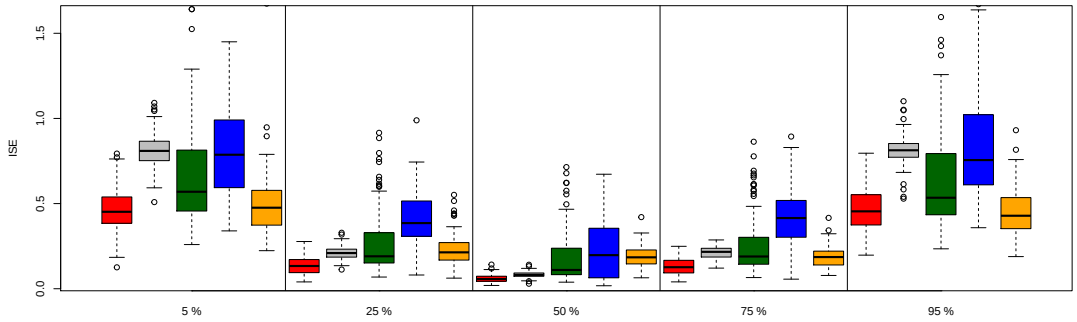
Table 2.2: Error statistics and computation times in the one-dimensional case

This table summarizes error statistics and computation times for five estimators and three models. The models $\mathcal{M}_1, \mathcal{M}_2$, and $\mathcal{M}_3$ are defined in Equation (2.18). The considered algorithms are the leveraging, quantization, kNN, smoothing splines, and xgboost estimator, see Section 2.5.1 for details. We report several error statistics: The mean integrated squared error (MISE) is defined as the average of the ISEs (see Equation 2.16) over the quantiles $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$ and 100 random samples of sizes $n = 500$ (above) and $n = 1500$ (below in brackets). Analogously, ME denotes the median squared error (per random sample and quantile level) averaged over 100 random samples and the quantile levels $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$. With SD we denote the standard deviation of the ISEs averaged over all values of α. With CPU we report the average computation time for estimating the quantile curves in seconds. Calculations are performed on an Intel(R) Core(TM) i7-4770 CPU with 3.4 GHz and 32 GB of RAM. Note that the reported times encompass the calculation of the optimal parameter(s) for each of the estimators, see Sections 2.4.2 and 2.5.1. The lowest values for each of the statistics are printed in bolt type.

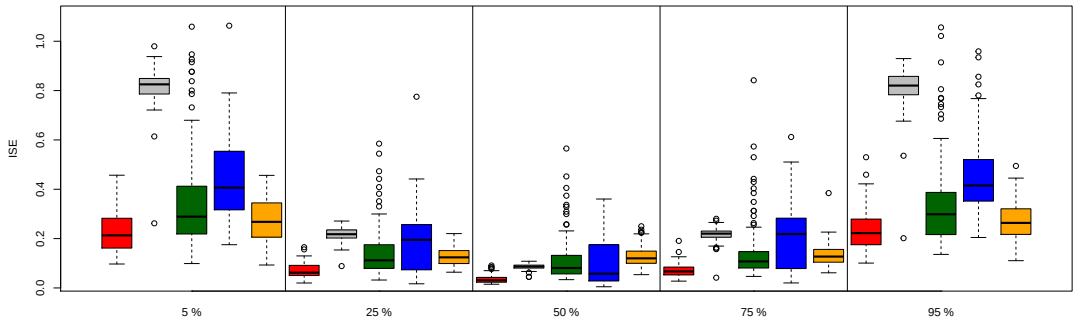
		leveraging	**quantization**	**kNN**	**smoothing splines**	**xgboost**
$\mathcal{M}_1$	MISE	**0.0622 (0.0277)**	0.0714 (0.0470)	0.2257 (0.1567)	0.1630 (0.0729)	0.3157 (0.1871)
	ME	**0.0274 (0.0119)**	0.0333 (0.0216)	0.1165 (0.0801)	0.0708 (0.0313)	0.0813 (0.0576)
	SD	**0.0218 (0.0089)**	0.0624 (0.0578)	0.1408 (0.1195)	0.0598 (0.0284)	0.1402 (0.0705)
	CPU	37.6823 (106.0321)	16.0802 (44.0168)	**3.2378** (20.5344)	8.2016 **(15.7793)**	111.4673 (121.9950)
$\mathcal{M}_2$	MISE	**0.2514 (0.1283)**	0.4265 (0.4293)	0.4334 (0.2334)	0.5383 (0.2838)	0.3201 (0.1866)
	ME	**0.0683 (0.0271)**	0.2190 (0.2272)	0.1290 (0.0675)	0.1653 (0.0667)	0.1348 (0.0719)
	SD	0.0770 (0.0466)	**0.0575 (0.0454)**	0.2847 (0.1772)	0.2268 (0.1500)	0.1978 (0.0594)
	CPU	34.9473 (106.2039)	17.9633 (46.4797)	**3.2264** (20.5632)	8.1842 **(15.7689)**	111.2531 (122.0514)
$\mathcal{M}_3$	MISE	**0.3265 (0.1694)**	0.6187 (0.6122)	1.0415 (0.9082)	0.6797 (0.3106)	1.0139 (0.6314)
	ME	0.1638 (0.0771)	**0.0839 (0.0458)**	0.6890 (0.5639)	0.2894 (0.1288)	0.6539 (0.3306)
	SD	0.1196 **(0.0557)**	**0.0932** (0.0591)	0.2487 (0.2279)	0.2522 (0.0871)	0.5882 (0.2130)
	CPU	36.3767 (108.5655)	16.0995 (44.0475)	**3.1296** (20.1224)	8.2248 **(15.8837)**	109.8797 (122.7161)

Figure 2.10: ISEs of the proposed and competing estimators for model $\mathcal{M}_2$

This figure presents the same plots as Figure 2.9 but for model $\mathcal{M}_2$.



(a) model $\mathcal{M}_2, n = 500$



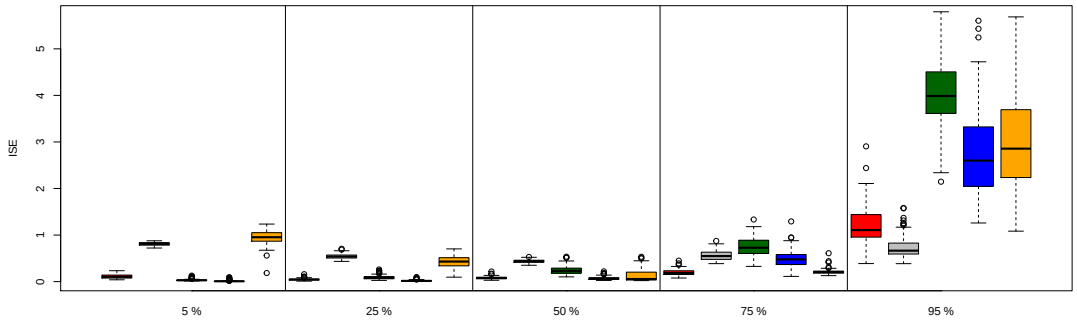
(b) model $\mathcal{M}_2, n = 1500$

is competitive. The smoothing splines estimator produces stable and low ISE values for models $\mathcal{M}_1$ and $\mathcal{M}_3$ but produces less satisfactory results for model $\mathcal{M}_2$. The xgboost estimator has been included in the simulation study to provide a comparison to an estimator based on boosting, a concept very similar to leveraging. However, the estimator has difficulties with both model $\mathcal{M}_1$ and model $\mathcal{M}_3$ with problems being more pronounced for high and low values of α.

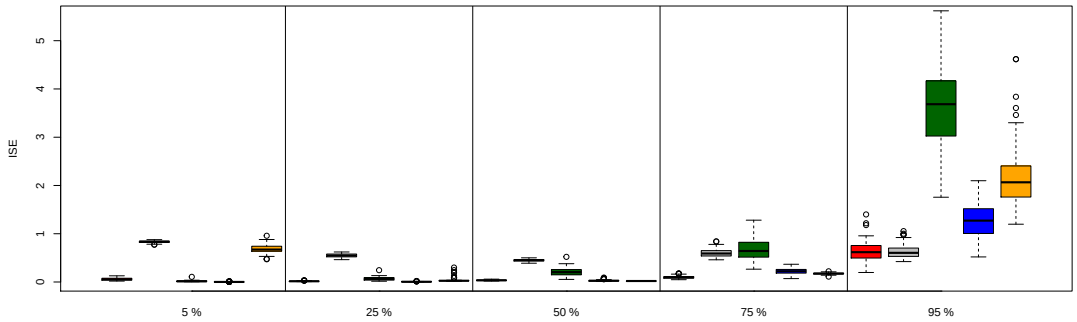
We report averaged results in Table 2.2 and provide two efficiency measures. The mean integrated squared error (MISE) is defined as the mean of the ISEs over all 100 repetitions and the five different values of α. The median error (ME) is obtained by calculating the median of the summands in Equation (2.16) and averaging these values over all 100 repetitions and five different α values. In terms of MISEs, the leveraging estimator performs the best for all of the three models $\mathcal{M}_1, \mathcal{M}_2$, and $\mathcal{M}_3$ and sample sizes $n = 500$ and $n = 1500$. In terms of the ME the estimator produces

Figure 2.11: ISEs of the proposed and competing estimators for model $\mathcal{M}_3$

This figure presents the same plots as Figure 2.9 but for model $\mathcal{M}_3$.



(a) model $\mathcal{M}_3, n = 500$



(b) model $\mathcal{M}_3, n = 1500$

either the best (models $\mathcal{M}_1$ and $\mathcal{M}_2$) or second best results (model $\mathcal{M}_3$) among the considered estimators. The conclusions derived from the boxplots are in line with the two efficiency measures.

We additionally report the standard deviation of the ISEs averaged over $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$. Again, the leveraging estimator provides either the lowest or the second lowest deviations, indicating that it reliably produces good estimates. However, these results come at the price of a higher computational burden. In Table 2.2 we also report the average computation time in seconds necessary for parameter selection and model estimation. The leveraging estimator requires approximately ten (five) times the computation time of the fastest algorithm (kNN) for $n = 500$ ($n = 1500$). However, most of the time is required for performing the parameter selection. When computations have to be done repeatedly without the need to determine new parameters every time, estimation could be performed much faster. Apart from this, by perform-

ing the hyperparameter selection procedure in parallel, computation times could also be reduced significantly. Another very appealing feature of the leveraging estimator is that conditional quantile estimates for new values of x or additional values of α can be calculated at almost no additional computational costs based on the already estimated grids from each iteration step.

The proposed estimator also exhibits many attractive features for multivariate covariates. In Section 2.5.3 we therefore study the behavior of the leveraging and two competing estimators for two-, three-, and four-dimensional covariates. Before, we shortly present some empirical results on the estimation error for increasing sample sizes (in one dimension).

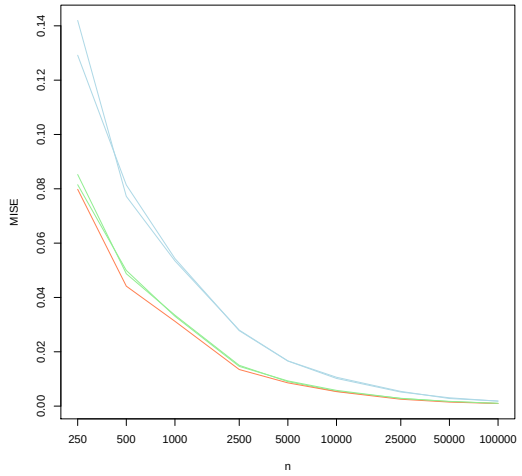
Results for increasing sample sizes

To provide some anecdotal evidence on the convergence of the leveraging estimator, we conduct an additional experiment. We generate random samples of sizes $n = 250, 500, 1000, 2500, 5000, 10000, 25000, 50000, 100000$ according to models $\mathcal{M}_1$ and $\mathcal{M}_3$. For each sample size we perform 50 independent repetitions yielding 50 ISE estimates for the proposed estimator and each $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$. We consider $N = 10, 20, 30, 40, 50, 75, 100, 125, 150, 175, 200$ for the number of quantizers and determine the optimal value via the hyperparameter selection procedure introduced in Section 2.4.2. For simplicity, λ and γ are set to 0.5. Panels 2.12(a) and 2.12(c) report the MISEs for models $\mathcal{M}_1$ and $\mathcal{M}_3$ and various sample sizes. Additionally, Panels 2.12(b) and 2.12(d) present the average number of quantizers per sample size.

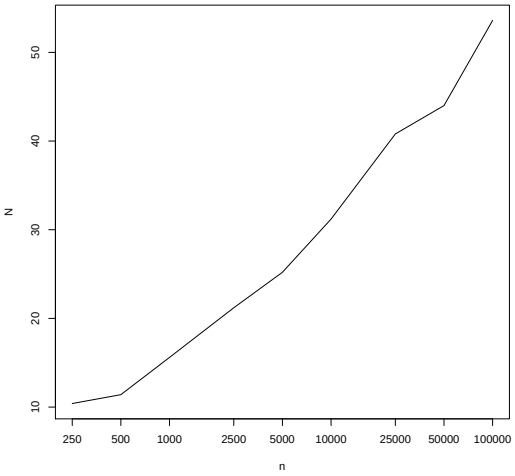
In line with Theorem 2, Figure 2.12 demonstrates that MISEs decrease when both the sample size n and the number of quantizers N increase. Furthermore, as N is chosen according to the proposed hyperparameter selection procedure, this experiment provides anecdotal evidence for the appropriateness of the parameter selection procedure in finding suitable values of N given a random sample of size n.

Figure 2.12: MISE and average number of quantizers depending on the sample size n

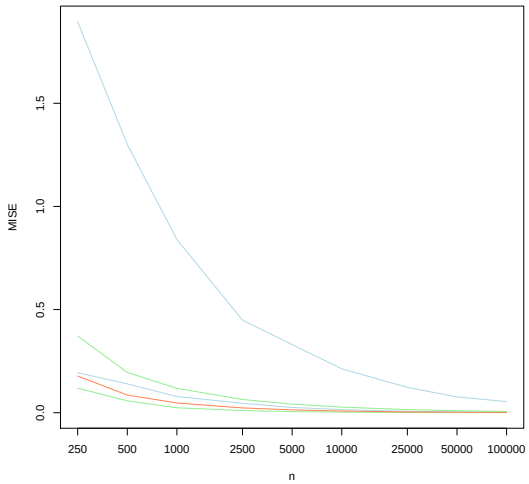
Panels 2.12(a) and 2.12(c) report the MISEs for random samples of sizes $n = 250, 500, 1000, 2500, 5000, 10000, 25000, 50000, 100000$ generated according to models $\mathcal{M}_1$ and $\mathcal{M}_3$. For each sample size, 50 independent repetitions are performed yielding 50 ISE estimates for the leveraging estimator and $\alpha = 0.05$ (blue), 0.25 (green), 0.5 (red), 0.75 (green), and 0.95 (blue). The ISE values rely on the number of quantizers N which is determined for each of the random samples based on the hyperparameter selection procedure proposed in Section 2.4.2. We consider $N = 10, 20, 30, 40, 50, 75, 100, 125, 150, 175, 200$ as possible values, λ and γ are set to 0.5. Panels 2.12(b) and 2.12(d) report the average values of the selected parameter N depending on the sample size n.



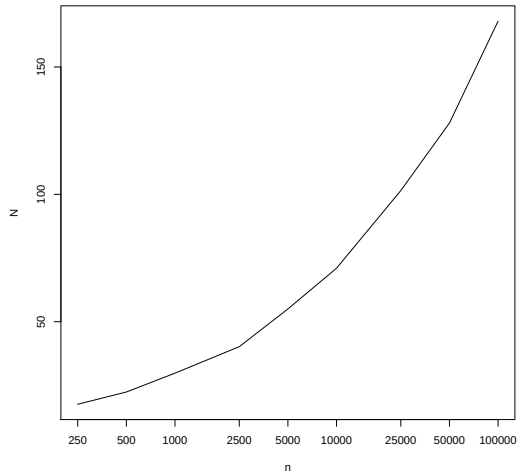
(a) MISE for model $\mathcal{M}_1$



(b) average number of quantizers for model $\mathcal{M}_1$



(c) MISE for model $\mathcal{M}_3$



(d) average number of quantizers for model $\mathcal{M}_3$

2.5.3 Analysis of the multi-dimensional case

In many real world applications one wants to analyze the dependency of a one-dimensional response variable on multiple covariates. As before, we are not interested in modeling the mean of the response. Instead, we want to capture the conditional distribution by estimating conditional quantile curves. Luckily, the leveraging estimator extends naturally to multiple covariates. In this section we therefore investigate the behavior of the estimator as dimension grows from two to four.

The model and competitors considered

We extend model $\mathcal{M}_1$ from Equation (2.18) by setting

$$(\mathcal{M}_1') \quad Y = |X|^2 + \epsilon, \tag{2.19}$$

where again $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$, ϵ is standard normally distributed and statistically independent from X, and X follows a continuous uniform distribution over the hypercube $[-2, 2]^d$, with $d = 2, 3, 4$ denoting the dimension. By choosing $n = 5000$ for the size of each random sample regardless of the dimension, we are able to analyze the effect of an increase in the dimension on the estimators' performance. The set of parameters considered in the estimation process of the estimators is reported in Table 2.1 while the parameters chosen for the leveraging estimator by the hyperparameter selection procedure are provided in Table A.1 in the Appendix.

In this part of the simulation study we consider the quantization algorithm and the kNN estimator as competing algorithms. We exclude the smoothing splines estimator, as it cannot easily be extended to dimensions greater than two. We further exclude the xgboost estimator because of the poor estimation results we observed in pre-tests. For more details we refer to Section 2.5.1.

Analysis of the error statistics

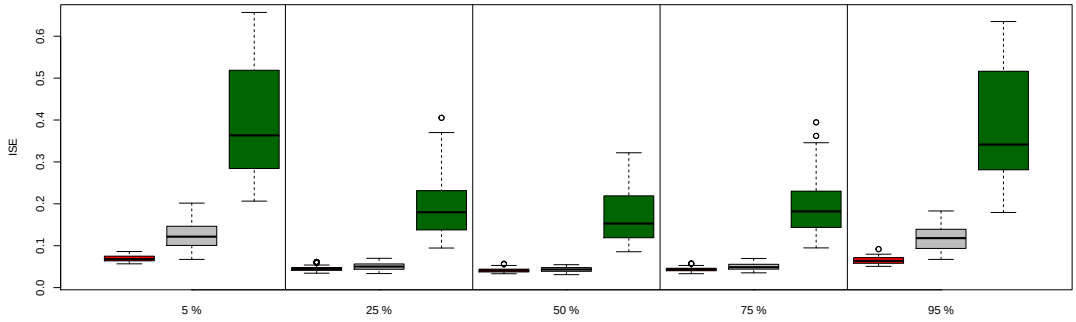
The estimation accuracy is measured in terms of the ISE for 50 random samples of size 5000 generated independently according to model $\mathcal{M}_1'$. The results for two to four dimensions and $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$ are illustrated in boxplots 2.13(a), 2.13(b) and 2.13(c). In each of the dimensions, the leveraging estimator provides the best results, which is especially pronounced for $\alpha = 0.05$ and $\alpha = 0.95$. The quantization estimator performs very well for two-dimensional covariates but fails for three- and four-dimensional covariates.[33] For two-dimensional covariates the kNN estimator performs the worst with relatively high ISEs and a large interquartile range of ISE estimates. Relatively to the competing estimators, the results improve in dimensions three and four. In all dimensions we fixed the sample size n to 5000. As the volume of the hypercubes from which observations for the covariates are generated increases with the dimension, the curse of dimensionality leads to an increase of the average ISEs. However, the proposed estimator still produces good results via adapting to the higher dimension by choosing a larger number of quantizers and increasing the ratio γ/λ, see Table A.1 for details. Additionally, the interquartile range of ISE estimates remains relatively low, indicating that the leveraging estimator provides stable estimates.

We report results for the MISE and ME in Table 2.3. In all of the considered dimensions, the leveraging estimator yields the lowest MISE and ME values. In dimensions two and three the standard deviation of the ISEs is the lowest for the leveraging estimator and is very close to that of the kNN estimator in dimension four. The good results come at the price of higher computational costs. However, due to an efficient implementation of the leveraging estimator, the computation time is only slightly raised by an increase of the dimension as opposed to the quantization and the kNN estimator. As a consequence, the leveraging estimator requires only a little more computation time than the kNN estimator. As we have pointed out previously, performing the hyperparameter selection in parallel could significantly reduce computation times.

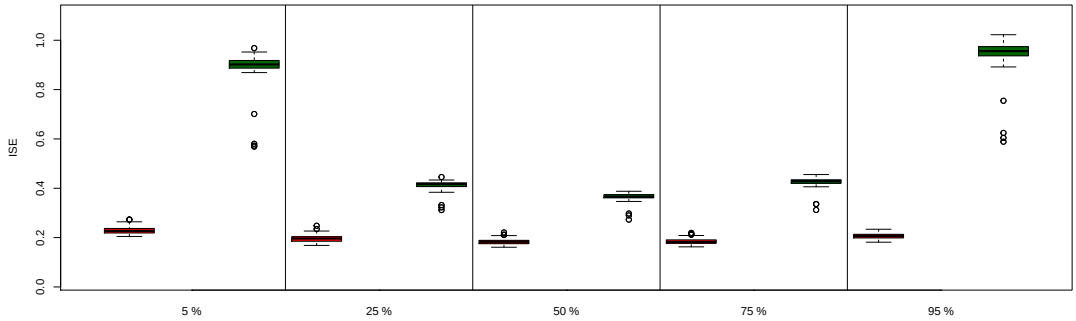
[33] The inferior results in dimensions three and four are most likely due to the hyperparameter selection procedure proposed in Charlier et al. (2015a) determining (too) small numbers of quantizers.

Figure 2.13: ISEs of the proposed and competing estimators for model $\mathcal{M}_1'$ and multivariate covariates

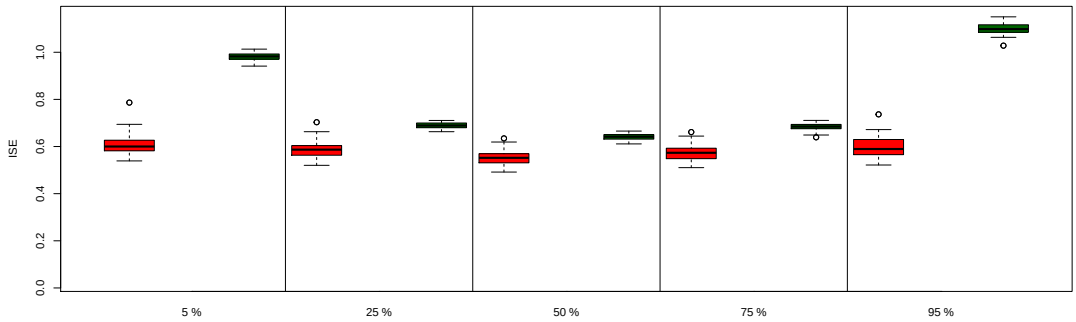
The panels show boxplots of the ISEs for random samples of size $n = 5000$ for two-dimensional (top), three-dimensional (middle), and four-dimensional (bottom) covariates generated according to model $\mathcal{M}_1'$. For each dimension, 50 independent repetitions are performed resulting in 50 ISE estimates for each estimator and $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$. The estimators are the leveraging estimator (red), the quantization estimator (grey), and the kNN estimator (green), see Section 2.5.1 for details. For dimensions three and four, ISEs for the quantization estimator are quite high. We therefore restrict the y-axis to lower values to enable a more detailed comparison between the leveraging estimator and the kNN estimator. Results for the quantization estimator in dimensions three and four are provided in Table 2.3.



(a) model $\mathcal{M}_1'$, dim = 2



(b) model $\mathcal{M}_1'$, dim = 3



(c) model $\mathcal{M}_1'$, dim = 4

Table 2.3: Error statistics and computation times in the multi-dimensional case

This table summarizes error statistics and computation times for the proposed and two competing estimators for two- to four-dimensional covariates. Therefore, 50 random samples of size 5000 are generated according to model $\mathcal{M}_1'$, which is the natural generalization of model $\mathcal{M}_1$ to the multi-dimensional case. We consider the quantization and the kNN estimator as competing algorithms. We report some error statistics: The mean integrated squared error (MISE) is defined as the average of the ISEs over the quantile levels $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$ and 50 random samples. Analogously, ME denotes the median squared error (per random sample and quantile level) averaged over 50 random samples and the quantile levels $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$. With SD we denote the standard deviation of the ISEs averaged over all values of α. CPU reports the average computation time in minutes for estimating the quantile curves including the selection of the parameter(s). Calculations are performed on an Intel(R) Core(TM) i7-4770 CPU with 3.4 GHz and 32 GB of RAM. Note that the reported times encompass the calculations of the optimal parameters for each of the estimators, see Sections 2.4.2 and 2.5.1. For the quantization estimator in three and four dimensions we determined the computation time based on ten (three dimensions) and only one (!) (four dimensions) repetitions of the implementation in the *QuantifQuantile* R-package by Charlier et al. (2015c). The error statistics, however, are based on 50 repetitions of an own implementation of the quantization estimator that provides the same results but is more efficient in terms of the computation times. The lowest values for each of the statistics are printed in bolt type.

		leveraging	**quantization**	**kNN**
dim = 2	MISE	**0.0529**	0.0773	0.2781
	ME	**0.0223**	0.0397	0.142
	SD	**0.0066**	0.0164	0.1148
	CPU	6.8759	228.19	**2.156**
dim = 3	MISE	**0.2005**	4.2899	0.6026
	ME	**0.0821**	2.3169	0.3131
	SD	**0.0143**	1.8172	0.0505
	CPU	8.0403	477.82	**2.5784**
dim = 4	MISE	**0.5844**	8.2904	0.8187
	ME	**0.2506**	4.4411	0.3715
	SD	0.0383	0.1763	**0.0159**
	CPU	9.3433	16384.30	**8.1136**

2.6 Empirical application

In this empirical study we apply the proposed leveraging estimator to one day ahead VaR and ES forecasts. The forecasts are obtained by various GARCH-type models and are thus subject to two types of uncertainty. First, the true data generating process for the stock returns is unknown and might not be adequately reflected by the GARCH-

type models giving rise to model risk. Secondly, the models require parameters that have to be estimated from the data. This gives rise to estimation risk (cf. Lönnbark, 2013). In this empirical application we focus on the latter.

2.6.1 Data

The sample period is from January 2000 until March 2021. The stock price of all S&P Composite 1500 Index constituents[34] are retrieved from Datastream. Subsequently, we calculate daily log-returns and remove outliers (returns with an absolute z-score above 10).[35]

The log-returns are used to produce one day ahead VaR and ES forecasts on the 99 % and 97.5 % confidence level, respectively.[36] The predictions are obtained from ARMA(1,1)-GARCH(1,1)-type models where the GARCH model by Bollerslev (1986), the EGARCH model by Nelson (1991), the GJR-GARCH model by Glosten et al. (1993), and the T-GARCH model by Zakoian (1994) are considered. These models are all nested within the fGARCH model by Hentschel (1995). For details on the models we refer to Bollerslev (2010) and the original papers. For the innovations we assume a skewed Student-t distribution, which can account for skewness and excess kurtosis in the data.[37] The models are fitted via the ugarchfit method from the *rugarch* R-package by Ghalanos (2020) over a moving window of 250 days corresponding to roughly one year of observations.[38] The models yield one day ahead mean and variance forecasts which are used to calculate the VaR as the 1 % quantile of the corresponding skewed Student-t distribution while the ES is obtained via numerical integration (cf. Cardona et al., 2019).[39] Additionally, we calculate non-parametric

[34]The index constituents are determined at the end of the sample period and held fixed over the whole period.

[35]For a normal distribution, a z-score with absolute value above 10 corresponds to a probability $< 10^{-16}$.

[36]These Var and ES levels are standard, see Basel Committee on Banking Supervision (2013, 2014).

[37]See Fernandez and Steel (1998) for details on the skewed Student-t distribution.

[38]We employ the "hybrid" solver that first uses the "solnp" solver and when failing to converge continues with the "nlminb", the "gosolnp", and the "nloptr" solvers. More details are provided in the package documentation.
All computations for the empirical application are performed on the Big-Data-Cluster Galaxy provided by the University Computing Center at Leipzig University.

[39]We observe some extreme forecasts that are due to a failure in convergence, e.g., VaR and ES pre-

estimates for VaR and ES via historical simulation, again over a moving window of 250 days. We compute the historical simulation VaR as the empirical 1 % quantile of the return distribution and the historical simulation ES as the average of the returns below the empirical 2.5 % quantile. For further analyses, we also include (annualized) realized volatility (RV) as well as the third (skewness) and fourth moment (kurtosis) of the empirical return distribution, again computed over a moving window.

We present summary statistics for the VaR and ES estimates obtained from the GARCH-type models and via historical simulation in Table 2.4. We first calculate summary statistics for the cross-section of risk estimates at a particular time point and subsequently average these over the sample period. Consequently, the values can be interpreted as average summary statistics of the cross-section of risk forecasts. While on average the four GARCH-type models produce very similar risk forecasts, there are noticeable differences in the range of predictions. For example, the ES forecasts by the EGARCH model lie on average between -48.55 % and -0.31 % while the GARCH model produces more moderate forecasts that on average assume values between -40.14 % and -0.73 %. The non-parametric risk forecasts obtained via historical simulation on average lie within an even closer range (between -33.73 % and -1.77 % for the ES). Average risk estimates by historical simulation are, however, similar to the GARCH-type models. While on average the absolute levels of VaR forecasts are lower than that of the ES forecasts, differences among the models and historical simulation are similar to the case of ES.

2.6.2 Deriving conditional quantiles

Computing VaR and ES estimates via historical simulation does not involve any parameters (other than the length of the moving window).[40] Hence, by using these values as a benchmark we can analyze the differences in *parametric* risk predictions (for a

dictions below -10^{99} for the EGARCH model. Of course, VaR and ES for stocks cannot fall below -100 % which is why we remove all such predictions.

[40] To obtain completely non-parametric risk estimates we refrain from using weighted historical simulation (Boudoukh et al., 1998), filtered historical simulation (Barone-Adesi et al., 1998, 1999), or similar approaches.

Table 2.4: Average summary statistics for the cross-section of risk forecasts

This table reports average summary statistics for the cross-section of one day ahead forecasts of 97.5 % ES and 99 % VaR by various models over the period December 2000 until March 2021. The values are obtained by first calculating summary statistics over the cross-section of risk forecasts for the S&P Composite 1500 Index constituents at a particular date. These values are subsequently averaged over the sample period. We report the minimum (min), median, mean, maximum (max), and standard deviation (sd). For more details on the calculation of the forecasts, see Section 2.6.1.

		Summary statistics (in %)				
		min	median	mean	max	sd
	model					
97.5 % ES	GARCH	-40.1381	-5.6008	-6.3885	-0.7343	3.4025
	EGARCH	-48.5467	-5.5936	-6.4970	-0.3128	3.8506
	GJR-GARCH	-40.7828	-5.5297	-6.3174	-0.5847	3.4254
	T-GARCH	-45.3770	-5.6850	-6.4749	-0.5131	3.5411
	historical simulation	-33.7273	-5.7410	-6.4006	-1.7660	2.9420
99 % VaR	GARCH	-35.6565	-5.2928	-5.9685	-0.6827	3.0288
	EGARCH	-42.5566	-5.2936	-6.0505	-0.2729	3.3697
	GJR-GARCH	-36.3738	-5.2400	-5.9233	-0.5252	3.0702
	T-GARCH	-39.2896	-5.3935	-6.0778	-0.4073	3.1508
	historical simulation	-33.4710	-5.4681	-6.0832	-1.5291	2.8058

given GARCH-type model at a given day) for stocks with very similar *non-parametric* risk predictions. For this purpose, we condition the parametric risk forecasts on their non-parametric counterparts and calculate conditional quantile curves via the proposed leveraging estimator (for a given day, GARCH-type model, and risk measure (VaR or ES)). In the hyperparameter selection procedure, we consider 1 %, 2 %, ..., 10 % of the observations as the number of quantizers and otherwise use the same parameters as in the simulation study. When considering multiple covariates we scale them to zero mean and unit variance to ensure that all variables contribute equally to the computation of the respective grids. The quantile curves are estimated separately for each GARCH-type model and risk measure on a daily basis.

2.6.3 Results

Estimation risk

Exemplary conditional quantile curves obtained by the leveraging estimator are presented in Figure 2.14. The figure illustrates conditional quantile curves for one day ahead ES forecasts by the GARCH model at two time points, before the 2020 stock market crash due to the COVID-19 pandemic (January 2, 2020) and at its peak (March 16, 2020, "Black Monday II"). It is clear from the figure that the disagreement between parametric GARCH risk forecasts of similar stocks (in terms of their non-parametric historical simulation ES) are substantially larger at the second time point. To condense this disagreement in a single figure and obtain a measure of estimation risk (for a given model and the whole stock market) we compute the differences between the conditional 75 % quantile and the 25 % quantile (the iqr) and average them over all stocks at a given date. This yields a measure of estimation risk for the US stock market.

There is substantial variation in estimation risk over time. Figure 2.15 shows the estimation risk for ES forecasts by the GARCH model over time. Estimation risk is especially high in the aftermath of the bursting dotcom bubble, during the great financial crisis, and during the 2020 stock market crash.[41] Estimation risk can also be determined on the stock level by evaluating the iqr at the historical simulation ES of a particular stock.[42]

There are also differences in estimation risk between the four GARCH-type models and the two risk measures, see Table 2.5. Average estimation risk for ES is higher than that for VaR with all pairwise comparisons being significant at the 1 % level.[43] This

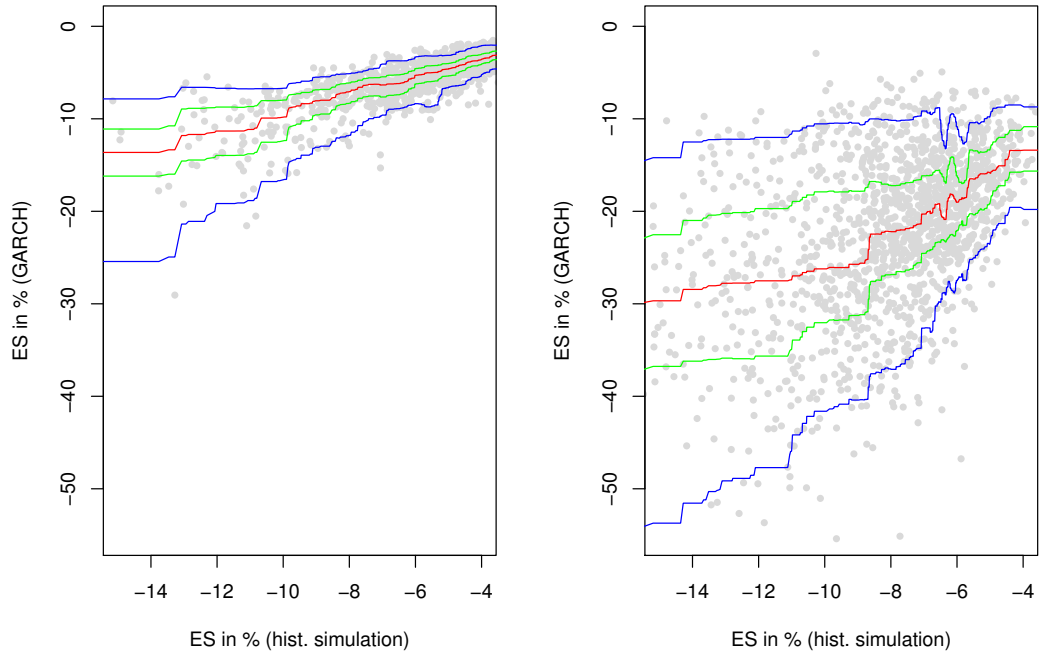
[41] Estimation risk also rises when the level of the historical simulation ES is increased. This is because the conditional quantile curves are in general wider for more extreme values of the non-parametric ES, see Figure 2.14. However, when reproducing Figure 2.15 for a fixed level of non-parametric ES (by evaluating the iqr at a constant value of non-parametric ES each time), the picture looks similar (not included here for brevity). Consequently, varying levels of the non-parametric ES do not explain the differences in estimation risk over time. We conclude that not only the level of risk itself rises during times of financial turmoil but also the uncertainty regarding its estimation.

[42] We also refer to Figure 2.17 providing the 90 % confidence bands of ES forecasts for the apple stock. Uncertainty about the estimation of risk forecasts corresponds to the width of the confidence band.

[43] Statistical significance is determined based on t-tests with standard errors corrected for serial correlation and heteroskedasticity according to Newey and West (1987) with the automatic bandwidth

Figure 2.14: Parametric vs. non-parametric risk forecasts

This figure compares one day ahead 97.5 % ES forecasts by a GARCH(1,1) model to the corresponding risk forecasts obtained via historical simulation at two different time points (before and at the peak ("Black Monday II") of the 2020 stock market crash due to the COVID-19 pandemic). Each point corresponds to a constituent of the S&P Composite 1500 Index. Additionally, the corresponding conditional quantile curves estimated by the leveraging estimator for the 5 % (blue), 25 % (green), 50 % (red), 75 % (green), and 95 % (blue) quantile level are included. For details on the calculation of the quantities, see Sections 2.6.1 and 2.6.2.



(a) January 2, 2020

(b) March 16, 2020

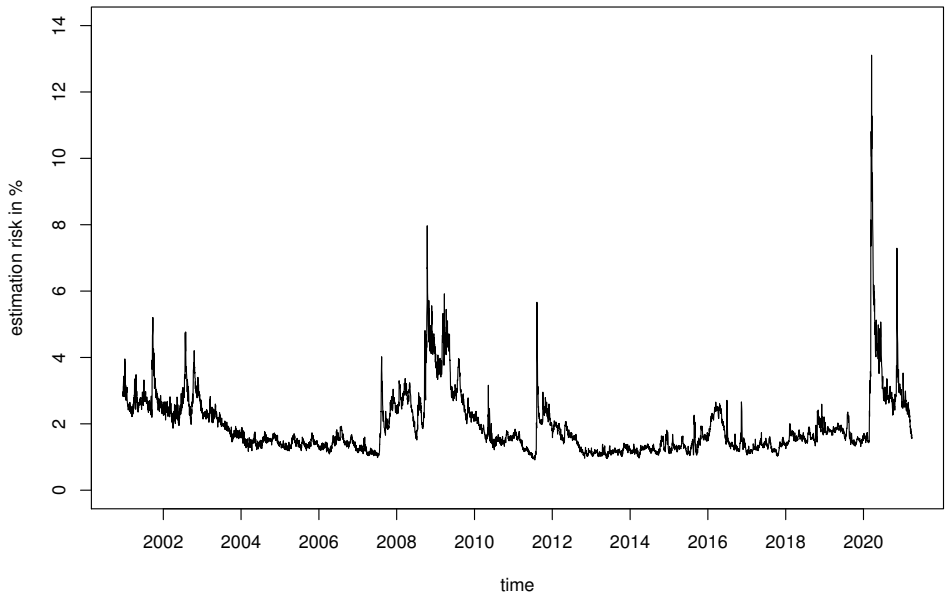
is in line with Christoffersen and Gonçalves (2005), who find that ES predictions are typically less accurate than VaR predictions. Among the four considered models, the GARCH model exhibits the lowest average estimation risk (1.94 % for ES), followed by the T-GARCH model (1.97 % for ES), the GJR-GARCH model (2.01 % for ES), and the EGARCH model (2.23 % for ES). The order of the models is the same for the VaR.[44] Of course, the GARCH-type models assign more weight to recent observations via some kind of exponential weighting scheme while historical simulation assumes

selection procedure described in Newey and West (1994).

[44]The estimation risk associated to the GARCH model is significantly lower at the one percent level than that of all the other models in pairwise comparisons for both ES and VaR. The only exception is the comparison between the GARCH and the T-GARCH model for ES. On the other hand, estimation risk for the EGARCH model is significantly larger at the 1 % level than that of all the other models in pairwise comparisons for both ES and VaR.

Figure 2.15: Estimation risk of the GARCH model for the US equity market over time

This figure shows the daily estimation risk of one day ahead 97.5 % ES forecasts by a GARCH(1,1) model for the constituents of the S&P Composite 1500 Index between December 2000 and March 2021. Estimation risk is calculated as the average interquartile range of GARCH ES forecasts conditional on the corresponding forecasts obtained via historical simulation. More detailed, at a given date the 75 % and the 25 % conditional quantile curves of the parametric (GARCH) risk forecasts conditional on the non-parametric (historical simulation) forecasts are calculated based on the constituents of the S&P Composite 1500 Index using the leveraging estimator (see Figure 2.14). Estimation risk at a particular day is then computed as the difference between these two conditional quantiles averaged over all index constituents.



equal weights for all observations. However, all GARCH-type models in our study are of order (1,1) such that this difference between GARCH-type models and historical simulation cannot explain the differences in the average iqr between the GARCH-type models. We conclude that the models indeed exhibit significant differences in estimation risk.

So far, we have conditioned the parametric VaR and ES forecast on their non-parametric counterparts obtained via historical simulation. In a further analysis for the GARCH model we instead condition the parametric forecasts on the realized volatility (RV) of the respective stock. This is done to study the robustness of our measure of estimation risk with regard to the variable we condition on. To provide a more complete picture we also consider different lengths of moving windows (20 and 250 days

Table 2.5: Estimation risk of various GARCH-type models for the US equity market

This table compares the estimation risk for one day ahead forecast of 97.5 % ES and 99 % VaR by various GARCH(1,1)-type models to each other. Estimation risk is measured in terms of the interquartile range of the conditional quantiles of the parametric risk forecasts (by the GARCH-type models) conditional on the corresponding non-parametric forecasts (obtained via historical simulation). At a given date, conditional quantile curves are estimated by the leveraging estimator for the cross section of risk estimates of the S&P Composite 1500 Index constituents. Subsequently, the associated interquartile ranges are averaged over all constituents. This yields our measure of estimation risk. The table presents summary statistics (minimum (min), median, mean, maximum (max), standard deviation (sd)) of the resulting time series over the period December 2000 until March 2021.

		Estimation risk (in %)				
		min	median	mean	max	sd
	model					
97.5 % ES	GARCH	0.9113	1.6098	1.9387	13.1087	0.9340
	EGARCH	0.9681	1.7994	2.2323	12.8004	1.1648
	GJR-GARCH	0.9502	1.6733	2.0054	15.3035	0.9961
	T-GARCH	0.8916	1.5774	1.9668	11.5562	1.0007
99 % VaR	GARCH	0.8243	1.4884	1.8041	12.7326	0.9119
	EGARCH	0.9344	1.6749	2.0634	12.1411	1.0611
	GJR-GARCH	0.8516	1.5628	1.8806	15.0015	0.9591
	T-GARCH	0.8786	1.5051	1.8674	11.3164	0.9550

for the calculation of RV, 500 days for historical VaR and ES). Furthermore, we include skewness and kurtosis of a particular stock's return distribution. The results are presented in Table 2.6. The table provides summary statistics of the estimation risk for one day ahead VaR and ES forecast by a GARCH model when conditioning on various covariates. The baseline case is conditioning on the historical simulation VaR and ES calculated over a moving window of 250 trading days, respectively.

For the ES, we observe that average iqr substantially increases when extending the moving window to 500 trading days (from 1.94 % (250 days) to 2.13 % (500 days)). This increase is not surprising as the GARCH model assigns more weight to recent observations while in the historical simulation all 500 observations of the longer moving window enter into the calculation with equal weight. These differences in the weights implicitly assigned to the observations manifest themselves in deviations between the GARCH and historical simulation risk forecast, finally leading to higher values of our

Table 2.6: Estimation risk of GARCH risk forecasts when conditioning on further variables

This table provides summary statistics (minimum (min), median, mean, maximum (max), and standard deviation (sd)) for the time series of estimation risk obtained by conditioning on various covariates. The quantities are calculated as in Table 2.5 but only for the GARCH(1,1) model and when conditioning on further covariates. For comparison, we again include the results obtained by conditioning on the non-parametric 97.5 % ES and 99 % VaR estimates obtained via historical simulation over a moving window of 250 days (d). Realized volatility (RV) is calculated as the standard deviation of daily log-returns, skewness and kurtosis denote the standardized third and fourth moment of the empirical return distribution over a moving window of 250 days.

		Estimation risk (in %)				
		min	median	mean	max	sd
	conditioning variable					
97.5 % ES	ES (hist. simulation, 250 d)	0.9113	1.6098	1.9387	13.1087	0.9340
	ES (hist. simulation, 500 d)	1.1189	1.8945	2.1272	14.5957	0.9360
	RV (20 d)	0.9517	1.7777	1.9545	10.6529	0.6833
	RV (250 d)	0.7694	1.6049	1.9237	14.3572	1.0109
	RV, skewness, kurtosis (250 d)	0.8331	1.5857	1.9145	11.3020	0.9396
99 % VaR	VaR (hist. simulation, 250 d)	0.8243	1.4884	1.8041	12.7326	0.9119
	VaR (hist. simulation, 500 d)	1.0045	1.6847	1.9395	13.8710	0.9019
	RV (20 d)	0.8778	1.5523	1.7127	9.8173	0.6172
	RV (250 d)	0.6874	1.3718	1.6885	13.3811	0.9498
	RV, skewness, kurtosis (250 d)	0.7298	1.3709	1.6921	10.5958	0.8913

measure of estimation risk. This highlights that the absolute values of the measure should not be interpreted on their own but rather in comparison over time or with other models. However, when conditioning on RV over a 20 day or a 250 day moving window, or when conditioning jointly on RV, skewness, and kurtosis (over a 250 day moving window) the average level of estimation risk is relatively stable (values between 1.91 % and 1.95 %) and similar to the baseline case (1.94 %).

For VaR, the average iqr also increases substantially when conditioning on the historical simulation VaR obtained from a moving window of 500 trading days (from 1.80 % (250 days) to 1.94 % (500 days)). When instead conditioning on the RV over a 20 day or a 250 day moving window, or when conditioning jointly on RV, skewness, and kurtosis (over a 250 day moving window) the average level of estimation risk varies only slightly (between 1.69 % and 1.71 %) but this time is considerably smaller

than in the baseline case (1.80 %). Again, this highlights that the estimation risk measure should be interpreted in comparison over time or with other models and not on its own. Finally, we find that regardless of the variable we condition on ES always exhibits a higher estimation risk than VaR. Again, this is in line with Christoffersen and Gonçalves (2005).

Confidence intervals

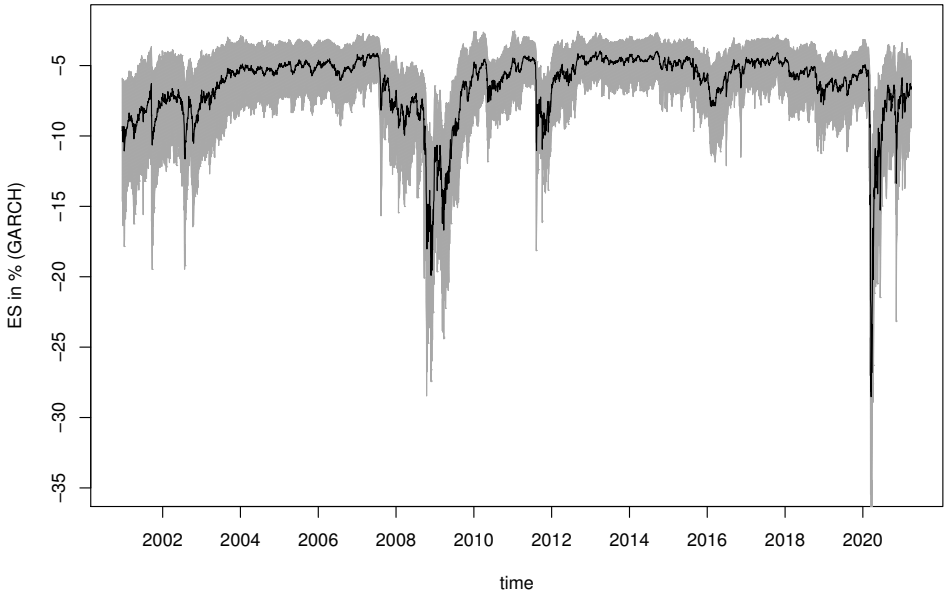
Instead of deriving a measure of estimation risk, the conditional quantile estimates can directly be used to determine confidence intervals for the risk forecasts of a particular stock at a particular day. Over time, this yields confidence bands. Figure 2.16 provides average ES forecasts along with the average 90 % confidence band obtained from the 5 % and 95 % quantile of GARCH ES forecasts conditional on historical simulation ES estimates averaged over all constituents of the S&P Composite 1500 Index.

Figure 2.17 illustrates risk forecasts along with the 90 % confidence band for a single stock (Apple Inc.). Again, the confidence band is constructed from the conditional 5 % and 95 % quantile over time. The ES forecasts for the Apple stock lie outside the confidence band several times.[45] In these cases, the ES forecasts for the Apple stock exceed the 95 % quantile or fall below the 5 % quantile of the risk forecasts for *comparable stocks* (in terms of their non-parametric ES predictions). This might be an indicator for erroneous risk forecasts for the Apple stock at these days. However, as discussed previously, in GARCH-type models more recent observations obtain more "weight" than earlier ones. Consequently, values outside the confidence band are most probably a result of recent evolutions in the Apple stock price that are not reflected to the same extent in the risk forecasts obtained via historical simulation. Either way, the confidence bands can provide valuable signals that a particular risk forecast (at a given

[45]The conditional quantiles are determined over the cross-section of stocks at a particular day such that the risk forecasts (VaR or ES) of 90 % of the stocks lie within their respective confidence interval. However, for a single stock over the time dimension there is no guarantee that risk forecasts lie inside the confidence band 90 % of the time (this is only fulfilled on average over all stocks). Indeed, the ES forecasts for the Apple stock exceed the upper bound in 7.0 % and fall below the lower bound in 7.7 % of the cases.

Figure 2.16: Average ES forecasts with average 90 % confidence band

This figure shows average one day ahead forecasts of 97.5 % ES obtained from a GARCH(1,1) model for the constituents of the S&P Composite 1500 Index between December 2000 and March 2021. The gray-shaded area is the average 90 % confidence band. More detailed, at a particular date the upper (lower) boundary of the confidence interval is obtained by taking the average of the 95 % (5 %) quantile of GARCH ES forecasts conditional on forecasts obtained via historical simulation over all constituents of the S&P Composite 1500 Index (see also Figure 2.14). The conditional quantile estimates are based on the leveraging estimator.



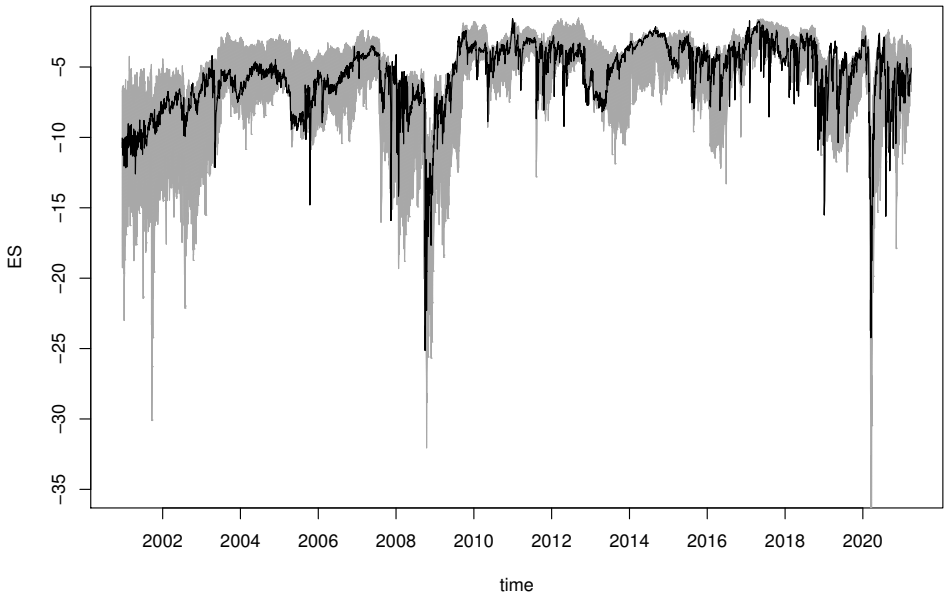
day for a given stock) should be given special attention by carefully interpreting its value or even reestimating it. At the same time, the width of the confidence interval at a particular time point can provide information on the estimation risk for a particular risk forecast.

2.7 Conclusion

This paper proposes a new estimator of conditional quantiles that is based on optimal quantization and leveraging, two approaches from the field of machine learning. Therefore, we build an ensemble of quantization-based estimators (Charlier et al., 2015b) by iteratively combining the ensemble members such that the performance of the aggregated estimator is improved in each step. This yields an estimator with variable

Figure 2.17: ES forecasts with 90 % confidence band for the Apple stock

This figure provides the one day ahead GARCH(1,1) forecasts of the 97.5 % ES for the Apple stock between December 2000 and March 2021 along with a 90 % confidence band. At a particular date, the upper and lower boundary is obtained from the 95 % and 5 % quantile of the ES (GARCH) forecasts of similar stocks (in terms of their non-parametric ES calculated via historical simulation). More detailed, we estimate the 95 % and 5 % conditional quantile curves of GARCH forecasts conditional on historical simulation forecasts with the leveraging estimator in the cross section of all S&P Composite 1500 Index constituents. As at a particular date the confidence interval is determined over the cross section of index constituents, ES forecasts for the Apple stock can exceed (fall below) the upper (lower) boundary. For more information, see Section 2.6.3.



bandwidth that adapts both to the distribution of the covariates and of the response variable. We introduce a data-driven procedure for determining the hyperparameters that is based on the empirical check-loss calculated via cross-validation. Furthermore, we provide convergence results for the proposed leveraging estimator.

In an extensive simulation study we compare the leveraging estimator to the quantization-based estimator and various competitors. In the univariate case, the proposed estimator produces smooth quantile curves that adapt well to the true curves, even in the edges of the support of the covariate. The estimator generalizes naturally to multiple dimensions. We study up to four-dimensional covariates and again yield competitive ISEs.

In an empirical study, we analyze the estimation risk associated with VaR and ES models in the broad US equity market (S&P Composite 1500 Index constituents) based on return data from January 2000 until March 2021. For this purpose, we apply the leveraging estimator to VaR and ES forecasts obtained by various GARCH-type models and condition these estimates on their non-parametric counterparts obtained via historical simulation. The estimation risk for a given model is defined as the average iqr of the conditional quantile curves. This approach of non-parametrically determining estimation risk without relying on Monte Carlo methods is new to the literature. It also yields non-parametric confidence bands for VaR and ES predictions at the stock level. We find that there is substantial variation in estimation risk over time with especially high values during times of financial turmoil. Furthermore, the results suggest that among the considered models the GARCH model exhibits the lowest estimation risk while the EGARCH model is associated with the highest. In general, estimation risk for ES is higher than for VaR, both across models and across different conditioning variables. The results for ES are robust to conditioning on RV instead of the non-parametric risk measures obtained via historical simulation while for the VaR we obtain somewhat lower values of the average iqr when conditioning on the RV. This reliance on a non-parametric benchmark constitutes the main weakness of our approach for measuring estimation risk. However, our proceeding provides a new approach for capturing estimation risk from the cross-section of stock returns and illustrates the applicability of the proposed leveraging estimator.

The key features of the leveraging estimator, being non-parametric and applicable in multiple dimensions, make it an interesting choice for many other applications. More generally, conditional quantile estimation can be used to replace standard methods like (conditional) portfolio sorts to use the available data more efficiently. Moreover, conditional quantiles can provide valuable insights into the relationship between dependent and explaining variables that go well beyond the conditional mean and therefore should become a standard tool in empirical research.

Chapter 3

Cross-Section of Option Returns and the Volatility Risk Premium

3.1 Introduction

Volatility is the single most important characteristic of a stock driving the prices of corresponding option contracts. Returns on stock options should thus carry a risk premium for changes in volatility. Likewise, any misestimation of an underlying stock's volatility and its dynamics should lead to a mispricing of options which traders can exploit. Yet, despite its ubiquity in option pricing models, the role volatility and its mispricing as well as volatility risk play for the cross-section of option returns remains unclear: while some studies have found no evidence for the existence of a volatility risk premium (see, e.g., Carr and Wu, 2009, Driessen et al., 2009), others have shown that volatility risk as well as volatility itself significantly affect the cross-section of option returns (see, e.g., Bollerslev et al., 2009, Goyal and Saretto, 2009, Cao and Han, 2013, Cao et al., 2019, Hu and Jacobs, 2020).

In this paper, we empirically test for the existence of a volatility risk premium (VRP) in the cross-section of option returns. We start our analysis by first documenting potential biases in analyses of the volatility-return relation that arise when relying on standard methods in asset pricing: portfolio sorts and cross-sectional regressions. As

an alternative, we propose to use non-parametric methods from the field of machine learning to estimate conditional quantile curves of implied stock option volatilities. We condition on a number of characteristics that would otherwise cloud the effect of implied volatility on option returns. Most importantly, we control for the stock's realized volatility and option moneyness. Doing so helps us to carve out the volatility risk premium in the cross-section of option returns.

Using the cross-section of option returns for US equities between January 1996 and June 2019, we find that call and put option portfolio returns exhibit a strong relation with the volatility risk premium. We sort options on their implied volatility *conditional* on their realized volatility. This yields portfolios with increasing deviations between realized and implied volatilities with average levels of realized volatility remaining constant. We use this to proxy for the volatility risk premium. A strategy that is long (short) in high (low) deviations between realized and implied volatilities yields returns that are both economically and statistically significant. This result holds for call and put delta-hedged and raw option returns for both at the money (ATM) options and options of arbitrary moneyness. For example, average monthly delta-hedged returns of 1-month ATM options are 2.0 % for call and 1.7 % for put contracts with (monthly) Sharpe ratios of 0.842 and 0.796, respectively.

By sorting options on their implied volatility conditional on realized volatility *and* option moneyness, we can easily extend our trading strategy to options of arbitrary moneyness while eliminating potential biases arising from systematic differences in realized volatility or option moneyness (and thus option liquidity). Again, this yields delta-hedged and raw option returns that are highly economically and statistically significant. For example, average monthly delta-hedged returns from a long-short trading strategy based on 1-month options of arbitrary moneyness are 2.4 % for call and 2.5 % for put contracts with (monthly) Sharpe ratios of 0.816 and 0.844, respectively. Our results are robust to controlling for further moments of the underlyings' return distribution, alternative estimators of conditional quantiles, reasonable transaction costs, different levels of trading volume, the expansion of the long-short portfolios to less

extreme options, and the inclusion of options on dividend-paying stocks.

Why do we find such strong evidence for the existence of a VRP when results of previous empirical studies have been ambiguous at best? One possible answer lies in our proposed use of non-parametric methods to form factor-mimicking portfolios based on characteristics while conditioning on a set of control variables. The standard technique in empirical finance for this purpose has been the sorting of portfolios on certain characteristics of assets. It is frequently used to test the assumption of pricing models that expected asset returns are a monotonic function in one or more idiosyncratic characteristics. This common practice of forming uni- or multivariate fractiles dates back to seminal papers on the cross-section of equity returns which, among others, include the works of Basu (1977), Banz (1981), de Bondt and Thaler (1985), Jegadeesh (1990), Fama and French (1992), and Jegadeesh and Titman (1993). Since then, portfolio sorting has been a methodological mainstay in empirical asset pricing, because it does not require the assumption of a linear relation between expected returns and characteristics, and because differences in the returns on the top and bottom fractile portfolios are easily interpreted as the profits from an implementable trading strategy. As intuitive as it may be, however, portfolio sorting does not come without shortcomings. While univariate portfolio sorts on one characteristic do not allow the economist to control for other asset characteristics, multivariate (conditional) sorts quickly become unfeasible for more than two characteristic-based factors due to the curse of dimensionality.

Our approach, in contrast, uses non-parametric methods from the field of machine learning to estimate the conditional quantile curves of implied volatilities while at the same time controlling for several characteristics. Thus it possesses several appealing features that should make it favorable to standard portfolio sorts and non-/semiparametric regression methods alike. First, using machine learning algorithms to estimate quantile curves allows for the data-efficient non-parametric modeling of the multivariate distribution of asset returns and characteristics. In contrast to standard (conditional) portfolio sorts, our method should alleviate at least in part the concern of "empty portfolios" which eventually arises when sorting on too many characteristics

(see, e.g., Goyal, 2012).

Second, as a remedy to the "empty portfolio" problem, many researchers have additionally performed multivariate regressions to test whether a certain characteristic is priced. While these cross-sectional regressions allow the inclusion of a large number of covariates, they also suffer from two drawbacks that make them less appealing in our setting. Standard as well as semiparametric regression methods (see, e.g., Connor and Linton, 2007, Connor et al., 2012, Cattaneo et al., 2020) assume additive separability between the explanatory variables in asset pricing models (in addition to a linear relation between characteristics and returns). Employing such standard models leads to a severe bias in the measurement of the VRP due to the nonlinear nature of the relation between implied and realized volatility. Moreover, results from cross-sectional regressions only yield information on long-short strategies that involve trading in *all securities* with potentially highly varying portfolio weights. Our proposed non-parametric approach circumvents both these problems: we make no assumption on the functional form of the relation between implied and realized volatility, and our approach yields a trading strategy that can easily be implemented.

Our paper is related to an increasing number of empirical studies on the relation between option returns and characteristics of the underlying stocks.[46] Coval and Shumway (2001) were among the first to look at the cross-section of expected option returns. Studying index options, they find that systematic stochastic volatility is priced in option returns. In a related study, Driessen et al. (2009) show that correlation risk is priced in both index and individual options but find no evidence for the existence of a VRP. Conversely, using model-free estimates of implied volatilities, Bollerslev et al. (2009) show in their study that stock returns include a VRP. Finally, Huang et al. (2019) study the pricing of volatility of volatility risk in index options. All of the studies, however, do not test for the existence of a VRP in expected option returns and usually concentrate on index options rather than options on individual equities.

[46] Recent studies on the pricing of stock and option characteristics in the cross-section of expected option returns include, but are not limited to, the studies by Baele et al. (2019), Cao et al. (2019), Andreou and Ghysels (2020), Eisdorfer et al. (2020), Cao et al. (2021).

More recent studies have concentrated on the effects of volatility, volatility risk, and volatility mispricing on the cross-section of expected option returns. In one of the first studies in this field of research, Goyal and Saretto (2009) show that large differences between realized and implied volatilities for at-the-money options are associated with economically and statistically significant monthly returns. Their use of linear differences to proxy for potential volatility mispricing, however, ultimately leads to a portfolio strategy that also (at least partially) invests in realized volatility. Our approach to measure the volatility risk premium by the use of conditional quantile curves builds on their study. After controlling for realized volatility and allowing for arbitrary moneyness, our results confirm the initial findings by Goyal and Saretto (2009) and make an even stronger case for the existence of a VRP. This is important and reassuring at the same time, as more recent studies have shown that idiosyncratic (see Cao and Han, 2013) and realized volatility (see Hu and Jacobs, 2020) by themselves drive expected option returns, questioning previous findings on the effect of volatility mispricing on option returns.

Methodologically, our paper is related to a small but growing number of papers that aim to improve standard methods in empirical asset pricing. For example, Patton and Timmermann (2010) were among the first to point out the shortcomings of portfolio sorts and standard tests of monotonicity in asset pricing. Similarly, Connor and Linton (2007), Connor et al. (2012), and Cattaneo et al. (2020) propose semi- and non-parametric models as alternatives for portfolio sorts and cross-sectional regressions. In contrast to our study, however, their models usually concentrate on nonlinear relations between returns and characteristics, and not between covariates. Moreover, none of these studies look at the cross-section of option returns. We complement this field of research by proposing the use of conditional quantile curves as an alternative to traditional portfolio sorts and applying it for the first time to expected option returns. Finally, our paper is also related to a growing number of studies that propose the use of machine learning algorithms in empirical asset pricing. For example, Moritz and Zimmermann (2016) use tree-based conditional portfolio sorts and model-averaging to

identify the most relevant factors of the famous "factor zoo" (cf. Cochrane, 2011) that drive stock returns, while Gu et al. (2020) employ trees and neural networks to forecast returns. Our work complements these studies by proposing the use of data-efficient machine learning algorithms to form conditional portfolio sorts in high dimensions.

The rest of the paper is organized as follows. The next section 3.2 discusses the measurement of the volatility risk premia in option returns as well as our methodology. Section 3.3 presents our empirical study. We discuss robustness checks in Section 3.4. Section 3.5 concludes.

3.2 Capturing the volatility risk premium

3.2.1 Volatility risk premium and volatility mispricing

We start our analysis by revisiting common definitions for the volatility risk premia of individual stocks from the related literature. For example, Cao and Han (2013) define the volatility risk premium of stock i in time period t as

$$VRP_{i,t} = RV_{i,t} - IV_{i,t} \tag{3.1}$$

where $RV_{i,t}$ is realized return volatility and $IV_{i,t}$ is the implied volatility of the stock extracted from corresponding options (see also Jiang and Tian, 2005, Bollerslev et al., 2009, Carr and Wu, 2009, Driessen et al., 2009). Similarly, Goyal and Saretto (2009) consider the log difference

$$VMP_{i,t} = \log RV_{i,t} - \log IV_{i,t} \tag{3.2}$$

between realized (historical or current) and implied volatility and interpret large values of $VMP_{i,t}$ as indicative of volatility mispricing.

From the definitions in Equations (3.1) and (3.2) it becomes clear that tests of the existence of a VRP in option returns critically depend on whether one is able to con-

trol for the level of realized volatility.[47] To this end, previous studies have traditionally relied on (conditional) portfolio sorts and cross-sectional regressions. However, controlling for additional asset characteristics (such as realized volatility) via conditional portfolio sorts quickly becomes infeasible due to the curse of dimensionality (Stone, 1980).[48] Beyond that, portfolio sorts exhibit further shortcomings, some of which we want to illustrate in the following Section 3.2.2. As a remedy, we advocate for replacing conditional portfolio sorts with conditional quantiles, of which we discuss the details in Section 3.2.3.

3.2.2 Replacing portfolio sorts by conditional quantiles

To illustrate the potential weaknesses of conditional portfolio sorts, we simulate 200 observations (x_i, y_i) according to $Y = \frac{1}{2} \cdot X + \frac{1}{10} \cdot X \cdot \epsilon$, where ϵ denotes the standard normal distribution and X follows a $Beta(5, 5)$ distribution with X and ϵ being statistically independent from each other. That is, we assume a linear relation between X and Y with heteroskedastic errors. Figure 3.1 illustrates quintile portfolios from a conditional double-sort of the data (first on x, then on y).

We first sort the observations into five portfolios according to their x-values, highlighted in Figure 3.1 as the strips bounded by the blue lines, and then sort observations on y-values within each of those five portfolios, indicated by the areas between the grey and dashed horizontal lines.[49] Based on this double-sort, we next form long and short portfolios of observations with low and high y-values, respectively, while at the same time controlling for x. Within each of portfolio 1 to 5, we choose the observations corresponding to the y-values in the lower quintile for the long and in the upper

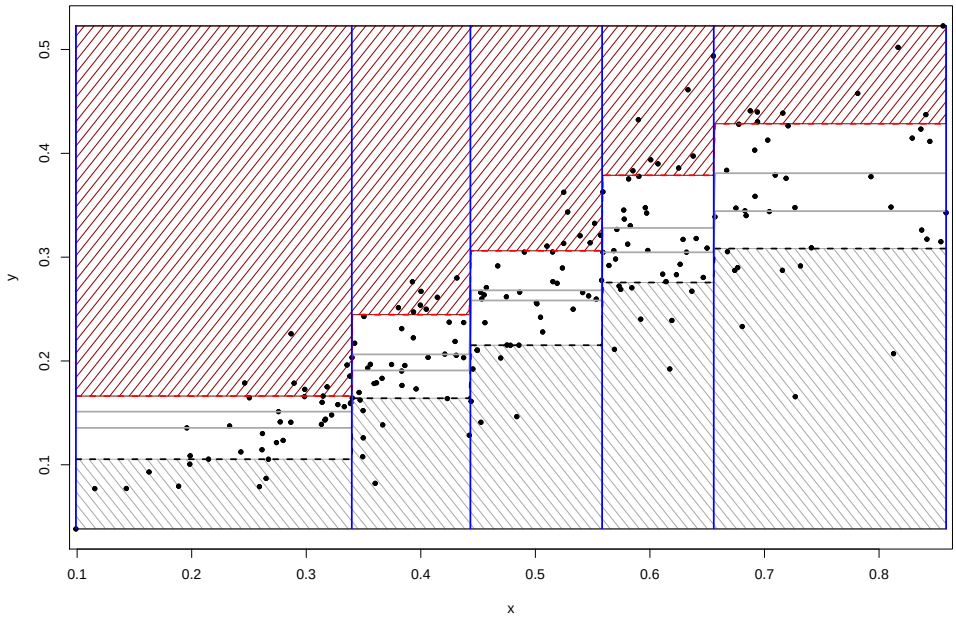
[47] The findings of Hu and Jacobs (2020) show that option raw returns are significantly affected by realized volatility.

[48] For example, while sorting on one characteristic and controlling for another via a double-sort might still be practicable, controlling for two or more covariates via triple-, quadruple-sorts etc. is almost always not possible. Note that independent sorts are not suited to control for characteristics.

[49] To obtain portfolios of equal size, we assign the observations with x-values below the (unconditional) empirical 20 % quantile of all x_i's into portfolio 1, observations between the 20 % and 40 % quantile into portfolio 2, etc. In each of these five portfolios, we then sort on y according to the (unconditional) empirical 20 %, 40 %, etc. quantile of all y-values *within* portfolio 1. This is subsequently repeated for portfolios 2 to 5.

Figure 3.1: Illustration of conditional portfolio sorts

This panel illustrates a conditional double-sort and derived conditional quantile curves based on 200 simulated observations according to $Y = \frac{1}{2} \cdot X + \frac{1}{10} \cdot X \cdot \epsilon$ where ϵ and X follow a standard normal and a $Beta(5, 5)$ distribution, respectively. We first sort observations into 5 bins based on their x-values (blue lines according to the 20 %, 40 %, 60 %, and 80 % (unconditional) quantile of all x-values). Subsequently, in each bin we further sort on the y-value. The dashed black and red lines mark the 20 % and 80 % quantile of y conditional on each of the 5 bins. The grey lines mark the 40 %, 60 %, and 80 % quantiles. Observations in the shaded areas are those that lie in the lower quintile in each of the five bins (i.e., below the black dashed line) and in the upper quintile (i.e., above the red dashed line), respectively.



quintile for the short portfolio, which are the observations that lie in the shaded areas of Figure 3.1.

Figure 3.2 compares the quantile curves implied by this conditional double-sort sort to the true quantiles of Y conditional on X. For example, combining the 20 % quantiles of each of the five portfolios, we obtain a step function (black dashed line) which can be understood as an approximation to the true conditional 20 % quantile curve of Y *given* X (black solid line). The same holds for the conditional 80 % quantile curves (red lines). It is evident from the figure that portfolio sorts provide a quite coarse approximation to the true conditional quantile curves. As a consequence, the long and

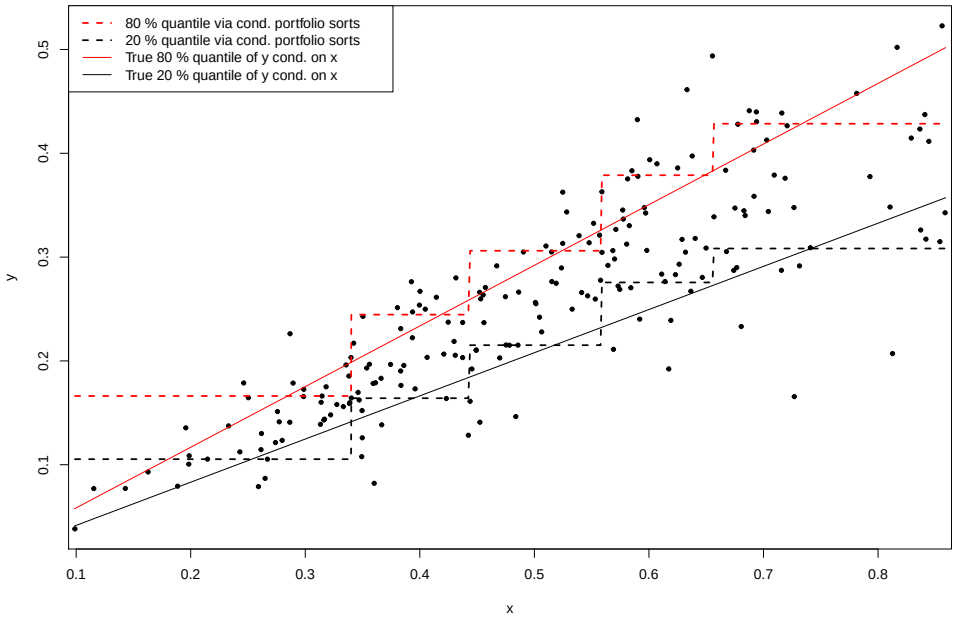
short portfolios derived from the double-sort exhibit systematic differences in x, which we illustrate in Figure 3.3. The black dashes on the x-axis of Figure 3.3 correspond to observations in the long portfolio while the red dashes correspond to observations in the short portfolio. If controlling for x had been successful, the black and red dashes would be mixed randomly along the x-axis, as this would indicate that x is no longer directly influencing the long-short portfolio construction. However, there are various clusters of red and black dashes with a clear trend of observations in the long portfolio tending to lower and observations in the short portfolio tending to higher x-values. That is, although long and short portfolios were build on a double-sort to control for x, the results of a long (short) strategy that is low (high) in y might still be biased by systematic differences in x.

One possibility to mitigate this bias is to simply increase the number of portfolios and sort the x- and y-values into, e.g., 10 portfolios each yielding 100 double-sorted portfolios. However, in such case, we still would approximate the true conditional quantile curves (which are linear in this example) with piecewise constant step functions, which is far from optimal. In addition, if we wanted to control for an additional covariate z, we would be required to do a triple-sort leading to 1000 portfolios. However, this is infeasible based on only 200 observations as this would lead to a large number of empty portfolios.

Based on the true conditional quantile function, as illustrated in Figure 3.2, we are able to form long (short) portfolios with low (high) y-values while at the same time *perfectly* controlling for x. Of course, in practice, we do not know the true conditional quantile curves. Instead we have to estimate them based on the data at hand. However, conditional portfolio sorts do not use the data efficiently when deriving conditional quantile curves. This is highlighted in Figure 3.2, where the true quantile curves are approximated with piecewise constant step functions requiring a large number of portfolios to achieve an acceptable fit. As a consequence, one can usually control for only one or at most two characteristics. As an alternative to portfolio sorts, there exist various methods specifically designed for the estimation of conditional quantiles providing

Figure 3.2: Comparison of double-sorts versus conditional quantile curves

This figure compares the 20 % and 80 % quantiles (black and red dashed line) of Y given X according to a double-sort based on 200 simulated observations according to $Y = \frac{1}{2} \cdot X + \frac{1}{10} \cdot X \cdot \epsilon$ where ϵ and X follow a standard normal and a $Beta(5,5)$ distribution, respectively, to the *true* conditional quantile curves of Y given X. Note that the illustrative example is based on simulated data for which the true conditional quantile curves are known. For example, the 20 % conditional quantile function is linear and given by $q_{20\%}(x) = (0.5 + 0.1 \cdot q^{norm}_{20\%}) \cdot x$ where $q^{norm}_{20\%}$ denotes the (unconditional) 20 % quantile of the standard normal distribution.

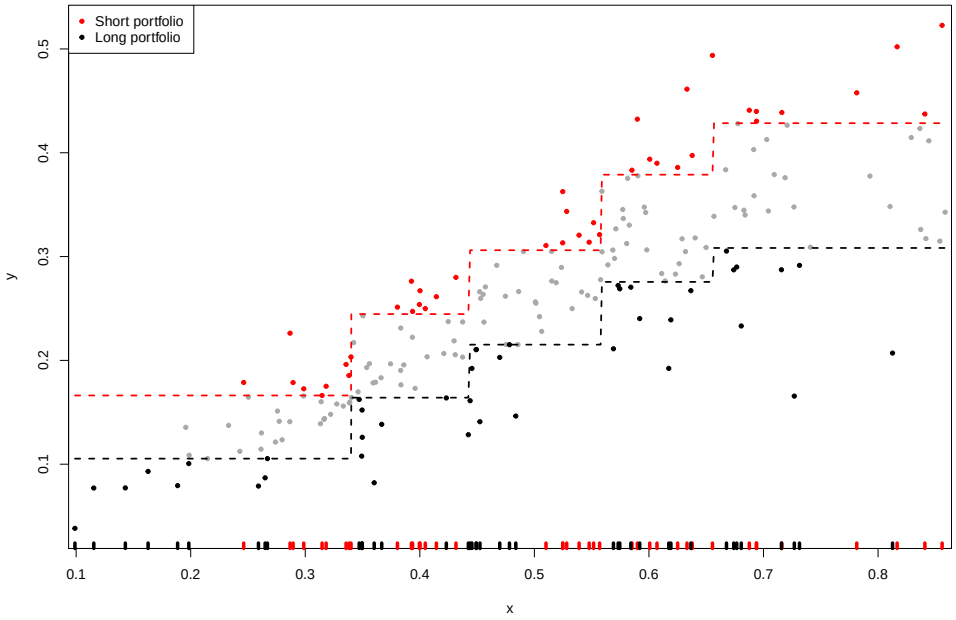


a more data-efficient way of estimation without imposing restrictions on the functional form of the quantile functions. In addition to the advantages mentioned above, this allows for the inclusion of more control variables.[50]

[50] Of course, due to the curse of dimensionality the number of variables one can efficiently control for with our non-parametric approach will still be limited to a low single figure in most applications. However, our approach can be seen as a way of extending the applicability of portfolio sorts to control for more covariates. For example, we later want to condition on options' implied volatility as well as realized volatility, moneyness, and skewness of the underlyings' return distribution, which would require the usage of quadruple sorts. This is infeasible for the major part of our option sample in the empirical study, but we can easily consider these control variables within our proposed methodology.

Figure 3.3: Long-short portfolio derived from using a double-sort to control for x

This figure illustrates limitations on forming a long-short portfolio via double-sorts that is long (short) in low (high) values of y while controlling for x. The black and red dashed lines are derived from conditional portfolio sorts (first on x, then on y) based on 200 simulated observations according to $Y = \frac{1}{2} \cdot X + \frac{1}{10} \cdot X \cdot \epsilon$ where ϵ and X follow a standard normal and a $Beta(5, 5)$ distribution, respectively. Securities corresponding to the observations below (above) the black (red) dashed line enter into the long (short) portfolio.



3.2.3 Estimation of conditional quantiles

For later reference, we now provide a brief introduction to the estimation of conditional quantiles. It is standard practice to (linearly) approximate the conditional *mean* function $x \mapsto E(Y|X = x)$. This is done by minimizing the squared errors. By instead minimizing the absolute errors one can derive the conditional *median* function, i.e., the conditional 50 % quantile function. This result generalizes naturally to conditional quantiles at arbitrary confidence levels. The conditional α quantile function is obtained by minimizing the so-called check-loss of the residuals, where the check-function ρ_α

is given as

$$\rho_\alpha(z) := \begin{cases} -(1-\alpha)z & \text{for } z \leq 0, \\ \alpha z & \text{for } z > 0. \end{cases} \tag{3.3}$$

That is, depending on the confidence level α, the residuals (deviations of observations from the estimated conditional quantiles) enter into the error term that has to be minimized asymmetrically with weights $1-\alpha$ and α, respectively. Building on this result Koenker and Bassett (1978) introduce linear conditional quantile estimators. Since the appearance of this seminal paper various other estimators have been proposed. In particular, non-parametric estimators appear promising as they do not require to make any assumptions about the functional form of the quantile curves, see, e.g., Kraus and Czado (2017) and the references therein.[51]

The problem of estimating conditional quantiles non-parametrically has been addressed with different techniques. For example, Kraus and Czado (2017) propose an estimator based on likelihood optimal D-vine copulas, in the following referred to as the *copula estimator*. The estimator models multivariate dependencies based on so-called pair-copula constructions.

Starting with the k-nearest neighbor (kNN) estimator, conditional quantile estimation has also been addressed within the realm of (unsupervised) machine learning (see Bhattacharya and Gangopadhyay, 1990). More recently, Charlier et al. (2015b) introduced an estimator that is derived from the concept of optimal quantization. Loosely speaking, this algorithm efficiently uses the data at hand by identifying clusters of observations of covariates via unsupervised machine learning and deriving empirical quantiles of the response variable within the clusters (see Charlier et al., 2015b,a, for details). In the following we will refer to this estimator as the *quantization estimator*.

Building on the quantization estimator, we employ a new estimator by employing a machine learning technique called *leveraging*. Leveraging is an ensemble technique

[51] Results from our empirical study highlight the necessity to account in particular for non-linear dependencies between implied and realized volatility as well as heteroskedasticity. Note that while in general we speak of conditional quantiles, in the case of conditioning on only one variable, we get conditional quantile *curves*. Therefore, we will use these terms interchangeably.

very similar to boosting[52] which according to Meir and Rätsch (2003) "combine[s] simple 'rules' to form an ensemble such that the performance of the single ensemble member is improved". For this purpose we define our *leveraging estimator* in an iterative manner such that in each iteration step we give more weight to those observations for which the latest conditional quantile estimates produce a higher estimation error and less weight to those observations associated with a lower estimation error. Errors are calculated based on the check-function from Equation (3.3).[53]

We use the leveraging estimator for our main analysis but also include results for the quantization and copula estimator as robustness checks.[54] Results from the leveraging estimator are based on our own implementation while for the quantization and the copula estimator we rely on the QuantifQuantile and vinereg (with non-parametric pair copulas) R-package by Charlier et al. (2015c) and Nagler (2020), respectively. All computations were performed on the Big-Data-Cluster Galaxy provided by the University Computing Center at Leipzig University.

3.3 Empirical study

As discussed above in Section 3.2.2, conditional quantile based portfolio sorts have several advantages over simple (conditional) portfolio sorts. Therefore, we build on this approach to study the VRP in the cross-section of options while controlling for the level of realized volatility. We compare our results to those obtained by sorting on the log-difference of RV and IV. In the construction of our data sample as well as in the corresponding trading strategies we closely follow Goyal and Saretto (2009).

[52] As the concepts of boosting and leveraging are very similar both terms are often used interchangeably in the literature. However, we follow Duffy and Helmbold (2002) and restrict usage of the term boosting to algorithms proved to fulfill a so-called *Probably Approximately Correct (PAC)* learning - property and use the term leveraging for all other related ensemble learning techniques.

[53] More details on the construction of the estimator along with an extensive simulation study can be found in Chapter 2.

[54] Although there is a variety of conditional quantile estimators, the number of methods that can account for two or more covariates is substantially lower. In addition, our findings indicate that conditional quantile curves in an equity option sample within our empirical study are best tackled by non-parametric methods.

3.3.1 Sample construction

The sample period is from January 1996 to June 2019. Data on US equities (including prices, closing bid and ask quotes, and returns) are retrieved from the Center for Research in Security Prices (CRSP). Option data are obtained from the OptionMetrics IvyDB US database. The data include information on the entire US equity option market (American options) covering in particular closing bid and ask quotes along with option implied volatilities (IV) and greeks (delta, gamma, vega).[55]

For our main empirical analysis we focus on the cross-section of equity options that are at the money (ATM) and one month away from expiration since they are the most frequently traded ones (cf. Goyal and Saretto, 2009). In further analyses we also include in the money and out of the money options. Every month, we form portfolios based on information from the first trading day after monthly option expiration.[56]

To minimize the impact of recording errors, we apply several standard filters to the data. Following Goyal and Saretto (2009) we exclude all observations with an ask price lower than the bid price, a bid price equal to zero, or a bid-ask spread below the minimum tick size.[57] We further remove prices that violate arbitrage bounds. Following Hu and Jacobs (2020), we exclude all call options where the ask price exceeds the price of the underlying (S) or where the ask price is below $S - K$ with K denoting the exercise price of the option. Additionally, we exclude all put options with a bid price above the exercise price or a bid price below $K - S$. To avoid errors due to stock splits and re-capitalizations, we remove all options for which the adjustment factor for the exercise price does not coincide with the adjustment factor for the share price. In order to eliminate options with no liquidity, we exclude options with zero open interest (cf.

[55] Implied volatility estimates as well as option greeks are derived from a binomial tree model based on Cox et al. (1979) For further details we refer to the OptionMetrics IvyDB US reference manual.

[56] The expiration day for standard exchange-traded options is the third Friday of the expiration month or the following Saturday.

[57] Before 2007, the minimum tick size is equal to \$0.05 (\$0.10) for options trading below (above) \$3. On January 26, 2007, the SEC introduced the industry wide Penny Pilot Program reducing the minimum tick size for certain equities to \$0.01 (\$ 0.05). This program, today know as Penny Interval Program, has subsequently been extended to cover more equities. For simplicity, we therefore consider a minimum tick size of \$0.05 (\$0.10) before January 26, 2007, and \$0.01 (\$0.05) for all options below (above) \$3 after January 26, 2007, respectively.

Driessen et al., 2009). All equity options in our sample are American. We therefore follow Hu and Jacobs (2020) and remove all options with an ex-dividend date during the remaining life of the option contract to reduce the impact of early exercise.[58] Finally, following Cont and da Fonseca (2002) we exclude all options with moneyness values (defined as the ratio K/S) outside of the interval $[0.5, 1.5]$ to limit numerical uncertainty in computing implied volatilities.

This constitutes our option sample for arbitrary moneyness consisting of 2,280,558 calls and 1,758,895 puts on 9,069 and 8,802 different stocks, respectively, over 282 points in time between January 1996 and June 2019. The number of option contracts varies substantially over time. For example, the number of calls ranges between 1,206 (May 1996) and 16,054 (December 2017) with the number of contracts increasing over time.

In our baseline analysis, we focus on ATM options. Therefore, for every month and each underlying we select the call and put contracts that are closest to ATM but according to Goyal and Saretto (2009) only consider options with moneyness values in the interval $[0.975, 1.025]$. This constitutes our ATM option sample consisting of 267,147 calls and 244,892 puts. There is substantial variation in the number of option contracts over time. For example, the number of calls in the ATM sample varies between 171 (June 1996) and 1683 (January 2018).

We complement our data sample with stock related characteristics. Following Goyal and Saretto (2009), for each month and each stock we calculate the realized volatility (RV) as the standard deviation of the realized daily stock returns over the preceding 12 months.[59] Additionally, we include the third (skewness) and forth (kurtosis) moment

[58] We acknowledge that this controls for early exercise of calls while American puts might still exhibit a premium (Goyal and Saretto, 2009, Barraclough and Whaley, 2012). However, there are several studies arguing that the empirical implications of adjustments for early exercise are small, see, e.g., Boyer and Vorkink (2014).

[59] Volatility is highly mean-reverting. Therefore, large deviations between current realized volatility (e.g., calculated over the current month) and the long-term average (calculated over a 12 month period) are unlikely to persist. Therefore, we consider the 12 month RV to be a realistic estimate of volatility over the remaining life of the respective options, see also Goyal and Saretto (2009) and the discussion therein. Apart from this, building on the RV over the preceding 12 months allows us to compare the results from our study to those obtained by Goyal and Saretto (2009).

of the underlyings' return distribution (over the most recent 12 months).

3.3.2 Summary statistics

We provide summary statistics for implied (IV) and realized volatility (RV) of ATM calls and puts as well as calls and puts of arbitrary moneyness in Table 3.1. The volatilities are annualized. We also include summary statistics for option greeks (delta, gamma, vega) as well as skewness and kurtosis of the underlyings' return distribution. The means are obtained by first taking time-series averages of IV and RV for each stock and then computing the cross-sectional averages of these average volatilities. For the other statistics (median, minimum, maximum, standard deviation, skewness, and kurtosis) we proceed analogously so that the provided statistics can be interpreted as summary statistics of an average stock.

For ATM calls and puts, IV and RV are on average very close to each other. For calls the average IV is 48.6 % compared to an average RV of 49.7 % while for puts the average IV is 50.3 % and the average RV is 50.0 %. However, the distribution of IV is more volatile than the distribution of RV. Additionally, IV is on average more positively skewed and more leptokurtic than RV. The other variables are, on average, very similar to each other, with the exception of options' delta (0.536 for calls and -0.465 for puts). For example, average values for call options are 0.212 (gamma), 3.476 (vega), 0.303 (skewness), and 10.669 (kurtosis).

The differences between ATM options and options of arbitrary moneyness are quite substantial, especially for IV. First of all, the average level of IV is much higher (59.9 % vs. 48.6 % for calls and 59.7 % vs. 50.0 % for puts). In addition, IV is on average more volatile, more positively skewed, and more leptokurtic than in the case of ATM options. This is due to the fact that implied volatilities are in general not constant for different moneynesses (for a given underlying at a given date) but rather exhibit the pattern of a volatility skew, smile, or smirk (see, e.g., Toft and Prucyk, 1997). The variation of IV across moneyness highlights the necessity to control for moneyness

Table 3.1: Summary statistics for the option samples

This table presents summary statistics for implied (IV) and realized volatilites (RV) of ATM calls and puts as well as option contracts of arbitrary moneyness. All options are American and have a maturity of one month. Our ATM sample consists of 267,147 calls and 244,892 puts while the sample of arbitrary moneyness is composed of 2,280,558 calls and 1,758,895 puts. The sample period is from January 1996 to June 2019. IVs and option greeks (delta, gamma, vega) are retrieved from the OptionMetrics IvyDB US database and calculated based on a binomial tree model (cf. Cox et al., 1979). The volatilities are annualized. Option underlyings' returns are retrieved from CRSP. Skewness (skew) and kurtosis (kurt) are computed from realized returns over the most recent 12 months. Means are obtained by first taking the time-series average of IV and RV for each stock and then computing the cross-sectional average of these average volatilities. For the other statistics (median, mininum (min), maximum (max), standard deviation (sd), skewness (skew), and kurtosis (kurt)) we proceed analogously.

			mean	median	min	max	sd	skew	kurt
ATM	**Calls**	IV	0.486	0.465	0.317	0.808	0.143	0.791	3.943
		RV	0.497	0.477	0.365	0.750	0.123	0.684	3.386
		delta	0.536	0.537	0.438	0.629	0.058	-0.094	2.360
		gamma	0.212	0.198	0.117	0.397	0.087	0.687	3.506
		vega	3.476	3.286	1.654	6.467	1.383	0.363	2.703
		skew	0.303	0.290	-1.223	1.909	0.931	0.028	3.513
		kurt	10.669	8.738	5.237	27.044	6.978	1.143	4.830
	Puts	IV	0.503	0.480	0.341	0.823	0.144	0.807	3.926
		RV	0.500	0.479	0.369	0.748	0.124	0.680	3.352
		delta	-0.465	-0.463	-0.557	-0.381	0.055	-0.137	2.274
		gamma	0.201	0.188	0.114	0.365	0.080	0.648	3.356
		vega	3.513	3.322	1.705	6.473	1.405	0.359	2.678
		skew	0.280	0.274	-1.213	1.836	0.936	0.021	3.401
		kurt	10.695	8.880	5.286	26.337	6.950	1.116	4.684
Arbitrary moneyness	**Calls**	IV	0.599	0.551	0.281	1.455	0.224	1.356	6.274
		RV	0.524	0.498	0.357	0.825	0.133	0.629	3.217
		delta	0.585	0.606	0.128	0.968	0.272	-0.148	1.763
		gamma	0.139	0.126	0.023	0.442	0.089	0.990	5.514
		vega	2.301	2.090	0.274	6.347	1.354	0.519	3.214
		skew	0.293	0.278	-1.582	2.406	0.977	0.119	4.261
		kurt	10.953	8.664	4.694	33.465	7.535	1.400	6.300
	Puts	IV	0.597	0.563	0.330	1.210	0.188	1.001	4.752
		RV	0.531	0.506	0.365	0.823	0.134	0.603	3.161
		delta	-0.405	-0.382	-0.808	-0.092	0.221	-0.222	1.921
		gamma	0.153	0.139	0.051	0.403	0.078	0.950	5.106
		vega	2.576	2.352	0.739	6.375	1.297	0.568	3.167
		skew	0.282	0.272	-1.529	2.233	0.958	0.066	4.030
		kurt	10.933	8.734	4.816	31.563	7.307	1.320	5.864

when building portfolios for options of arbitrary moneyness. On average, the calls and puts in our sample of arbitrary moneyness also exhibit a slightly higher RV then ATM calls and puts. This is explained by the fact that at a given date, on average, about half of the underlyings do not enter into the ATM sample because there is no corresponding option contract with moneyness in the interval [0.975, 1.025]. These are stocks with a higher RV. Due to the inclusion of out of the money as well as in the money options into the sample of arbitrary moneyness, there are also substantial differences in option greeks. For example, call options of arbitrary moneyness have an average delta of 0.585 (0.536 ATM), gamma of 0.139 (0.212 ATM), and vega of 2.301 (3.476 ATM).

3.3.3 Portfolio formation

Our portfolio formation is closely related to Goyal and Saretto (2009) who show that ATM delta-hedged call returns and straddle returns increase as a function of the volatility risk premium (VRP), measured as the log-difference of RV and IV (see Hu and Jacobs, 2020).[60] Goyal and Saretto (2009) determine large deviations between RV and IV based on their log-difference. The underlying assumption is that by applying the transformation $t : \mathbb{R}^2 \to \mathbb{R} : (RV, IV) \mapsto \log\left(\frac{RV}{IV}\right)$ deviations between RV and IV of all (ATM) options at a particular date can be adequately compared to each other. The trading strategy is then simply derived by investing long (short) in the options within the lowest (highest) decile of the log-differences, i.e., low and large deviations are identified based on a one-dimensional portfolio sort. This strategy is subsequently shown to earn a statistically and economically significant average monthly return.

Identifying options with large deviations between RV and IV is equivalent to deter-

[60] Goyal and Saretto (2009) argue that volatility is highly mean-reverting and therefore large deviations of the current volatility from the long-term average are unlikely to persist. As IV incorporates expectations on future volatility this implies that large deviations between RV (as long-term average) and IV (as forecast on future volatility) are likely to reduce in magnitude. The authors conclude that options with an IV much lower than the corresponding RV are cheap while options with a much higher IV than RV are expensive. This raises the question how large deviations between RV and IV should be quantified. Note that the authors do not take a clear stand on the question if their observed returns are abnormal or arise as compensation for some aggregate risk. On the one hand the authors argue that high deviations between RV and IV are indicative of option mispricing. On the other hand they highlight that deviations between RV and IV will be more (less) pronounced for equities with higher (lower) volatility of volatility.

mining options where IV is particularly high or low *given* a particular level of RV. The log-transformation can then be seen as an intriguingly simple approach that essentially translates the two-dimensional problem of identifying value pairs (RV, IV) of options where IV is abnormally high or low *given* their RV to a one-dimensional problem that can be tackled by simple portfolio sorts. However, Goyal and Saretto (2009) report that RV increases when proceeding from the first to the last decile, which is illustrated in Table 3.2. As a consequence, the proposed strategy is not only long (short) in large (low) deviations between RV and IV but also long (short) in high (low) realized volatility.

This is due to the implicit assumption of a linear relationship between RV and IV. As an illustrative example we visualize the value pairs (RV, IV) of all ATM call options in January 2010 along with the conditional quantile curves implied by the log-difference of RV and IV in Figure 3.4.[61] The red dashes (mostly on the left) and black dashes (mostly on the right) on the x-axis illustrate that there are systematic differences in RV between the long and the short portfolio.

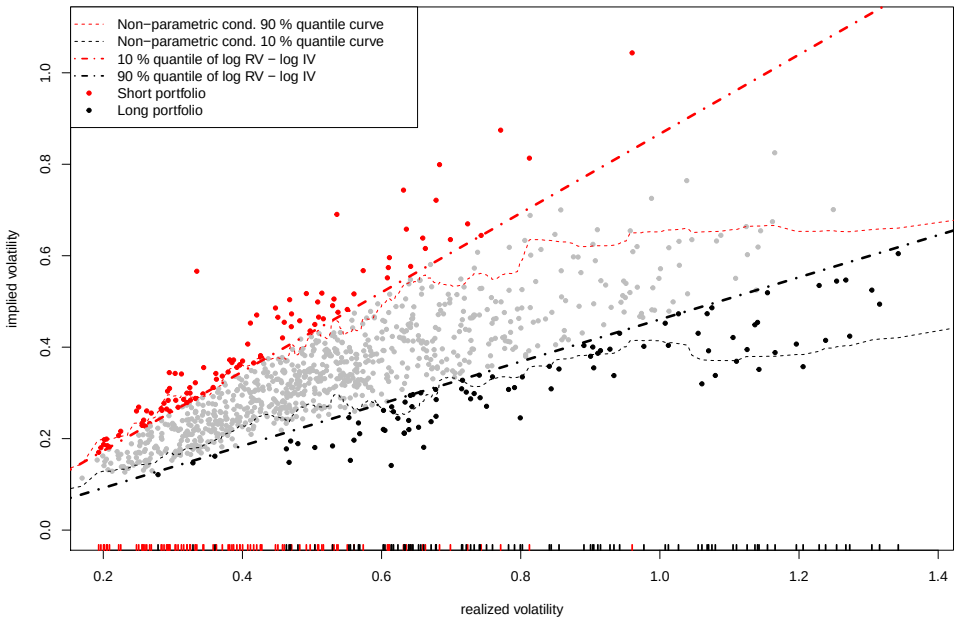
As a further illustration we provide the conditional quantile curves implied by measuring the VRP as the simple difference $RV - IV$ according to Cao and Han (2013). The main difference is that large deviations between RV and IV are no longer determined based on the relative deviation $\frac{RV}{IV}$ in case of the log-differences, but rather by their absolute deviation $RV - IV$. Consequently, the derived quantile curves are linear and parallel to each other. Again, this causes systematic differences in RV between the long and the short portfolio, which is illustrated in Figure 3.5.

These differences in realized volatility are problematic against the background of the recent literature. For example, Cao and Han (2013) show that delta-hedged equity option returns decrease when the underlying stock's idiosyncratic volatility increases. Furthermore, Hu and Jacobs (2020) find that (total) realized volatility drives raw option

[61] The long portfolio consists of those options fulfilling the inequality $\log RV_i - \log IV_i \geq q_{90\%}$ where $q_{90\%}$ is defined as the *unconditional* empirical 90 % quantile of the log-differences. Analogously, options in the short portfolio fulfill $\log RV_i - \log IV_i \leq q_{10\%}$. This is equivalent to requiring $IV_i \leq RV_i \cdot e^{-q_{90\%}}$ in the long portfolio and $IV_i \geq RV_i \cdot e^{-q_{10\%}}$ in the short portfolio, i.e., the conditional quantile curves are implicitly assumed to be linear functions in RV.

Figure 3.4: Sorting options on the *log* difference of RV and IV

This figure shows the value pairs (RV, IV) of all ATM call options in our sample in January 2010. According to Goyal and Saretto (2009) options are sorted on the log-differences of RV and IV into decile portfolios. Sorting on the log-differences (unconditionally) and choosing the highest and lowest 10 % of the value pairs (RV, IV) is equivalent to requiring $IV_i \leq RV_i \cdot e^{-q_{90\%}}$ in the long portfolio and $IV_i \geq RV_i \cdot e^{-q_{10\%}}$ in the short portfolio, where $q_{10\%}$ and $q_{90\%}$ denote the (unconditional) empirical 10 % and 90 % quantiles of the log-differences of RV and IV, respectively. This translates into the black and red dash-dotted straight lines in the scatter plot. Consequently, options in decile 1 (red data points) constitute the short portfolio, while options in decile 10 enter into the long portfolio (black data points). The red and black dashes on the x-axis correspond to the value pairs (RV, IV) in the long and short portfolio and illustrate that there are systematic differences in RV between the two portfolios. The faint dashed lines correspond to the non-parametric 10 % (black) and 90 % (red) quantile curves of IV *conditional* on RV.

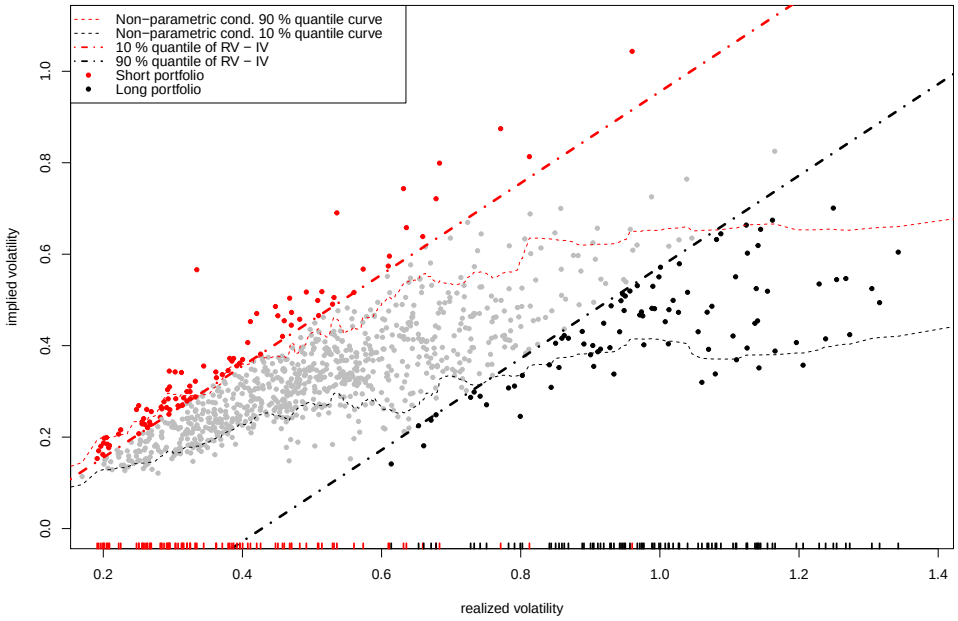


returns. In the light of these findings it is unclear to which extent the positive returns of the long-short strategy (high minus low VRP) are attributable to differences in the VRP or differences in the average level of RV. This complicates an interpretation of the obtained returns.

As application of conditional quantile curves allows us to derive long (short) portfolios with high (low) differences between RV and IV while controlling for RV. This is

Figure 3.5: Sorting options on the difference of RV and IV

This figure shows the value pairs (RV, IV) of all ATM call options in our sample in January 2010. Sorting on the difference (unconditionally) and choosing the highest and lowest 10 % of the value pairs is equivalent to requiring $IV_i \leq RV_i - q_{90\%}$ in the long portfolio and $IV_i \geq RV_i - q_{10\%}$ in the short portfolio, where $q_{10\%}$ and $q_{90\%}$ denote the (unconditional) empirical 10 % and 90 % quantiles of the differences of RV and IV, respectively. This translates into the black and red dash-dotted straight lines in the scatter plot. We form a long-short portfolio where options in decile 1 according to the difference of RV and IV (red data points) constitute the short portfolio, while options in decile 10 enter into the long portfolio (black data points). The red and black dashes on the x-axis correspond to the value pairs (RV, IV) in the long and short portfolio and illustrate that there are systematic differences in RV between the two portfolios. The faint dashed lines correspond to the non-parametric 10 % (black) and 90 % (red) quantile curves of IV *conditional* on RV.



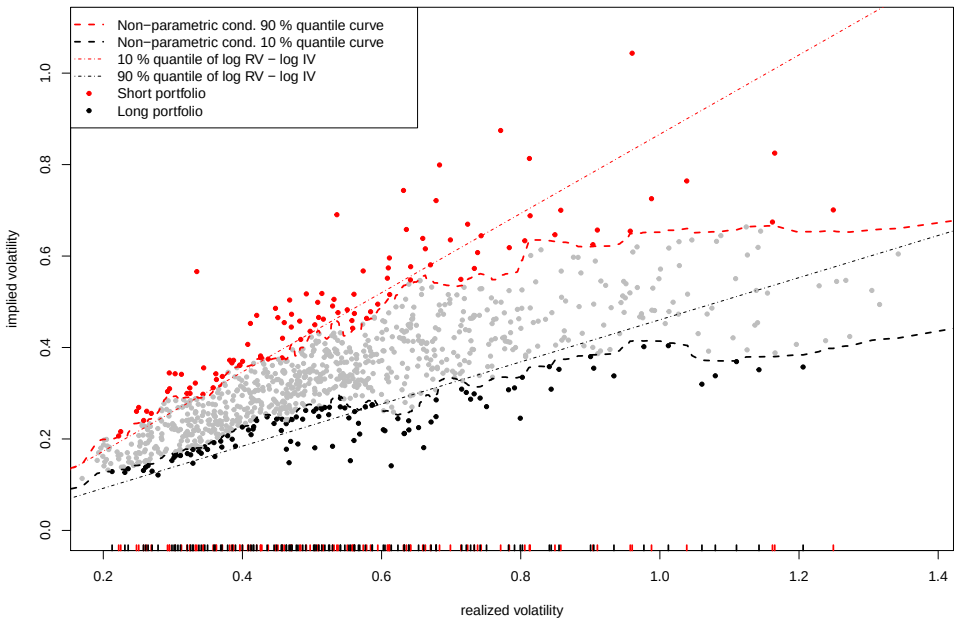
done by forming long (short) portfolios consisting of options with low (high) IV *conditional* on their underlying's RV.[62] Furthermore, our non-parametric procedure enables us to model the relation between RV and IV without making any assumptions about the functional form, as can be seen in Figure 3.6, where we provide conditional quantile

[62] As long and short portfolios are formed on a fixed date based on the cross-section of ATM options with maturity of one month, we automatically control for moneyness, maturity, and the risk-free interest rate, too. In further analyses on our option sample of arbitrary moneyness we additionally control for option moneyness.

curves of IV given RV for ATM call options in January 2010.[63] The Figure illustrates that the conditional quantile curves are in fact non-linear. Furthermore, especially for option with high RV, the gradients of the conditional quantile curves are lower than the ones implied by the log-difference.

Figure 3.6: Sorting options on IV conditional on RV

This figure shows the value pairs (RV, IV) of all ATM call options in our sample in January 2010. Options are sorted according to IV *conditional* on RV into decile portfolios. Options in decile 10 (red data points) constitute the short portfolio, while options in decile 1 enter into the long portfolio (black data points). That is, the long (short) portfolio is constituted of the 10 % of the options with the highest (lowest) IV conditional on RV. More exactly, for option i in the long portfolio we require $F(IV_i|RV_i) \leq 10$ % while options in the short portfolio have to fulfill $F(IV_i|RV_i) \geq 90$ %, where $F(\cdot|RV_i)$ denotes the conditional cumulative distribution function of implied volatility given a specific level of realized volatility (RV_i). This translates into the red and black dashed conditional quantile curves. The faint black and red dash-dotted lines correspond to the log-difference of RV and IV and are included for comparison. The red and black dashes on the x-axis correspond to the value pairs (RV, IV) in the long and short portfolio.



Analogously to the trading strategy by Goyal and Saretto (2009), our short portfolio

[63] Our approach can easily be extended to control for more covariates. In Section 3.3.4 we also report results for our option sample of arbitrary moneyness. To avoid confounding effects due to high or low moneyness ("volatility skew") we additionally condition on option moneyness. In Section 3.4.1 we provide results when conditioning on further moments of the underlyings' return distribution.

is constituted of options with high IVs relative to the RVs of their underlyings. However, the criterion of choosing options in the lowest decile of log RV - log IV is replaced by selecting options with IV above the *conditional* 90 % quantile *given* the RV of the corresponding underlying. More precisely: for option i with realized volatility RV_i and implied volatility IV_i being in the short portfolio, we require $F(IV_i|RV_i) \overset{!}{\geq} 90\ \%$ where $F(\cdot|RV_i)$ denotes the conditional cumulative distribution function of IV given a specific level of realized volatility (RV_i). Intuitively, our short portfolio is constituted of the options where the value-pair (RV, IV) is above the 90 % conditional quantile curve in Figure 3.6. Analogously, the long portfolio is constituted of the options below the 10 % conditional quantile curve.

Figure 3.6 illustrates that there is no systematic difference in the average RV between the short and long portfolio as visualized by the red and black dashes on the x-axis. In comparison with Figure 3.4 it can further be seen that there are noteworthy disagreements in which options enter into the short and long portfolios, especially for options with a large RV.

While these figures reflect the relation between RV and IV on a particular date only, Table 3.2 presents evidence that by employing conditional quantile curves systematic differences in RV between the decile portfolios can be reduced significantly. The table compares average values of IV and RV in the 10 decile portfolios obtained by sorting ATM options on the log-difference between RV and IV to our approach of sorting options according to their IV conditional on their RV.[64] The comparison is done for calls and puts separately. We further include option greeks (delta, gamma, vega) as well as further moments of the underlyings' return distribution (skewness, kurtosis) in the table.[65]

Results for call and put options are very similar. For brevity, we therefore focus on

[64] Decile portfolios based on our conditional quantile approach are obtained by including options with $F(IV|RV) \geq 90\ \%$ in the first portfolio, options with $80\ \% \leq F(IV|RV) < 90\ \%$ in the second portfolio, etc. The portfolios are equal-weighted. Note that we sort on the conditional IV in descending order to obtain decile portfolios where the differences between RV and IV are increasing.

[65] Means are first calculated (equally weighted) for each month and each portfolio and are then averaged over time.

Table 3.2: Decile portfolios for ATM options

This table provides information on the average values of various covariates within decile portfolios that are formed according to two approaches: This paper proposes sorting options (in descending order) on their IV conditional on their underlyings' RV into decile portfolios to proxy for the Volatility Risk Premium (VRP). This yields portfolios where, on average, the differences RV-IV increase monotonically while RV remains nearly constant. We compare this to portfolios obtained by sorting options on the log-difference of RV and IV (according to Goyal and Saretto, 2009). This yields portfolios that are monotonically increasing in the average RV (due to an implicit linear assumption on the conditional quantile curves). Within the decile portfolios, we provide means of IV, RV, option greeks (delta, gamma, vega) as well as skewness (skew), and kurtosis (kurt) of option underlyings' return distribution. The means are calculated by first computing averages for each portfolio and each month and then taking the time-series averages. Results are based on ATM call and put samples from January 1996 to June 2019.

			Decile portfolios								
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Calls	**log-differences**	IV	0.438	0.423	0.417	0.413	0.410	0.410	0.411	0.413	0.414	0.418
		RV	0.423	0.421	0.421	0.421	0.421	0.428	0.431	0.439	0.447	0.467
		RV-IV	-0.015	-0.002	0.004	0.008	0.011	0.018	0.020	0.026	0.033	0.049
		delta	0.534	0.534	0.533	0.533	0.531	0.532	0.533	0.531	0.531	0.532
		gamma	0.151	0.150	0.149	0.152	0.150	0.152	0.151	0.153	0.155	0.164
		vega	4.857	4.941	5.037	4.983	4.986	4.921	4.956	4.952	4.861	4.753
		skew	0.160	0.169	0.165	0.170	0.172	0.166	0.156	0.175	0.179	0.242
		kurt	8.758	8.619	8.548	8.652	8.647	8.674	8.740	8.990	9.136	10.679
	cond. quantiles	IV	0.451	0.437	0.425	0.421	0.413	0.410	0.406	0.401	0.398	0.392
		RV	0.438	0.438	0.432	0.432	0.428	0.430	0.428	0.427	0.431	0.434
		RV-IV	-0.013	0.001	0.007	0.011	0.015	0.020	0.022	0.026	0.033	0.042
		delta	0.534	0.535	0.534	0.534	0.533	0.533	0.532	0.531	0.531	0.531
		gamma	0.153	0.151	0.150	0.152	0.151	0.151	0.151	0.154	0.154	0.163
		vega	4.770	4.840	4.957	4.865	4.947	4.981	4.991	4.923	5.094	4.987
		skew	0.171	0.175	0.171	0.181	0.175	0.161	0.159	0.171	0.177	0.218
		kurt	8.870	8.770	8.634	8.829	8.624	8.740	8.732	8.910	9.181	10.280
Puts	**log-differences**	IV	0.457	0.439	0.431	0.426	0.426	0.425	0.426	0.423	0.424	0.430
		RV	0.430	0.427	0.426	0.426	0.430	0.433	0.438	0.440	0.450	0.475
		RV-IV	-0.027	-0.012	-0.005	0.000	0.004	0.008	0.012	0.017	0.026	0.045
		delta	-0.469	-0.470	-0.470	-0.470	-0.473	-0.473	-0.472	-0.472	-0.473	-0.473
		gamma	0.146	0.143	0.145	0.145	0.145	0.145	0.146	0.148	0.148	0.154
		vega	4.818	4.990	5.002	5.036	5.032	5.005	5.001	4.989	5.001	4.866
		skew	0.170	0.161	0.163	0.163	0.183	0.167	0.147	0.161	0.169	0.228
		kurt	8.850	8.716	8.577	8.529	8.478	8.683	8.921	8.773	9.283	11.018
	cond. quantiles	IV	0.475	0.457	0.442	0.438	0.427	0.427	0.420	0.415	0.408	0.399
		RV	0.449	0.446	0.440	0.439	0.434	0.437	0.434	0.433	0.432	0.431
		RV-IV	-0.026	-0.011	-0.002	0.001	0.007	0.010	0.014	0.018	0.024	0.032
		delta	-0.470	-0.469	-0.469	-0.470	-0.471	-0.472	-0.471	-0.472	-0.474	-0.475
		gamma	0.147	0.145	0.145	0.145	0.145	0.146	0.145	0.147	0.148	0.153
		vega	4.702	4.853	4.901	4.920	4.973	5.014	5.091	5.011	5.112	5.150
		skew	0.191	0.175	0.177	0.178	0.165	0.170	0.151	0.160	0.153	0.196
		kurt	9.026	8.817	8.820	8.617	8.658	8.749	8.635	8.956	9.208	10.412

the analysis of call results. Like Goyal and Saretto (2009) we find that when sorting on log RV - log IV, IV decreases when proceeding from decile 1 to decile 10 by 2.0 percentage points while RV increases by 4.4 percentage points. Thus, the differences in RV between decile 1 and 10 are more than double the corresponding differences in IV. Furthermore, while the decile portfolios are (almost) monotonic in RV, there is no clear pattern for IV. It is therefore unclear to which extent a corresponding long-short strategy (decile 10 - decile 1) is driven by systematic differences in RV rather than differences in the VRP.

In our approach, we control for RV when forming the decile portfolios by sorting option contracts on their IV *conditional* on their RV. This leads to average differences in IV of nearly 6 percentage points that are by construction ensured to decrease monotonically from decile 1 to decile 10. More importantly, differences in RV between decile 1 and 10 are very small (about 0.4 percentage points). Consequently, the influence of different levels of RV on the returns from a long-short strategy is significantly reduced.[66] For both strategies there is not much variation in option greeks (delta, gamma, vega) across portfolios. However, for both strategies the returns of the underlyings in portfolio 10 are more positively skewed and more leptokurtic than those in portfolio 1, with the differences being more pronounced in the portfolios formed on the log-differences. For example, the average skewness in portfolio 10 is 0.218 compared to 0.171 in portfolio 1 for the approach based on conditional quantile curves (0.242 vs. 0.160 for portfolios 10 and 1 formed on the log-differences).[67] To account for the fact that results from our long-short strategy might be biased by systematic differences in skewness and kurtosis, we control for these characteristics in a robustness check in Section 3.4.1.

[66] For puts, the average difference in RV between decile 1 and 10 is a little bit higher (1.8 percentage points). This might be due to the smaller size of our put option sample. However, this difference is relatively small compared to the average difference in IV (7.6 percentage points). When turning to the option sample of arbitrary moneyness (not covered in Table 3.2), the average differences in RV between decile 1 and 10 are 1.1 percentage points for calls and 0.1 percentage points for puts.

[67] Eisdorfer et al. (2020) provide evidence that the nominal stock price level matters for option returns. However, for both approaches (log-differences, conditional quantiles) we do not find much variation in the average log stock price across the decile portfolios which is why we do not include the log stock price in further analyses.

3.3.4 Trading strategy

At the money option contracts

We start with a trading strategy based on ATM options only so that we are able to compare the results of a trading strategy based on conditional quantile curves to those obtained by sorting according to the criterion by Goyal and Saretto (2009). First of all, in each of the 10 decile portfolios we calculate monthly returns from a raw option strategy.[68] While the portfolios themselves are formed every month on the first trading day (typically a Monday) after the option expiration, we follow Goyal and Saretto (2009) and start trading the day after (typically a Tuesday) to mitigate microstructure biases. We use the mid-point of bid and ask quotes to proxy for the market price of the option at the beginning of the monthly trade (cf., e.g., Coval and Shumway, 2001, Driessen et al., 2009, Goyal and Saretto, 2009, Cao and Han, 2013, Hu and Jacobs, 2020). We hold all options until expiration.[69] For an option expiring in the money, the return is given by the terminal payoff divided by the price of the option contract minus 1. For an option that expires out of the money we set the return to -100 %.

In addition to raw option returns, we also calculate returns from a delta-hedged strategy to reduce the directional exposure to the underlying stocks.[70] Returns for these trading strategies are calculated for calls and puts separately and based on equal-weighted portfolios.

Results for both approaches (log-differences and conditional quantiles) are reported in Table 3.3 (delta-hedged returns) and Table 3.4 (raw option returns). The tables provide summary statistics on the monthly returns in each of the decile portfolios (long positions) as well as for a long-short strategy (high minus low VRP). The tables illus-

[68] Goyal and Saretto (2009) do not implement a raw option strategy but rather provide results for a delta-hedged option strategy as well as for a strategy based on straddles (built of pairs of calls and puts with the same exercise price).

[69] The issue of early exercise is discussed in Section 3.3.1.

[70] Delta-hedged option positions are formed by buying one option contract and buying (for puts) or short-selling (for calls) delta shares of the underlying stock. We follow the conservative approach of Goyal and Saretto (2009) and do not rebalance the portfolio during the holding period, see the discussion ibidem.

trate that returns in the portfolios sorted on the VRP increase (almost) monotonically for both puts and calls in the delta-hedged and in the raw option strategy, respectively. This is true for portfolios formed based on conditional quantile curves and portfolios formed according to the log-difference of RV and IV.

Table 3.3: Delta-hedged returns of ATM options

This table provides summary statistics on monthly delta-hedged returns of decile portfolios for ATM call and put options. Decile portfolios are formed by sorting on the log-difference of RV and IV according to Goyal and Saretto (2009) and by sorting on options' IV *conditional* on their RV, respectively. We additionally provide returns from a long-short strategy, that is long in decile 10 (highest) and short in decile 1 (lowest). We report the mean, standard deviation (sd), minimum (min), maximum (max), and Sharpe ratio (SR) of the monthly returns. Since the long-short strategy is a zero investment strategy, the Sharpe ratio is simply calculated as the ratio between mean and standard deviation. We calculate the Sharpe ratios in the decile portfolios accordingly for easy comparison. The sample period is from January 1996 to June 2019.

			Decile portfolios									
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**10-1**
Calls	**log-differences**	mean	-0.015	-0.007	-0.006	-0.005	-0.003	-0.003	-0.002	-0.001	0.000	0.005	0.020
		sd	0.027	0.026	0.026	0.028	0.029	0.028	0.029	0.029	0.030	0.032	0.025
		min	-0.078	-0.086	-0.073	-0.068	-0.069	-0.056	-0.081	-0.062	-0.069	-0.051	-0.045
		max	0.150	0.171	0.203	0.184	0.187	0.191	0.192	0.186	0.194	0.196	0.161
		SR	-0.571	-0.257	-0.243	-0.188	-0.106	-0.102	-0.068	-0.046	0.015	0.142	0.786
	cond. quant.	mean	-0.016	-0.008	-0.006	-0.005	-0.002	-0.003	-0.001	0.000	0.001	0.003	0.020
		sd	0.030	0.029	0.029	0.028	0.031	0.028	0.028	0.027	0.025	0.025	0.023
		min	-0.088	-0.076	-0.072	-0.070	-0.079	-0.071	-0.061	-0.060	-0.056	-0.038	-0.057
		max	0.161	0.199	0.183	0.180	0.230	0.185	0.191	0.184	0.179	0.169	0.113
		SR	-0.541	-0.273	-0.209	-0.178	-0.073	-0.103	-0.047	-0.018	0.047	0.129	0.842
Puts	**log-differences**	mean	-0.012	-0.006	-0.004	-0.003	-0.001	0.000	0.000	0.001	0.002	0.003	0.015
		sd	0.026	0.024	0.025	0.026	0.025	0.027	0.027	0.027	0.028	0.031	0.025
		min	-0.078	-0.065	-0.087	-0.078	-0.053	-0.064	-0.075	-0.088	-0.054	-0.057	-0.040
		max	0.132	0.169	0.145	0.177	0.171	0.161	0.203	0.160	0.170	0.202	0.209
		SR	-0.477	-0.256	-0.166	-0.112	-0.058	0.002	-0.001	0.026	0.086	0.104	0.612
	cond. quant.	mean	-0.014	-0.007	-0.003	-0.003	-0.001	0.000	0.001	0.001	0.002	0.003	0.017
		sd	0.028	0.027	0.027	0.026	0.026	0.026	0.026	0.025	0.025	0.025	0.021
		min	-0.078	-0.071	-0.075	-0.080	-0.066	-0.059	-0.061	-0.070	-0.053	-0.065	-0.058
		max	0.147	0.172	0.193	0.161	0.190	0.162	0.181	0.142	0.161	0.155	0.081
		SR	-0.492	-0.265	-0.120	-0.120	-0.026	-0.016	0.020	0.043	0.085	0.130	0.796

A long-short delta-hedged strategy based on conditional quantile curves yields average monthly returns of 2.0 % for calls and 1.7 % for puts with a monthly Sharpe ratio of 0.842 (2.917 annualized) and 0.796 (2.757 annualized), respectively. In comparison, the delta-hedged strategy based on the log-difference of RV and IV earns monthly Sharpe ratios of 0.786 (calls) and 0.612 (puts). That is, we confirm the existence of a

volatility risk premium in the cross-section of option returns and find an even higher effect for delta-hedged returns when controlling for the level of realized volatility.

Table 3.4: Raw returns of ATM options

This table provides the same information on monthly returns from decile portfolios of ATM options as Table 3.3 but for a raw option strategy.

			Decile portfolios									
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**10-1**
Calls	**log-differences**	mean	0.006	0.098	0.084	0.104	0.112	0.129	0.120	0.129	0.159	0.185	0.179
		sd	0.512	0.591	0.566	0.622	0.646	0.648	0.657	0.659	0.681	0.687	0.413
		min	-0.919	-0.945	-0.998	-0.998	-0.987	-0.975	-1.000	-1.000	-0.990	-0.982	-1.307
		max	2.407	1.951	1.836	2.009	2.106	2.175	2.206	2.441	2.751	2.792	2.317
		SR	0.012	0.166	0.149	0.168	0.174	0.199	0.183	0.195	0.234	0.269	0.434
	cond. quant.	mean	-0.006	0.058	0.078	0.088	0.111	0.113	0.103	0.137	0.155	0.207	0.213
		sd	0.511	0.544	0.568	0.597	0.643	0.634	0.643	0.670	0.676	0.687	0.407
		min	-0.918	-0.982	-0.980	-0.992	-0.989	-1.000	-1.000	-1.000	-1.000	-0.975	-0.721
		max	2.375	1.674	1.737	1.786	2.322	2.093	2.294	2.819	2.157	2.475	1.885
		SR	-0.012	0.107	0.138	0.148	0.173	0.179	0.161	0.205	0.229	0.301	0.523
Puts	**log-differences**	mean	-0.187	-0.164	-0.143	-0.155	-0.103	-0.090	-0.076	-0.081	-0.062	-0.062	0.125
		sd	0.615	0.695	0.756	0.767	0.791	0.815	0.841	0.801	0.827	0.838	0.419
		min	-0.980	-0.978	-0.987	-1.000	-0.974	-0.990	-0.951	-0.976	-0.979	-0.985	-1.184
		max	3.825	4.728	4.338	4.980	4.971	4.575	5.665	4.699	4.665	4.827	1.635
		SR	-0.304	-0.235	-0.188	-0.202	-0.131	-0.111	-0.090	-0.101	-0.075	-0.074	0.298
	cond. quant.	mean	-0.176	-0.159	-0.122	-0.143	-0.105	-0.100	-0.087	-0.084	-0.077	-0.077	0.099
		sd	0.600	0.661	0.740	0.743	0.800	0.788	0.838	0.810	0.850	0.888	0.467
		min	-0.965	-0.982	-0.964	-0.992	-0.976	-0.947	-0.976	-0.990	-0.990	-0.958	-1.295
		max	3.794	4.017	5.054	4.481	5.425	4.738	5.136	4.439	5.013	5.054	2.741
		SR	-0.294	-0.240	-0.164	-0.192	-0.131	-0.127	-0.103	-0.103	-0.090	-0.087	0.212

Higher absolute returns can be earned with raw option strategies. A long-short strategy based on conditional quantile curves yields average monthly returns of 21.3 % (calls) and 9.9 % (puts) with Sharpe ratios of 0.523 and 0.212, respectively. All trading strategies based on conditional quantiles yield returns that are both economically and statistically significant with t-statistics of at least 3.5. While for calls the Sharpe ratio for a strategy based on conditional quantile curves is higher than for a strategy based on the log-differences of RV and IV (0.523 vs. 0.434), the opposite is true for puts (0.212 vs. 0.298).[71] Except for delta-hedged call returns, the differences in the Sharpe ratios are statistically significant at the 5 % level when testing according to Ledoit and Wolf (2008). We attribute these differences between delta-hedged returns

[71] These differences are robust to the choice of the method for estimating conditional quantiles, see Section 3.4.3.

and raw option returns as well as between calls and puts to the systematic differences in RV when sorting options on the log-difference of RV and IV. While we control for the level of RV across the decile portfolios by using conditional quantile estimates, the long-short strategy based on the log-differences is long (short) in high (low) RV (see Table 3.2). This differing influence of RV on the returns from various strategies highlights the importance of controlling for the influence of the level of RV when empirically analyzing the VRP.

Options of arbitrary moneyness

So far, we have analyzed our sample of ATM calls and puts. As ATM calls (puts) account for only 11.7 % (13.9 %) of all calls (puts) in our sample, we extend our analysis to options with moneyness (defined as the ratio K/S) between 0.5 and 1.5. In doing so, we can analyze if there is still a volatility risk premium in the cross-section of option returns when additionally including in the money and out of the money options. Furthermore, this allows us to increase the number of observations by a factor of about 8. Finally, by requiring ATM options to have a moneyness in the interval [0.975, 1.025] (see Section 3.3.1) we also exclude many underlyings from our analysis. This happens when there are no options with appropriate moneyness available at a specific date. Not restricting moneyness to the small interval [0.975, 1.025] therefore doubles the number of underlyings available in our sample on average over all months. In particular, the minimum number of different underlyings at a given day increases from 171 to 734 (for calls) and 105 to 387 (for puts). This facilitates controlling for further moments of the underlyings' return distribution, see Section 3.4.1.

Our approach of using conditional quantiles for deriving decile portfolios generalizes naturally to options of arbitrary moneyness. Therefore, we form portfolios based on extreme values of IV conditional on RV *and* options' moneyness.[72] We condition

[72]To put this into perspective: If one were to replicate this based on portfolio sorts, one would be required to apply a conditional triple-sort to the data. This is something that is impracticable if not infeasible against the background of only 826 put options at the beginning of our data sample. In Section 3.4.1 we additionally control for skewness and kurtosis in the underlyings' return distributions. This would be, at the latest, impossible to achieve with portfolio sorts.

on options' moneyness to avoid results from such a trading strategy to be biased by systematic differences in options' moneyness (volatility skew, see, e.g., Toft and Prucyk (1997)). Summary statistics for monthly delta-hedged and raw option returns for both calls and puts are reported in Table 3.5. The results are in line with our previous findings and confirm that the VRP is also priced in the cross-section of returns of options with arbitrary moneyness. For example, a delta-hedged strategy that is long (short) in options with a high (low) VRP yields an average monthly return of 2.4 % for calls and 2.5 % for puts with a monthly Sharpe ratio of 0.816 (calls) and 0.844 (puts), respectively. The corresponding long-short raw option strategy exhibits an average monthly return of 20.1 % for calls and 13.1 % for puts with a monthly Sharpe ratio of 0.390 and 0.217, respectively. All trading strategies yield returns that are both economically and statistically significant with t-statistics above 3.6.

Table 3.5: Returns of the trading strategies for options with arbitrary moneyness

This table provides similar information on monthly delta-hedged as well as raw option returns from decile portfolios to Tables 3.3 and 3.4 but for options with arbitrary moneyness. Furthermore, decile portfolios are formed by sorting on options' IV conditional on their RV *and* their moneyness.

			Decile portfolios									
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**10-1**
delta-hedged returns	**Calls**	mean	-0.023	-0.014	-0.010	-0.008	-0.007	-0.005	-0.004	-0.004	-0.002	0.001	0.024
		sd	0.032	0.030	0.029	0.029	0.029	0.029	0.030	0.030	0.029	0.032	0.030
		min	-0.117	-0.093	-0.105	-0.072	-0.067	-0.065	-0.065	-0.081	-0.061	-0.056	-0.078
		max	0.186	0.178	0.178	0.182	0.194	0.192	0.192	0.223	0.226	0.271	0.290
		SR	-0.726	-0.469	-0.333	-0.285	-0.233	-0.163	-0.143	-0.119	-0.071	0.029	0.816
	Puts	mean	-0.023	-0.013	-0.010	-0.008	-0.006	-0.005	-0.004	-0.003	-0.001	0.002	0.025
		sd	0.032	0.038	0.036	0.036	0.036	0.036	0.035	0.034	0.036	0.039	0.029
		min	-0.103	-0.074	-0.083	-0.085	-0.069	-0.062	-0.067	-0.071	-0.057	-0.049	-0.084
		max	0.245	0.308	0.341	0.308	0.326	0.299	0.284	0.286	0.334	0.356	0.250
		SR	-0.712	-0.333	-0.273	-0.213	-0.160	-0.134	-0.099	-0.087	-0.038	0.054	0.844
raw returns	**Calls**	mean	-0.059	-0.014	0.018	0.030	0.039	0.058	0.061	0.066	0.093	0.142	0.201
		sd	0.411	0.444	0.476	0.498	0.524	0.559	0.576	0.600	0.635	0.726	0.515
		min	-0.855	-0.889	-0.950	-0.949	-0.961	-0.959	-0.963	-0.948	-0.945	-0.935	-1.429
		max	1.384	1.477	1.723	2.047	2.321	2.847	3.562	4.246	4.242	5.699	4.871
		SR	-0.143	-0.032	0.037	0.060	0.075	0.104	0.105	0.110	0.147	0.196	0.390
	Puts	mean	-0.232	-0.202	-0.194	-0.179	-0.181	-0.176	-0.166	-0.152	-0.139	-0.101	0.131
		sd	0.596	0.750	0.795	0.835	0.862	0.885	0.906	0.888	0.957	1.030	0.603
		min	-0.919	-0.919	-0.959	-0.853	-0.929	-0.911	-0.942	-0.957	-0.941	-0.944	-1.634
		max	4.801	6.321	7.279	7.175	7.429	7.135	6.847	6.002	6.918	6.872	4.643
		SR	-0.389	-0.269	-0.244	-0.214	-0.210	-0.199	-0.184	-0.171	-0.145	-0.098	0.217

3.4 Robustness checks

3.4.1 Controlling for higher moments of the underlyings' return distribution

In our baseline analysis, we sort options into decile portfolios based on their IV conditional on their RV. This allows us to identify large deviations between IV and RV to study the VRP without potential biases arising from different levels in RV. However, Table 3.2 indicates that after controlling for RV (the square root of the second moment of the underlyings' return distribution) via conditional quantiles there remain differences in the third (skewness) and fourth (kurtosis) moment of the underlyings' return distribution between the long (decile portfolio 10) and the short portfolio (decile portfolio 1). These differences might potentially bias returns from a strategy exploiting the VRP.

Fortunately, our approach of using conditional quantiles allows us to additionally condition on the underlyings' skewness and kurtosis. This is impossible via conditional portfolio sorts as this would require a quadruple sort on IV, RV, skewness, and kurtosis in our ATM option sample.[73] Additionally controlling for moneyness in our sample of arbitrary moneyness would even require a quintuple sort.

Summary statistics for the monthly returns from the delta-hedged and raw option strategies for puts and calls are reported in Table 3.6 for the ATM option sample as well as the sample of arbitrary moneyness. Skewness and kurtosis are calculated based on the realized returns from the option contracts' underlying over the most recent 12 months. The results are in line with our previous findings.

[73] This becomes even more apparent when considering the number of only 105 puts at the beginning of our sample period in January 1996. This makes a conditional portfolio sort in 10 bins for each variable ($10,000$ bins in total) infeasible. Instead performing a portfolio sort based on only 5 bins for each variable would still be infeasible (625 bins). Apart from this, it would be a very imprecise approximation of the true conditional quantile curves (see the illustrative example in Figure 3.2).

Table 3.6: Conditioning on higher moments of the underlyings' return distribution

This table presents summary statistics on the monthly returns from long-short portfolios based on the conditional 10 % and 90 % quantiles of option IV. In our baseline analysis, we condition on the square root of the second moment (RV) of the underlyings' return distribution (and additionally on moneyness for the sample of arbitrary moneyness). In this table, we additionally condition on the third (skewness) and fourth (kurtosis) moment of the underlyings' return distribution. Summary statistics for the monthly returns (mean, standard deviation (sd), minimum (min), maximum (max), Sharpe ratio (SR)) are reported when conditioning on skewness (skew) and kurtosis (kurt) separately and jointly, respectively. We provide results for delta-hedged and raw option strategies for puts and calls for both ATM options as well as options with arbitrary moneyness. The sample period is from January 1996 to June 2019.

				mean	sd	min	max	SR
ATM	**Delta-hedged**	**Calls**	skew	0.021	0.026	-0.088	0.104	0.802
			kurt	0.021	0.025	-0.058	0.148	0.855
			skew and kurt	0.021	0.025	-0.095	0.144	0.823
		Puts	skew	0.018	0.024	-0.093	0.124	0.752
			kurt	0.018	0.021	-0.046	0.091	0.850
			skew and kurt	0.018	0.023	-0.087	0.093	0.770
	Raw returns	**Calls**	skew	0.230	0.428	-0.982	1.954	0.537
			kurt	0.227	0.435	-0.864	2.302	0.522
			skew and kurt	0.232	0.421	-0.853	2.505	0.552
		Puts	skew	0.090	0.454	-0.751	2.812	0.197
			kurt	0.101	0.466	-1.166	2.826	0.216
			skew and kurt	0.081	0.456	-0.938	3.088	0.178
Arbitrary moneyness	**Delta-hedged**	**Calls**	skew	0.025	0.030	-0.104	0.291	0.840
			kurt	0.025	0.029	-0.090	0.275	0.864
			skew and kurt	0.025	0.029	-0.094	0.252	0.863
		Puts	skew	0.026	0.028	-0.103	0.174	0.919
			kurt	0.026	0.029	-0.101	0.221	0.902
			skew and kurt	0.026	0.027	-0.086	0.150	0.951
	Raw returns	**Calls**	skew	0.213	0.505	-1.122	4.872	0.421
			kurt	0.209	0.511	-1.171	4.565	0.408
			skew and kurt	0.206	0.489	-0.975	4.419	0.422
		Puts	skew	0.143	0.571	-1.421	4.664	0.251
			kurt	0.154	0.570	-1.280	4.144	0.270
			skew and kurt	0.149	0.549	-1.335	4.487	0.272

3.4.2 Transaction costs

Transaction costs in option markets can be quite large and might in part explain some pricing anomalies such as put-call parity violations (see Goyal and Saretto (2009) and the references therein). For example, in our ATM option sample the average bid-ask spread relative to the mid-prices is 23.7 % for calls and 22.7 % for puts. We therefore study limitations of investors in exploiting profits from the long-short strategies based on the VRP.

So far, we assumed investors to trade options at their mid-point price. However, to account for transaction costs, we recalculate returns from our strategies when incorporating bid-ask spreads. This not only concerns the option contracts but also the underlying stocks. For the raw option strategy transaction costs in the underlying stocks only occur at expiration as all equity options have to be delivered physically. In our delta-hedged strategy, stock related transaction costs additionally arise at the beginning of each monthly trading period.

As already mentioned, (quoted) bid-ask spreads can be quite high. However, although *effective* bid-asks spreads are usually still quite high in absolute terms, there is empirical evidence that they are small relative to the quoted spreads with ratios below 0.5 (see, e.g., Mayhew, 2002, de Fontnouvelle et al., 2003). We therefore recalculate returns for our raw and delta-hedged option strategies based on an effective spread of 50% relative to the quoted spreads. For example, for an option contract with quoted bid and ask equal to \$2 and \$3, respectively, we assume investors to buy at \$2.75 and sell at \$2.25, that is, at an effective spread of \$0.5 instead of the quoted spread of \$1. To provide a more complete picture, we also consider ratios between effective and quoted spreads of 25 %, 75 %, and 100 %.

Average returns for raw and delta-hedged long-short strategies for puts and calls, respectively, are reported in Table 3.7 along with their t-statistics. As expected, average monthly returns decrease substantially when accounting for transaction costs. For example, the average monthly return from the ATM delta-hedged call strategy decreases

from 2.0 % when trading at the mid-point prices to 0.3 % when considering an effective spread of 50 %. However, average delta-hedged returns remain positive for calls and puts both ATM as well as for arbitrary moneyness with t-statistics between 1.078 and 3.286. Nevertheless, while raw option returns are still positive after considering transaction costs (50 % ratio of effective to quoted spreads) for calls (both ATM and for arbitrary moneyness), raw option returns for puts are negative (both ATM and for arbitrary moneyness). When further increasing transaction costs to 75 % or even 100 % of effective to quoted spreads, returns deteriorate further and are negative for all reported strategies. However, when considering more moderate transaction costs like, e.g., Cao and Han (2013) and Bali et al. (2021) (25 % ratio of effective to quoted spreads), the returns of all trading strategies remain positive with t-statistics between 1.390 and 8.452. Overall, the results illustrate that transaction costs can substantially reduce the returns from our option portfolios, especially for the raw option trading strategies. Nevertheless, except for the raw option strategy for puts of arbitrary moneyness, we conclude that the profits from our trading strategies are not eliminated at reasonable levels of transaction costs (ratios of effective to quoted spreads of up to 50 %).

3.4.3 Other estimators of conditional quantiles

There are various methods for estimating conditional quantiles, see Section 3.2.3 for details on the estimation of conditional quantiles in general and specific estimators in particular. So far, we have employed the *leveraging estimator* to form decile portfolios and subsequently calculate returns from a long-short strategy. To ensure that our results are not driven by the specific choice of the estimator, we repeat our analyses based on the *quantization estimator* due to Charlier et al. (2015b) and the *copula estimator* due to Kraus and Czado (2017). Results on the returns from long-short strategies based on the three different estimators are reported in Table 3.8. For illustrative purposes, we also provide results obtained from a (conditional) double-sort (for ATM options) and

Table 3.7: Returns after accounting for transaction costs

This table reports average monthly returns (along with their t-statistics) from our long-short trading strategies when considering transaction costs. In our baseline analysis we assume investors to buy and sell options at their mid-point price (MidP). This table reports results when considering ratios of effective to quoted spreads of 25 %, 50 %, 75 %, and 100 %. Note that we also take transaction costs for the underlying stocks into account as all equity options in our sample have to be delivered physically at option exercise. Long-short portfolios are formed based on the lowest and highest deciles of options' IV *conditional* on their RV (for ATM options) or conditional on RV *and* moneyness (for options of arbitrary moneyness). We report average returns from the delta-hedged and raw option strategies separately for calls and puts. Additionally, we consider the sample of ATM options as well as the sample of arbitrary moneyness. The sample period is from January 1996 to June 2019.

			MidP	**25%**	**50%**	**75%**	**100%**
ATM	**Delta-hedged**	Calls	0.020	0.011	0.003	-0.006	-0.015
			(14.131)	(8.161)	(1.874)	(-4.707)	(-11.218)
		Puts	0.017	0.010	0.002	-0.005	-0.013
			(13.364)	(7.191)	(1.772)	(-3.815)	(-9.350)
	Raw	Calls	0.213	0.114	0.021	-0.082	-0.277
			(8.776)	(4.937)	(0.952)	(-3.642)	(-9.854)
		Puts	0.099	0.036	-0.032	-0.109	-0.252
			(3.565)	(1.390)	(-1.296)	(-4.509)	(-8.264)
Arbitrary moneyness	**Delta-hedged**	Calls	0.024	0.013	0.002	-0.010	-0.021
			(13.700)	(7.725)	(1.078)	(-5.810)	(-12.638)
		Puts	0.025	0.015	0.006	-0.004	-0.013
			(14.178)	(8.452)	(3.286)	(-2.136)	(-7.678)
	Raw	Calls	0.201	0.113	0.026	-0.072	-0.238
			(6.553)	(4.003)	(0.972)	(-2.727)	(-8.339)
		Puts	0.131	0.063	-0.010	-0.091	-0.225
			(3.644)	(1.924)	(-0.348)	(-3.345)	(-7.918)

triple-sort (for options of arbitrary moneyness) into 10 bins for each variable.[74]

The results are in line with our previous findings. In particular, all conclusions with regard to the comparison between our approach of calculating the VRP while controlling for RV and the approach based on the log-difference of RV and IV remain valid, see Section 3.3.4 for details. However, the variation between the different estimators

[74] As outlined in Section 3.2.2, portfolio sorts are inflexible with regard to the percentage of observations that are supposed to enter into the long and short portfolios, especially when the number of observations is low or when sorting on multiple variables. For example, in our put sample of arbitrary moneyness the percentage of observations entering into the long and short portfolios varies between 10.32 % and 18.62 % over all trading periods with a mean of 11.08 %. However, while conditional portfolio sorts might still be feasible in this case (with limitations), controlling for further characteristics as in Section 3.4.1 would be impossible.

Table 3.8: Returns when using different estimators of conditional quantile curves

This table provides information on the monthly returns from long-short portfolios formed based on the conditional 10 % and 90 % quantiles of options' IV conditional on RV (for ATM options) and conditional on RV *and* moneyness (for options of arbitrary moneyness). Conditional quantiles are computed based on three different estimators. We compare our baseline estimator (leveraging estimator) to the copula estimator by Kraus and Czado (2017) and the quantization estimator by Charlier et al. (2015b). For illustrative purposes, we also include results from a conditional double-sort (ATM) and triple-sort (arbitrary moneyness) into 10 portfolios for each variable. We report the mean, standard deviation (sd), and Sharpe ratio (SR) of monthly returns for the delta-hedged and raw option strategy separately for calls and puts. The strategies are evaluated for ATM options and for options of arbitrary moneyness between January 1996 and June 2019.

				Leveraging	**Copula**	**Quantization**	**Portfolio sort**
ATM	**Delta-hedged**	**Calls**	mean	0.020	0.021	0.019	0.018
			sd	0.023	0.025	0.022	0.022
			SR	0.842	0.836	0.874	0.835
		Puts	mean	0.017	0.018	0.016	0.015
			sd	0.021	0.022	0.020	0.020
			SR	0.796	0.813	0.808	0.755
	Raw returns	**Calls**	mean	0.213	0.223	0.208	0.201
			sd	0.407	0.447	0.408	0.414
			SR	0.523	0.499	0.510	0.487
		Puts	mean	0.099	0.113	0.094	0.097
			sd	0.467	0.483	0.446	0.432
			SR	0.212	0.235	0.210	0.224
Arbitrary moneyness	**Delta-hedged**	**Calls**	mean	0.024	0.026	0.023	0.023
			sd	0.030	0.032	0.030	0.028
			SR	0.816	0.796	0.786	0.803
		Puts	mean	0.025	0.023	0.028	0.022
			sd	0.029	0.032	0.030	0.028
			SR	0.844	0.730	0.955	0.786
	Raw returns	**Calls**	mean	0.201	0.221	0.190	0.197
			sd	0.515	0.556	0.510	0.493
			SR	0.390	0.398	0.372	0.399
		Puts	mean	0.131	0.147	0.198	0.114
			sd	0.603	0.611	0.636	0.564
			SR	0.217	0.241	0.312	0.202

is higher for the trading strategies involving options of arbitrary moneyness, which is mainly due to the quantization estimator.[75]

3.4.4 Including dividend-paying stocks

All equity options in our sample are American. In our baseline analysis we therefore follow Hu and Jacobs (2020) and exclude all options with an ex-dividend date during the remaining life of the contract. This is done to reduce the impact of early exercise, as mentioned in Section 3.3.1. However, to show robustness of our results, we recalculate the returns of our option strategies when not excluding dividend-paying stocks. This increases our ATM option sample to 326,224 calls and 299,924 puts while the sample of arbitrary moneyness expands to 2,682,492 calls and 2,145,435 puts. Summary statistics on the monthly returns of delta-hedged and raw option strategies for calls and puts both ATM and for arbitrary moneyness are reported in Table 3.9. The results are in line with our previous findings.

3.4.5 Options with low and high trading volume

Our ATM option sample covers a vast number of stocks with listed options.[76] However, most of these options only experience a small amount of trading volume.[77] For example, the median of the total volume of option contracts over all options at the beginning of each monthly trade is on average 25.16 for the calls and 10.28 for the puts with an average maximum of 27,774 and 26,849, respectively. While for heavily traded options, bid and ask quotes by market makers might largely reflect the supply and demand of actual investors, the data for rarely traded options might just mirror quotes

[75] In an unreported simulation study we find the leveraging and the copula estimator to be most appropriate. However, we include the quantization estimator to provide a more comprehensive picture.

[76] In this analysis, we only consider the ATM option sample. Sorting options on their trading volume in our sample of arbitrary moneyness would otherwise lead to groups that differ systematically in their moneyness as the trading volume of ATM options is usually the highest. These differences *between* the groups could not be eliminated by conditioning on moneyness when computing the conditional quantile curves as this approach can only control for differences in the moneyness *within* the groups.

[77] All options with zero open interest have already been excluded from our option samples. For more details see Section 3.3.1.

Table 3.9: Including options on dividend-paying stocks

In our baseline analyses, we exclude all options with an ex-dividend date during the remaining life of the contract. This is done to reduce the impact of early exercise. In this table, we present summary statistics (mean, standard deviation (sd), minimum (min), maximum (max), and Sharpe ratio (SR)) of monthly returns that we obtain when we do *not* exclude options on dividend-paying stocks. Long-short portfolios are formed based on the conditional 10 % and 90 % quantiles of options' IV conditional on RV (for ATM options) or conditional on RV *and* moneyness (for the option sample of arbitrary moneyness). We report returns from the delta-hedged and the raw option strategy separately for calls and puts. The ATM option sample consists of 326,224 calls and 299,924 puts while the sample of arbitrary moneyness comprises 2,682,492 calls and 2,145,435 puts. The sample period is from January 1996 to June 2019.

				mean	sd	min	max	SR
ATM		**Delta-hedged**	Calls	0.019	0.022	-0.045	0.128	0.873
			Puts	0.016	0.019	-0.042	0.075	0.830
		Raw returns	Calls	0.205	0.419	-0.762	2.641	0.490
			Puts	0.110	0.439	-0.855	2.649	0.251
Arbitrary	**moneyness**	**Delta-hedged**	Calls	0.023	0.028	-0.070	0.271	0.816
			Puts	0.023	0.027	-0.081	0.225	0.858
		Raw returns	Calls	0.193	0.487	-1.131	4.432	0.395
			Puts	0.139	0.572	-1.542	4.313	0.242

by market makers. We therefore analyze if the returns from our trading strategies differ when forming long-short portfolios separately based on a sample of options with low and high trading volume, respectively. For this purpose, at the beginning of each monthly trade, our ATM option sample is split into two groups of equal size according to the options' trading volume at that day. Conditional quantile curves of options' IV conditional on their RV are then calculated for each of the two groups separately.

Summary statistics for the returns from the strategies are provided in Table 3.10. Except for the raw option put strategy, the average returns from all trading strategies are very similar to the previously observed ones and statistically significantly greater than zero at the 1 % level. Overall, we conclude that our findings are overall robust to the options' trading volume.

Table 3.10: Returns for options with low and high trading volume

This table reports summary statistics on the monthly returns from our long-short trading strategies applied to a sample of options with low and high trading volume, respectively. Therefore, at the beginning of each monthly trade, our ATM option sample is split into two groups of equal size according to the options' trading volume. Subsequently, long and short portfolios are derived separately for these two groups as before. We report the mean, standard deviation (sd), minimum (min), maximum (max), and Sharpe ratio (SR) of the returns from the delta-hedged and the raw option strategy separately for calls and puts. The ATM option sample consists of 326,224 calls and 299,924 puts between January 1996 and June 2019.

			mean	sd	min	max	SR
Delta-hedged	**Calls**	low volume	0.020	0.026	-0.060	0.103	0.772
		high volume	0.020	0.035	-0.165	0.169	0.577
	Puts	low volume	0.018	0.023	-0.049	0.115	0.767
		high volume	0.015	0.027	-0.087	0.095	0.568
Raw returns	**Calls**	low volume	0.199	0.489	-1.081	2.567	0.407
		high volume	0.223	0.540	-1.775	2.530	0.412
	Puts	low volume	0.128	0.491	-1.280	2.277	0.261
		high volume	0.049	0.504	-1.438	2.340	0.097

3.4.6 Trading 50 % of the options

While at a particular date, all options are used to estimate the conditional 10 % and 90 % quantile curves, only 20 % of the options enter into the long-short portfolios. In this section, we study the returns from a trading strategy that exploits not only the most extreme options but instead trades 50 % of all options. Therefore, the long-short portfolios are formed based on the 25 % and 75 % quantile of options' IV conditional on their RV (for ATM options) or conditional on RV *and* moneyness (for the option sample of arbitrary moneyness).

Summary statistics of the returns from the delta-hedged and the raw option strategy for both ATM calls and puts as well as options of arbitrary moneyness are provided in Table 3.11. As expected, the average monthly returns when trading a larger fraction of options are lower than in the original strategy.[78] At the same time, the standard

[78] All mean returns are statistically significantly greater than zero at the 1 % level with t-statistics between 3.22 and 13.76.

deviation of the monthly returns is also lower. Taken together, we observe Sharpe ratios that are slightly lower than those of the original strategy except for the returns from the ATM raw option put strategy where the Sharpe ratio is even higher than in the original strategy. Summing up, the analysis shows that the high Sharpe ratios from our proposed trading strategies are not due to outliers but can be realized in a much broader sample.

Table 3.11: Returns when trading 50 % of the options

In our baseline strategy, only the most extreme 20 % of the options enter into the long-short portfolios. The table presents summary statistics on the monthly returns from a strategy that involves trading 50 % of the options. This trading strategy is based on the 25 % and 75 % quantiles of options' IV conditional on their RV (for ATM options) or conditional on RV *and* moneyness (for the option sample of arbitrary moneyness). We report the mean, standard deviation (sd), minimum (min), maximum (max), and Sharpe ratio (SR) of the returns from the delta-hedged and the raw option strategy separately for calls and puts. The sample period is from January 1996 to June 2019.

				mean	sd	min	max	SR
ATM		**Delta-hedged**	Calls	0.013	0.015	-0.044	0.079	0.820
			Puts	0.012	0.015	-0.054	0.063	0.783
		Raw returns	Calls	0.132	0.301	-0.632	1.300	0.440
			Puts	0.082	0.327	-0.949	2.044	0.249
Arbitrary	**moneyness**	**Delta-hedged**	Calls	0.016	0.021	-0.049	0.212	0.761
			Puts	0.016	0.023	-0.098	0.179	0.678
		Raw returns	Calls	0.135	0.360	-0.912	3.566	0.376
			Puts	0.083	0.434	-1.071	3.592	0.192

3.5 Conclusion

In this paper, we find new evidence that delta-hedged equity option returns include a volatility risk premium. We sort options on their implied volatility conditional on their realized volatility to proxy for the volatility risk premium. A strategy that is long (short) in high (low) deviations between realized and implied volatilities yields returns that are both economically and statistically significant. This result holds for

call and put delta-hedged and raw option returns for both at the money (ATM) options and options of arbitrary moneyness. Changing the type of estimator of conditional quantiles as well as controlling for additional characteristics of the underlying does not affect our main finding.

The key to our main finding, and the difference to previous work, is our use of *conditional quantiles* in contrast to standard portfolio sorts or regression techniques. Using conditional quantiles estimated via non-parametric machine learning algorithms allows us to capture the non-linear relation between implied and realized volatility while at the same time controlling for characteristics that are known to affect stock volatility as well as the cross-section of expected option returns. As our main result, we find that previous work on the existence of a risk premium for volatility and volatility mispricing were correct, but considerably underestimated the size of the effect. Exploiting the estimated quantiles of implied volatility conditional on realized volatility and moneyness leads to returns that are higher than those reported in previous work on similar volatility strategies.

Our proposed use of conditional quantiles should be seen as a good compromise between standard portfolio sorts and non-parametric cross-sectional regressions. While the former easily fail to control for more than two covariates, the latter do not come with an easy interpretability in empirical asset pricing where one is interested in ready-to-use trading strategies involving a reasonably small number of assets. While our empirical study is concerned with the cross-section of option returns, our method is sufficiently general and can easily be applied to other assets, most importantly stocks.

Chapter 4

Marginals Versus Copulas: Which Account for More Model Risk in Multivariate Risk Forecasting?

4.1 Introduction

Financial institutions employ quantitative models in almost all aspects of their risk management (e.g., for the pricing of derivatives, the modeling of credit risk portfolios, or the forecasting of market risk measures). The increase in the multiplicity and the complexity of these risk models is driven by both the occurrence of tail risks after the Great Financial Crisis as well as the incentives set by the Basel II/III and Solvency II regulations for institutions to develop and use internal quantitative models. However, any risk measurement that does not rely on a standard approach prescribed by regulators will require the risk manager to select a candidate risk model thus introducing potential *model risk*. In this paper, we study such model risk for a class of prominent models for multivariate risk forecasting: copula GARCH models. As our main result, we find that copulas account for considerably more model risk than marginals in multivariate models.

We study the model risk of multivariate risk models in a comprehensive empirical

study using copula GARCH models. Our ultimate goals are to quantify model risk in multivariate forecasts of portfolio Value-at-Risk (VaR) and Expected Shortfall (ES), and to propose ways how to reduce model uncertainty. To achieve these goals, we forecast the VaR and ES for a large number of portfolios using a variety of copula GARCH models. The first question which we want to answer is whether the model risk inherent in the forecasting of portfolio risk is caused by the candidate marginal or copula models. For this, we analyze different groups of models in which we fix either the marginals, the copula, or neither. We then propose the use of the model confidence set procedure by Hansen et al. (2011) to narrow down the set of available models and reduce model risk for copula GARCH risk models.

As our first main result, we find that model risk is economically significant for the set of candidate multivariate models that we consider. For a portfolio with a value of \$100,000 and a holding period of 10 days, model risk can account for a mean absolute deviation between the VaR (ES) forecasts by the candidate models of up to \$2,678 (\$2,264). Interestingly, these high levels of model risk are almost completely due to the choice of the copula with the choice of the marginal model having only a small effect on overall model uncertainty. Finally, and not surprisingly, periods of high market volatility lead to a surge in model risk. We then propose the use of the model confidence set procedure to narrow down the set of available models and reduce model risk for copula GARCH risk models. Our proposed approach leads to a significant improvement in the mean absolute deviation of one day ahead forecasts by our various candidate risk models.

Our paper contributes to several strands in the literature on both model risk and multivariate risk forecasting. First, our paper complements several studies on the importance of moral risk in financial risk forecasting. In this field, early studies focused on the choice of models for pricing option contracts. For example, Green and Figlewski (1999) and later Hull and Suo (2002) show that model risk in the form of inaccurate volatility forecasts and modeling errors in the implied volatility function can lead to significant risk exposure for option writers. In a similar setting, Cont (2006) proposes

a simple framework for quantifying model uncertainty in derivatives pricing. While our paper relies on similar definitions of model risk, we do no restrict our study to derivatives pricing but instead focus on the more general problem of studying model risk in the context of market risk forecasting.[79] After the Great Financial Crisis, interest in the study of model risk surged again as the failure of market and credit risk models to adequately capture tail risks was seen as a major driver of the global crisis. In this strand of the literature, Alexander and Sarabia (2012) and Glasserman and Xu (2014) both propose new methodologies for quantifying model risk with the former concentrating on Value-at-Risk models and the latter studying credit and counterparty risk. Similarly, Danielsson et al. (2016) study model risk of models used for forecasting systemic risk. In contrast to these related studies, we focus on *multivariate* risk models and disentangle the parts of model risk that are due to the separate modeling of the marginal behavior and the dependence structure in copula models.[80]

Moreover, our study is also related to a large strand of literature on copula modeling in quantitative risk management.[81] In this field of research, the majority of studies have been concerned with the model risk caused by the need of selecting the best parametric copula family in a multivariate risk model (see, e.g., Kole et al., 2007, Savu and Trede, 2008, Genest et al., 2009, Dißmann et al., 2013). In addition, several recent papers propose new copula models that are suitable for modeling dependence structures in high-dimensional financial data (e.g., large market risk portfolios) (see, e.g., Aas et al., 2009, Brechmann and Czado, 2013, Oh and Patton, 2017, 2018, Bassetti et al., 2018). Finally, several papers have looked at the superiority of copula models over competing one-dimensional or correlation-based risk models (see, e.g., Jondeau and Rockinger,

[79] It is interesting to see that the shortcomings of VaR-models were identified even as early as the mid 90s with Hendricks (1996) pointing our that even though "*[...]virtually all of the approaches produce accurate 95th percentile risk measures.*" the models considered in his study did not sufficiently capture extreme events.

[80] In this respect, our paper is related to the studies of Bernard and Vanduffel (2015) and Bernard et al. (2020) who develop a framework to allocate model risk to the different assumptions inherent in a risk model and try to incorporate information on the dependence of risks into the computation of risk bounds.

[81] Other fields in which copulas have been applied include (among others) decision trees (see Wang and Dyer, 2012), reliability modeling (see Wu, 2014), and systemic risk modeling (see, e.g., Jayech, 2016, Calabrese et al., 2017, Calabrese and Osmetti, 2019).

2006, Grundke and Polle, 2012, Weiß, 2013). Complementing these studies, our paper analyzes the economic significance and the source of model risk in copula risk models. As such, it is the first to quantify the extent to which model risk stemming from the choice of a parametric copula can lead to significant additional risk exposure for risk managers and investors.

The rest of the paper is structured as follows. In Section 4.2, we present the models and backtests employed in our empirical study as well as a description of the model confidence set. In Section 4.3, we shortly discuss the financial market data used in our study. Sections 4.4 and 4.5 present our main results of the analysis of model risk and the proposed use of the model confidence set procedure, respectively. Section 4.7 concludes.

4.2 Market risk models and model risk

This section provides details on the estimation of market risk via copula GARCH models, backtesting of the risk estimates as well as the calculation of model risk. In addition, we introduce the model confidence set (MCS) procedure due to Hansen et al. (2011) yielding a set of models that contains the best model with a certain probability.

4.2.1 Multivariate estimation of market risk

As a consequence of the theorem by Sklar (1959), one can separate the modeling of the marginals and the dependence of multivariate return series. This is usually done by using GARCH-type models to filter the univariate time series while copulas are subsequently applied to model the dependence structure between different assets in a given portfolio. We will now present the corresponding two step approach to multivariate time series modeling in more detail.

We start with modeling the marginals as ARMA-GARCH-type processes with various error distributions. We consider not only the standard GARCH model by Bollerslev (1986), but also employ the EGARCH model by Nelson (1991), GJR-GARCH

model by Glosten et al. (1993), T-GARCH model by Zakoian (1994), and aPARCH model by Ding et al. (1993). These models are all nested within the fGARCH model by Hentschel (1995). We provide the specifications for an ARMA(p,q)-GARCH(r,s) process in Section B.1.1 in more detail as this is a very popular representative of this class of models. For the remaining models we refer to the reference guide by Bollerslev (2010) as well as to the original papers.

We consider innovations following a normal distribution, a Student-t distribution, a skewed Student-t distribution, and a generalized error distribution.[82] Apart from the normal distribution, these distributions are able to account for skewness and/or fat tails in the data. Combining these four distributions with the five GARCH-type models yields in total 20 different (univariate) specifications that we use to fit ARMA(1,1)-GARCH(1,1)-type processes to the return series.[83]

In a second step, we apply copula dependence models to the GARCH filtered data from the first step. Copulas are multivariate distributions with all marginals being uniformly distributed in the interval $[0, 1]$. For a d-dimensional distribution with distribution function F and marginals $F_1, \ldots, F_d$, the copula associated with F is a function $C : [0, 1]^d \to [0, 1]$ satisfying

$$F(\mathbf{x}) = C(F_1(x_1), \ldots, F_d(x_d))$$

for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. On the basis of the two step approach lies the fact that copulas enable us to model the dependence structure of multivariate distributions separately from the marginal distributions. This is a consequence of the theorem by Sklar (1959), see Section B.1.2 for more details.

We consider various copula functions. We include the Gaussian and Student-t copulas because they are widely used in financial applications (cf. Cherubini et al., 2004, McNeil et al., 2005). We additionally employ the Archimedean copulas Clayton, Gum-

[82] For details on the skewed Student-t distribution and the generalized error distribution we refer to Fernandez and Steel (1998) and Nadarajah (2005).

[83] Estimation is performed based on the R-package *rugarch* by Ghalanos (2020).

bel, Frank, and Joe copula. This gives us further possibilities to model multivariate dependence as for example the Clayton copula allows for modeling a positive lower tail dependence. For more details on copulas in general and the copulas used in this paper we refer to the comprehensive books by by Joe (2001) and Nelsen (2006).

While the Gaussian and Student-t copulas can only model symmetric dependencies using correlation matrices, Archimedean copulas typically have only one (like the above mentioned copulas) or two parameters, which is very restrictive. So-called pair copula constructions (also referred to as vine copulas) offer a very convenient possibility for a highly flexible modeling of the dependence structure. Originally introduced by Joe (1996), further significant contributions have been made by Bedford and Cooke (2001, 2002), and Kurowicka and Cooke (2006). Additionally, we include the Gaussian mixture copula due to Tewari et al. (2011) that can capture multi-modal dependencies as well as asymmetric and tail dependencies. Another very popular model is the Dynamic Conditional Correlation (DCC) model due to Engle and Sheppard (2001) and Engle (2002). Though not a copula model, the DCC model fits perfectly into our framework. It allows for a two stage estimation of the model parameters where in a first stage GARCH-type models are fitted to each of the univariate return series. The conditional correlation matrix is derived in a second step to model the multivariate dependence between the return series. In the following, we will therefore subsume all models under the term copula GARCH models.[84]

For each combination of the considered copula and ARMA-GARCH-type models we now proceed as follows to obtain one day ahead VaR and ES forecasts for a given portfolio of K assets: First, we simulate 10,000 K-dimensional vectors of standardized residuals from the fitted copula. Based on the parameters and ex-ante mean and vari-

[84]Estimation for the Gaussian, Student-t, Clayton, Gumbel, Frank, and Joe copula is performed based on the *copula* R-package by Hofert et al. (2020). Inference for the vine copula is performed using the *VineCopula* R-package by Nagler et al. (2019) considering the Gaussian, Student-t, Frank, Clayton, Gumbel, and Joe copula (along with rotated and survival versions of the latter three copulas) as bivariate building blocks of a regular vine copula. Selection of a particular bivariate copula is performed based on the AIC value. We estimate a Gaussian mixture copula with three components (two for capturing potentially fat tails, one for "regular" returns) employing the *GMCM* R-package by Bilgrau et al. (2016). Inference for the DCC model is based on the *rmgarch* R-package by Ghalanos (2019) assuming a multivariate Student-t distribution.

ance forecasts from the ARMA and GARCH-type models that have already been fitted to the univariate return series, we transform the simulated standardized residuals into 10,000 return forecasts for each of the portfolio constituents. We finally derive 10,000 portfolio returns and calculate the VaR as the sample quantile and the ES as the conditional mean of the returns falling below this quantile (both values are subsequently multiplied by (-1)). This procedure is repeated on a daily basis based on a moving window of data. For more details we refer to Brechmann and Czado (2013), see also Aas and Berg (2009), Ausin and Lopes (2010), and Nikoloulopoulos et al. (2012). One important advantage of this approach is that copulas and marginals only have to be fitted once to consider various portfolio weights. This is because the weights enter into the procedure only at the end when calculating the portfolio returns.[85]

4.2.2 Backtests

Our aim is to analyze the risk associated with choosing an appropriate model from a variety of valid candidate models, not to identify the *optimal* model for forecasting VaR and ES estimates. Nonetheless, we backtest our risk forecasts before measuring the model risk itself for the following reasons: First, in order to determine a set of *valid* candidate models, we need to prevent our results from being biased by erroneous risk forecasts due to misspecified models. Second, the approach of calculating model risk after having applied backtests is also favorable from a more practical perspective. The Basel III regulation requires banks to backtest their (internal) market risk models. For banks, uncertainty on the model choice is, essentially, uncertainty on the choice of models that have not been rejected by backtests. This is in line with our calculation of model risk.

We measure the quality of the VaR estimates using the duration-based test of Christoffersen (2004). The test follows a more general approach of independence rather than focussing on the independence of VaR violations only. The test assumes

[85] We make use of this fact to include 100 portfolios with randomly generated portfolio weights to ensure robustness of our results.

that in the case of independent VaR violations, the time between violations must be independent of the time to a previous violation. In simple terms, this means that the probability that a VaR violation will occur in the next 10 days must be independent of whether the last one occurred in the last 10 or 100 days.[86] Christoffersen (2004) shows that his duration-based test tends to reveal VaR methods that violate the independence property in realistic situations more often and hence has better power properties than former tests. For more details on the duration-based test we refer to B.1.3 as well as to the original paper.[87]

Since the VaR has been replaced by the ES as the leading market risk measure in regulatory requirements, the debate on suitable backtesting for the ES has also been increasing in the literature. In view of the regulatory requirements and because the VaR does not have to be issued by financial institutions, ideally a backtest should only determine the quality of the ES model based on real data and the forecasts. In practice, however, many backtests require additional input variables or assumptions.[88] We evaluate ES estimates using a comparatively new test by Nolde and Ziegel (2017), which is based on the concept of conditional calibration (CC). For a brief description of the test, we refer to B.1.4 and the original paper. The CC test requires in its simple version both VaR and ES forecasts.[89] We find the use of a joint backtest of VaR and ES predictions sensible for the following reasons: First, in line with Bayer and Dimitriadis (2020b), we find that the ES is by definition closely related to the VaR and thus suitable ES forecasts are based on adequate VaR estimations. Secondly, we apply the MCS procedure at a later point of our analysis. This method is again based on functions that require both VaR and ES forecasts. Consequently, the use of a joint VaR and ES backtest is appropriate.[90]

[86] See Campbell (2007).

[87] In addition, we implement the dynamic quantile test of Engle and Manganelli (2004), which as a conditional coverage test examines not only the number and independence of VaR hits but also the independence of the estimators. The results can be found in Section 4.6.

[88] See Bayer and Dimitriadis (2020b).

[89] In the general version, volatility is also taken into account.

[90] As a supplement, we use the exceedance residual test of McNeil and Frey (2000), which is based on the ES-specified residuals that exceed the VaR. A comprehensive description of the test can also be found in Bayer and Dimitriadis (2020b). The exceedance residual test is a joint backtest as well,

By using a joint backtest for the ES, it should be noted that this is a possible reason why more models fail the backtest than in the case of the VaR. For this reason and because different confidence levels are considered in the baseline case, we separate the analysis of the VaR and the ES in terms of model risk.

4.2.3 Model risk

Financial risk cannot be measured directly, but only be estimated using statistical models. However, it is well known that distinct models may differ vastly in their risk predictions (see, e.g., Danielsson et al., 2016). In the literature there are various approaches for defining model risk. A large part of research focuses on the factors that control model risk within models such as the misspecification of the underlying theoretical models (Green and Figlewski, 1999) or assumptions made about unknown (or unobservable) parameters, distributions, or other model specifications (e.g., Hull and Suo, 2002, Alexander and Sarabia, 2012, Glasserman and Xu, 2014, Boucher et al., 2014). However, we focus on a more general approach: With a variety of standard VaR and ES models within the industry, uncertainty about the choice of such a model creates model risk per se. Especially, our aim is not to identify an *optimal* model for forecasting VaR or ES estimates, but to analyze the risk that emerges from the presence of a large set of valid candidate models. Our notion of model risk as uncertainty on the model choice itself in the presence of many possible alternative models is most closely related to Cont (2006) and Danielsson et al. (2016).

We consider all VaR and ES models that are not rejected in the respective backtest as viable candidates for measuring market risk. This leaves us with a large number of models and corresponding risk forecasts for the same quantity. As a measure of model risk, we quantify the level of disagreement between the individual estimates based on the mean absolute deviation (mad), the standard deviation (sd), and the interquartile range (iqr). The mean absolute and the standard deviation are intuitive and plausible

which is why the same reasons for use as for the conditional calibration test of Nolde and Ziegel (2017) apply. The results can be found in Section 4.6.

measures since both take into account how much the models deviate from the average forecast. The iqr of an observation variable is the difference of its 75% and 25% percentiles and consequently a measure of the maximum disagreement within the 50% of observations around the median. We choose the mad as our main measure of model risk for the following reasons: First, the mad is more robust against outliers than the sd. Second, it can easily be interpreted in absolute terms relative to a given portfolio value and reflects the deviation of a forecast by a randomly chosen model from the average risk forecast. Finally, since our main analysis relies on forming different groups of models, we need a measure of model risk that is as independent as possible from the number of models within a particular group. For these reasons, we focus on the mad to measure model risk and include the measures sd and iqr for robustness.[91]

4.2.4 Model confidence set

Researchers, practitioners, and regulators are often confronted with situations where a variety of models for computing a specific estimate, e.g., for VaR or ES forecasts, exist. Optimally, one would like to know which of the many available models is the *best*. However, in many situations this question cannot be answered, especially when the set of competing methods is large and the data are not sufficiently informative. Yet, one can try to reduce the set of available models to a smaller set of alternatives. This can be done by the model confidence set (MCS) procedure by Hansen et al. (2011).

The MCS procedure yields a set of models (the *model confidence set*) that contains the *best* model with a certain *probability*. That is, the procedure does in general not

[91] Another measure to determine model risk is the *risk ratio* by Danielsson et al. (2016) that is defined as the ratio of the highest to the lowest risk forecast within a set of candidate models. Disagreement between models is therefore captured by a risk ratio greater than 1. We do not include the risk ratio into our analysis for the following reasons: First, we want to focus on the *average* deviations of risk forecasts whereas the risk ratio tends to capture the extreme, *maximum* possible deviations within a set of candidate models. Moreover, the mad is more suitable in our context than the risk ratio, because risk estimates by the latter can depend on the number of models. For example, let us assume we have a group of models A with maximum $\max_A$ and minimum $\min_A$. For an arbitrary subgroup B of group A, we have $\min_B \geq \min_A$ and $\max_B \leq \max_A$. As a consequence, the risk ratio of group B is smaller than (or equal to) the risk ratio of the larger group of models A. As our approach of studying the importance of modeling the marginals and the multivariate dependence structure heavily relies on building subgroups, this property could potentially bias our results.

identify a best model nor does it assume that a particular model represents the true data generating process. Instead, the MCS can be seen as an analogue to a confidence interval that contains a parameter of interest with a specified probability. An important advantage of the MCS procedure over methods that choose a single model is that it accounts for the informativeness of the data at hand. When data are very informative one may obtain a MCS that consists of only the best model. Less informative data, on the other hand, may lead to a MCS containing several models as the data make it hard to distinguish between models. Additionally, the MCS procedure allows for valid statements about significance that are not hampered by multiple pairwise comparisons (Hansen et al., 2011). These attractive features make it interesting to apply the MCS procedure to our set of VaR and ES models. The MCS procedure might help in further narrowing down the set of valid candidate models (after applying the backtests) such that the remaining models exhibit a lower model risk.

The construction of the MCS procedure relies on an *equivalence test*, $\delta_{\mathcal{M}}$, and an *elimination rule*, $e_{\mathcal{M}}$.[92] First, the equivalence test is applied to the set of candidate models $\mathcal{M}^0$. If equivalence is rejected at a given confidence level α, this implies that the candidate models are not equally "good". Thus, the elimination rule is applied to remove a poorly performing model from the set $\mathcal{M}^0$. These two steps are repeated until the equivalence test is not rejected for the first time. The remaining elements of $\mathcal{M}^0$ are then considered as the model confidence set $\hat{\mathcal{M}}_{1-\alpha}$. As the same confidence level α is used in each iteration for the equivalence test, the procedure guarantees that $\lim_{n\to\infty} P(\mathcal{M}^* \subseteq \hat{\mathcal{M}}^*_{1-\alpha}) \geq 1-\alpha$, where $\mathcal{M}^*$ denotes the true set of best models and n is the number of observations per model. Additionally, the MCS procedure provides p-values for each model that can be interpreted as the probability that the respective model is among the best alternatives in $\mathcal{M}^0$. For more details on the procedure we refer to Section B.1.5 and to the original paper.

In the MCS approach, models are evaluated based on a user-defined loss function.

[92] This short introduction to the MCS procedure is based on the original paper by Hansen et al. (2011) and we refer to it for more details and proofs of the results.

For evaluating the forecast accuracy of risk models it is natural to compare the forecasts to the realized financial losses over a period of time. Nolde and Ziegel (2017) highlight that for comparing different models' risk forecasts elicitability of the risk measure is a desirable property. Generally speaking, a risk measure is elicitable if it minimizes the expected value of a *scoring function*.[93] Elicitability is a property that has been proven to be useful for forecast ranking, comparative backtesting, and for model selection (Nolde and Ziegel, 2017).

As the VaR represents a quantile of a probability distribution (multiplied by -1) it is well known to be elicitable (see, e.g., Koenker and Bassett, 1978). The associated scoring function, the so-called check-function, is given by

$$L_{VaR}(r_t, VaR_\alpha^t, \alpha) := (r_t - (-VaR_\alpha^t)) \cdot (\alpha - \mathbb{I}_{(-\infty,0)}(r_t - (-VaR_\alpha^t))),$$

where r_t and VaR_α^t denote the realized return and the VaR forecast with coverage level α (and confidence level $1-\alpha$) at day t and $\mathbb{I}_{(-\infty,0)}$ denotes the characteristic function of the open interval $(-\infty, 0)$. We choose this scoring function as the loss function for the VaR in the MCS framework.

As opposed to the VaR, the ES alone is not elicitable. Instead, Fissler et al. (2016) show that ES and VaR are *jointly* elicitable, see also Acerbi and Székely (2014). There is a growing body of literature building on this result, see, e.g., Fissler and Ziegel (2016), Nolde and Ziegel (2017) for forecast comparisons and Patton et al. (2019), Barendse et al. (2021), Bayer and Dimitriadis (2020b) for applications in a regression procedure. As the ES is only jointly elicitable with the VaR, we choose a loss function for the ES that is based on both VaR and ES forecasts as input into the MCS procedure. This is consistent with the conditional calibration backtest introduced in Section 4.2.2 that is used to determine the set of candidate models for the ES. As Nolde and Ziegel (2017) provide a joint scoring function for the VaR and the ES only in a general form,

[93] See Gneiting (2011) for a comprehensive literature review on elicitability as well as Frongillo and Ian A. Kash (2015), Fissler et al. (2016), Ziegel (2016) for more recent advances in the field.

we adopt the 0-homogeneous version introduced in Patton et al. (2019)

$$L_{ES}(VaR_\alpha^t, ES_\alpha^t, r_t, \alpha) := \frac{1}{\alpha ES_\alpha^t} \cdot \mathbb{I}_{(-\infty,0)}(r_t + VaR_\alpha^t) \cdot (-VaR_\alpha^t - r_t) + \frac{VaR_\alpha^t}{ES_\alpha^t} + \log(ES_\alpha^t) - 1,$$

where the notation is as above.[94] For an implementation of the MCS procedure we rely on the R-package *MCS* by Catania and Bernardi (2017). For more details we refer to Section 4.5.

4.3 Data

We form well diversified portfolios consisting of equity indices (developed and emerging markets), bond indices (governement, corporate, and high-yield bonds) as well as commodity and real estate indices. Therefore, we retrieve the total return indices (in US$) of the following set of indices from Datastream: Stoxx Europe 600, Dow Jones Industrial Average, FTSE Developed Asia Pacific Index, MSCI Emerging Markets Index, S&P U.S. Treasury Bond Index, S&P 500 Investment Grade Corporate Bond Index, S&P U.S. High Yield Corporate Bond Index, S&P Pan-Europe Developed Sovereign Bond Index, S&P GSCI, and Developed Markets Datastream Real Estate Index. The sample period is January 2001 to December 2018. Next, we calculate geometric returns that allow us to easily derive portfolio returns. For our main analysis we focus on an equally weighted portfolio. This corresponds to a portfolio consisting of 40% stocks, 40% bonds, 10% commodities, and 10% real estate. For robustness we also consider portfolios based on random portfolio weights that were drawn from a unit-simplex. Summary statistics on the equally weighted portfolio returns as well as on the individual index returns can be found in Table 4.1.

We calculate daily VaR and ES estimates based on 180 different model specifications of copula GARCH models, see Section 4.2.1 for details. Additionally, we derive risk estimates from univariate GARCH-type models applied to the portfolio return series.

[94] Note that throughout the paper we regard VaR and ES estimates as positive values.

Table 4.1: Summary statistics of index and portfolio returns

The table provides summary statistics for the returns of the indices that form the basis of our well-diversified portfolio as well as for the resulting portfolio returns with equal weighting. We include equity indices (developed and emerging markets), bond indices (governement, corporate, and high-yield bonds), a commodity index, and a real estate index. We retrieve the total return indices (in $) from *Datastream* for a period from January 2001 to December 2018. We first calculate geometric returns and derive equally weighted portfolio returns next. This corresponds to a portfolio consisting of 40% stocks, 40% bonds, 10% commodities and 10% real estate. We provide minimum (Min), median, mean, maximum (Max), and standard deviation (SD) of daily returns in percent as well as the kurtosis and skewness of the time series.

	Min (in %)	Median (in %)	Mean (in %)	Max (in %)	SD (in %)	Kurtosis	Skewness
Indices							
Stoxx Europe 600	-9.691	0.042	0.024	11.284	1.341	10.642	-0.004
Dow Jones Industrial Average	-7.873	0.033	0.032	11.080	1.107	12.686	0.100
FTSE Developed Asia Pacific Index	-8.524	0.047	0.024	9.928	1.185	8.097	-0.262
MSCI Emerging Markets Index	-9.484	0.095	0.040	10.598	1.178	11.372	-0.306
S&P U.S. Treasury Bond Index	-1.658	0.012	0.014	1.758	0.232	5.995	-0.120
S&P 500 Investment Grade Corporate Bond Index	-1.668	0.027	0.020	1.845	0.284	5.028	-0.232
S&P U.S. High Yield Corporate Bond Index	-3.715	0.042	0.027	2.226	0.246	32.336	-2.377
S&P Pan-Europe Developed Sovereign Bond Index	-3.705	0.020	0.021	5.028	0.617	5.897	0.138
S&P GSCI	-8.762	0.000	-0.003	7.483	1.434	5.787	-0.148
Developed Markets Datastream Real Estate Index	-6.849	0.061	0.036	7.223	0.939	11.411	-0.370
Equally weighted portfolio	-4.009	0.042	0.023	4.110	0.552	11.085	-0.516

Estimations are performed based on various confidence levels (99.9%, 99%, 97.5%, and 95%)[95] using a moving window of 500 days corresponding to approximately two years of daily observations[96]and a forecast horizon of one day. We clean the risk estimates from outliers that are due to convergence errors in fitting the copula GARCH models.[97]

Subsequently, we perform VaR or ES backtests to determine the set of candidate models that enter into the calculation of model risk on a daily basis. The backtests are based on a confidence level of 99% in line with Basel Committee on Banking Supervision (2019) and a moving window of 500 days. Note that by employing a moving window for the backtests, we avoid introducing a look-ahead bias into the selection of the set of candidate models. In our main analysis, we rely on the duration-based VaR backtest by Christoffersen (2004) and the conditional calibration ES backtest by Nolde and Ziegel (2017), see Section 4.2.2 for details.[98]

Afterwards, we calculate model risk on a daily basis for the risk models that have passed the respective backtest. Note that by using moving windows the composition of the set of candidate models may vary over time. Since VaR and ES estimations as well as backtests are performed based on a 500 day moving window in our baseline analysis, we obtain daily model risk estimates from day 1001 onwards. This corresponds to the time period from November 4, 2004 until December 31, 2018. We obtain model risk estimates for both VaR and ES forecasts for various confidence levels and portfolio weights. In our main analysis we focus on the model risk of risk forecasts for an

[95] In line with the Basel II and III market risk regulations we focus on the 99% VaR and the 97.5% ES.

[96] For robustness we also consider a moving window of 1000 days.

[97] We identify outliers based on the daily absolute changes of the risk forecasts. Therefore, we calculate z-scores based on per model standard deviation and mean calculated over the first 500 risk estimates. We then replace observations with a z-score above 25 with the value from the previous day. This affects on average 0.17% of all risk estimates. Note that by this procedure we do not introduce a look-ahead bias into our analysis.

[98] We use the *rugarch* R-package by Ghalanos (2020) for performing the duration-based backtest and the *esback* R-package by Bayer and Dimitriadis (2020a) for the conditional calibration backtest (simple version one-sided using Hommel's correction). For comparison, we also run the backtests using a moving window of 1000 days as well as a fixed window over the entire period. Additionally, we perform the dynamic quantile test by Engle and Manganelli (2004) with the *GAS* R-package by David Ardia et al. (2019) and the exceedance residual backtest by McNeil and Frey (2000) (one-sided) with 1000 bootstrap iterations implemented in the *esback* R-package by Bayer and Dimitriadis (2020a). For more details see Section 4.6.

equally weighted portfolio and the 99% VaR and 97.5% ES in line with the Basel II and Basel III market risk regulations.[99] Our main measure of model risk is the mean absolute deviation (mad) of risk forecasts.

4.4 Analysis of model risk

4.4.1 All multivariate models

We start with an analysis of the model risk over time and different market conditions of all multivariate VaR and ES models that passed the respective backtest.[100] Figure 4.1 presents the daily model risk associated with one day ahead forecasts of the 99% VaR and the 97.5% ES in terms of mad between November 4, 2004 and December 31, 2018. The figure reveals that model risk is normally quite moderate but increases significantly during and after the global financial crisis. Summary statistics are provided in Table 4.2. Over the entire time period, model risk is on average about 0.165% of the portfolio value for VaR forecasts and about 0.092% of the portfolio value for ES forecasts. Model risk is quite volatile over time ranging from 0.075% to 0.847% for the VaR and from 0.029% to 0.716% for the ES with a standard deviation of daily model risk estimates of more than half of the average model risk.

Model risk is especially pronounced during times of financial turmoil. During the years 2008-2009 (in the following referred to as the *crisis period*) the average model risk is 0.286% of the portfolio value for the VaR and 0.145% for the ES. That is, the average model risk more than doubles compared to the period before 2008 (the *pre-crisis period*) with an average VaR of 0.119% and ES of 0.063 %.[101] The maximum

[99] In the following, VaR will refer to the 99% VaR and ES will refer to the 97.5% ES unless specified differently.

[100] We start with 180 different multivariate model specifications. After applying the backtests we are left with on average 174 99% VaR 121 97.5% ES models.

[101] These differences between pre-crisis and crisis period are statistically significant at the 1% level where statistical significance throughout the paper is determined based on t-tests with standard errors corrected for serial correlation and heteroskedasticity according to Newey and West (1987) with the automatic bandwidth selection procedure described in Newey and West (1994). Note that although for robustness we include 100 randomly generated portfolio weights into the study, the portfolio bootstrap procedure according to Danielsson et al. (2016) is not applicable in our setting. This is, because

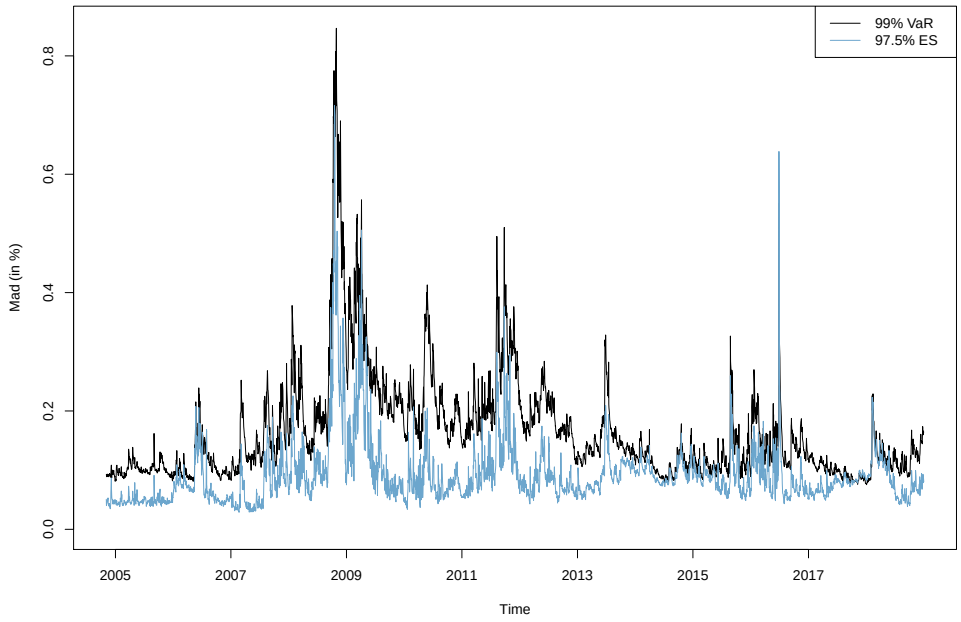
Table 4.2: Model risk and the great financial crisis

This table presents summary statistics for the time series of model risk associated with 99% VaR and 97.5% ES forecasts for a well diversified portfolio. Model risk is measured in terms of the mean absolute deviation (mad) of risk forecasts by various models from November 4, 2004 until December 31, 2018. With *crisis* we refer to the years 2008-2009 while *pre-crisis* and *after-crisis* denote the period before and after, respectively. Model risk is calculated based on all multivariate models that passed the respective backtest, see Section 4.2.2 for details. We report model risk estimates for one day ahead risk forecasts in percent of the portfolio value (first and second column) and in absolute terms (third and fourth column) for an equally weighted portfolio. The results in absolute terms are based on a portfolio value of $100,000 and a 10 day forecast horizon obtained by applying the square-root-of-time rule. We provide minimum (Min), median, mean, maximum (Max), and standard deviation (SD) of the daily model risk estimates. Further results for randomly generated portfolio weights as well as alternative measures of model risk can be found in Tables 4.3 and 4.4.

		Model risk (in %)		**Model risk (in $)**	
		99% VaR	97.5% ES	99% VaR	97.5% ES
Whole period	Min	0.075	0.029	237	92
	Median	0.138	0.080	436	253
	Mean	0.165	0.092	522	291
	Max	0.847	0.716	2,678	2,264
	SD	0.091	0.056	288	177
Pre-crisis	Min	0.081	0.029	256	92
	Median	0.105	0.050	332	158
	Mean	0.119	0.063	376	199
	Max	0.280	0.207	885	655
	SD	0.036	0.030	114	95
Crisis	Min	0.116	0.046	367	145
	Median	0.241	0.108	762	342
	Mean	0.286	0.145	904	459
	Max	0.847	0.716	2,678	2,264
	SD	0.139	0.101	440	319
Post-crisis	Min	0.075	0.034	237	108
	Median	0.138	0.082	436	259
	Mean	0.155	0.090	490	285
	Max	0.551	0.638	1,742	2,018
	SD	0.064	0.037	202	117

Figure 4.1: Daily model risk for all multivariate models

This figure shows the model risk associated with one day ahead 99% VaR and 97.5% ES forecasts for a well diversified portfolio. Model risk is measured in terms of the mean absolute deviation (*mad*) of one day ahead forecasts by various risk models. Values are calculated on a daily basis between November 4, 2004 until December 31, 2018 in percent of the portfolio value based on all multivariate models that passed the respective backtest, see Section 4.2.2 for details.



model risk values are realized in Q4 2008 in the follow-up of the Lehman Brothers bankruptcy on the peak of the financial crisis. The extraordinary impact of the financial crisis on model risk is further highlighted by the fact that nearly all VaR model risk values above the 99% quantile occurred in Q4 2008. The same is true for the majority of the ES model risk values.[102]

These results can only partly be explained by an increase in volatility. On the one

the portfolio weights enter into the risk forecast after the GARCH models and copula functions are fit.

[102] Further high model risk values occurred in particular in Q2 2009, Q3 2011, and Q3 2015. The highest model risk value for the ES was realized on October 17, 2008, just two days after the Dow Jones Industrial Average Index experienced its largest drop in relative terms since 1987. The highest model risk value for the VaR was realized on October 29, 2008. Two days earlier, the Nikkei 225 Index lost more than 6.4% while the Hang Seng Index decreased by 12.7% while the consecutive day world wide stock markets saw a huge rally in anticipation of rate cuts by central banks.

hand we calculate VaR and ES forecasts based on conditional volatility estimates derived from GARCH-type models. Consequently, whenever volatility is high, VaR and ES forecasts will on average also show increased levels resulting in a higher model risk. On the other hand, the average model risk in 2008-2009 is 140% higher than in the pre-crisis period for the VaR while the average level of VaR estimates is only 113% higher. Similarly, the model risk for the ES in 2008-2009 is increased by 130% while the average level of ES estimates is 119% higher. This disproportionately high increase of model risk in periods of crisis might be due to the fact that all models treat history and shocks quite differently such that a change in statistical regimes can be expected to lead to higher disagreements between risk forecasts (Danielsson et al., 2016). Following the great financial crisis, model risk does not decrease to the pre-crisis level (0.119% for the VaR and 0.063% for the ES) but remains elevated at 0.155% (VaR) and 0.090% (ES).[103] The findings are robust to considering randomly generated portfolio weights, see Table 4.3. Summary statistics for other measures of model risk (standard deviation, interquartile range) are provided in Table 4.4.

Risk models are embedded within the Basel accords and play a central role in the regulatory process to determine bank capital. That is, expensive decisions such as the amount of capital held or portfolio allocations depend on the outputs of risk forecasting models as input. We highlight this point by providing model risk estimates in absolute terms in Table 4.2. These values can be interpreted as average deviations in regulatory capital to be held according to different models (at a particular day). We calculate model risk estimates in $ by assuming a portfolio value of $100,000 and a holding period of 10 days. Therefore, we multiply the mad values (as percentage of the portfolio value) with $100,000 and $\sqrt{10}$.[104] We thereby emphasize that differences in model risk are not only statistically but also economically significant. This is also highlighted

[103] These differences between the pre-crisis and the after-crisis period are statistically significant at the 1% level.

[104] The square-root-of-time rule for scaling daily VaR forecasts is the industry standard although this approach might lead to underestimation (Danielsson and Zigrand, 2006, Wang et al., 2011) or overestimation of the 10-day VaR (Diebold et al., 1997), see Kole et al. (2017). A detailed analysis on the effects of different choices of temporal aggregation can be found ibid. For simplicity, we also rely on the square-root-of-time rule for scaling ES forecasts.

Table 4.3: Model risk for all multivariate models averaged over 100 random portfolios

This table provides the same results as Table 4.2 but for 100 random portfolios obtained by drawing portfolio weights from a unit-simplex. Therefore, we first calculate summary statistics for the time series of model risk for each of the portfolios. These statistics are then averaged over all 100 portfolios. The results can thus be interpreted as summary statistics for the model risk of an average portfolio.

		Model risk (in %)		**Model risk (in $)**	
		99% VaR	97.5% ES	99% VaR	97.5% ES
Whole period	Min	0.065	0.034	206	106
	Median	0.127	0.085	401	268
	Mean	0.153	0.099	484	313
	Max	0.813	0.763	2570	2413
	SD	0.086	0.057	272	181
Pre-crisis	Min	0.069	0.037	218	118
	Median	0.097	0.064	307	203
	Mean	0.111	0.074	351	233
	Max	0.281	0.233	890	735
	SD	0.036	0.029	113	91
Crisis	Min	0.108	0.046	342	146
	Median	0.221	0.114	697	361
	Mean	0.267	0.153	843	483
	Max	0.807	0.711	2551	2249
	SD	0.133	0.105	420	331
Post-crisis	Min	0.069	0.039	217	123
	Median	0.127	0.087	401	276
	Mean	0.143	0.096	451	303
	Max	0.593	0.626	1876	1980
	SD	0.060	0.039	189	123

by Figure 4.2 illustrating the extent of disparity between VaR forecasts. The average model risk (mad) in absolute terms over the entire time period is $522 for the VaR and $291 for the ES with maximum values of $2,678 (VaR) and $2,264 (ES).

Note that model risk of VaR and ES forecasts cannot directly be compared to each other. This is due to the fact that risk forecasts entering into the calculation of model

Table 4.4: Alternative measures of model risk

This table provides summary statistics for different measures of model risk for 99% VaR and 97.5% forecasts. Our baseline measure is the mean absolute deviation (mad). We additionally include the standard deviation (sd) and interquartile range (iqr) of the risk forecasts by various models in percent of the portfolio value, see Section 4.2.3 for more details. Model risk is calculated for all multivariate models that passed the respective backtest. We provide minimum (Min), median, mean, maximum (Max), and standard deviation (SD) of the daily model risk estimates (according to the different model risk measures) over the period November 4, 2004 until December 31, 2018.

		Model risk				
		Min	**Median**	**Mean**	**Max**	**SD**
	Measure					
VaR	mad (in %)	0.075	0.138	0.165	0.847	0.091
	sd (in %)	0.088	0.161	0.195	0.992	0.108
	iqr (in %)	0.130	0.269	0.325	1.631	0.180
ES	mad (in %)	0.029	0.080	0.092	0.716	0.056
	sd (in %)	0.037	0.099	0.117	0.867	0.072
	iqr (in %)	0.040	0.133	0.152	1.399	0.094

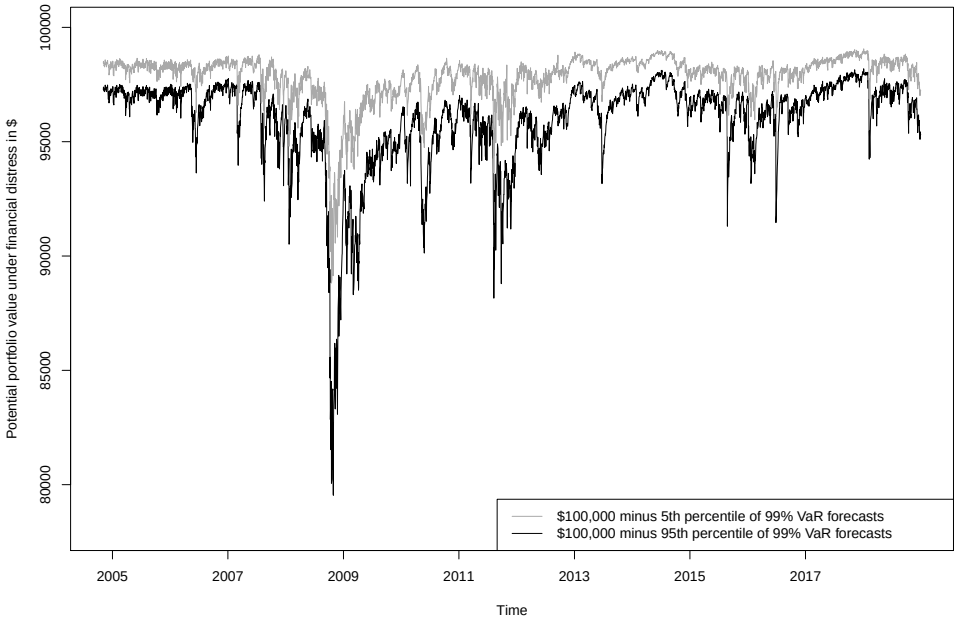
risk are determined based on different backtests for the VaR and the ES. As a result, on average 174 different model specifications enter into the model risk calculation for the VaR and only 121 specifications into the model risk calculation for the ES. The effect of choosing different backtests or no backtest at all are discussed in Section 4.6 in more detail.

When considering several confidence levels for VaR and ES forecasts we observe that model risk for the 99.9% VaR (0.306%) is approximately twice as high as for the 99% VaR (0.165%) which again is twice as high as for the 95% VaR (0.085%). For the 99.9% ES (0.309%), 99% ES (0.142%), 97.5% ES (0.092%), and 95% ES (0.067%) the proportions are similar, see Table 4.8 for more details.[105] This rise in model risk when increasing the underlying confidence level can, however, only partially be ex-

[105] All differences in pairwise comparisons of the same risk measure at different confidence levels are statistically significant at the 1% level. Note that the backtests are performed separately for each confidence level. However, when focusing on the model risk of either VaR or ES, differences in the average percentage of models that passed the respective backtest are quite low. The average percentage of models that passed the backtest is 99.3%, 97.6%, 96.7%, and 98.9% for the VaR and 60.7%, 67.2%, 75.5%, and 68.6% for the ES at the 95%, 97.5%, 99%, and 99.9% confidence level, respectively.

Figure 4.2: Potential portfolio value under financial distress

This figure illustrates the economic significance of model risk arising from the disparity between different VaR forecasts. Here, we focus on the 99% VaR for a well diversified portfolio ($100,000) and a 10 day holding period. We provide the portfolio value minus the 5th and the 95th percentile of VaR forecasts from all multivariate models that passed the duration-based backtest by Christoffersen (2004) on a daily basis. This corresponds to the potential portfolio value under financial distress according to the more (95th percentile) or less (5th percentile) conservative VaR models. The sample period is November 4, 2004 until December 31, 2018.



plained by an increase in the absolute level of the risk forecasts as a consequence of the higher confidence level. That is, even when relating the average model risk to the average level of risk forecasts, model risk is more pronounced for higher confidence levels. This can be seen when looking on the ratio of average model risk divided by the average level of risk forecasts which is 0.208, 0.163, 0.142 and 0.126 for the VaR and 0.201, 0.111, 0.086 and 0.075 for the ES at the 99.9%, 99%, 97.5% and 95% confidence level, respectively. These results are not surprising as they mainly reflect that the disagreements between different models in modeling the tails of the return distributions increase when considering more extreme quantiles.

4.4.2 Analysis of the subgroups

The VaR and ES estimates are obtained via copula GARCH models in a two step procedure. Therefore, we turn to the question if a greater portion of model risk is attributable to the statistical modeling of the univariate marginals (via GARCH-type models) or to the estimation of the multivariate dependence structure (via copulas). For this, we analyze different groups of models in which we fix either the marginals, the copula, or neither.

Figure 4.3 provides the average model risk for the 99% VaR and the 97.5% ES for on an equally weighted portfolio from November 4, 2004 to December 31, 2018 for four different groups: Group 1 covers the average model risk across the sets of models in which a copula is fixed while the marginal distribution is varied. Group 1 thus measures the impact of choosing a specific GARCH-type model for the marginals. Analogously, Group 2 captures the average model risk among sets of models with fixed marginal distributions and varying copulas. Additionally, Group 3 measures the average model risk of all multivariate and Group 4 of all univariate models. Table 4.5 presents the corresponding descriptive statistics.

Considering the 99% VaR, average model risk for model sets with fixed copulas is 0.052% of the portfolio value. For sets with fixed marginal distributions, on the other hand, model risk increases significantly and is three times higher with an average of 0.157% of the portfolio value. Consequently, model risk is higher when choosing a copula function compared to choosing the marginal distribution. Higher model risk in the choice of a copula function means that the risk forecasts in groups of models with fixed marginal distribution and varying copula functions differ more from each other than risk forecasts in groups where the marginal distributions vary while the copula is held constant. In absolute terms, this translates into an average model risk in terms of the mad of \$164 to \$498 for a portfolio with \$100,000 in value and a holding period of ten days. When focusing at the median of the model risk estimates, model risk is even 3.5 times higher when fixing the marginal distribution compared to fixing the copula

Table 4.5: Summary statistics of average model risk for all groups

This table presents summary statistics for the time series of average model risk associated with 99% VaR (Panel A) and 97.5% ES (Panel B) forecasts for a well diversified portfolio per group. Model risk is measured in terms of the mean absolute deviation (mad) of risk forecasts by various models within a model set from November 4, 2004 until December 31, 2018. All models passed the respective backtest, see Section 4.2.2 for details. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. *Group 3* (G3) consists of all multivariate and *Group 4* (G4) of all univariate models. We report model risk estimates for one day ahead risk forecasts in percent of the portfolio value (columns 1-5) and in absolute terms (column 6) for an equally weighted portfolio. The results in absolute terms are based on a portfolio value of $100,000 and a 10 day forecast horizon obtained by applying the square-root-of-time rule. We provide minimum (Min), median, mean, maximum (Max), and standard deviation (SD) of the daily averaged model risk estimates per group. Further results for randomly generated portfolio weights, various VaR and ES confidence levels as well as alternative measures of model risk (Groups 1 and 2 only) can be found in Tables 4.6, 4.8, and 4.7. *** denotes a statistically significant increase in average model risk at the 1% level compared to Group 1.

Panel A: 99% VaR

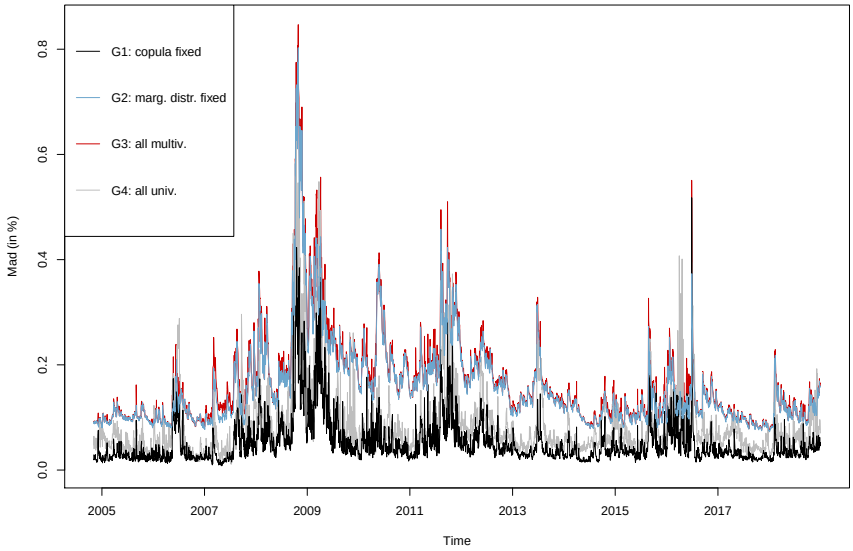
	Average model risk (in %)					**Average model risk (in $)**
	Min	Median	Mean	Max	SD	Mean
Group						
G1: copula fixed	0.008	0.037	0.052	0.518	0.044	164
G2: marg. distr. fixed	0.073	0.130	0.157***	0.803	0.084	498
G3: all multivariate	0.075	0.138	0.165	0.847	0.091	523
G4: all univariate	0.011	0.065	0.085	0.651	0.069	269

Panel B: 97.5% ES

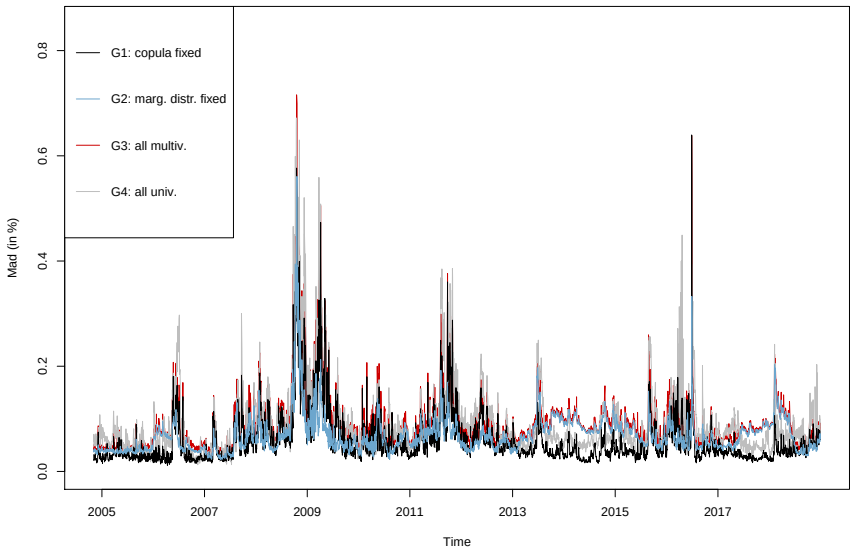
	Average model risk (in %)					**Average model risk (in $)**
	Min	Median	Mean	Max	SD	Mean
Group						
G1: copula fixed	0.012	0.042	0.058	0.640	0.049	184
G2: marg. distr. fixed	0.022	0.061	0.068***	0.561	0.036	216
G3: all multivariate	0.029	0.080	0.092	0.716	0.056	290
G4: all univariate	0.013	0.070	0.090	0.672	0.070	285

Figure 4.3: Average model risk for all groups

This figure shows the average model risk associated with one day ahead 99% VaR (first panel) and 97.5% ES (second panel) forecasts for a well diversified portfolio per group. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. *Group 3* (G3) consists of all multivariate and *Group 4* (G4) of all univariate models. Model risk is measured in terms of the mean absolute deviation (mad) of one day ahead forecasts by various risk models within a model set. Values are calculated on a daily basis between November 4, 2004 until December 31, 2018 in percent of the portfolio value based on all models that passed the respective backtest, see Section 4.2.2 for details.



(a) Average model risk (99% VaR)



(b) Average model risk (97.5% ES)

function.

The result of significant higher model risk due to the choice of a copula function is robust when considering model risk of VaR forecasts with respect to randomly generated portfolio weights. Following Table 4.6, the average model risk of such model sets with varying copulas is 0.142% of the portfolio in contrast to 0.057% when using fixed copulas. Besides, the result is robust with respect to both the choice of a model risk measure (see Table 4.7 and Figure B.2) and the confidence level (see Table 4.8 and Figure B.3). Figure B.3 shows, in addition to the significant rise in model risk due to the choice of a copula, that increasing the confidence level from 95% to 99.9% triples the model risk, in case of model sets with fixed copula from 0.034% to 0.108% and for fixed marginal distributions from 0.078% to 0.293%.

As stated before, we do not compare VaR and ES results to each other as the results depend on risk measure specific confidence levels as well as backtests and consequently on different sets of models (see also Section 4.6). For the 97.5% ES, the results show that the model risk for choosing a copula function is again significantly higher than for choosing a marginal distribution. More detailed, the average model risk is 0.068% (0.058 %) of the portfolio value for varying (fixed) copulas (see Table 4.5). For the ES, the finding that model risk for choosing a copula function is significantly higher than for choosing a marginal distribution is again robust with respect to the portfolio weighting (see Table 4.6), the measure of model risk (see Table 4.7 and Figure B.2), and the confidence level (see Table 4.8 and Figure B.4). Only for a confidence level of 95% the average model risk shows identical values of 0.045% of the portfolio value for fixed and varying copulas. However, when considering the median, the model sets with fixed marginal distributions exhibit a higher model risk (0.041%) than the model sets with fixed copulas (0.033%). Again, increasing the confidence level from 95% to 99.9% results in a higher model risk. Here, the model risk triples for sets with fixed copulas (from 0.045% to 0.136%), while the model risk increases sixfold for fixed marginal distributions (from 0.045% to 0.270%), see Table 4.8.

We illustrate the impact of the choice of the multivariate dependence structure or of

Table 4.6: Summary statistics of average model risk for all groups over 100 random portfolios

This table presents summary statistics for the time series of average model risk per group associated with 99% VaR (Panel A) and 97.5% ES (Panel B) forecasts for 100 random portfolios obtained by drawing portfolio weights from a unit-simplex. Therefore, we first calculate summary statistics for the time series of model risk for each of the portfolios. These statistics are then averaged over all 100 portfolios. The results can thus be interpreted as summary statistics for the model risk of an average portfolio. Model risk is measured in terms of the mean absolute deviation (mad) of risk forecasts by various models within a model set from November 4, 2004 until December 31, 2018. All models passed the respective backtest, see Section 4.2.2 for details. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. *Group 3* (G3) consists of all multivariate and *Group 4* (G4) of all univariate models. We report model risk estimates for one day ahead risk forecasts in percent of the portfolio value. We provide minimum (Min), median, mean, maximum (Max) and standard deviation (SD) of the daily averaged model risk estimates per group. *** denotes a statistically significant increase in average model risk at the 1% level compared to Group 1.

Panel A: 99% VaR

	Average model risk (mad in %)				
	Min	Median	Mean	Max	SD
Group					
G1: copula fixed	0.013	0.041	0.057	0.588	0.048
G2: marg. distr. fixed	0.056	0.118	0.142***	0.753	0.078
G3: all multivariate	0.065	0.127	0.153	0.813	0.086
G4: all univariate	0.012	0.066	0.087	0.751	0.071

Panel B: 97.5% ES

	Average model risk (mad in %)				
	Min	Median	Mean	Max	SD
Group					
G1: copula fixed	0.013	0.045	0.062	0.692	0.052
G2: marg. distr. fixed	0.024	0.069	0.075***	0.485	0.035
G3: all multivariate	0.034	0.085	0.099	0.763	0.057
G4: all univariate	0.015	0.071	0.093	0.762	0.072

Table 4.7: Summary statistics of average model risk for alternative model risk measures (model sets with fixed and varying copula only)

This table presents summary statistics for the time series of average model risk associated with 99% VaR (Panel A) and 97.5% ES (Panel B) forecasts for a well diversified portfolio per group. Model risk is captured by different measures based on risk forecasts by various models within a model set from November 4, 2004 until December 31, 2018. All models passed the respective backtest, see Section 4.2.2 for details. Our baseline measure is the mean absolute deviation (mad). We additionally include the standard deviation (sd) and interquartile range (iqr), see Section 4.2.3 for more details. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. We report model risk estimates for one day ahead risk forecasts in percent of the portfolio value. We provide minimum (Min), median, mean, maximum (Max) and standard deviation (SD) of the daily averaged model risk estimates per group. *** (**) denotes a statistically significant increase in average model risk at the 1% (5%) level compared to Group 1.

Panel A: 99% VaR

		Average model risk (in %)				
		Min	Median	Mean	Max	SD
Group	**Measure**					
G1: Copula fixed	mad	0.008	0.037	0.052	0.518	0.044
G1: Copula fixed	sd	0.016	0.048	0.069	0.636	0.063
G1: Copula fixed	iqr	0.011	0.057	0.075	0.869	0.060
G2: Marg. distr. fixed	mad	0.073	0.130	0.157***	0.803	0.084
G2: Marg. distr. fixed	sd	0.089	0.158	0.191***	0.975	0.101
G2: Marg. distr. fixed	iqr	0.103	0.255	0.311***	1.631	0.181

Panel B: 97.5% ES

		Average model risk (in %)				
		Min	Median	Mean	Max	SD
Group	**Measure**					
G1: Copula fixed	mad	0.012	0.042	0.058	0.640	0.049
G1: Copula fixed	sd	0.015	0.055	0.077	0.734	0.067
G1: Copula fixed	iqr	0.017	0.065	0.085	1.261	0.068
G2: Marg. distr. fixed	mad	0.022	0.061	0.068***	0.561	0.036
G2: Marg. distr. fixed	sd	0.031	0.081	0.090***	0.720	0.046
G2: Marg. distr. fixed	iqr	0.023	0.080	0.094**	0.857	0.057

Table 4.8: Summary statistics of average model risk for all groups under various confidence levels

This table presents summary statistics for the time series of average model risk associated with 99.9%, 99%, 97.5% and 95% VaR (Panel A) and 99.9%, 99%, 97.5% and 95% ES (Panel B) forecasts for a well diversified portfolio per group. Model risk is measured in terms of the mean absolute deviation (mad) of risk forecasts by various models within a model set from November 4, 2004 until December 31, 2018. All models passed the respective backtest, see Section 4.2.2 for details. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. *Group 3* (G3) consists of all multivariate and *Group 4* (G4) of all univariate models. We report model risk estimates for one day ahead risk forecasts in percent of the portfolio value. We provide minimum (Min), median, mean, maximum (Max) and standard deviation (SD) of the daily averaged model risk estimates per group. *** denotes a statistically significant increase in average model risk at the 1% level compared to Group 1.

Panel A: VaR

		Average model risk (mad in %)				
		Min	Median	Mean	Max	SD
Conf. level	**Group**					
99.9%	G1: copula fixed	0.031	0.088	0.108	0.747	0.070
	G2: marg. distr. fixed	0.122	0.248	0.293***	1.507	0.156
	G3: all multivariate	0.127	0.261	0.306	1.553	0.162
	G4: all univariate	0.031	0.169	0.197	1.454	0.124
99.0%	G1: copula fixed	0.008	0.037	0.052	0.518	0.044
	G2: marg. distr. fixed	0.073	0.130	0.157***	0.803	0.084
	G3: all multivariate	0.075	0.138	0.165	0.847	0.091
	G4: all univariate	0.011	0.065	0.085	0.651	0.069
97.5%	G1: copula fixed	0.007	0.028	0.040	0.391	0.036
	G2: marg. distr. fixed	0.049	0.090	0.109***	0.581	0.061
	G3: all multivariate	0.050	0.097	0.117	0.626	0.067
	G4: all univariate	0.009	0.045	0.062	0.506	0.056
95.0%	G1: copula fixed	0.006	0.023	0.034	0.308	0.031
	G2: marg. distr. fixed	0.032	0.065	0.078***	0.453	0.047
	G3: all multivariate	0.032	0.070	0.085	0.484	0.054
	G4: all univariate	0.006	0.035	0.050	0.412	0.046

Panel B: ES

		Average model risk (mad in %)				
		Min	Median	Mean	Max	SD
Conf. level	**Group**					
99.9%	G1: copula fixed	0.023	0.119	0.136	0.875	0.083
	G2: marg. distr. fixed	0.054	0.245	0.270***	1.435	0.125
	G3: all multivariate	0.068	0.279	0.309	1.506	0.139
	G4: all univariate	0.000	0.221	0.244	1.521	0.161
99.0%	G1: copula fixed	0.019	0.060	0.079	0.731	0.062
	G2: marg. distr. fixed	0.047	0.105	0.117***	0.682	0.054
	G3: all multivariate	0.056	0.124	0.142	0.820	0.073
	G4: all univariate	0.017	0.104	0.126	0.848	0.085
97.5%	G1: copula fixed	0.012	0.042	0.058	0.640	0.049
	G2: marg. distr. fixed	0.022	0.061	0.068***	0.561	0.036
	G3: all multivariate	0.029	0.080	0.092	0.716	0.056
	G4: all univariate	0.013	0.070	0.090	0.672	0.070
95.0%	G1: copula fixed	0.008	0.033	0.045	0.543	0.039
	G2: marg. distr. fixed	0.012	0.041	0.045	0.369	0.025
	G3: all multivariate	0.018	0.058	0.067	0.651	0.044
	G4: all univariate	0.009	0.052	0.069	0.521	0.059

the univariate marginals based on a portfolio with a value of \$100,000 and the 99% VaR. Figure 4.4 features the sets of models with fixed copula functions where the model risk is induced by the choice of a model for the marginals. More detailed, at a given date the 20th and 80th percentile of the risk estimates are subtracted from the portfolio value to represent the potential portfolio value under financial distress (averaged over the various copulas). Analogously, we illustrate the impact of the choice of a copula model in Figure 4.5. The potential portfolio values in Figure 4.5 based on the 20th and 80th percentile of risk estimates differ more widely from each other than in Figure 4.4, illustrating that choosing a copula function generates higher model risk than choosing a model for the marginals.

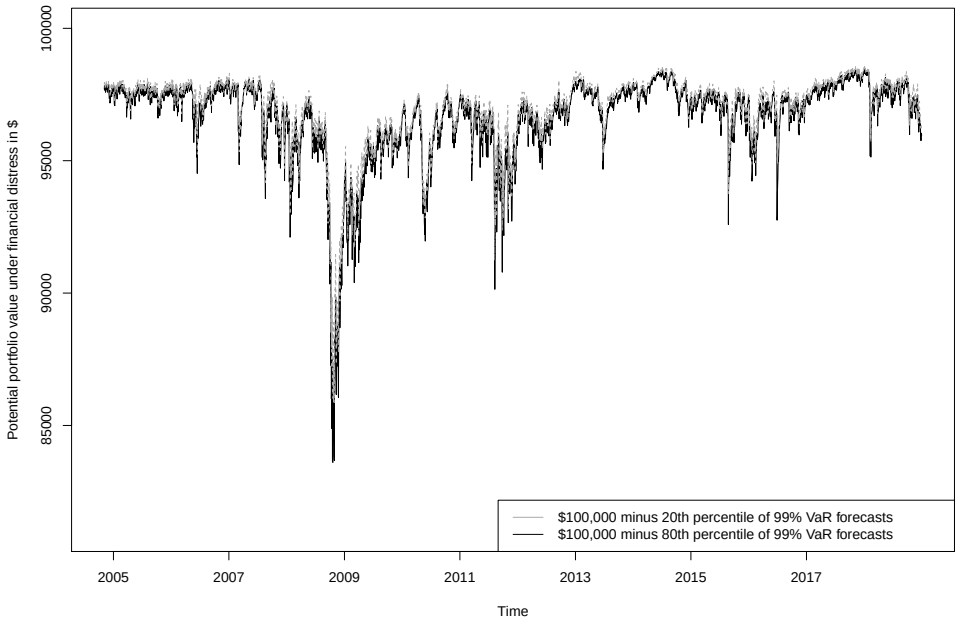
4.5 Model risk for models in the model confidence set

The MCS procedure by Hansen et al. (2011) yields a set of models that contains the best model with a given confidence. That is, the MCS procedure does not assume a particular model to be the true or best one. Instead, it yields a set of models that can be seen as an analogue to a confidence interval for parameters, see Section 4.2.4 for details. As in this paper we study the model risk of risk models that are valid ex-ante and but provide differing forecasts, the MCS procedure fits perfectly into our study. We apply the MCS procedure to the set of models that passed the (daily) backtests to further narrow the set of candidate models.[106] We then determine the model risk corresponding to the models in the MCS to analyze if model risk can be reduced by

[106]To employ the MCS procedure outlined in Section 4.2.4 we use the *MCS* R-package by Catania and Bernardi (2017). We mainly rely on the default parameters. In particular, we adopt the choice of 15% for the confidence level α. We use the test statistic T_R, see Section B.1.5 for more details. The MCS procedure is computationally very expensive, especially for our large set of up to 180 models. We therefore employ the MCS procedure only every 20 days to the set of models that has not been rejected by the respective backtest on that day. Computations are performed for all confidence levels (95%, 97.5%, 99%, and 99.9%) based on a moving window of 500 days and 1000 bootstrapped samples. We additionally employ the MCS procedure on a daily basis for the 99% VaR and the 97.5% ES estimates based on 100 bootstrapped samples. The results remain qualitatively unchanged. Furthermore, we apply the MCS procedure not only to the equally weighted portfolio but also to 10 portfolios with randomly generated portfolio weights (again on a 20 day basis). The results stay qualitatively the same. All computations are performed on the Big-Data-Cluster Galaxy provided by the University Computing Center at Leipzig University.

Figure 4.4: Potential portfolio value under financial distress based on the 99% VaR for fixed copula functions

This figure illustrates the economic significance of model risk arising from the choice of a model for the marginals. Here, we focus on the 99% VaR for model sets with fixed copula functions and varying univariate marginal distributions for a well diversified portfolio ($100,000) and a 10 day holding period. We provide the portfolio value minus the 20th and the 80th percentile of VaR forecasts from all models within each model set that passed the duration-based backtest by Christoffersen (2004) on a daily basis. The values are averaged over the various copula specifications. This corresponds to the potential portfolio value under financial distress according to the more (80th percentile) or less (20th percentile) conservative VaR models. The sample period is November 4, 2004 until December 31, 2018.



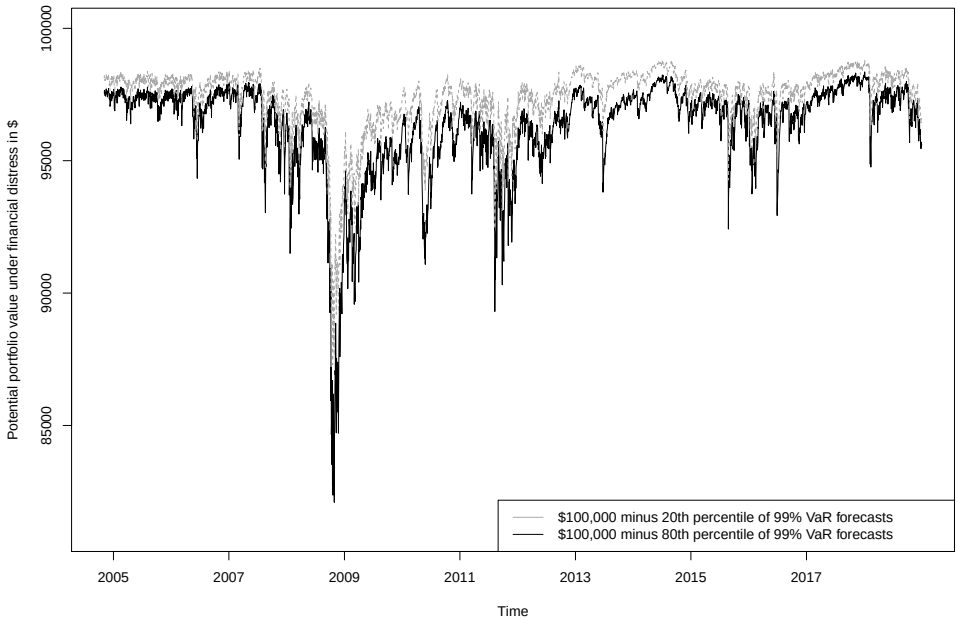
these means.[107]

The main results are summarized in Table 4.9 which compares model risk values before and after applying the MCS procedure in various periods of time. For the whole period, model risk before applying the MCS procedure is on average 0.165% for the VaR and 0.092% for the ES. These values are statistically significant reduced by 23% to 0.127% for the VaR and by 3% to 0.089% for the ES when considering only the models

[107] Note that opposed to Santos et al. (2013) our aim is not to determine the *best* VaR or ES model nor to rank models by their forecasting accuracy. Instead we quantify the extent of non-conformity of the risk forecasts.

Figure 4.5: Potential portfolio value under financial distress based on the 99% VaR for fixed marginal distributions

This figure illustrates the economic significance of model risk arising from the choice of the copula function. Here, we focus on the 99% VaR for model sets with fixed univariate marginal distribution and varying copula functions for a well diversified portfolio ($100,000) and a 10 day holding period. We provide the portfolio value minus the 20th and the 80th percentile of VaR forecasts from all models within each model set that passed the duration-based backtest by Christoffersen (2004) on a daily basis. The values are averaged over the various marginal distributions. This corresponds to the potential portfolio value under financial distress according to the more (80th percentile) or less (20th percentile) conservative VaR models. The sample period is November 4, 2004 until December 31, 2018.



in the MCS.[108] For both VaR and ES this approach reduces model risk not only in the whole period but also in all sub-periods (*pre-crisis, crisis*, and *post-crisis*).[109] The reduction of model risk for ES is, however, very small with declines ranging between 2% and 5%. Opposed to this, we can achieve substantial reductions in VaR model risk (8% in the pre-crisis, 25% in the crisis, and even 27% in the post-crisis period). A graphical representation of the model risk of VaR forecasts before and after applying the MCS procedure can be found in Figure 4.6.

[108] When considering other measures of model risk (*sd* and *iqr*) results are simlar and reductions range between 20% and 30% for VaR and 1% and 3% for ES forecasts.

[109] All reductions are statistically significant at the 1% level.

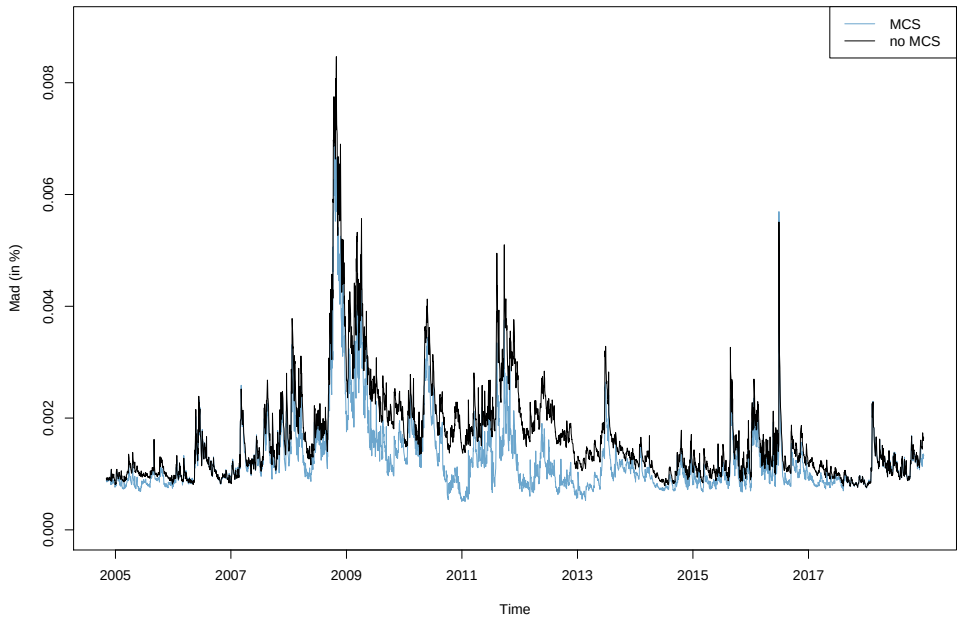
Table 4.9: Model risk before and after applying the MCS procedure

This table compares the model risk associated with one day ahead forecasts of the 99% VaR and the 97.5% ES before and after applying the model confidence set (MCS) procedure by Hansen et al. (2011). Details on the method and its implementation can be found in Sections 4.2.4 and 4.5, respectively. Model risk is measured in terms of the mean absolute deviation (mad) of risk forecasts by various models in percent of the portfolio value. Model risk is calculated on a daily basis from November 4, 2004 until December 31, 2018 for all multivariate models that passed the respective backtest. The term *crisis* refers to the years 2008-2009 while *pre-crisis* and *after-crisis* denote to the period before and after, respectively. We provide minimum (Min), median, mean, maximum (Max) and standard deviation (SD) of the daily model risk estimates. *** denotes statistically significant reductions of mean model risk at the 1% level by applying the MCS procedure.

		No MCS		**MCS**	
		Model risk (in %)		Model risk (in %)	
		99% VaR	97.5% ES	99% VaR	97.5% ES
Whole period	Min	0.075	0.029	0.051	0.022
	Median	0.138	0.080	0.106	0.078
	Mean	0.165	0.092	0.127***	0.089***
	Max	0.847	0.716	0.696	0.703
	SD	0.091	0.056	0.070	0.054
Pre-crisis	Min	0.081	0.029	0.069	0.022
	Median	0.105	0.050	0.098	0.049
	Mean	0.119	0.063	0.110***	0.062***
	Max	0.280	0.207	0.259	0.214
	SD	0.036	0.030	0.034	0.030
Crisis	Min	0.116	0.046	0.087	0.045
	Median	0.241	0.108	0.172	0.099
	Mean	0.286	0.145	0.215***	0.138***
	Max	0.847	0.716	0.696	0.703
	SD	0.139	0.101	0.120	0.097
Post-crisis	Min	0.075	0.034	0.051	0.034
	Median	0.138	0.082	0.099	0.081
	Mean	0.155	0.090	0.113***	0.088***
	Max	0.551	0.638	0.569	0.647
	SD	0.064	0.037	0.045	0.037

Figure 4.6: Model risk of the VaR before and after applying the MCS procedure

This figure compares the model risk associated with one day ahead 99% VaR forecasts before and after applying the MCS procedure by Hansen et al. (2011). Model risk is measured in terms of the mean absolute deviation (*mad*) of risk forecasts by all multivariate VaR models that passed the duration-based test by Christoffersen (2004) (*no MCS*). Subsequently, we apply the MCS procedure to those models and recalculate model risk (*MCS*). Details on the method and its implementation can be found in Sections 4.2.4 and 4.5, respectively. Values are given in percent of the portfolio value between November 4, 2004 and December 31, 2018.



The reduction of model risk that can be achieved corresponds largely to the percentage of models that is excluded by the MCS procedure additionally to the models that have already been removed from the set of candidate models due to the backtests.[110] On average, only 8% of ES models that have passed the respective backtest are excluded by the MCS procedure while the same is true for 21% of the VaR models.[111] However, after applying the MCS procedure on average less ES models (62%) than VaR models (77%) are left over. This is due to the fact that only those models that

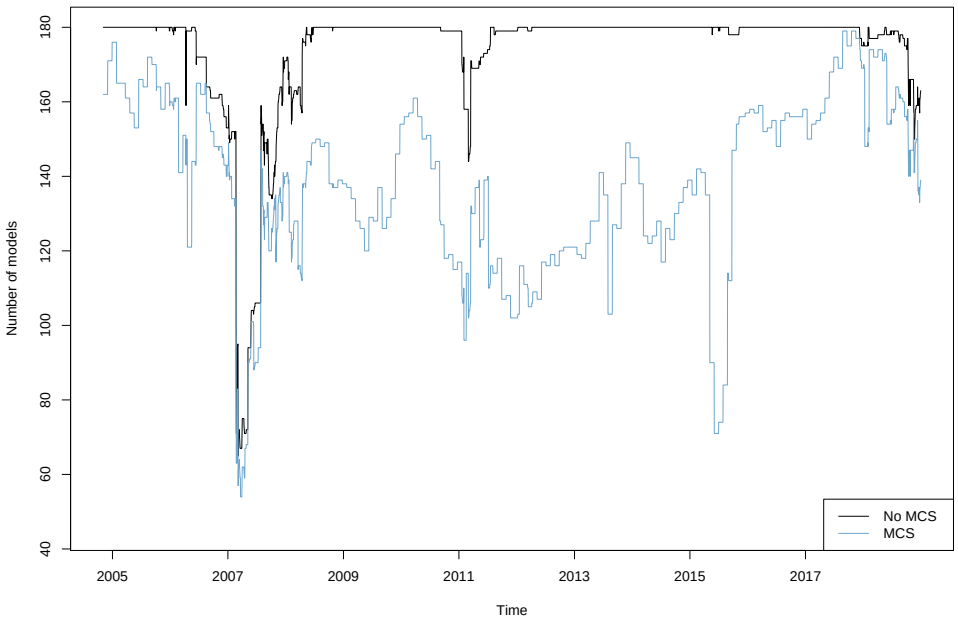
[110] Note that the set of models that passed the backtests varies over time as the backtests are performed based on a moving window.

[111] The difference in the percentages of VaR and ES models that are excluded via the MCS procedure can partially be explained by the fact that the loss function for the ES in the MCS framework is closely related to the conditional calibration backtest by Nolde and Ziegel (2017) while the same is not true in case of the VaR, see Sections 4.2.2 and 4.2.4 for more details.

have not been rejected in the respective backtest enter into the MCS procedure and a lower fraction of ES models (67%) has not been rejected (97% for the VaR). While the percentage of models excluded is relatively stable over time for ES models, it varies for VaR models. In the pre-crisis period about 11% of VaR models are excluded (additionally to the ones rejected by the backtest) while in the crisis and post-crisis period 24% and 23% are excluded, respectively. A graphical representation of the number of models before and after applying the MCS procedure can be found in Figure 4.7. By excluding those models that are inferior[112] to the ones remaining in the MCS, model risk in the post-crisis period can be reduced substantially to the level of the pre-crisis period.

Figure 4.7: Number of VaR models before and after applying the MCS procedure

This figure shows the number of 99% VaR models before and after applying the MCS procedure by Hansen et al. (2011), see Section 4.2.4. In both cases, only those models (out of 180) that were not rejected by the duration-based VaR backtest by Christoffersen (2004) enter into the MCS procedure. The sample period is November 4, 2004 until December 31, 2018.



[112]Comparisons between different models are based on loss functions, see Section 4.2.4 for details.

4.6 Model risk and backtesting

In line with our notion of model risk as uncertainty on the model choice itself when having to choose among many possible alternative models (cf. Cont, 2006, Danielsson et al., 2016), we include many different risk forecasts into the calculation of our model risk measure. We consider various GARCH-type models for the marginals and copula functions for the dependence structure yielding 180 different multivariate model specifications. However, to prevent our results from being biased by erroneous risk forecasts due to misspecified models, we first perform backtests to determine a set of valid candidate models on a daily basis.

In our main analysis, we rely on the duration-based VaR backtest by Christoffersen (2004) and the conditional calibration ES backtest by Nolde and Ziegel (2017). Table 4.10 provides summary statistics on the number of models passing these backtests from November 4, 2004 until December 31, 2018.[113] After having applied the backtest for the 99% VaR, there remain on average 174 models while only 121 97.5% ES models pass the ES backtest. This difference might be explained by the fact that the conditional calibration test is a joint VaR and ES backtest while the duration-based VaR backtest is not (see Section 4.2.2 for details). Figure 4.8 illustrates the number of VaR and ES models that are not rejected by the backtests over time.

For robustness, we provide summary statistics on the number of models that enter into the calculation of model risk when considering alternative backtests or when using different specifications in Table 4.11.

When relying on the dynamic quantile test by Engle and Manganelli (2004), on average 71 VaR models are not rejected while the same is true for 127 ES models in the exceedance residual test by McNeil and Frey (2000). Using a moving window of 1000 days instead of 500 days reduces the average number of models to 145 for the

[113] The backtests are performed based on a moving window of 500 days and a confidence level of 99%. For VaR and ES estimates themselves, we consider the confidence levels 95%, 97.5%, 99%, and 99.9%.

Table 4.10: Summary statistics of market risk models passing the backtest

This table presents summary statistics on the number of multivariate market risk models passing the respective backtest, see Section 4.2.2 for details. We consider VaR (Panel A) and ES (Panel B) forecasts for a well diversified portfolio from November 4, 2004 until December 31, 2018 and confidence levels (*Conf. level*) ranging from 95% to 99.9%. We provide minimum (Min), median, mean, maximum (Max) and standard deviation (SD) of the daily number of models.

Panel A: VaR

	Models passing backtest				
	Min	Median	Mean	Max	SD
Conf. level					
99.9%	137	180	178	180	5
99.0%	65	180	174	180	17
97.5%	89	180	176	180	11
95.0%	158	180	179	180	3

Panel B: ES

	Models passing backtest				
	Min	Median	Mean	Max	SD
Conf. level					
99.9%	45	113	123	180	46
99.0%	97	130	136	180	25
97.5%	63	120	121	179	27
95.0%	24	119	109	168	31

VaR and 82 for the ES while a fixed window spanning the entire sample period[114] leads on average to 113 VaR and 17 ES models. In the following, we focus on the impact of different backtesting specifications on our model risk estimates.

4.6.1 All multivariate models

Summary statistics for our main measure of model risk (mad) are presented in Table 4.12. There, we compare the results of our main analysis to model risk estimates

[114] Note that the usage of a fixed window is not feasible in practice as it introduces a look-ahead-bias into the evaluation of risk models. However, we include results to provide a more complete picture.

Table 4.11: Summary statistics of market risk models passing alternative backtests

This table presents summary statistics on the number of multivariate market risk models passing a set of backtesting alternatives. For our main analysis, we rely on the duration-based VaR backtest by Christoffersen (2004) and the conditional calibration ES backtest by Nolde and Ziegel (2017) with a moving window of 500 days (*Baseline*). For robustness, we also consider the dynamic quantile VaR backtest by Engle and Manganelli (2004) and the exceedance residual ES backtest by McNeil and Frey (2000) (*Alternative backtests*), for more details we refer to Section 4.2.2. Additionally, we provide results when not applying any backtest (*No backtest*) or when using the baseline backtest with a fixed window (*Fixed window*) or a moving window of 1000 days (*1000 days mov. window*). We consider VaR (Panel A) and ES (Panel B) forecasts for a well diversified portfolio from November 4, 2004 until December 31, 2018 and confidence levels (*Conf. level*) ranging from 95% to 99.9%. We provide minimum (Min), median, mean, maximum (Max) and standard deviation (SD) of the daily number of models.

Panel A: 99% VaR

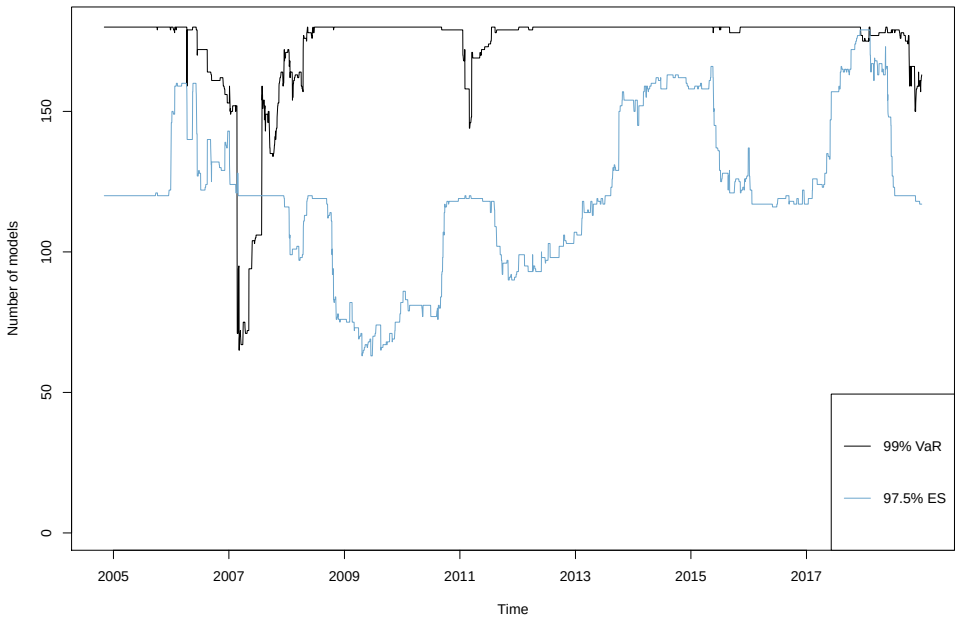
	Models passing backtest				
	Min	Median	Mean	Max	SD
Baseline	65	180	174	180	17
No backtest	180	180	180	180	0
Alternative backtest	0	65	71	160	42
Fixed window	113	113	113	113	0
1000 days mov. window	0	179	145	180	62

Panel B: 97.5% ES

	Models passing backtest				
	Min	Median	Mean	Max	SD
Baseline	63	120	121	179	27
No backtest	180	180	180	180	0
Alternative backtest	99	119	127	180	21
Fixed window	17	17	17	17	0
1000 days mov. window	0	95	82	120	39

Figure 4.8: Number of market risk models passing the backtest

This figure shows the daily number of multivariate market risk models passing the respective backtest, see Section 4.2.2 for details. We consider risk forecasts associated with the one day ahead 99% VaR and 97.5% ES for a well diversified portfolio from November 4, 2004 to December 31, 2018.



obtained by using no backtests or by employing alternative backtests. For robustness, we also add results when using our main backtests with a fixed estimation window and a moving window of 1000 days (instead of 500 days), respectively.

Over the entire sample period model risk (mad) for the 99% VaR is on average 0.165% when determining the set of candidate models via the duration-based backtest.[115] When calculating model risk based on all 180 risk forecasts (i.e., without employing a backtest), model risk for the VaR is on average slightly lower (0.156%). When instead relying on the dynamic quantile test by Engle and Manganelli (2004), average model risk is substantially lower (0.079%).[116] Average model risk for the

[115] In our main analysis we rely on a moving estimation window of 500 days. When instead using a moving window of 1000 days we obtain an average model risk of 0.202% and when building on a fixed estimation window an average model risk of 0.129%. Note that although employing a fixed estimation window substantially decreases average model risk, doing so introduces a look-ahead bias to our model risk estimates and is therefore not feasible in practice.

[116] This can be explained by the fact that the dynamic quantile test on average rejects a larger fraction

Table 4.12: Model risk and alternative backtests (all multivariate models)

This table provides summary statistics on the model risk associated with one day ahead forecasts of 99% VaR and 97.5% ES after having different (or differently specified) backtests. For our main analysis, we rely on the duration-based VaR backtest by Christoffersen (2004) and the conditional calibration ES backtest by Nolde and Ziegel (2017) with a moving window of 500 days (*Baseline*). For robustness, we also consider the dynamic quantile VaR backtest by Engle and Manganelli (2004) and the exceedance residual ES backtest by McNeil and Frey (2000) (*Alternative backtests*), for more details we refer to Section 4.2.2. Additionally, we provide results when not applying any backtest (*No backtest*) or when using the baseline backtest with a fixed window (*Fixed window*) or a moving window of 1000 days (*1000 days mov. window*). Model risk is measured in terms of the mean absolute deviation (mad) of the risk forecasts that passed the respective backtest on a daily basis from November 4, 2004 until December 31, 2018. We report model risk estimates in percent of the portfolio value (first and second column) and in absolute terms (third and fourth column) for an equally weighted portfolio. The results in absolute terms are based on a portfolio value of $100,000 and a 10 day forecast horizon. We provide minimum (Min), median, mean, maximum (Max), and standard deviation (SD) of the daily model risk estimates.

			Model risk (in %)		**Model risk (in $)**	
			99% VaR	97.5% ES	99% VaR	97.5% ES
	Baseline	Min	0.075	0.029	237	92
		Median	0.138	0.080	436	253
		Mean	0.165	0.092	522	291
		Max	0.847	0.716	2,678	2,264
		SD	0.091	0.056	288	177
	No backtest	Min	0.056	0.058	177	183
		Median	0.127	0.133	402	421
		Mean	0.156	0.163	493	515
		Max	0.847	0.876	2678	2770
		SD	0.089	0.092	281	291
Alternative	**backtests**	Min	0.000	0.032	0	101
		Median	0.065	0.095	206	300
		Mean	0.079	0.114	250	360
		Max	0.672	0.864	2125	2732
		SD	0.054	0.077	171	243
Fixed	**window**	Min	0.035	0.011	111	35
		Median	0.097	0.068	307	215
		Mean	0.129	0.078	408	247
		Max	1.081	0.712	3418	2252
		SD	0.103	0.049	326	155
1000 days	**mov. window**	Min	0.048	0.025	152	79
		Median	0.170	0.072	538	228
		Mean	0.202	0.091	639	288
		Max	1.118	0.962	3535	3042
		SD	0.113	0.067	357	212

97.5% ES over the entire sample period is 0.092% in our main analysis[117] (conditional calibration backtest) and 0.163% when not using any backtest. For the exceedance residual test by McNeil and Frey (2000) average model risk (0.114%) is higher than model risk for the conditional calibration backtest, but still much lower than in the case of not using any backtest.

4.6.2 The subgroups

Table 4.13 presents summary statistics of average model risk when using alternative backtests or no backtest at all. Again, the focus is on the comparison of model sets with fixed and model sets with varying copula function. For the 99% VaR we find that the average model risk remains identical when we do not backtest. This is true for model sets with fixed (0.052%) as well as for model sets with varying (0.157%) copula function and can be explained by the fact that the duration-based backtest by Christoffersen (2004) rejects on average only about 3.3% of VaR market risk models. In contrast, the conditional calibration backtest by Nolde and Ziegel (2017) rejects on average about 33% of models forecasting ES estimates. For model sets with fixed copula function, the model risk is almost similar (deterioration of 0.003 percentage points) if no backtest is performed. On the other hand, for model sets with varying copula and consequently fixed marginal distribution, the average model risk is reduced from 0.164% to 0.068%. In addition, Figure 4.9 shows that our main result of increasing model risk by choosing a copula function is robust to no prior backtesting.

For the 99% VaR using the dynamic quantile test by Engle and Manganelli (2004), the average model risk is almost not reduced for groups with fixed copula (from 0.052% to 0.051%) and more than three times reduced for groups with varying copula (from 0.157% to 0.042%).[118] In contrast to our baseline backtest, 60.6% of the market

of models (60.6%) than the duration-based backtest (3.3%). However, there are periods of time (in total 14.0% of all days corresponding to roughly 2 years of our sample period) with none of the 180 models passing the dynamic quantile test, which is the main reason for not using this backtest in our main analysis. Further details can be found in Table 4.11.

[117] When using a moving window of 1000 days (instead of 500 days in our main analysis), we obtain an average model risk of 0.091%. For a fixed estimation window average model risk is 0.078%.

[118] When using a moving window of 1000 days (instead of 500 days in our main analysis), we obtain

Table 4.13: Model risk and alternative backtests (model sets with fixed and varying copula only)

This table presents summary statistics for the time series of average model risk per group associated with 99% VaR (Panel A) and 97.5% ES (Panel B) forecasts for a well diversified portfolio and a set of backtesting alternatives. For our main analysis, we rely on the duration-based VaR backtest by Christoffersen (2004) and the conditional calibration ES backtest by Nolde and Ziegel (2017) with a moving window of 500 days (*Baseline*). For robustness, we also consider the dynamic quantile VaR backtest by Engle and Manganelli (2004) and the exceedance residual ES backtest by McNeil and Frey (2000) (*Alternative backtests*), for more details we refer to Section 4.2.2. Additionally, we provide results when not applying any backtest (*No backtest*) or when using the baseline backtest with a fixed window (*Fixed window*) or a moving window of 1000 days (*1000 days mov. window*). Model risk is measured in terms of the mean absolute deviation (mad) of risk forecasts by various models within a model set that passed the respective backtest on a daily basis from November 4, 2004 until December 31, 2018. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. We report model risk estimates for one day ahead risk forecasts in percent of the portfolio value. We provide minimum (Min), median, mean, maximum (Max), and standard deviation (SD) of the daily averaged model risk estimates per group.

Panel A: 99% VaR

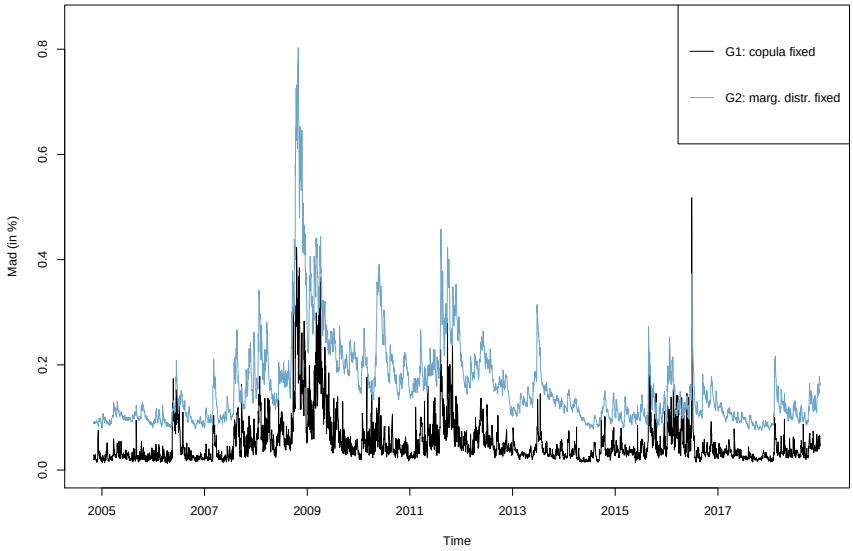
		Average model risk (in %)				
		Min	Median	Mean	Max	SD
Group						
G1: Copula fixed	Baseline	0.008	0.037	0.052	0.518	0.044
	No backtest	0.012	0.037	0.052	0.518	0.044
	Alternative backtest	0.000	0.039	0.051	0.524	0.041
	Fixed window	0.011	0.033	0.048	0.516	0.043
	1000 days mov. window	0.014	0.040	0.053	0.518	0.041
G2: Marg. distr. fixed	Baseline	0.073	0.130	0.157	0.803	0.084
	No backtest	0.075	0.130	0.157	0.803	0.084
	Alternative backtest	0.000	0.038	0.042	0.439	0.024
	Fixed window	0.058	0.103	0.126	0.717	0.072
	1000 days mov. window	0.075	0.137	0.159	0.669	0.073

Panel B: 97.5% ES

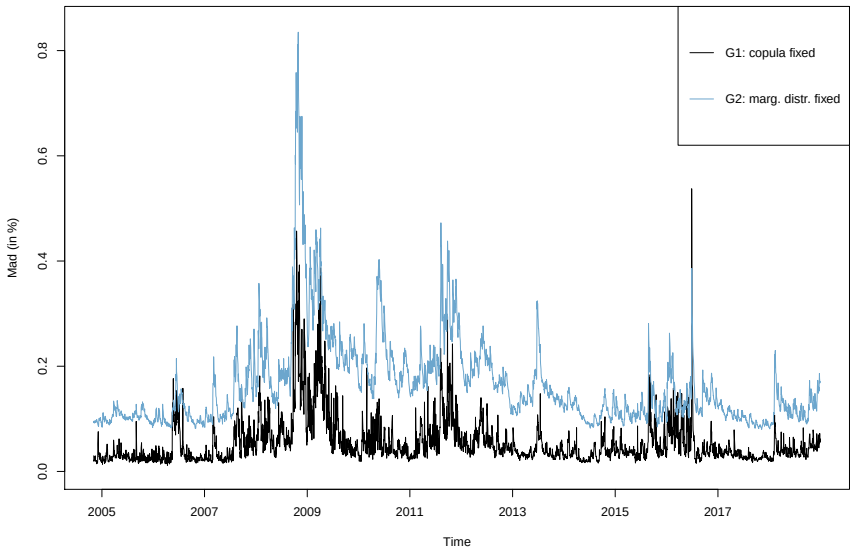
		Average model risk (in %)				
		Min	Median	Mean	Max	SD
Group						
G1: Copula fixed	Baseline	0.012	0.042	0.058	0.640	0.049
	No backtest	0.012	0.040	0.055	0.537	0.045
	Alternative backtest	0.012	0.042	0.058	0.581	0.050
	Fixed window	0.006	0.027	0.039	0.546	0.039
	1000 days mov. window	0.017	0.047	0.065	0.640	0.053
G2: Marg. distr. fixed	Baseline	0.022	0.061	0.068	0.561	0.036
	No backtest	0.080	0.137	0.164	0.835	0.087
	Alternative backtest	0.029	0.069	0.077	0.561	0.045
	Fixed window	0.003	0.029	0.034	0.423	0.026
	1000 days mov. window	0.015	0.039	0.047	0.546	0.035

Figure 4.9: Average model risk without applying backtests (model sets with fixed and varying copula only)

This figure shows the average model risk associated with one day ahead 99% VaR (first panel) and 97.5% ES (second panel) forecasts for a well diversified portfolio per group. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. Model risk is measured in terms of the mean absolute deviation (mad) of one day ahead forecasts by various risk models within a model set. Values are calculated on a daily basis between November 4, 2004 until December 31, 2018 in percent of the portfolio value. In contrast to our baseline analysis, the models did not have to pass a backtest.



(a) Average model risk (99% VaR)



(b) Average model risk (97.5% ES)

risk models are discarded on average. When using the exceedance residual ES backtest by McNeil and Frey (2000), average model risk for sets with fixed copulas is robust to our baseline analysis (0.058%). The alternative ES backtest rejects 29.7% of market risk models on average. For the group of model sets with fixed marginal distributions, the alternative backtest leads to a slightly higher average model risk (0.077%) than our baseline (0.068%). Again, the average model risk is reduced in comparison to no prior backtest (0.164%).[119] These results highlight that model risk depends on the choice of the backtests, some of which are able to reduce average model risk (compared in particular to using no backtest). As a consequence, employing backtests can be seen as additional means for reducing model risk. A further analysis of the choice of backtests with regard to model risk is, however, beyond the scope of this paper.

4.7 Conclusion

In this paper, we study the model risk inherent in copula GARCH models used for forecasting financial risk. More precisely, we forecast the VaR and ES for a large number of portfolios using a variety of copula GARCH models. We then analyze different groups of models in which we fix either the marginals, the copula, or neither in a comprehensive empirical study to identify the main source of model risk in multivariate risk forecasting. As our first main result, we find that copula GARCH models come with considerable model risk that is economically significant. Interestingly, and as our second main result, we find that copulas account for considerably more model risk than marginals in multivariate models with the choice of the marginal model having only a small effect on overall model uncertainty. We then propose the use of the model confidence set procedure to narrow down the set of available models and reduce model risk for copula GARCH risk models using ready-to-use backtests for VaR and ES, respec-

an average model risk of 0.053% (0.159%) for model sets with fixed (varying) copula. For a fixed estimation window average model risk is 0.048% (0.126%).

[119] When using a moving window of 1000 days (instead of 500 days in our main analysis), we obtain an average model risk of 0.065% (0.047%) for model sets with fixed (varying) copula. For a fixed estimation window average model risk is 0.039% (0.034%).

tively. Our proposed approach leads to a significant improvement in the mean absolute deviation of one day ahead forecasts by our various candidate risk models.

The findings of our analysis stress the importance of an adequate modeling of the dependence structure inherent in financial portfolios. While the choice of marginal models is not negligible, it is however of lesser importance than selecting the right parametric copula model. In this respect, our findings are reassuring as the majority of previous papers in this field have solely concentrated on copula modeling and have relied on standard GARCH(1,1)-models for the marginals. Our quantification of the degree of model risk caused by the large set of candidate parametric copula families, however, shows that multivariate models include an economically significant amount of model risk. Using the model confidence set approach seems to alleviate this danger to some degree. Finally, our findings are of high relevance for supervisors in the banking and insurance sector as we illustrate the need for carefully checking the adequacy of a multivariate copula-based risk model.

Chapter 5

Estimating the Relation Between Digitalization and the Market Value of Insurers

5.1 Introduction

Digitalization has already massively transformed many industries. The insurance industry, however, has yet to take advantage of the full potential of digital technologies. This becomes even more important as rising customer expectations, the effects of the financial markets crisis, and the zero interest rate policy lead to an increased competitive pressure. In general, there is no doubt on the strong impact digitalization will have on the insurance ecosystem (see, e.g., Cappiello, 2020). It is considered to affect the whole insurance value chain, from product development to pricing/underwriting, sales and distribution, policy and claims management, and asset and risk management (Eling and Lehmann, 2018). However, in contrast to other megatrends such as urbanization or aging societies, the precise scope of digitalization is difficult to grasp. Although we know that digitalization clearly manifests itself in cloud computing, Internet of Things, mobile communication, blockchain technology, artificial intelligence, etc. (Schmidt, 2018), evidence on the question of how to measure digitalization and its relation to

firm outcomes is still scarce (see, e.g., Scott et al., 2017, Bohnert et al., 2019, Hanelt et al., 2020).

In this paper, we fill this gap by proposing a new method to measure digitalization in the insurance sector. Our method exploits the prevalence of different topics in standard annual reports. Based on the assumption that digitally innovative insurers report their progress more extensively, Latent Dirichlet Allocation (LDA) helps to assess the importance of digitalization for the particular insurance company. At the same time, the method enables us to separate digitalization from mere firm innovation. In a second step, we use our text-based measure on digitalization to investigate its relation with the market valuation of a large set of publicly-listed US insurance companies. Finally, we account for potential confounding issues related to the construction of the digitalization measure, the reference document used for LDA, and the sentiment in which annual reports are written.

Our results provide first evidence for a positive association between digitalization efforts and market valuation in the US insurance sector. We find that an increase in digitalization is strongly related to an increase in market value and market-to-book value of US insurance companies. Put differently, market participants associate a more digitalized insurance company with higher future profitability and consequently a higher firm value. Although LDA is by design subject to some discretion, we show that our results are robust to several variations of our model parameters. Most importantly, our findings are robust to different numbers of topics used to structure the annual reports and to isolate digitalization from general innovation. Furthermore, the results do not depend on the discretionary choice of the reference document and are not confounded by annual reports' sentiment.

The topic model LDA by Blei et al. (2003) has only recently been introduced to the finance literature.[120] In general, topic models can be used to analyze large datasets of texts that are often unstructured (Roberts et al., 2016). These probabilistic models pro-

[120] One of the first applications of this approach in accounting and finance is due to Huang et al. (2018), who study topical differences between conference calls and subsequent analyst reports.

vide a finite set of common topics which optimally reflect a collection of documents. By applying a topic model to a specific document, we obtain a vector of topic loadings representing how intensively each topic is discussed in the respective document. One of the main advantages of LDA over simple word-list approaches is that the topics and corresponding word distributions arise endogenously from the data and do not have to be specified by the researcher. That is, the underlying machine learning algorithm determines the terms that are most important to discriminate between documents and topics in an unsupervised fashion.

We then apply this powerful tool to the annual reports of 86 publicly-listed US insurance companies available in Thomson Reuters Datastream from 2006 to 2015 and derive a distribution of topics for each of the annual reports. This yields a low-dimensional representation of the document (cf. Blei et al., 2003) that we exploit to construct our text-based measure of digitalization. For this purpose, we compare the extent to which each topic is discussed in the respective report to a reference document about digitalization in the insurance sector. Specifically, we use the paper by Bohnert et al. (2019) since it is closely related to our work. To the best of our knowledge, it is one of the few studies trying to establish an empirical relation between the expression of a digital agenda and the market valuation of insurance companies.[121] More exactly, based on these topic distributions we calculate a measure of similarity between the digitalization document and the insurers' annual reports using the Kullback Leibler (KL) divergence (Kullback and Leibler, 1951). This measure is then used to proxy for the extent of digitalization in our sample insurance companies.

Our paper is related to a growing body of literature on textual analysis and machine learning in finance. Starting with Frazier et al. (1984), Antweiler and Frank (2004), and Tetlock (2007), researchers have studied the effect of qualitative information on equity valuations. More recent papers (e.g., Hanley and Hoberg, 2010, Jegadeesh and Wu, 2013, Hoberg et al., 2014, Hoberg and Maksimovic, 2015, Jegadeesh and Wu, 2017,

[121] However, we also consider further reference documents in the robustness checks, e.g., Cappiello (2020) and Nicoletti (2016) as well as Bohnert et al. (2019) with the empirical study being removed.

Ke et al., 2019) conduct text-based analyses to examine a wide variety of finance research questions.[122] Intriguingly, within the field of text analysis and machine learning, LDA is becoming more and more popular (see, e.g., Goldsmith-Pinkham et al., 2016, Ganglmair and Wardlaw, 2017, Hoberg and Lewis, 2017, Huang et al., 2018, Lopez-Lira, 2019). Within this growing strand of literature, our paper is most closely related to Bellstam et al. (2020) and Lowry et al. (2020), who derive topics based on LDA and employ the KL divergence as a measure of similarity between probability distributions.

At the same time, our paper is related to a growing body of literature on the effect of digitalization in the insurance sector. This literature is basically centered around the impact of digitalization on the business model of insurers (see, e.g., Desyllas and Sako, 2013, Cappiello, 2020), new forms of online marketing and sales activities (Seitz, 2017), and the overall transformation of insurance companies (Barkur et al., 2007). The various facets of digitalization such as Big Data, artificial intelligence, predictive modelling, telematics, and Internet of Things are considered to have a tremendous impact on the whole insurance value chain. In detail, product design and development, underwriting/pricing, sales and distributions, as well as policy and claims management are all subject to fundamental change in the future (see, e.g., Rayport and Sviokla, 1995, van Rossum et al., 2002, Meier and Stormer, 2012, Cappiello, 2020). Numerous opportunities like a facilitated interaction with customers via mail, chatbots, and social media or cost reduction via automation and standardization of business processes are challenged by few risks like the depersonalization of the insurer-customer relationship (Cappiello, 2020). However, there is only little empirical evidence on the relation between digitalization and the market valuation of insurance companies.[123] Our work contributes to the current literature on digitalization and firm valuation in several ways. First, we propose a novel approach to quantify the impact of digitalization on the insurance sector that can also be applied in any other empirical studies based on textual analyses. Second, by employing LDA, we overcome standard pitfalls that arise from

[122] An excellent review of this literature can be found in Loughran and McDonald (2016).
[123] One of the few empirical studies in this field, as mentioned above, is Bohnert et al. (2019).

text mining approaches. In fact, we do not rely on dictionary methods where the precise word lists depend on the researchers' discretion. Instead, the thematic structure within our collection of annual reports is identified via an unsupervised machine learning algorithm. Finally, we add to the new strand of literature on LDA by making use of the whole distribution of topics instead of just focusing on a particular topic. This allows us to differentiate between digitalization and innovation based on a medium number of topics. At the same time, our approach is less prone to topic splits (cf. Bellstam et al., 2020).

The remainder of the paper is structured as follows. Section 5.2 explains in detail the theoretical background of LDA and its implementation in the context of digitalization in the insurance sector. In Section 5.3, we describe and analyze our data and present the empirical strategy using multivariate OLS. Section 5.4 reports the estimation results including alternative specifications based on different topic distributions, sentiment subsamples, different calculations of our digitalization measure, and other reference documents on digitalization. Finally, Section 5.5 summarizes and gives concluding remarks.

5.2 Measuring digitalization using LDA

We use the LDA method due to Blei et al. (2003) as well as natural language processing techniques to automatically analyze annual reports of insurance companies in the United States. In LDA, unsupervised machine learning is used to obtain a finite set of topics frequently discussed in the annual reports along with the fraction of time each topic is covered in each of the reports. We use these information to derive a measure of digitalization. While Sections 5.2.1 to 5.2.4 are optional and can be skipped by the reader who is familiar with the techniques, the construction of our measure is described beginning from Section 5.2.5.

5.2.1 Data preprocessing

To obtain meaningful topics it is very important to preprocess the raw text data before applying any model to them. This is done to reduce the vocabulary to a set of meaningful words that are likely to provide information about the topics and concepts of interest. This facilitates the derivation of meaningful topics that best fit the context of the annual reports. The preprocessing steps we conduct are standard (cf. Hansen et al., 2018, Lopez-Lira, 2019, Bellstam et al., 2020) and shortly outlined below.

After having extracted the plain text from the annual reports,[124] we start by lowercasing all letters. We then remove common stopwords, i.e., words that are commonly used but do not bear a contextual meaning (e.g., "and","or","the","of").[125] We continue by removing all one-letter words like "a" and "i" because these words are frequently used to itemize lists (cf. Loughran and McDonald, 2016). We also exclude all numbers as our focus is on a qualitative analysis of the annual reports. Additionally, we remove all special characters and email addresses.

There are many different words that have the same meaning, e.g., "technological" and "technology", but might be treated differently by the topic modeling algorithm. To avoid this we use a standard technique called *stemming* to derive groups of words with a similar meaning. We rely on Porter's algorithm (Porter, 1980) as the most common algorithm for stemming English documents that has proven to be empirically very effective (Manning et al., 2009). The algorithm essentially consists of five phases of word reduction which are applied sequentially to a text corpus. Word reduction is achieved by applying predefined rules, e.g., removing "ing" at the end of words. In each of the phases there are different conventions on how to select rules. We use the implementation of Porter's algorithm readily available in the *tm* R-package (Feinerer and Hornik, 2020).

Finally, we exclude all words that appear less than 25 times in the whole data set.

[124] We retrieved the plain text via the Xpdf extraction engine.

[125] We thank Bill McDonald for providing lists of common stopwords on his website `https://sraf.nd.edu/textual-analysis/resources/`.

These words are on average used about once every 30 annual reports and are therefore very unlikely to contribute meaningfully to any of the discovered topics. Furthermore, these words might be due to errors in parsing the PDF documents, e.g., a missing letter or whitespace character. By removing those rarely used words we are able to reduce noise in our derivation of topics and sparsity of the document term matrix, see Section 5.2.2.

We now have a list of meaningful words for every document. In the language of text mining these collections of words are referred to as a text corpus. For further analyses this text corpus is transformed into a document term matrix as outlined in the following subsection.

5.2.2 Document term matrix

To apply the LDA methodology to the preprocessed text corpus we have to transform it into a structure that can be utilized by a statistical model. We therefore make use of the so-called *bag of words* approach. The underlying assumption is that the ordering of the words in a text is negligible so that it can be represented by a vector of word counts. That is, the bag of words approach is only concerned about how often specific words occur in a document while the place of occurrence is not considered. Furthermore, the specific ordering of the documents in a text corpus is assumed to be insignificant (Blei et al., 2003). At the cost of losing the word ordering we gain the possibility to apply powerful statistical models to our annual reports that are able to derive context not only within a document but even across documents (cf. Bellstam et al., 2020). The representation of documents as vectors of word counts is, however, not only fundamental to topic modeling but lies on the basis of a variety of information retrieval algorithms such as document scoring in a query, document classification and document clustering (Manning et al., 2009).

By combining the individual vectors of word counts, we obtain a so-called document term matrix, where each row corresponds to a specific document and each column to

Figure 5.1: A simplified example of a document term matrix

This figure, adapted from Lopez-Lira (2019), shows a simplified example of a document term matrix with 4 documents and 12 terms, see Section 5.2.2 for details. Before construction of the matrix, stopwords are removed and words are stemmed. For more information on the preprocessing steps we refer to Section 5.2.1.

impact	digit	financi	insur	affect	valu	develop	manag	mobil	firm	communic	technolog
1	0	0	1	1	0	0	0	1	2	0	0
0	1	1	0	1	1	0	0	0	1	0	2
1	0	0	1	0	0	1	2	0	1	1	0
1	1	0	0	0	2	1	0	1	0	2	1

a specific word. A simplified example can be found in Figure 5.1. For our sample of annual reports we obtain a matrix with 11,440 columns highlighting the importance of the preprocessing steps especially in reducing the dimensionality of the document term matrix. Document term matrices are typically very sparse because they contain the counts of words being used across the whole sample of annual reports (cf. Hansen et al., 2018). However, the whole vocabulary is not used in each individual report. Therefore, even after removing some of the most rarely used words in the preprocessing steps, our document term matrix only has 17 % non-zero entries. In this context, topic models such as LDA can be understood as very powerful dimensionality reduction techniques.

5.2.3 LDA

Topic models can be used to analyze large datasets of texts that are often unstructured (Roberts et al., 2016). They are probabilistic models that produce a finite set of com-

mon topics that best represent a collection of documents whereby each topic itself can be represented as a distribution over words. Each document typically covers multiple topics. By applying a topic model to a specific document we obtain a vector of topic loadings representing how intensively each topic is discussed in the respective document. This yields essentially a probabilistic representation of the document.

In this paper we use LDA, a topic model developed by Blei et al. (2003).[126] The unsupervised machine learning algorithm LDA models each document in a text corpus as a finite mixture over an underlying set of latent topics. The topics are derived from the sets of words that group together in and across the annual reports. To derive the latent structure (the discussed topics) from the observed data (the words) generative models like LDA postulate a complex latent structure being responsible for the observed data. By employing statistical inference this latent structure can then be recovered (Griffiths and Steyvers, 2004). While the topics arise endogenously from the LDA algorithm, the number of topics has to be specified in advance. The selection of an optimal number of topics is discussed in Section 5.2.6.

Thinking of the documents as discrete distributions over topics which themselves are distributions over words can be seen as a matrix factorization of the document term matrix (Arun et al., 2010). That is, the document term matrix, containing per document counts of specific words, is factorized in a matrix mapping topics to documents and a matrix mapping words to topics. From this perspective, LDA can be understood as a type of principal component analysis (cf. Blei, 2012) that reduces the dimensionality of each document from a vector of thousands of words to a vector of the number of topics. However, as in the LDA model topics are understood as probability distributions over words, most of the information can be preserved in the factorized matrices.

More formally, LDA is a three-level hierarchical Bayesian model that relies on a generative process for each document D in a text corpus. First, the length N of the document is determined according to a Poisson distribution, then the parameter θ is

[126] Algorithms for topic modeling can be adapted to other kinds of discrete data and have been successfully applied to genetic data and social networks (cf. Blei, 2012).

chosen from a Dirichlet distribution. This parameter governs the distribution of topics in the document and is used to specify the particular topic z_n from which an individual word ω_n is generated. More exactly, z_n is chosen according to a multinomial distribution with parameter θ. Finally, a word ω_n is drawn from the multinomial distribution of the topic z_n. This procedure is repeated for all words in a document and for all documents in a text corpus.[127] This is essentially a two-stage process. First, generate a distribution over topics. Second, choose a specific topic from this distribution over topics and generate a word from the corresponding distribution over words. The intuition behind this is that a document usually covers multiple topics and that different topics use certain words in different frequencies.[128] A representation of the underlying generative process in the language of graphical models can be found in Figure 5.2.

The parameters underlying the model can be determined for instance by variational expectation-maximization (VEM) or Gibbs Sampling.[129] We rely on the R-package *topicmodels* that provides an interface to the C code by Blei et al. (2003) for estimating a LDA model based on the VEM algorithm.

For our digitalization measure we make use of the fact that each document has its individual distribution of topics providing a low dimensional representation of each document. To infer topic loadings we are not restricted to the set of documents that we have applied the LDA to. Instead, we are able to first derive the topics that are common over all annual reports of the insurance companies in our sample by applying the LDA algorithm only to these documents. Afterwards, we can utilize this set of topics as well as the representation of the topics (probability distributions over words) to infer how intensively these topics are discussed in previously unseen documents. These might be for instance additional annual reports or in this paper the reference document on digitalization.

The LDA model has many advantages, especially over simple dictionary methods.

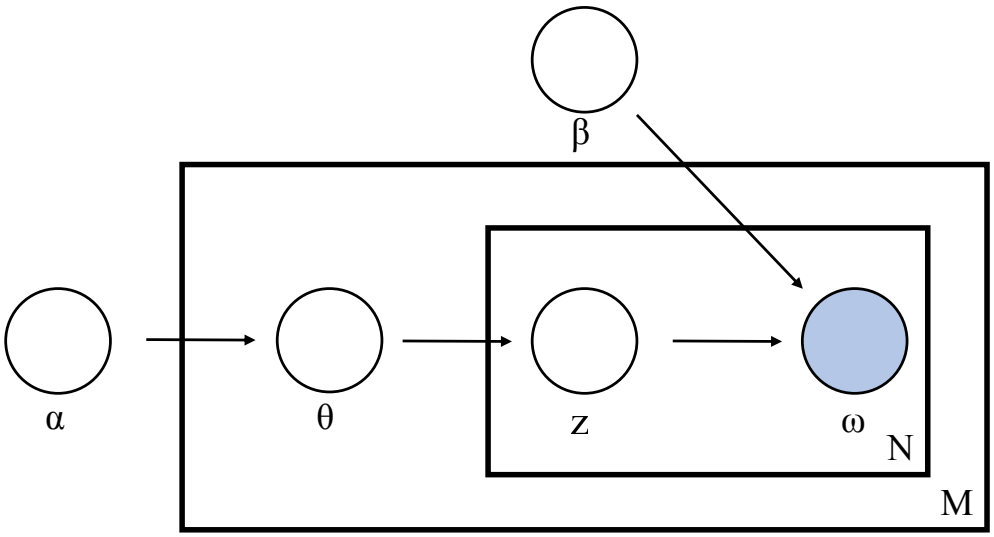
[127] For all technical details see the original paper by Blei et al. (2003).

[128] Note that each topic consists of the same vocabulary. However, conditional on the topic, each word occurs in a different frequency.

[129] An introduction to VEM and Gibbs Sampling can be found in Wainwright and Jordan (2008) and Gelfand and Smith (1990), respectively.

Figure 5.2: Graphical representation of LDA

This figure presents a graphical model of the LDA adapted from Blei et al. (2003). M is the set of all documents in a collection. The inner rectangle represents the i'th document in the collection where N denotes the number of words of the specific document. The j'th word in this document is generated from the random topic z where the topic is chosen from the specific topic distribution for that document specified by the parameter θ. The parameters α and β are priors that control the sparsity of topics within a document and the sparsity of words within a topic, respectively. Essentially, these parameters specify how much topics are needed to describe a document and how much words are needed to describe a topic. For more details on the LDA we refer to Section 5.2.3.



First of all, there is no need to provide any lists of words as the underlying machine learning algorithm determines the terms that are most important to discriminate between documents and topics in an unsupervised fashion. That is, the topics and corresponding word distributions arise endogenously from the data. Additionally, in contrast to the word list approach each topic derived via LDA consists of the same set of words. However, the topics differ in the probabilities they assign to each word. The possibility that multiple topics can be responsible for the words in a single document gives a lot more flexibility in modeling textual data. It corresponds to the intuition that the same word might be used in different contexts and thus also in different topics. Hansen et al. (2018) give the example of the term "growth" being used to describe economic activity while at the same time "wage growth" might appear in a context of

inflationary pressure. By modeling the interaction between documents and topics with the probabilistic generative process underlying the LDA model, one can account for the usage of the same words in different contexts.

There are, however, also some disadvantages that are inherent to the LDA approach. Like most other topic models, the LDA model is based on the bag of words approach, i.e., the order of words within a document is not considered but only the number of occurrences. Therefore, the context of specific words can only be derived based on the words that frequently occur within the same document while the distance between words is neglected. Nevertheless, although the exchangeability assumption behind the bag of words approach is unrealistic, it is reasonable in a context of assessing the coarse semantic structure of documents (see Blei, 2012). Another disadvantage is that, unlike the topics themselves, the number of topics does not arise endogenously from the data but has to be specified by the researcher in advance.[130] Other potential weaknesses are the exchangeability assumption on the documents within a collection and that relationships between topics are not considered (see Blei et al., 2010). In an overall view of the advantages and disadvantages, we are convinced that the LDA model as the *"simplest"* (Blei, 2012) and *"most common topic model currently in use"* (Lopez-Lira, 2019) is most appropriate in our context where we use the topics and corresponding probabilities as a low dimensional representation of the documents.

5.2.4 Other text modeling methods

In this section, we will shortly discuss other methods for the analysis of textual data in finance and insurance. Among the most common and simple approaches to automated content analysis are dictionary methods, where a list of words related to a specific topic is defined by the researcher. The prevalence of this topic in a document is then simply derived based on the number of occurrences of the list entries. This approach can of course be extended to account for multiple topics. Word lists are an intriguingly simple approach to textual analysis. This comes, however, at the price of subjectivity

[130] The choice of the optimal number of topics is discussed in Section 5.2.6.

as the words in the dictionary have to be specified by the researcher. Additionally, a broad concept like digitalization can hardly be captured by just counting words like "computer" or "IT" and providing a complete list of words poses a very complicated if not impossible task for the researcher. Furthermore, in dictionary methods all entries in a word list are assumed to be of equal importance.

A more sophisticated word list based approach is provided by the key word in context (KWIC) concordances (see Gries and Newman, 2013) where not only the counts of specific words but also the words in their direct proximity are considered. The main advantage of this approach is that the immediate context in which words of interest are used is taken into account. For example, Bohnert et al. (2019) employ this method to analyze to which extent companies address digitalization in the context of external and internal stakeholders. However, the main issue with dictionary based approaches that lists of words have to be specified by the researchers is not resolved.[131] Apart from that, we do not study the context in which digitalization is addressed in companies but rather the amount to which they digitalize making the LDA method more appropriate in our study.

Topic models, with LDA being the most common one, provide another approach to automated text analysis. An early and rather simple topic modeling method is Latent Semantic Analysis (LSA, Deerwester et al., 1990) that at its core is a principal component analysis performing a singular value decomposition on the document term matrix to extract the most informative dimensions. The probabilistic nature of LDA and the flexibility that comes from it is an important distinguishing feature from this dimensionality reduction technique. In fact, LDA was introduced to fix an issue with a probabilistic extension of the LSA method (probabilistic LSA, Hofmann, 1999).

LDA can be used as a module in more complicated models to relax some of its assumptions discussed in the previous section. For example, the topic model due to Wallach (2006) generates words inside a topic conditional on the previous word

[131] Actually, by requiring not only to specify keywords but also additional lists of words to be analyzed in the direct proximity of the words of interest and the maximum distance between words to be considered, this method necessitates even more discretionary choices by the researcher.

and Griffiths et al. (2005) propose a composite model combining LDA with a Hidden Markov Model (HMM) to account for short-range dependencies between words. Another extension is the hierarchical LDA (hLDA) by Blei et al. (2010). It introduces a hierarchy of topics by including the nested Chinese restaurant process in the generative model. However, the internal nodes of the resulting topic tree are not summaries of their children, i.e., high probability words of a node do not necessarily coincide with high probability words of its children making interpretation difficult. Additionally, as in the LDA model, the number of topics still has to be specified by the researcher in advance.[132] In the structural topic model (STM) due to Roberts et al. (2016) covariates external to the respective document can enter into the model to allow for interactions between covariates and the topics. Grace (2019) uses this feature to analyze how topics in the 10-Ks vary with information about the company as well as over SIC code sector and time. However, as opposed to Grace (2019) we do not study how digitalization evolves depending on exogenous variables. Instead, we analyze how digitalization affects firm valuation. A summary of advantages and disadvantages of selected models can be found in Table 5.1. Of course, there exist more extensions of the LDA model,[133] see Blei (2012) for a review.

Overall, these extensions relax assumptions of the LDA model, allow for the inclusion of additional covariates or model relationships between the topics. However, although some of these methods exhibit improved language modeling performance especially (which is especially important in language generation), this comes at the price of an expanded parameter space and more complexity which can complicate interpretation. From a more practical perspective, there exist implementations of LDA in many different programming languages. Unfortunately, the same is not true for many of its extensions. Finally, Blei (2012) argue that the bag of words assumption underlying

[132] The hLDA model should not be confused with the Hierarchical Dirichlet Process (HDP, Teh et al., 2006) where the number of topics can be unbounded and learned from the data. However, in the HDP model the term "hierarchical" refers to the generative process and not to the topics which are flat clusterings.

[133] There even exist extensions to the computer vision field like the spatial LDA model by Wang and Grimson (2007) that are able to account for the spatial and temporal dependence of "visual words" in pictures and videos.

Table 5.1: Advantages and disadvantages of selected text modeling methods

This table summarizes advantages and disadvantages of selected text modeling methods discussed in Sections 5.2.3 and 5.2.4. Of course the strengths and weaknesses of the models have to be weighted against the background of the specific application.

Method	Advantages	Disadvantages
Word list approach	• Simple • Easy to interpret	• Context in which words appear is not considered • Word lists depend on the researcher's discretion
KWIC concordances	• Accounts for the context in which words of interest are used • Still simple and easy to interpret	• Requires the researcher to specify additional words to be studied in the proximity of the words of interest • Introduces more subjectivity than the word list approach
LDA	• "*Simplest*" (Blei, 2012) and "*most common topic model currently in use*" (Lopez-Lira, 2019) • Dimensionality reduction technique that provides low dimensional representations of documents • Accounts for the usage of the same words in different contexts by allowing multiple topics to be responsible for the same word in a single document • Can be used to construct more complicated models • Implementations available in many different programming languages	• Accounts for context only via words that frequently occur in the same documents while the order of words within a document is neglected (bag of words approach) • Number of topics must be specified by the researcher • Interpretation is more complicated than in word list based approaches but easier than in many other topic models • Topics are assumed to be flat clusterings • Higher computational effort, but feasible on modern computers
hLDA	• Provides a hierarchy of topics • Topics are organized in a tree with more general topics being located near the root and more specialized topics near the leaves	• Most disadvantages of LDA remain valid • High probability words of a node do not coincide with high probability words of its children making interpretation difficult • Generative process is more complex • More hyperparameters have to be specified by the researcher
STM	• Covariates external to the respective document can enter into the model • Allows for interactions between covariates and topics	• Most disadvantages of LDA remain valid • Expands the parameter space and the complexity of the generative process • Introduces subjectivity in the choice of covariates

LDA is reasonable when uncovering the coarse thematic structure of documents on which our measure of digitalization is build.

5.2.5 A text-based measure of digitalization

This section outlines how our measure of digitalization is derived from the annual reports of US insurance companies. The construction of our text-based measure is motivated by the approach of Bellstam et al. (2020).

By employing LDA we obtain 45 topics[134] that best describe the distribution of empirical word groupings across our sample of annual reports of US insurance companies. As has been outlined before, LDA is a dimensionality reduction technique that essentially reduces the dimension of a document from the number of different words to the number of topics. The topics arise endogenously from the data based on an unsupervised machine learning algorithm. As a result, we obtain a discrete probability distribution over 45 topics for each of the reports. This distribution corresponds to how intensively each topic is covered in the respective annual report. According to Blei et al. (2003) *"the topic probabilities provide an explicit representation of a document."* We build on this low dimensional representation of our annual reports to derive a text-based measure of digitalization.

Therefore, we compare the topic distribution of each annual report, i.e., the extent to which each topic is covered in the respective report, to a reference document about digitalization in the insurance sector (Bohnert et al., 2019)[135]. Although the reference document about digitalization has not been presented to the LDA algorithm, the structure of the underlying generative process allows us to derive the distribution of the previously identified 45 topics in the digitalization document. Based on the distribution of topics in the reference document and in the annual reports we can calculate a measure of similarity between the digitalization document and the reports. The intu-

[134] The number of topics is a hyperparameter that has to be specified by the researcher in advance. The optimal choice of the number of topics is discussed in Section 5.2.6.

[135] In Section 5.4.4 we also consider different documents about digitalization. The results remain qualitatively unchanged.

ition behind this is that an annual report covering similar topics as the digitalization document extensively is more likely about digitalization than an annual report that discusses these topics only marginally.

We take a similar approach to Lowry et al. (2020) and Bellstam et al. (2020) and quantify the similarity between different topic distributions based on the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence measures how much information is lost when one uses the distribution of topics in a particular report to proxy for the distribution of topics in the digitalization document. The reports with a low KL divergence are therefore most likely discussing topics related to digitalization whereas the ones with a large divergence are not. This reasoning is supported by Figure 5.3 where we present the KL divergence between our reference document on digitalization and various other documents on digitalization (Cappiello, 2020, Bohnert et al., 2019 with the empirical section being excluded, Eling and Lehmann, 2018, Nicoletti, 2016 and the industry white papers by McKinsey, 2017 and Deloitte, 2016), on innovation (Tidd and Bessant, 2018), and on financial statement analysis (Subramanyam, 2014). To provide additional robustness we also include three randomly chosen papers from the oldest available issue in the public archive of The Journal of Risk and Insurance (Grosen and Jørgensen, 2002, Doherty and Richter, 2002, Lee and Yu, 2002) that are completely unrelated to the topic at hand. The analysis is based on 45 topics.

The mean KL divergences between the topic distributions of our reference document on digitalization and some other documents on digitalization (Cappiello, 2020, the paper by Bohnert et al., 2019 with the empirical section being removed, and the book by Nicoletti, 2016) are the lowest. The KL divergence between our reference document on digitalization and the document on innovation (Tidd and Bessant, 2018) is also relatively low. This is not surprising as digitalization and innovation are related concepts. However, while the higher level of dissimilarity between Deloitte (2016) and our reference document on digitalization can be explained by the fact that this white paper also covers the more general subjects disruption and innovation, the higher mean

dissimilarity between our reference document on digitalization and McKinsey (2017) and Eling and Lehmann (2018), respectively, illustrates that discriminating between digitalization and innovation is difficult and has limitations.[136] The KL divergence between the topic distributions in our reference document on digitalization and the document on financial statement analysis by Subramanyam (2014) is quite large. That is, even though the topics are derived from annual reports, the topic distributions are well suited to differentiate between digitalization on the one and general finance language on the other hand. To provide a more complete picture, we added three papers from The Journal of Risk and Insurance covering completely unrelated topics. As expected, we obtain very high KL divergences signaling a large dissimilarity.

For a specific annual report the text-based digitalization measure is calculated as the reciprocal of the KL divergence between the topic distribution in the digitalization document and the topic distribution in the particular annual report. We calculate the reciprocal to obtain a measure that is high for more digitalized and low for less digitalized companies.[137] For convenience, the measure is then linearly scaled to the interval $[0, 1]$, i.e., the observation with the lowest value of the digitalization value assumes the value 0 while the observation with the highest digitalization value takes on the value 1.[138] A summary of the whole process for deriving the measure of digitalization can

[136] When we base the analysis on the five most pronounced topics (out of 45) in Bohnert et al. (2019) accounting for nearly 90 % of the paper, the mean KL divergences between our reference document on digitalization and the other documents on digitalization are the lowest, followed by the document on innovation and the remaining documents. However, to avoid the introduction of an additional parameter (the number of the most prominent topics to consider), our main measure is based on the whole topic distribution. Nevertheless, we include regression results for a digitalization measure based on the five most prominent topics in Bohnert et al. (2019) in Table 5.8.

[137] The main measure in Bellstam et al. (2020) is based on a fourth-root transformation. For parsimony, we do not further transform our measure. However, the empirical results remain qualitatively unchanged when we replace our original measure by its fourth root.

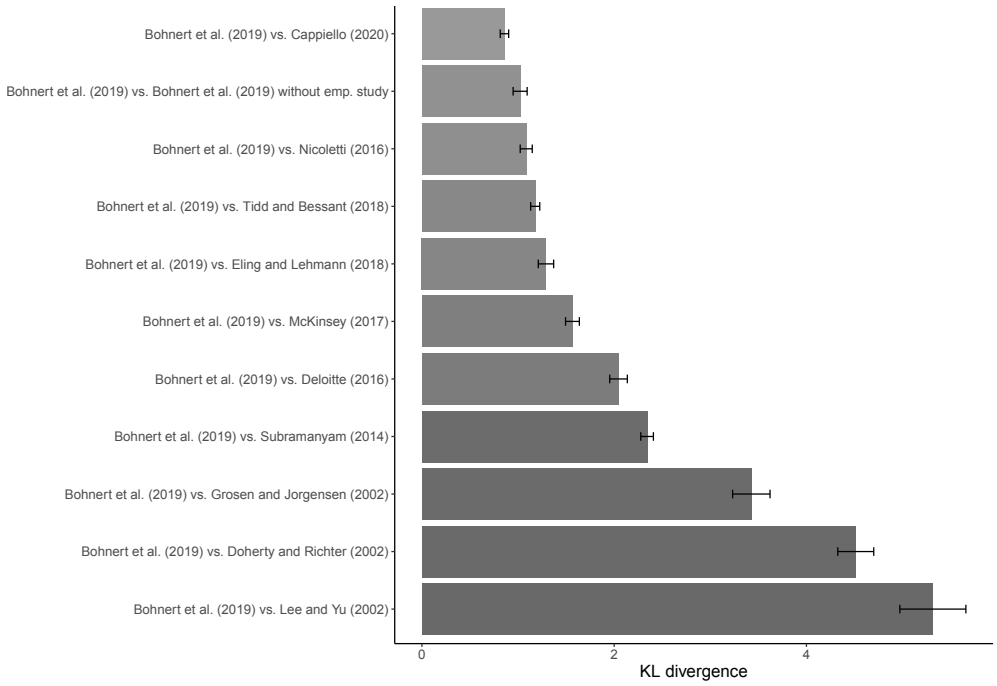
[138] In the following we provide some exemplary excerpts from the annual reports within the top decile of firm-year observations according to our text-based digitalization measure:

"*Responding to higher customer expectations, we recently introduced our mobile app, which makes it easier for our customers to find their nearest agent or repair shop from their mobile device, and a new online environment for policyholders to manage their policies. We have also made improvements to our online quoting interface to make the process of buying insurance easier.*" (Infinity Property and Casualty Corporation, annual report 2012)

"*Expanding use of third-party data and analytics will identify profitable growth opportunities. With the amount of consumer data available, tapping in to "Big Data" supports targeted marketing efforts based on educator household characteristics. This is the initial step in a multi-year strategy to*

Figure 5.3: The topic distribution as a representation of a document

This figure compares the distribution of topics in our reference document on digitalization (Bohnert et al., 2019) to various other documents on digitalization (Cappiello, 2020, Bohnert et al., 2019 with the empirical section being excluded, Eling and Lehmann, 2018, Nicoletti, 2016 and the white papers McKinsey, 2017 and Deloitte, 2016), on innovation (Tidd and Bessant, 2018), on financial statement analysis (Subramanyam, 2014) as well as randomly chosen papers from the oldest available issue in the public archive of The Journal of Risk and Insurance (Grosen and Jørgensen, 2002, Doherty and Richter, 2002, Lee and Yu, 2002) covering unrelated subjects. The dissimilarity between the respective topic distributions is measured by the KL divergence where more similar documents exhibit a lower KL divergence value. We provide the means of 100 bootstrap samples that we obtain by repeatedly sampling 90 % of the annual reports at random and applying the LDA method with 45 topics. Based on the topics obtained in each bootstrap sample, we derive the distribution of topics in the various reference documents that have not entered into the estimation process of the LDA. The bands provide 95 % confidence intervals for the mean computed based on the bootstrap samples.



be found in Figure 5.4.

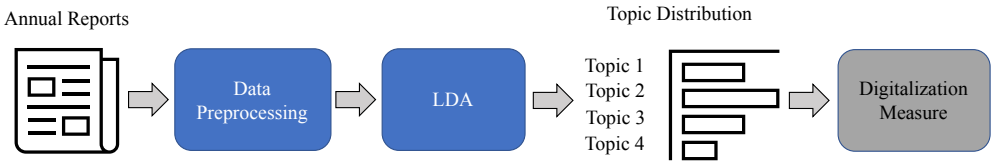
We depart from Bellstam et al. (2020) particularly by deriving our text-based measure of digitalization from the *distribution of topics* within each document while the

identify the most efficient ways to access preferred segments of the educator market." (Horace Mann Educators Corporation, annual report 2014)

"*More than 23 million consumers have access to Rally, our online digital health portal. Users are steadily advancing in selecting primary care physicians, making better use of urgent care over emergency care and more readily adopting personal health and condition management programs.*" (UnitedHealth Group, annual report 2015)

Figure 5.4: Construction of the text-based measure of digitalization

This figure, freely adapted from Lopez-Lira (2019), shows the main steps to derive our text-based measure of digitalization (see Section 5.2 for details). First, we preprocess the annual reports of US insurance companies (removing stopwords, stemming, etc.) and construct a document term matrix. Then, we employ LDA to derive a distribution over topics for each of the reports. These topic distributions are subsequently used to calculate the measure of digitalization.



measure of innovation in Bellstam et al. (2020) is based on the *loadings on a specific topic* (the "innovation topic") across all documents. Bellstam et al. (2020) identify the "innovation topic" by calculating the KL divergence between the distribution of *words* in a reference document on innovation (Tidd and Bessant, 2018) and the distribution of words in the topics derived via LDA. The topic with the lowest KL divergence is assumed to be the "innovation topic" and the innovation measure is defined as the loadings on the innovation topic across documents. For illustrative purposes, we provide a word cloud for the "digitalization topic"[139] derived analogously to this approach but based on the reference document on digitalization (Bohnert et al., 2019) in Figure 5.5.

We can, however, not simply apply this approach analogously to measure digitalization. When we replace the document on innovation by our reference document on

[139]Note that our proposed digitalization measure is not based on a specific topic (the "digitalization topic") and that the word cloud is provided for illustrative purposes only. Additionally, we want to point out that when deriving topics via LDA, the same word appears in multiple topics. More exactly, all topics obtained by a LDA consist of the same set of words. However, the topics differ in the probabilities they assign to each word. It is therefore not surprising that (stemmed) words like "compani", "insur" or "financi" appear frequently in the "digitalization topic". This does not mean that these words are more influential in the "digitalization topic" but merely reflects the overall commonality of these terms in the annual reports of US insurance companies.

The relative importance of single words can be better assessed based on the ratio between the frequency of a particular word in the "digitalization topic" and the overall frequency across all reports. To provide some examples, these ratios are smaller than 1 for "compani" (0.96), "insur" (0.90), and "financi" (0.90). This illustrates that even though these terms appear frequently in the "digitalization topic", they are less frequently used than in the average annual report. On the contrary, terms like "digit" (7.74), "mobil" (5.44), "internet" (5.06), and "autom" (5.13) occur clearly more often in the "digitalization topic" than in the average annual report.

Figure 5.5: Word cloud of the "digitalization topic"

This figure shows the 50 most frequent (stemmed) terms in the "digitalization topic". This topic is determined according to Bellstam et al. (2020). That is, the word distributions in each topic obtained via a 45 topic LDA are compared to the word distribution in the reference document on digitalization by Bohnert et al. (2019). We use the KL divergence as a measure of similarity between these distributions and choose the topic with the lowest KL divergence as the "digitalization topic". Note that our proposed text-based digitalization measure is not based on a specific topic (the "digitalization topic"). Instead, we derive our measure by exploiting the whole topic distribution, see Section 5.2.5 for details. The word cloud is provided for illustrative purposes only.

digitalization we identify the same topic. That is, when adapting the approach by Bellstam et al. (2020), the "innovation topic" and the "digitalization topic" are identical. Again, this is not surprising as innovation and digitalization are related topics. However, choosing the same topic as digitalization and innovation topic would imply both measures to be defined as the loadings across documents on the *same* topic. Consequently, the digitalization and the innovation measure would be identical. Of course, one could simply increase the number of topics until the innovation and the digitalization topic are different.[140] However, that causes another problem apart from adding more complexity to the model. The vector of loadings of all documents on a specific topic obtained via LDA is approximately orthogonal to the vector of loadings on any

[140] Even for 30 and 45 topics the digitalization and the innovation topic stay the same. The issue remains when we identify the digitalization topic based on another reference document on digitalization (Cappiello, 2020).

other topic. This stays true when the number of topics is increased (cf. Arun et al., 2010). This means that even if one could identify different innovation and digitalization topics, the innovation and the digitalization measures according to Bellstam et al. (2020) would exhibit a correlation of approximately 0. This would, however, contradict the intuition that innovation and digitalization are related concepts and would essentially represent a measure of digitalization while controlling for innovation.

Another problem that arises from measuring a concept of interest based on a single topic is the following: When increasing the number of topics, at some point a particular topic might split into two aspects of the same concept. This is, for instance, observed in Bellstam et al. (2020) in the case of 50 topics. The measures derived from the split topics would again have a correlation of nearly 0. Consequently, a measure based on a single topic might completely miss a part of the concept of interest. This problem does not emerge with our measure based on the distribution of topics rather than words.

5.2.6 Optimal number of topics

While the topics and the topic loadings arise endogenously from the LDA algorithm, the number of topics plays the role of a hyperparameter. In machine learning, this is a parameter that controls the learning process and has to be specified in advance. In contrast, the topic distributions within each document and the word distributions within each topic are parameters that are derived by the unsupervised machine learning algorithm via the training process. Therefore, prior to employing the LDA method, the researcher has to provide the number of topics to be extracted from the documents by the algorithm. On the one hand, the number of topics has to be sufficient to distinguish between different themes in the document, but on the other hand should not be too high to ensure interpretability of the topics (Lopez-Lira, 2019).

There exist several methods for determining the number of topics in a data driven manner, e.g., approaches introduced by Griffiths and Steyvers (2004), Cao et al. (2009), Arun et al. (2010), and Deveaud et al. (2014). These technical methods rely on differ-

ent objective functions to be minimized or maximized to best describe the underlying text corpus. Not surprisingly, these methods do not agree on the optimal number of topics in our application. Furthermore, some of them suggest even more than 100 topics whereby interpretability of the topics would be lost and overfitting issues might arise. According to Blei (2012), *"develop[ing] evaluation methods that match how the algorithms are used"*, particularly for determining the optimal number of topics, is still an open direction for research in topic modeling. In contrast to the technical methods for topic number selection, we do not intend to fully capture every aspect of the documents. Instead, we employ LDA to measure the extent to which digitalization is discussed in the reports. Although there is no clear guidance in the literature on how to select the optimal number of topics as most researchers apply different rules based on the task at hand, in many applications a number of topics between 10 and 50 seems to be appropriate (see, e.g., Bao and Datta, 2014, Israelsen, 2014, Ganglmair and Wardlaw, 2017, Bellstam et al., 2020, Weiss Hanley and Hoberg, 2019, Lopez-Lira, 2019).

Bellstam et al. (2020) use 15 topics to derive a text-based measure of innovation. Digitalization can be understood as a more specific concept than the broader notion of innovation. We therefore also consider 30 and 45 topics to resemble the thematic structure of the collection of annual reports more granularly. As innovation and digitalization are related topics we choose the number of topics such that we are best able to differentiate between these two concepts. Our measure of digitalization is based on the distribution of topics within each annual report. This distribution of each particular report is compared to the topic distribution in a reference document on digitalization (Bohnert et al., 2019). To differentiate between digitalization and innovation, we choose the number of topics such that the mean dissimilarity between the topic distribution in the reference document on digitalization and the document on innovation (Tidd and Bessant, 2018) is maximized. This leads to a number of 45 topics being most appropriate for measuring digitalization even when we additionally analyze 60

topics.[141] While Bellstam et al. (2020) rely on 15 topics to measure innovation, this number is too low to sufficiently capture the thematic structure of the annual reports in our sample in order to measure digitalization. This is manifested in the fact that for 15 topics the mean KL divergence between the documents on digitalization is not statistically significantly lower than the mean KL divergence between the document on digitalization and innovation. It is therefore not surprising that we partly yield insignificant results when including this measure based on 15 topics in our regression framework. However, additional to our baseline measure derived from 45 topics we also consider a measure obtained by a more parsimonious LDA based on 30 and even 15 topics in our regression framework, see Section 5.4.2.

5.2.7 Sentiment

Our text-based measure of digitalization as introduced in the previous sections does not take sentiment into account. For a report covering digitalization in a rather negative tone, our procedure might nevertheless assign a high value to the digitalization proxy although this report is less likely to represent more digitalization efforts by the firm (cf. Bellstam et al., 2020). This might induce measurement errors to our measure of digitalization.

Sentiment is most frequently measured by counting the occurrence of specific "positive" or "negative" words see (see Henry and Leone, 2016). Of course, the drawbacks related to word list based approaches (see Section 5.2.4) also apply in this case. First of all, lists of positive and negative words have to be provided by the researcher. Secondly, the sentiment of specific words can depend on the context in which it is used. For example, Loughran and McDonald (2011) find that more than 70 % of the nega-

[141] To ensure robust results, we calculate the similarity measure (KL divergence) based on 100 bootstrap samples which are computed by a LDA with 15, 30, 45, and 60 topics based on 100 subsamples obtained by repeatedly choosing 90 % of the reports at random. Based on the topics we derive for each of these samples we can calculate the distribution of topics in the documents on digitalization and innovation. We then compute the similarity measure between these topic distributions. The number of topics is then selected based on the mean over the 100 bootstrap samples. These calculations are computationally very expensive and were performed on the Big-Data-Cluster Galaxy provided by the University Computing Center at Leipzig University.

tive words in the Harvard Psychosociological Dictionary[142] are typically not negative in a financial context (e.g., board, capital or liability). We therefore use a dictionary of positive and negative words provided by Loughran and McDonald (2011) and Bodnaruk et al. (2015) that has been specifically adjusted for financial language.[143] We follow Bellstam et al. (2020) and measure sentiment as the difference between positive and negative words divided by the total number of words. Of course, there are other dictionaries as well as other methods to measure sentiment. For more details we refer to Loughran and McDonald (2011) and Loughran and McDonald (2016).

To overcome potential measurement errors introduced by not considering sentiment, we exclude all firm-year observations with sentiment below the 25 % quantile in a robustness check (see Section 5.4.3). Note that the LDA is performed based on all annual reports in our sample and firm-year observations are removed according to sentiment *after* the construction of the digitalization measure. Consequently, topics still arise endogenously from the data and are not affected by the word lists on which measurement of sentiment is based on.

5.3 Financial data & empirical strategy

5.3.1 Sample construction

We start the construction of our sample by selecting all insurance companies in the US with stock market data available in Thomson Reuters Datastream. We use market value based measures due to their ability to capture short-term performance and long-term prospects (Lubatkin and Shrieves, 1986, Allen, 1993). The focus on publicly-listed US insurance companies is motivated by the strong distortionary effect different regulatory and accounting standards in different countries would have in our setting. Thus, we collect annual reports for as many publicly-listed US insurers as possible. Annual

[142] The Harvard Psychosociological Dictionary, more specifically the Harvard-IV-4 TagNeg file, is a commonly used source for word classifications.

[143] See Loughran and McDonald Sentiment Word Lists at `https://sraf.nd.edu/textual-analysis/resources`.

reports have already been subject to textual analysis (e.g., Li and Racine, 2008, Yekini et al., 2016, Gatzert and Heidinger, 2020). In contrast to Bellstam et al. (2020), we prefer annual reports over analyst reports since the former provide first-hand information about a company's status quo, current projects, and upcoming trends. Furthermore, we consider digitalization issues to be less obvious to externals because digitalization efforts are often aimed at improving internal procedures rather than developing new insurance products.

We complement our data set with accounting data (i.e., total assets, ROA, total investment, solvency ratio, and current liquidity ratio) from Orbis Insurance Focus covering the period from 2006 to 2015. As will be described in the following Section 5.3.3, we lag all explanatory variables by one year. Thus, our initial sample consists of 86 insurance companies from 2006 to 2015 resulting in 748 observations in an unbalanced panel.

5.3.2 Summary statistics

Table 5.2 presents the summary statistics of the sample. Our first main measures of interest, *market value*, exhibits mean and median values of $US 7.69 bn. and $US 2.33 bn. indicating a right-skewed distribution. We account for the skewness by taking the log of *market value*. The mean and median of the second dependent variable, *market-to-book value*, are 1.25 and 1.10, respectively. Our main explanatory variable, *digitalization* based on a 45 topic distribution, ranges between 0 and 1 with mean and median of 0.15 and 0.05, respectively. Values of the digitalization measure close to 1 indicate a very digitalized insurance company whereas values close to 0 indicate the opposite. Consequently, a mean value of 0.15 emphasizes that the insurance industry on average has yet to take advantage of the full potential of digital technologies. Companies that exhibit particularly high values of digitalization over time are Horace Mann Educators, Argo, and Aflac, for instance. Their digitalization values continuously rank among the top 10 %. In contrast, companies that digitalize comparatively

little over time are MetLife, Enstar, or Alleghany. Here, digitalization appears to be less pronounced in annual reports with values below 0.025. Statistical moments of the three alternative digitalization measures based on 15, 30, and 60 topic distributions are similar to those of our main digitalization measure.

Table 5.2: Summary statistics

This table presents summary statistics on all variables used in the multivariate OLS analyses. The panel spans from 2007 to 2016. The following columns present the number of observations, mean, median, standard deviation, as well as minimum and maximum value. The variable definitions and data sources are given in Appendix I.

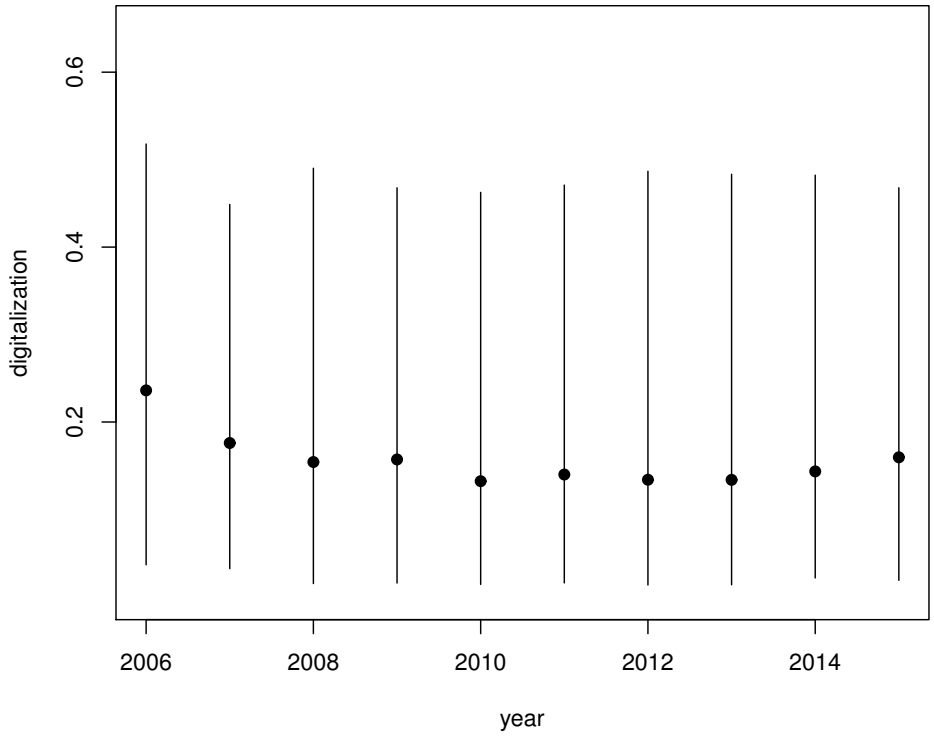
	Obs.	Mean	Median	Std. Dev.	Min	Max
Market Value (bn.)	748	7.69	2.33	15.45	0.00	186.30
Market-to-Book	735	1.25	1.10	0.76	-0.79	8.52
$\text{Digitalization}_{15,t-1}$	748	0.15	0.12	0.15	0.00	1.00
$\text{Digitalization}_{30,t-1}$	748	0.16	0.07	0.18	0.00	1.00
$\text{Digitalization}_{45,t-1}$	748	0.15	0.05	0.20	0.00	1.00
$\text{Digitalization}_{60,t-1}$	748	0.19	0.08	0.22	0.00	1.00
$\text{Total Assets}_{t-1}$ (bn.)	748	54.60	8.95	140.84	0.01	1048.36
ROA_{t-1}	748	3.42	2.96	4.60	-36.37	20.23
Investment_{t-1} (bn.)	748	32.78	5.84	87.85	0.00	806.04
Solvency R._{t-1}	748	25.69	25.73	13.08	1.92	69.46
Current R._{t-1}	748	93.40	92.12	34.09	3.42	289.77
Foreign Assets R._{t-1}	527	2.81	0.00	12.35	0.00	87.38

The time dimension of our digitalization variable is depicted in Figure 5.6 along with the 10th and 90th percentiles indicating a high cross-sectional variation. At first sight, the fairly stable annual mean values just below 0.2 with a peak above 0.2 in 2006 appear to be counterintuitive since one might expect digitalization to become more important over time, especially in recent years. However, taking into consideration the time span of our sample from 2006 to 2015 it rather covers what could be described as

the first surge of digitalization. This first surge is characterized by improvements like e-mail alerts or investments in online resources.[144]

Figure 5.6: Digitalization measure over time

This figure shows the mean of the digitalization measure (based on 45 topics) between 2006 and 2015 along with the 10th and 90th percentiles. For details on the construction of our text based measure of digitalization we refer to Section 5.2.5.



To get a better idea of what is causing this rather constant digitalization trend on average, we next examine the average relative frequency of particular words related to digitalization in the annual reports over time.[145] The results presented in Figure 5.7

[144] In the following we provide exemplary excerpts from the 2006 annual reports of Aflac Inc. and Principal Financial Group:

"In 2006, we also tested AflacAnywhere^{SM}, a new technology that greatly improves communications between headquarters and our sales force. AflacAnywhere enables sales associates and coordinators to receive notification on important information via ***e-mail alert****, text message on a cell phone or PDA, or computer-generated voice message to any phone number."* (Aflac Inc., annual report 2006)

"For customers, we've ***invested in online resources****, improved technology in our contact centers and simplified communication materials."* (Principal Financial Group, annual report 2006)

[145] Note that our text based measure of digitalization is not based on particular words but on the distri-

show that basic digital issues often expressed through terms stemming from “internet”, “online”, and “web” are more prominent in the first years of our sample period. In the last years of our sample period, one can recognize the beginning of a second surge of digitalization expressed through terms stemming from “data”, “mobile”, or “digital”. In between, the average value of the digitalization proxy decreases (see Figure 5.6) which might be due to consequences of the financial crisis in 2007/2008 for the insurance industry including regulatory issues. Taken together, we can infer that the rather constant average trend of our measure is probably caused by the balancing effects of different digital issues across time.

In addition to the variance of the digitalization measure across time, we also investigate its cross-sectional variation with respect to business line, size, profitability, and market orientation. Table 5.3 presents the results of mean comparisons for life and non-life, small and large, profitable and less profitable, and (inter-)national insurance firms.

Classification into life and non-life stems from Thomson Reuters Datastream. Since some insurers cover both business lines we excluded them from this comparison. The mean values do not differ significantly which also applies to the median values (see Figure 5.8(a)). There seems to be no significant difference in the extent of digitalization across business lines.

In contrast, size seems to be of higher importance. A comparison of the first and the fourth quartile of insurance companies with respect to size shows a significant difference (see Table 5.3). On average, large firms exhibit higher values of digitalization (mean = 0.163) than small firms (mean = 0.093). However, comparing the median values in Figure 5.8(b) gives a different picture since median digitalization in the smallest firms lies above the median digitalization of large firms. Hence, this evidence does not support a clear positive relation between size and the extent of digitalization in an insurance company. Furthermore, we do not find clear evidence that profitability affects the extent of digitalization in a company. Although from Figure 5.8(c) one might

bution of topics in the annual reports. Figure 5.7 is provided for illustrative purposes only.

Figure 5.7: Relative frequency of words related to digitalization

This figure presents the relative frequency of words related to digitalization in the annual reports in our sample between 2006 and 2015. Note that our measure of digitalization is derived from the distribution of topics within the annual reports (see Section 5.2.5).

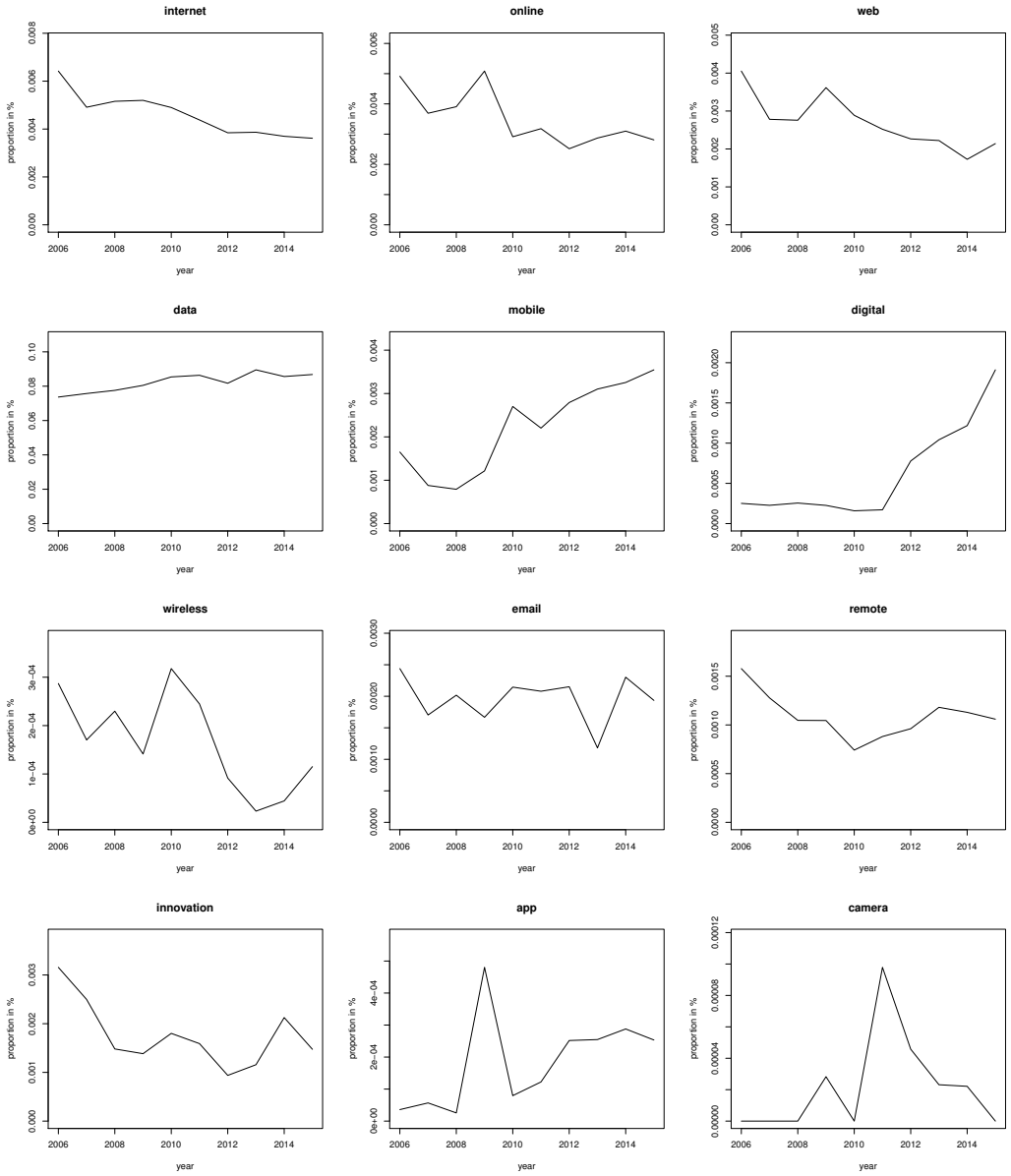


Table 5.3: Descriptive statistics (subsamples)

This table presents summary statistics on four characteristic variables of our sample: Sector/Business Line, Size, Profitability, and Market Orientation. The panel spans from 2007 to 2016. The following columns present the number of observations, mean, standard deviation, as well as the difference in means including the p-value. The division into life and non-life stems from Thomson Reuters Datastream. The subsample statistics for size and profitability only consider the first (small) and fourth (large) quartile of the respective variable.

Sector	Life		Non-Life			
	Mean	SD	Mean	SD	Difference	p-value
$Digitalization_{45,t-1}$	0.156	0.196	0.140	0.204	0.016	(0.374)
Observations	159		494		653	

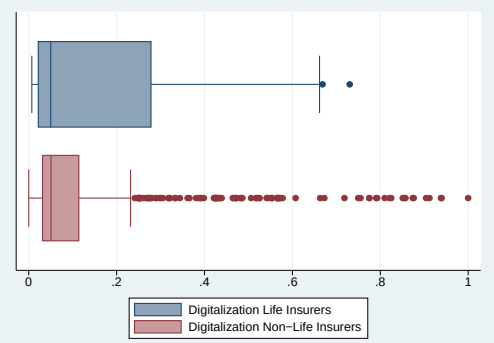
Size	Small		Large			
	Mean	SD	Mean	SD	Difference	p-value
$Digitalization_{45,t-1}$	0.093	0.109	0.163	0.200	-0.070***	(0.000)
Observations	187		187		374	

Profitability	Small		Large			
	Mean	SD	Mean	SD	Difference	p-value
$Digitalization_{45,t-1}$	0.132	0.202	0.162	0.179	-0.030	(0.132)
Observations	187		186		373	

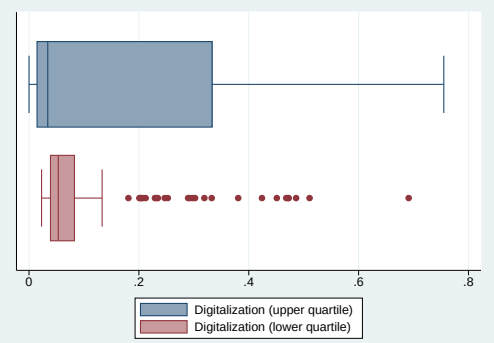
Orientation	International		National			
	Mean	SD	Mean	SD	Difference	p-value
$Digitalization_{45,t-1}$	0.123	0.167	0.163	0.215	0.031*	(0.080)
Observations	197		551		677	

Figure 5.8: Digitalization distributions according to different criteria

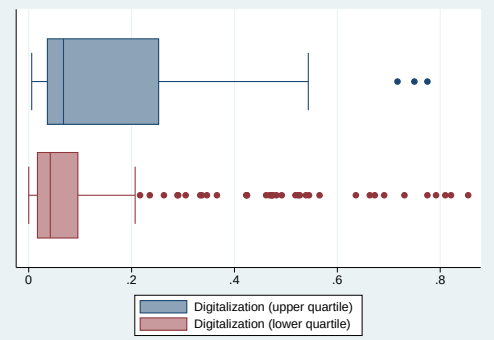
This figure presents the digitalization measures of insurance companies included in our sample with respect to five categories: business line (a), size (b), profitability (c), market orientation (d), and geographical origin (e). Classification into business line, i.e., life and non-life, stems from Thomson Reuters Datastream. We excluded insurers covering both business lines from the comparison. Classification according to size (large vs. small) and profitability (high vs. low) is made using the upper and lower quartile of companies with respect to total assets and ROA, respectively. Market orientation (international vs. national) is assessed based on the ratio of foreign assets to total assets. A positive ratio is considered to indicate an international orientation of the firm. For the comparison of geographical origin we contrast the digitalization measures of eight large European insurance companies, namely Allianz, Aviva, AXA, CNP, Generali, Mapfre, Prudential, and Zurich with those of eight similarly large US insurance companies, namely AIG, MetLife, Hartford, Lincoln National, Principal, Allstate, Ameriprise, and Aflac.



(a) Life vs. Non-Life

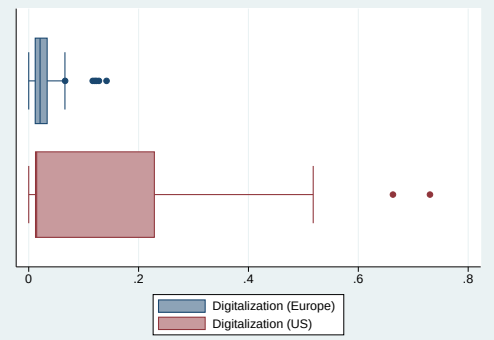


(b) Large vs. small



(c) Profitability high vs. low

(d) International vs. national insurance companies



(e) European vs. US insurance companies

get the impression that the most profitable firms (top 25 %) exhibit higher values of digitalization on average, there is no significant difference in means (see Table 5.3).

Digitalization might also be related to the market orientation of a company (national vs. international). However, using the ratio of foreign assets to total assets from Thomson Reuters Datastream as a proxy for the market orientation of an insurer, we do not find strong empirical evidence for such a relation. According to Table 5.3, the difference in means between the international and national insurance companies in our sample is only weakly significant at the 10 percent level with national insurers exhibiting a higher mean (0.163). However, the median values appear to be fairly equal (see Figure 5.8(d)).

In a further illustrative example, we compare digitalization in the US insurance sector with the European insurance market. Therefore, we apply our LDA to the annual reports of eight large European insurance companies, namely Allianz, Aviva, AXA, CNP, Generali, Mapfre, Prudential, and Zurich. The choice of the European insurance companies is motivated by their international focus as well as their size and hence relevance for the European insurance sector. To maintain comparability, we also choose a subsample of large US insurance companies for the descriptive analysis of the digitalization measures. These are AIG, MetLife, Hartford, Lincoln National, Principal, Allstate, Ameriprise, and Aflac. The results are presented in Figure 5.8(e) and Table 5.4.

Although median values of the two groups are relatively similar, mean values differ substantially. On average, large US insurance companies exhibit much higher values of digitalization than their European counterparts. In detail, the mean digitalization value of the eight large US insurance companies (0.129) is four times higher than the one of the European insurance companies (0.030). This leads to the conclusion that digitalization might be of higher importance for American insurance companies whereas European insurers might lag behind.

Table 5.4: Summary statistics European vs. US subsample

This table presents summary statistics on the digitalization measure based on a LDA (45 topic distribution) applied to the annual reports of eight large US and European insurance companies, respectively. The annual reports cover the period from 2006 to 2015. The European insurance companies are Allianz, Aviva, AXA, CNP, Generali, Mapfre, Prudential, and Zurich. The US insurance companies are AIG, MetLife, Hartford, Lincoln National, Principal, Allstate, Ameriprise, and Aflac. The companies are chosen on the basis of size and international focus (for the European insurers). For further explanation see Section 5.3.2. The following columns present the number of observations, mean, median, standard deviation, as well as minimum and maximum value.

	Obs.	Mean	Median	Std. Dev.	Min	Max
$\text{Digitalization}_{\text{European}}$	74	0.030	0.021	0.031	0.000	0.142
$\text{Digitalization}_{\text{US}}$	80	0.129	0.014	0.195	0.000	0.730

Notes: Data from 2006 to 2015.

European Insurance Companies: Allianz, Aviva, AXA, CNP, Generali, Mapfre, Prudential, Zurich

US Insurance Companies: AIG, MetLife, Hartford, Lincoln National, Principal, Allstate, Ameriprise, Aflac

5.3.3 Empirical strategy

We analyze how digitalization of the business model is associated with the firm value of an insurance company using panel data regressions. However, estimating the relation is not straightforward due to issues of endogeneity. In fact, one could imagine that large and profitable insurance companies have more capacities to invest in digitalization than small insurers. This would lead to a situation in which digitalization not just affects market valuation, but vice versa. As a result, our estimated coefficients would be biased because of reverse causality. In addition, there is also a problem of omitted variable bias. More specifically, other independent variables omitted in the regression are likely to be correlated with both our proxies for market valuation and the main explanatory variable digitalization. If they become part of the error term, the OLS assumption of conditional mean independence will be violated resulting in biased estimates for the effect of digitalization on market valuation. We address the potential endogeneity by making use of two basic econometric means. First, we lag all our explanatory variables by one year to make sure that they will not be affected by the current firm

value and hence not be subject to estimation biases due to reverse causality. Second, we add company and year fixed effects to the regressions to account for unobserved heterogeneity across firms and time.[146] In particular, we estimate the model:

$$FirmValue_{i,t} = \beta_0 + \beta_1 Digitalization_{i,t-1} + \beta_2 X_{i,t-1} + \delta_t + u_i + \epsilon_{i,t}.$$

The index i represents a particular insurance company whereas index t denotes the year. Firm value is approximated by the *market value* or the *market-to-book value*, respectively. $Digitalization_{i,t-1}$ represents the digitalization measure derived from our LDA. As already mentioned, we find 45 topics to maximize the mean dissimilarity between the topic distribution in the reference document on digitalization and the document on innovation, thus reflecting the digitalization approach most accurately. However, in subsequent robustness checks we also run the regression with digitalization measures derived from 15 and 30 topic distributions. $X_{i,t-1}$ is a vector of control variables that are commonly used. In detail, we add the natural logarithm of total assets to account for the size of an insurer, return on assets (ROA) to account for its profitability, the natural logarithm of total investment, and the solvency as well as the current liquidity ratio to account for the short- and long-term obligations of an insurance company.

5.4 Estimation results

5.4.1 Baseline estimation

Having laid out the estimation strategy to identify the relation between digitalization and firm outcome, Table 5.5 presents the results of our main estimation using our whole sample of insurance companies and a distribution over 45 topics to calculate

[146] Another method to address endogeneity is the use of instrumental variables. However, since empirical evidence on the relation between firm value and digitalization is scarce, we were not able to find a commonly accepted instrument in the literature. In separate regressions, we tried to instrument our digitalization measure using dummy variables for the presence of a digital officer and a CEO change, respectively. Unfortunately, both of them turned out to be invalid. Therefore, we leave the issue of instrumentation up for future research.

our digitalization measure. As already mentioned, we use two proxies for firm value, i.e., *market value* and *market-to-book value*. Furthermore, we control for unobserved heterogeneity using firm fixed effects (columns 1 to 4) and also time fixed effects (column 2 and column 4).[147] In all specifications, the relation between digitalization and market value and market-to-book value, respectively, is positive and statistically significant. However, as already mentioned in Section 5.3.3, our estimated coefficients might be subject to endogeneity. Consequently, establishing a causal link between digitalization and firm value is not possible unequivocally. The estimated coefficient for digitalization in the OLS regression on market value including company fixed effects is 0.374. Hence, an increase in digitalization by one standard deviation is associated with an increase in market value by about 7.48 % (0.20×0.374) in the subsequent year. Similarly, an increase in digitalization by one standard deviation is related to an increase in the market-to-book value by about 8.04 % (0.20×0.402) in the next year.[148] The relation becomes slightly weaker when we add time fixed effects but remains significant. Our results provide evidence for a strong positive relation between digitalization and firm valuation. Put differently, market participants might expect insurance companies making progress in digitalization to attain higher future cashflows and become more profitable. As a result, the market value and the market-to-book value increase.

5.4.2 Alternative number of topics

To check for the robustness of our main estimation results, we replace our main explanatory variable by digitalization measures derived from LDA with different numbers of topics. The results are presented in Table 5.6. In detail, we use a 15 topic, a 30 topic, and a 60 topic LDA model, respectively. The 15 topic model has already

[147] Pointing towards the R^2 of each regression, it can be seen that the explanatory power of our estimation equation is fairly high which is probably due to firm fixed effects. They obviously capture a lot of variation in the data. Consequently, adding time fixed effects does not add much to the explanatory power leading to the conclusion that time trends do not play an important role in our setting.

[148] Instead of considering an increase of the digitalization measure by one standard deviation, one could also examine the effect of an increase from the first to the third quartile. Such an increase of about 0.1771 is associated with an increase in market value of approximately 6.62 % (0.1771×0.374). For the market-to-book value the increase is about 7.12 % (0.1771×0.402).

Table 5.5: The relation between digitalization and firm valuation

This table presents the results of the panel regressions that examine the relation between digitalization based on a 45 topic distribution and firm value proxied by the log of *market value* and *market-to-book value* (MtB), respectively. The topic distribution in each annual report has been compared to the one in Bohnert et al. (2019). Column (1) and column (2) show the estimation results for the OLS regressions using the log of *market value* as the dependent variable. Column (3) and column (4) report the estimation results of *market-to-book value* as the dependent variable. The panel has one observation for each company-year combination, and spans the time period 2007-2016. We include company fixed effects in all specifications. Column (2) and column (4) additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) and test statistics for the joint significance of regressors (F-test) are reported at the bottom of the table. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively.

Dependent	(1) Market Value	(2) Market Value	(3) MtB	(4) MtB
Digitalization$_{45,t-1}$	0.374***	0.238***	0.402***	0.188*
	(0.091)	(0.084)	(0.135)	(0.110)
Total Assets$_{t-1}$	0.506**	0.207	0.901***	0.726***
	(0.231)	(0.171)	(0.255)	(0.278)
ROA$_{t-1}$	0.029***	0.007	0.044***	0.014*
	(0.010)	(0.008)	(0.008)	(0.007)
Investment$_{t-1}$	0.430**	0.590***	-0.597***	-0.348*
	(0.208)	(0.134)	(0.201)	(0.201)
Solvency Ratio$_{t-1}$	0.037***	0.035***	-0.028**	-0.017*
	(0.008)	(0.005)	(0.012)	(0.010)
Current Ratio$_{t-1}$	-0.009***	-0.009***	0.012**	0.011**
	(0.003)	(0.002)	(0.005)	(0.005)
Company fe	Yes	Yes	Yes	Yes
Year fe	No	Yes	No	Yes
Observations	748	748	735	735
R^2	0.97	0.98	0.60	0.71
F-Test	53.04	39.79	8.26	3.95

been employed by Bellstam et al. (2020). However, in our context it rather serves as a reference since a lower number of topics might blur the line we want to establish between innovation in general and digitalization in particular.[149] Taking this into account, it is not surprising that the statistical significance of the estimated coefficients for the digitalization variable based on 15 topics (line 1) is not as strong as in Table 5.5. Therefore, we can infer that a more granular topic distribution is necessary to capture digitalization instead of general firm innovation. However, the estimated coefficient in the model using firm fixed effects (column 1) is of the same magnitude.

In contrast to the 15 topic model specification, the estimated coefficients for our models based on 30 and 60 topic distributions (line 2 and line 3) are predominantly statistically significant. Especially OLS estimation results using firm fixed effects exhibit estimation coefficients for the main explanatory variable that are similar to those in our main estimations (Table 5.5). Hence, we can conclude that our main estimations are robust to alternative calculations of the digitalization measure using different topic distributions as long as the number of topics exceeds those used to capture general innovation (see, e.g., Bellstam et al., 2020).

5.4.3 Sentiment analysis

As we argue in Section 5.2.7, a particular report with a high value of the digitalization value is less likely to represent more digitalization by the firm when the report is written in a negative or neutral tone. In a subsample analysis, we therefore only consider firm-year observations with a sentiment above the 25 % quantile.

The results presented in Table 5.7 for market value and the market-to-book value, respectively, are in line with those presented in Table 5.5 and Table 5.6. Again, there is a positive and (highly) statistically significant relation between digitalization based on a 45 topic distribution and market value and market-to-book value, respectively. The estimated coefficients are of the same magnitude. According to this specification, an increase in digitalization by one standard deviation is associated with an increase

[149] See Section 5.2.6 for details.

Table 5.6: The relation between digitalization and firm valuation (altered topic distribution)

This table presents the results of the panel regressions that examine the relation between digitalization based on alternative topic distributions on firm value proxied by the log of *market value* and *market-to-book value* (MtB), respectively. The topic distribution in each annual report has been compared to the one in Bohnert et al. (2019). Columns (1) to (6) show estimation results for *market value* as dependent variable. In Columns (7) to (12), regression results for *market-to-book value* are reported. In column (1) and (2) as well as in column (7) and (8) the digitalization measure is based on a 15 topic distribution, whereas columns (3) and (4) as well as columns (9) and (10) were calculated on the basis of a 30 topic distribution LDA. Columns (5) and (6) as well as columns (11) and (12) report the results based on a 60 topic distribution LDA. The panel has one observation for each company-year combination, and spans the time period 2007-2016. We include company fixed effects in all specifications. Even columns additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) and test statistics for the joint significance of regressors (F-test) are reported at the bottom of the table. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively.

Dependent	(1) Market Value	(2) Market Value	(3) Market Value	(4) Market Value	(5) Market Value	(6) Market Value	(7) MtB	(8) MtB	(9) MtB	(10) MtB	(11) MtB	(12) MtB
Digitalization$_{15,t-1}$	0.480***	0.325***					0.497**	0.108				
	(0.127)	(0.121)					(0.220)	(0.167)				
Digitalization$_{30,t-1}$			0.426***	0.297***					0.535***	0.267*		
			(0.104)	(0.082)					(0.190)	(0.142)		
Digitalization$_{60,t-1}$					0.371***	0.246***					0.477***	0.238**
					(0.091)	(0.070)					(0.144)	(0.106)
Total Assets$_{t-1}$	0.514**	0.207	0.524**	0.214	0.502**	0.200	0.910***	0.732***	0.921***	0.731***	0.894***	0.717**
	(0.232)	(0.173)	(0.231)	(0.171)	(0.232)	(0.172)	(0.257)	(0.280)	(0.256)	(0.279)	(0.255)	(0.279)
ROA$_{t-1}$	0.029***	0.007	0.028***	0.007	0.028***	0.007	0.045***	0.014*	0.044***	0.014*	0.044***	0.014*
	(0.011)	(0.008)	(0.011)	(0.008)	(0.010)	(0.008)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)
Investment$_{t-1}$	0.422**	0.584***	0.423**	0.583***	0.437**	0.593***	-0.606***	-0.350*	-0.604***	-0.353*	-0.585***	-0.344*
	(0.208)	(0.134)	(0.208)	(0.133)	(0.209)	(0.134)	(0.203)	(0.202)	(0.201)	(0.202)	(0.201)	(0.202)
Solvency Ratio$_{t-1}$	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***	-0.028**	-0.017*	-0.027**	-0.017*	-0.027**	-0.017*
	(0.008)	(0.005)	(0.008)	(0.005)	(0.008)	(0.005)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)
Current Ratio$_{t-1}$	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	0.013**	0.011**	0.012**	0.011**	0.012**	0.011**
	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Company fe	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fe	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	748	748	748	748	748	748	735	735	735	735	735	735
R^2	0.97	0.98	0.97	0.98	0.97	0.98	0.60	0.71	0.60	0.71	0.60	0.71
F-Test	49.47	39.48	52.69	40.44	53.09	41.40	8.16	3.98	8.17	4.01	8.62	4.24

Table 5.7: The relation between digitalization and firm valuation above the 25 % sentiment

This table presents the results of the panel regressions that examine the relation between digitalization and *market value* (Columns 1 to 6) and *market-to-book value* (Columns 7 to 12), respectively, in sentiment subsamples. Only those company-year observations that exhibit a sentiment above the 25%-percentile have been selected. Column (1) and column (2) as well as Column (7) and column (8) show the estimation results using a 30 topic distribution to construct the digitalization regressor. Column (3) and column (4) as well as column (9) and column (10) report the estimation results including a digitalization regressor based on a 45 topic distribution. Column (5) and (6) as well as column (11) and (12) report the results based on a 60 topic distribution LDA. The panel has one observation for each company-year combination, and spans the time period 2007-2016. We include company fixed effects in all specifications. Even columns additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) an test statistics for the joint significance of regressors (F-test) are reported at the bottom of the table. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively.

Dependent	(1) Market Value	(2) Market Value	(3) Market Value	(4) Market Value	(5) Market Value	(6) Market Value	(7) MtB	(8) MtB	(9) MtB	(10) MtB	(11) MtB	(12) MtB
Digitalization$_{30,t-1}$	0.446***	0.318***					0.443**	0.232				
	(0.113)	(0.090)					(0.196)	(0.159)				
Digitalization$_{45,t-1}$			0.381***	0.265***					0.329**	0.169		
			(0.096)	(0.091)					(0.128)	(0.110)		
Digitalization$_{60,t-1}$					0.345***	0.222***					0.409***	0.216*
					(0.092)	(0.069)					(0.144)	(0.115)
Total Assets$_{t-1}$	0.522**	0.277*	0.502**	0.268	0.492**	0.257	0.749***	0.571**	0.730***	0.565**	0.717***	0.552**
	(0.227)	(0.164)	(0.226)	(0.163)	(0.228)	(0.166)	(0.232)	(0.248)	(0.232)	(0.248)	(0.232)	(0.249)
ROA$_{t-1}$	0.024**	0.002	0.024**	0.002	0.023**	0.002	0.046***	0.015*	0.046***	0.015*	0.046***	0.015*
	(0.012)	(0.009)	(0.012)	(0.008)	(0.011)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Investment$_{t-1}$	0.372*	0.543***	0.379*	0.551***	0.391*	0.557***	-0.513***	-0.270	-0.508***	-0.264	-0.490***	-0.258
	(0.205)	(0.130)	(0.205)	(0.130)	(0.207)	(0.131)	(0.176)	(0.181)	(0.176)	(0.181)	(0.176)	(0.182)
Solvency Ratio$_{t-1}$	0.030***	0.032***	0.029***	0.032***	0.030***	0.032***	-0.025***	-0.014*	-0.025***	-0.014*	-0.024**	-0.014*
	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.009)	(0.009)	(0.010)	(0.009)	(0.009)	(0.009)
Current Ratio$_{t-1}$	-0.008**	-0.008***	-0.008**	-0.008***	-0.008**	-0.008***	0.010**	0.008*	0.010**	0.008*	0.010**	0.008*
	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Company fe	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fe	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	572	572	572	572	572	572	559	559	559	559	559	559
R^2	0.97	0.98	0.97	0.98	0.97	0.98	0.69	0.79	0.69	0.79	0.70	0.79
F-Test	38.98	38.80	39.29	38.12	39.16	39.17	8.26	4.37	8.50	4.68	8.57	4.83

in market value by about 7.62 % (0.20×0.381) in the model with firm fixed effects (column 3). The estimated coefficients for the digitalization proxies based on 30 and 60 topic distributions are also similar to those reported in Table 5.6.

5.4.4 Alternative specifications and reference documents

In the last part of our empirical investigation, we control for potential confounding effects due to the construction of our digitalization measure or the choice of our reference document. According to Bellstam et al. (2020), we consider a measure based on only 98 % of the annual reports with the longest 2 % of the reports being excluded.[150] Furthermore, we construct a measure based on the five most prominent topics (out of 45) in Bohnert et al. (2019) representing nearly 90 % of the paper to account for potential confounding effects by the less pronounced topics. Finally, we consider a fourth-root transformation of the original measure.

The choice of Bohnert et al. (2019) as our main reference paper is motivated by its particular focus on digitalization in the insurance sector. However, the paper contains an empirical section which might complicate a proper analysis of the topic distribution. As a robustness check, we therefore compare the topic distributions in our annual reports to those in Bohnert et al. (2019) without the empirical section. Furthermore, we choose another insurance-related paper (Cappiello, 2020) as well as a textbook (Nicoletti, 2016) and a consultancy report (McKinsey, 2017) on digitalization in the insurance sector to justify that our estimation results do not depend on the particular choice of the reference document on digitalization.[151]

The results are presented in Table 5.8 for the market value as the dependent variable and in Table 5.9 for the market-to-book ratio as the dependent variable. The estimated coefficients (rows 1 to 7) are in line with those resulting from our baseline regression. The relation between digitalization and firm value remains positive and statistically

[150] In Bellstam et al. (2020) all analyst reports with less than 100 words are dropped as well. However, as all annual reports in our sample consist of more than 100 words (after preprocessing), we do not exclude any short reports.

[151] We also compared our results to other insurance-related papers, e.g., Eling and Lehmann (2018). The results are in line with our findings.

Table 5.8: The relation between digitalization and firm market value (alternative measure construction and reference documents)

This table presents the results of the panel regressions that examine the relation between digitalization based on a 45 topic distribution and firm value proxied by the log of *market value*. For robustness, we consider digitalization measures that have been constructed in an alternative way or are based on a different reference document on digitalization: A measure based on only 98 % of the annual reports with the longest 2 % of the reports being excluded (columns 1 and 2), a measure based on the five most prominent topics (out of 45) in Bohnert et al. (2019) (columns 3 and 4), a fourth-root transformation of the original measure (columns 5 and 6), and measures based on the reference documents Bohnert et al. (2019) without the empirical section (columns 7 and 8), Cappiello (2020) (columns 9 and 10), Nicoletti (2016) (columns 11 and 12), and McKinsey (2017) (columns 13 and 14).

The panel has one observation for each company-year combination, and spans the time period 2007-2016. We include company fixed effects in all specifications. Even columns additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) and test statistics for the joint significance of regressors (F-test) are reported at the bottom of the table. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively.

Dependent	(1) MV	(2) MV	(3) MV	(4) MV	(5) MV	(6) MV	(7) MV	(8) MV	(9) MV	(10) MV	(11) MV	(12) MV	(13) MV	(14) MV
Digitalization$_{BNT98,t-1}$	0.352***	0.272***												
	(0.105)	(0.095)												
Digitalization$_{BNT5,t-1}$			0.404***	0.255***										
			(0.101)	(0.096)										
Digitalization$_{B4R,t-1}$					0.365***	0.228***								
					(0.083)	(0.074)								
Digitalization$_{BNTtxt,t-1}$							0.436***	0.271***						
							(0.107)	(0.097)						
Digitalization$_{CPLO,t-1}$									0.305***	0.207***				
									(0.080)	(0.072)				
Digitalization$_{NCLT,t-1}$											0.357***	0.235***		
											(0.090)	(0.075)		
Digitalization$_{McK,t-1}$													0.310***	0.213***
													(0.085)	(0.074)
Total Assets$_{t-1}$	0.526**	0.237	0.506**	0.209	0.502**	0.202	0.510**	0.209	0.510**	0.207	0.510**	0.206	0.512**	0.208
	(0.230)	(0.172)	(0.231)	(0.171)	(0.230)	(0.171)	(0.231)	(0.171)	(0.231)	(0.171)	(0.230)	(0.171)	(0.230)	(0.171)
ROA$_{t-1}$	0.030***	0.009	0.029***	0.007	0.028***	0.007	0.029***	0.008	0.029***	0.007	0.028***	0.007	0.028***	0.007
	(0.012)	(0.009)	(0.011)	(0.008)	(0.010)	(0.008)	(0.011)	(0.008)	(0.010)	(0.008)	(0.010)	(0.008)	(0.010)	(0.008)
Investment$_{t-1}$	0.413**	0.577***	0.429**	0.590***	0.438**	0.592***	0.427**	0.588***	0.426**	0.587***	0.427**	0.587***	0.424**	0.586***
	(0.208)	(0.135)	(0.208)	(0.134)	(0.208)	(0.134)	(0.208)	(0.133)	(0.208)	(0.133)	(0.208)	(0.133)	(0.208)	(0.133)
Solvency Ratio$_{t-1}$	0.036***	0.036***	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***
	(0.007)	(0.005)	(0.008)	(0.005)	(0.007)	(0.005)	(0.008)	(0.005)	(0.008)	(0.005)	(0.008)	(0.005)	(0.008)	(0.005)
Current Ratio$_{t-1}$	-0.010***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***
	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
Company fe	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fe	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	732	732	748	748	748	748	748	748	748	748	748	748	748	748
R^2	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.98
F-Test	53.44	44.99	52.46	39.58	54.79	40.15	52.33	39.62	52.56	40.19	53.49	40.58	52.46	40.34

Table 5.9: The relation between digitalization and firm market-to-book value (alternative measure construction and reference documents)

This table presents the results of the panel regressions that examine the relation between digitalization based on a 45 topic distribution and firm value proxied by *market-to-book value* (MtB). For robustness, we consider digitalization measures that have been constructed in an alternative way or are based on a different reference document on digitalization: A measure based on only 98 % of the annual reports with the longest 2 % of the reports being excluded (columns 1 and 2), a measure based on the five most prominent topics (out of 45) in Bohnert et al. (2019) (columns 3 and 4), a fourth-root transformation of the original measure (columns 5 and 6), and measures based on the reference documents Bohnert et al. (2019) without the empirical section (columns 7 and 8), Cappiello (2020) (columns 9 and 10), Nicoletti (2016) (columns 11 and 12), and McKinsey (2017) (columns 13 and 14).

The panel has one observation for each company-year combination, and spans the time period 2007-2016. We include company fixed effects in all specifications. Even columns additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) and test statistics for the joint significance of regressors (F-test) are reported at the bottom of the table. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively.

Dependent	(1) MtB	(2) MtB	(3) MtB	(4) MtB	(5) MtB	(6) MtB	(7) MtB	(8) MtB	(9) MtB	(10) MtB	(11) MtB	(12) MtB	(13) MtB	(14) MtB
Digitalization$_{BNT98,t-1}$	0.487***	0.254*												
	(0.173)	(0.135)												
Digitalization$_{BNT5,t-1}$			0.422***	0.206*										
			(0.146)	(0.121)										
Digitalization$_{B4R,t-1}$					0.428***	0.197*								
					(0.133)	(0.108)								
Digitalization$_{BNTtxt,t-1}$							0.520***	0.256*						
							(0.178)	(0.142)						
Digitalization$_{CPLO,t-1}$									0.321***	0.144				
									(0.117)	(0.094)				
Digitalization$_{NCLT,t-1}$											0.426***	0.199*		
											(0.143)	(0.110)		
Digitalization$_{McK,t-1}$													0.308**	0.119
													(0.120)	(0.096)
Total Assets$_{t-1}$	0.921***	0.744**	0.902***	0.728***	0.896***	0.721***	0.905***	0.726***	0.906***	0.728***	0.905***	0.725***	0.908***	0.730***
	(0.261)	(0.290)	(0.255)	(0.278)	(0.254)	(0.277)	(0.255)	(0.278)	(0.255)	(0.278)	(0.255)	(0.278)	(0.256)	(0.279)
ROA$_{t-1}$	0.046***	0.015**	0.045***	0.014*	0.044***	0.014*	0.045***	0.014**	0.044***	0.014*	0.044***	0.014*	0.044***	0.014*
	(0.009)	(0.008)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)
Investment$_{t-1}$	-0.618***	-0.363*	-0.600***	-0.348*	-0.587***	-0.346*	-0.599***	-0.350*	-0.603***	-0.350*	-0.599***	-0.350*	-0.605***	-0.350*
	(0.205)	(0.205)	(0.202)	(0.201)	(0.200)	(0.201)	(0.201)	(0.202)	(0.202)	(0.202)	(0.202)	(0.202)	(0.202)	(0.202)
Solvency Ratio$_{t-1}$	-0.029**	-0.017*	-0.028**	-0.017*	-0.027**	-0.017*	-0.028**	-0.017*	-0.028**	-0.017*	-0.027**	-0.017*	-0.028**	-0.017*
	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)
Current Ratio$_{t-1}$	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Company fe	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fe	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	719	719	735	735	735	735	735	735	735	735	735	735	735	735
R^2	0.61	0.71	0.60	0.71	0.60	0.71	0.60	0.71	0.60	0.71	0.60	0.71	0.60	0.71
F-Test	7.81	4.19	8.24	3.96	8.33	3.89	8.21	3.94	8.12	3.90	8.20	3.97	8.08	3.88

significant. Furthermore, the results are in most cases robust to the inclusion of time fixed effects (even columns).

5.5 Conclusion

Although there is ample evidence on the fundamental impact digitalization will have on the business model of insurance companies, it is not yet clear whether it represents "only" a megatrend insurance companies need to follow or powerful means to increase their profitability. Unfortunately, the answer to that question is not straightforward since digitalization is difficult both to quantify and to differentiate from general innovation. In this study, we present a novel approach to tackle these two problems using unsupervised machine learning algorithms.

In detail, we exploit the prevalence of different topics in insurers' annual reports to construct a text-based measure of digitalization. By employing LDA, we determine the distribution over topics in each report and compare it to a reference document on digitalization based on the KL divergence. We then use this measure of similarity between the reports and the reference document to proxy for the extent of digitalization in an insurance company.

The digitalization proxy is then used as main explanatory variable to investigate the relation between digitalization and firm market valuation in a multivariate OLS model including firm and time fixed effects. Our results show that digitalization efforts are positively rewarded by market participants. An increase in digitalization is related to an increase in the market value of the insurers in our sample. Put differently, market participants expect efforts in digitalization to result in higher future profits. The estimation results are robust to different specifications of the LDA model employing other topic distributions or sentiment analysis. Furthermore, they neither depend on the particular calculation of the digitalization measure nor the choice of the digitalization reference document.

Of course, there are also limitations to our approach. A first limitation of this study

concerns the assumptions of the LDA model, most notably the bag of words assumption. There exist several extensions of the LDA model that address some of its shortcomings. However, the question which topic model to use when being confronted with a new set of texts and a new task is still an open direction for topic modeling (Blei, 2012). Additionally, while we are doing our best in constructing a measure to discriminate between digitalization and innovation, our approach has limitations in disentangling both concepts because they are closely related to each other. A third limitation concerns the question of causation between digitalization and firm market valuation. The estimated effect of our digitalization measure might be subject to endogeneity issues if larger or more profitable insurers implement digital systems quicker and more extensively than their competitors simply because of greater capacities. Whereas descriptive analyses of our sample do not provide clear evidence in favor of these channels, we refrain from interpreting our estimation results causally and instead consider them as profound confirmation of a positive relation between digitalization and firm market valuation. Establishing a causal link is an interesting avenue for future research.

Digitalization is a complex concept that cannot easily be captured empirically. However, with the rise of machine learning algorithms in the field of textual analysis and massive gains in computational power, the researcher is provided with a new set of powerful tools to analyze large amounts of textual data and retrieve the underlying thematic structure. In this sense, our approach can be considered a first step towards a new empirical analysis of the impact of digitalization not just in the insurance sector, but also in any other sector that will be disrupted by digitalization. Moreover, it could also be applied in various other settings where it is hard to retrieve empirical data due to the indefinite subject of research, e.g., corporate sustainability/social responsibility and its impact on corporate performance.

Appendix A

Supplementary Material for Chapter 2

A.1 Proofs

In this section we proof Theorems 1 and 2 from Section 2.3. Therefore, we first introduce some assumptions and establish auxiliary results needed to establish the theorems. For the proof of Theorem 1 we fix an arbitrary $p \geq 1$. For the proof of Theorem 2 however, we assume $p = 2$ as the theorem relies on assumptions about the convergence of the empirical grids obtained by the CLVQ algorithm.

The following assumptions are adopted from Charlier et al. (2015b):

Assumption A.3 *For the random variable (X, Y) on the probability space $(\Omega, \mathfrak{A}, P)$ we have $Y = f(X) + g(X) \cdot \epsilon$, where the d-dimensional random variable of covariates X is stochastically independent from the one-dimensional error term ϵ and $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$ and $g(\cdot) : \mathbb{R}^d \to \mathbb{R}_0^+$ are Lipschitz continuous functions. We further assume that $\|X\|_p < \infty$, $\|\epsilon\|_p < \infty$, and that the distribution P_X of X does not charge any hyperplane.*

Assumption A.4 *(i) The support S_X of P_X is compact. (ii) There exists a continuous density $f_\epsilon : \mathbb{R} \to \mathbb{R}_0^+$ of the distribution P_ϵ with respect to the Lebesgue measure on $\mathbb{R}$.*

The following lemma gives an estimate of the quantization error of the grids introduced in Section 2.3.2. As these grids are not necessarily L_p-optimal, the lemma and the derived corollary are needed in order to proof Lemma A.7.

Lemma A.5 *Fix $\alpha, \lambda, \gamma \in (0,1)$, and $N \in \mathbb{N}$ with $\lceil N \cdot \gamma \rceil < N$. Let further $\Gamma_N \in \{\Gamma_N^{(0)}, \ldots, \Gamma_N^{(\kappa_N)}\}$ be one of the grids obtained in the κ_N iteration steps of the algorithm proposed in Section 2.3.2. Then the following holds:*

$$\|Proj_{\Gamma_N}(X) - X\|_p \leq \|Proj_{\Gamma^*_{N'}}(X) - X\|_p,$$

where $N' := \min\{\lceil N \cdot \gamma \rceil, N - \lceil N \cdot \gamma \rceil\}$ and $\Gamma^*_{N'}$ denotes an L^p-optimal N'-grid for X.

Proof. If $\Gamma_N = \Gamma_N^{(0)}$ then by construction Γ_N is an L^p-optimal N-grid for X and the lemma follows from the fact that $N > N'$. Hereafter, we will therefore assume $\Gamma_N \in \{\Gamma_N^{(1)} \ldots \Gamma_N^{(\kappa_N)}\}$.

By construction of the grids in Section 2.3.2 we have that $\Gamma_N = \Gamma_A \cup \Gamma_{\Omega \setminus A}$, where A is a measurable set and Γ_A is an L^p-optimal $\lceil N \cdot \gamma \rceil$-grid for the restriction $X_{|A}$ of X to A and $\Gamma_{\Omega \setminus A}$ is an L^p-optimal $N - \lceil N \cdot \gamma \rceil$-grid for $X_{|\Omega \setminus A}$. It follows:

$$\begin{aligned}
\left\|Proj_{\Gamma_N}(X) - X\right\|_p^p &= \int_A \left|Proj_{\Gamma_A \cup \Gamma_{\Omega \setminus A}}(X) - X\right|^p dP + \int_{\Omega \setminus A} \left|Proj_{\Gamma_A \cup \Gamma_{\Omega \setminus A}}(X) - X\right|^p dP \\
&\leq \int_A \left|Proj_{\Gamma_A}(X) - X\right|^p dP + \int_{\Omega \setminus A} \left|Proj_{\Gamma_{\Omega \setminus A}}(X) - X\right|^p dP \\
&\leq \int_A \left|Proj_{\Gamma^*_{N'}}(X) - X\right|^p dP + \int_{\Omega \setminus A} \left|Proj_{\Gamma^*_{N'}}(X) - X\right|^p dP \\
&= \left\|Proj_{\Gamma^*_{N'}}(X) - X\right\|_p^p.
\end{aligned}$$

■

Corollary A.6 *With the notation and assumptions of Lemma A.5, the following holds:*

$$\sup_{0 \leq k \leq \kappa_N} \left\|Proj^{(k)}_{\Gamma_N}(X) - X\right\|_p \to 0 \text{ as } N \to \infty.$$

Proof. Let $(\xi_j)_{j \in \mathbb{N}}$ denote an everywhere dense sequence in $\mathbb{R}^d$ and set $\Gamma'_N := \{\xi_1, \ldots, \xi_N\}$ for $N \in \mathbb{N}$. It follows from Lebesgue's dominated convergence theorem that

$$\|Proj_{\Gamma'_N}(X) - X\|_p \to 0 \text{ as } N \to \infty. \tag{A.1}$$

As Γ_N^* is an L_p-optimal N-grid for X we have $\|Proj_{\Gamma_N^*}(X) - X\|_p \leq \|Proj_{\Gamma_N'}(X) - X\|_p$. Because of $\gamma \in (0,1)$, as N goes to ∞ so does $N' = \min\{\lceil N \cdot \gamma \rceil, N - \lceil N \cdot \gamma \rceil\}$. The corollary follows then directly from (A.1) and Lemma A.5. ■

The following lemma is adopted from Charlier et al. (2015b) with only slight changes:

Lemma A.7 *Fix $\alpha \in (0,1)$ and $x \in S_X$. Let further $(\Gamma_N)_{N\in\mathbb{N}}$ denote a sequence of N-grids such that*

$$\left\|Proj_{\Gamma_N}(X) - X\right\|_p \to 0 \text{ as } N \to \infty.$$

For any $N \in \mathbb{N}$ let $\tilde{x}_N := Proj_{\Gamma_N}(x)$, $\tilde{X}_N := Proj_{\Gamma_N}(X)$, and $C_{x,N} := \{z \in S_X : Proj_{\Gamma_N}(z) = \tilde{x}_N\}$. Define $G_a(x) := E(\rho_\alpha(Y-a)|X = x)$ and the corresponding quantity $\tilde{G}_a(\tilde{x}_N) := E(\rho_\alpha(Y-a)|\tilde{X}_N = \tilde{x}_N)$. Then under Assumptions A.3 and A.4, the following holds:

(i) $\sup_{x\in S_X} |x - \tilde{x}_N| \to 0$ *as* $N \to \infty$,

(ii) $\sup_{x\in S_X} R(C_{x,N}) \to 0$ *as* $N \to \infty$, *where* $R(C_{x,N})$ *is given as* $\sup_{z\in C_{x,N}} |z - \tilde{x}_N|$,

(iii) $\sup_{x\in S_X} \sup_{a\in\mathbb{R}} |\tilde{G}_a(\tilde{x}_N) - G_a(x)| \to 0$ *as* $N \to \infty$,

(iv) $\sup_{x\in S_X} |\min_{a\in\mathbb{R}} \tilde{G}_a(\tilde{x}_N) - \min_{a\in\mathbb{R}} G_a(x)| \to 0$ *as* $N \to \infty$.

Proof. (i) follows from the fact that by assumption $(\Gamma_N)_{N\in\mathbb{N}}$ is a sequence of N-grids such that $\|Proj_{\Gamma_N}(X) - X\|_p \to 0$ as $N \to \infty$, see the proof of Lemma A.2 in Charlier et al. (2015b). The proofs of statements (ii) - (iv) can be adopted unchanged from ibidem. ■

We are now able to proof Theorem 1.

Proof of Theorem 1. Choose an arbitrary $\epsilon > 0$. By employing the iterative procedure

introduced in Section 2.3.2, we obtain for a sufficiently large[152] $N \in \mathbb{N}$ the sequence

$$\begin{aligned} &\Gamma_N^{(0)}, \Gamma_N^{(1)}, \Gamma_N^{(2)}, \ldots, \Gamma_N^{(\kappa_N)}, \\ &\Gamma_{N+1}^{(0)}, \Gamma_{N+1}^{(1)}, \Gamma_{N+1}^{(2)}, \ldots, \Gamma_{N+1}^{(\kappa_{N+1})}, \\ &\ldots \end{aligned} \tag{A.2}$$

of grids as well as the sequence of corresponding base estimators

$$\begin{aligned} &\bar{q}_{\alpha,N}^{(0)} := \tilde{q}_{\alpha,N}^{(0)},\ \bar{q}_{\alpha,N}^{(1)}, \bar{q}_{\alpha,N}^{(2)}, \ldots, \bar{q}_{\alpha,N}^{(\kappa_N)}, \\ &\bar{q}_{\alpha,N+1}^{(0)} := \tilde{q}_{\alpha,N+1}^{(0)},\ \bar{q}_{\alpha,N+1}^{(1)}, \bar{q}_{\alpha,N+1}^{(2)}, \ldots, \bar{q}_{\alpha,N+1}^{(\kappa_{N+1})}, \\ &\ldots . \end{aligned} \tag{A.3}$$

As by Corollary A.6 the sequence of grids in (A.2) meets the assumptions of Lemma A.7, it follows from Charlier et al. (2015b)[153] that Theorem 1 holds for the sequence of estimators from (A.3) instead of the sequence $(\tilde{q}_{\alpha,N})_{N\in\mathbb{N}}$. Consequently, there is an $M \in \mathbb{N}$ such that for all $N > M$ and $k \in \{0, 1, \ldots, \kappa_N\}$

$$\sup_{x \in S_X} \left| \bar{q}_{\alpha,N}^{(k)}(x) - q_\alpha(x) \right| < \epsilon \tag{A.4}$$

is fulfilled. Now, we choose an arbitrary $N > M$. By construction $\tilde{q}_\alpha = \tilde{q}_{\alpha,N}$ is given as a convex combination of the base estimators $\bar{q}_{\alpha,N}^{(0)}, \bar{q}_{\alpha,N}^{(1)}, \ldots, \bar{q}_{\alpha,N}^{(\kappa_N)}$, say

$$\tilde{q}_\alpha = \sum_{j=0}^{\kappa_N} c_j \bar{q}_{\alpha,N}^{(j)},$$

where $c_j \geq 0$, $j = 0, \ldots, \kappa_N$ and $\sum_{j=0}^{\kappa_N} c_j = 1$. From Inequality (A.4) and the choice of

[152] In order that all the mathematical expressions in the iterative procedure in Section 2.3.2 are defined, $\lceil N \cdot \gamma \rceil < N$ is required.

[153] See the proof of Theorem 3.2 in conjunction with Lemma A.7 of that paper.

$N > M$ follows

$$\begin{aligned}
\sup_{x \in S_X} \left| \tilde{q}_{\alpha,N}(x) - q_\alpha(x) \right| &= \sup_{x \in S_X} \Big| \sum_{j=0}^{\kappa_N} c_j (\bar{q}^{(j)}_{\alpha,N}(x) - q_\alpha(x)) \Big| \\
&\leq \sup_{x \in S_X} \sum_{j=0}^{\kappa_N} c_j \left| \bar{q}^{(j)}_{\alpha,N}(x) - q_\alpha(x) \right| \\
&\leq \sum_{j=0}^{\kappa_N} c_j \sup_{x \in S_X} \left| \bar{q}^{(j)}_{\alpha,N}(x - q_\alpha(x)) \right| \\
&< \epsilon.
\end{aligned}$$

This completes the proof of Theorem 1. ■

In order to prove Theorem 2 we need the following two additional assumptions:

Assumption A.8 *The distribution P_X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.*

Assumption A.9 *For $N \in \mathbb{N}$, $n >> N$, and $\lambda, \gamma \in (0,1)$ fixed, we assume that the number of iteration steps in the finite sample case equals almost surely the number of iteration steps κ_N in the full sample case and that the empirical quantization of X almost surely converges to the population one at each iteration step $0 \leq j \leq \kappa_N$, that is*

$$Proj_{\hat{\Gamma}^{(j)}_{N,n}}(X) \to Proj_{\Gamma^{(j)}_N}(X) \; a.s. \; as \; n \to \infty.$$

The assumption concerning the number of iteration steps is unproblematic as Theorems 1 and 2 are independent of the number of iterations. Consequently, one could simply fix the number ex-ante. The strong assumption on the convergence of the involved grids at each iteration step, however, is necessary to proof Lemma A.10. Convergence results for the CLVQ algorithm justifying the assumption that the empirical quantization of X almost surely converges to the population one can be found in Pagès (1998, Theorem 27 et seq.) along with a discussion of those results in Charlier et al. (2015a).

We can now proof the following lemma:

Lemma A.10 *Fix $\alpha, \lambda, \gamma \in (0,1)$, and $x \in S_X$. Given that the grids are obtained in the quadratic case ($p = 2$), we have under Assumptions A.3, A.4 (i), A.8, and A.9:*

$$p - \lim_{N\to\infty} p - \lim_{n\to\infty} \left|\hat{q}_{\alpha,N,n}(x) - \tilde{q}_{\alpha,N}(x)\right| = 0, \tag{A.5}$$

where $\tilde{q}_{\alpha,N}$ and $\hat{q}_{\alpha,N,n}$ denote the estimators introduced in Sections 2.3.2 and 2.3.3, respectively.

Proof of Lemma A.10. For $N \in \mathbb{N}$ and $n >> N$ there exists by construction a representation

$$\hat{q}_{\alpha,N,n}(x) = \sum_{j=0}^{\kappa_N} c_{j,N,n} \bar{q}^{(j)}_{n,N}(x),$$

where the $c_{j,N,n}$ are random variables satisfying $c_{j,N,n} \geq 0$ and $\sum_{j=0}^{\kappa_N} c_{j,N,n} = \mathbb{I}$, that is, $\hat{q}_{\alpha,N,n}$ is a (pointwise) convex combination of the estimators obtained in each iteration step, see Section 2.3.3. Analogously (see Section 2.3.2) we obtain $\tilde{q}_{\alpha,N}(x)$ as a (pointwise) convex combination, too, say

$$\tilde{q}_{\alpha,N}(x) = \sum_{j=0}^{\kappa_N} c'_{j,N} \bar{q}'^{(j)}_{N}(x).$$

It follows:

$$\begin{aligned} 0 &\leq \Big| \sum_{j=0}^{\kappa_N} c_{j,N,n} \bar{q}^{(j)}_{n,N}(x) - \sum_{j=0}^{\kappa_N} c'_{j,N} \bar{q}'^{(j)}_{N}(x) \Big| \\ &\leq \underbrace{\Big| \sum_{j=0}^{\kappa_N} c_{j,N,n} (\bar{q}^{(j)}_{n,N}(x) - \bar{q}'^{(j)}_{N}(x)) \Big|}_{=:T_1} + \underbrace{\Big| \sum_{j=0}^{\kappa_N} (c_{j,N,n} - c'_{j,N}) \bar{q}'^{(j)}_{N}(x) \Big|}_{=:T_2}. \end{aligned} \tag{A.6}$$

It follows from Charlier et al. (2015b)[154] that under Assumptions A.3, A.4 (i), A.8, and A.9 we have

$$p - \lim_{n\to\infty} T_1 = 0. \tag{A.7}$$

[154] See the proof of Theorem 4.1 therein.

Because $\sum_{j=0}^{\kappa_N}(c_{j,N,n} - c'_{j,N}) = 0$, we have

$$\begin{aligned} T_2 &= \Big| \sum_{j=0}^{\kappa_N} (c_{j,N,n} - c'_{j,N})(\bar{q}_N'^{(j)}(x) - \bar{q}_N'^{(0)}(x)) \Big| \\ &\leq \sum_{j=0}^{\kappa_N} \left| c_{j,N,n} - c'_{j,N} \right| \max_{j \in \{1,\dots,\kappa_N\}} \left| \bar{q}_N'^{(j)}(x) - \bar{q}_N'^{(0)}(x) \right| \\ &\leq 2 \cdot \max_{j \in \{1,\dots,\kappa_N\}} \left| \bar{q}_N'^{(j)}(x) - \bar{q}_N'^{(0)}(x) \right|. \end{aligned} \tag{A.8}$$

It follows from the proof of Theorem 1 that $p - \lim_{N\to\infty} \max_{j\in\{1,\dots,\kappa_N\}} \left| \bar{q}_N'^{(j)}(x) - \bar{q}_N'^{(0)}(x) \right| = 0$. Because of Inequality (A.8) we therefore have

$$p - \lim_{N\to\infty} p - \lim_{n\to\infty} T_2 = 0.$$

By combining this result with (A.6) and (A.7) we obtain

$$0 \leq p - \lim_{N\to\infty} p - \lim_{n\to\infty} \Big| \sum_{j=0}^{\kappa_N} c_{j,N,n} \bar{q}_{n,N}^{(j)}(x) - \sum_{j=0}^{\kappa_N} c'_{j,N} \bar{q}_N'^{(j)}(x) \Big| \leq p - \lim_{N\to\infty} p - \lim_{n\to\infty} \; T_1 + T_2 = 0.$$

This completes the proof. ■

We can now proof Theorem 2.

Proof of Theorem 2. We have by the triangle inequality

$$0 \leq |\hat{q}_{\alpha,N,n}(x) - q_\alpha(x)| \leq |\hat{q}_{\alpha,N,n}(x) - \tilde{q}_{\alpha,N}(x)| + |\tilde{q}_{\alpha,N}(x) - q_\alpha(x)|.$$

As the convergence in Theorem 1 in particular implies convergence in probability, we have

$$p - \lim_{N\to\infty} p - \lim_{n\to\infty} |\tilde{q}_{\alpha,N}(x) - q_\alpha(x)| = p - \lim_{N\to\infty} |\tilde{q}_{\alpha,N}(x) - q_\alpha(x)| = 0.$$

The theorem now follows from Lemma A.10. ■

A.2 Additional figures and tables

Table A.1: Hyperparameters for the leveraging estimator chosen by 5-fold cross-validation

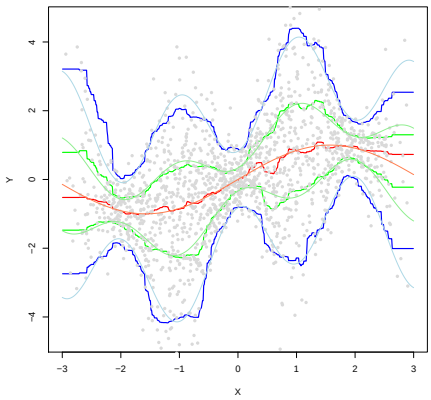
This table reports the number of quantizers N as well as the ratio γ/λ chosen via 5-fold cross-validation. The values are averaged over 100 random samples in the one-dimensional and 50 random samples in the multi-dimensional case. The results are reported for random samples of sizes 500 and 1500 for the univariate models $\mathcal{M}_1, \mathcal{M}_2$, and $\mathcal{M}_3$ and for random samples of size 5000 for the multivariate model $\mathcal{M}_1'$. Instead of reporting the parameters λ and γ separately, we include the ratio γ/λ. The ratio describes how much the "weight" for data examples associated with a high approximation error is increased in the next iteration step, see Section 2.3.3 for details.

	dim = 1					
	model $\mathcal{M}_1$		model $\mathcal{M}_2$		model $\mathcal{M}_3$	
	$n = 500$	$n = 1500$	$n = 500$	$n = 1500$	$n = 500$	$n = 1500$
N	12.28	16.74	10.70	17.32	21.66	33.88
γ/λ	1.20	1.20	1.10	1.03	1.18	1.15

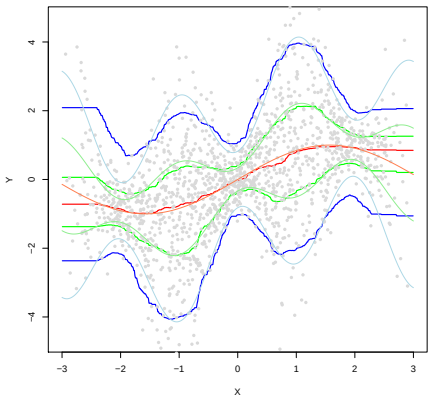
	dim = 2	**dim = 3**	**dim = 4**
	model $\mathcal{M}_1'$	model $\mathcal{M}_1'$	model $\mathcal{M}_1'$
	$n = 5000$	$n = 5000$	$n = 5000$
N	147.00	367.00	564.00
γ/λ	1.35	1.56	1.60

Figure A.1: Estimated conditional quantile curves for model $\mathcal{M}_2$ and $n = 1500$

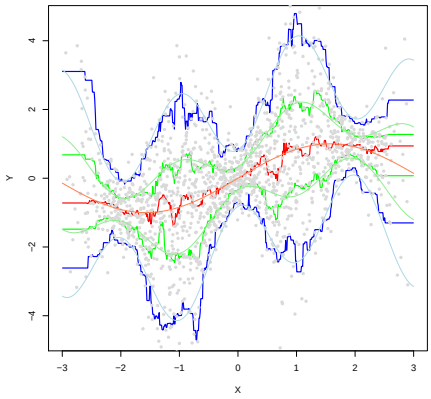
This figure presents the same plots as Figure 2.5 but for a random sample of size $n = 1500$ generated according to model $\mathcal{M}_2$.



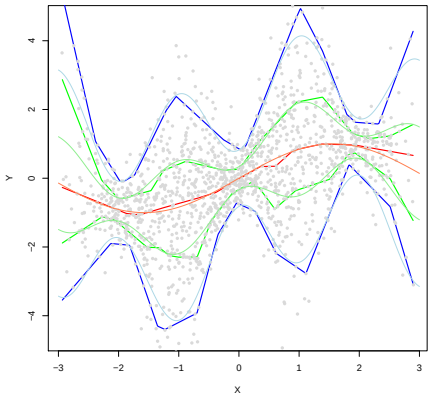
(a) leveraging estimator



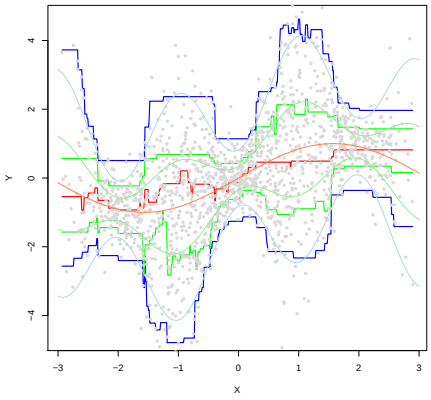
(b) quantization estimator



(c) kNN estimator



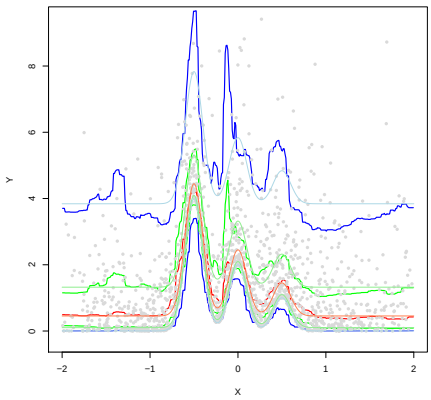
(d) smoothing splines estimator



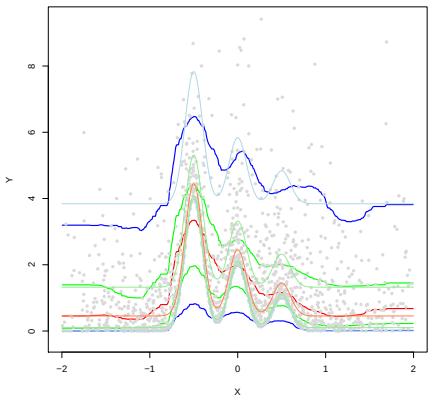
(e) xgboost estimator

Figure A.2: Estimated conditional quantile curves for model $\mathcal{M}_3$ and $n = 1500$

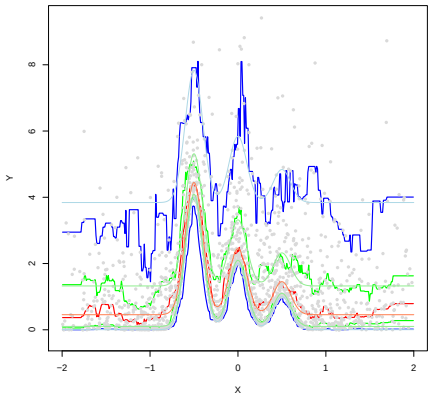
This figure presents the same plots as Figure 2.5 but for a random sample of size $n = 1500$ generated according to model $\mathcal{M}_3$.



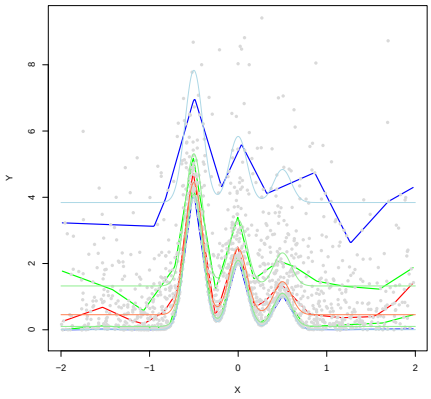
(a) leveraging estimator



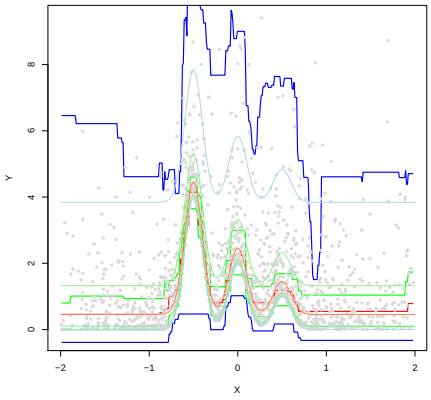
(b) quantization estimator



(c) kNN estimator



(d) smoothing splines estimator



(e) xgboost estimator

Table A.2: Parameters associated with the quantile plots

The parameters in the table correspond to Figures 2.5, 2.6, 2.7, 2.8, A.1, and A.2. For the leveraging estimator the parameters are presented in the order N, λ, γ. For the quantization estimator we report the number of quantizers, for the kNN estimator the number of neighbors, for the smoothing splines estimator the values of the smoothing parameter λ (for $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$), and for the xgboost estimator the parameter gamma. For details on the parameter selection procedures see Sections 2.4.1 and 2.5.1.

	model $\mathcal{M}_1$		**model $\mathcal{M}_2$**		**model $\mathcal{M}_3$**	
	$n = 500$	$n = 1500$	$n = 500$	$n = 1500$	$n = 500$	$n = 1500$
leveraging	10, 0.35, 0.5	16, 0.3, 0.5	14, 0.3, 0.4	20, 0.5, 0.5	22, 0.35, 0.5	24, 0.3, 0.5
quantization	18	40	6	10	8	10
kNN	28	62	46	66	32	58
smoothing splines	0.3, 0.4, 0.8, 1, 0.3	0.3, 0.3, 0.3, 0.4, 0.3	0.3, 0.3, 0.5, 0.8, 0.3	0.3, 0.3, 0.5, 0.3, 0.3	0.3, 0.4, 0.3, 0.3, 0.3	0.3, 0.4, 0.3, 0.3, 0.3
xgboost	2	3	1.5	2.5	1	2

Appendix B

Supplementary Material for Chapter 4

B.1 Theoretical foundations

B.1.1 ARMA-GARCH process

In an ARMA(p,q)-GARCH(r,s) process, the conditional mean of a (univariate) time series is modeled by the ARMA part while the conditional volatility is captured by the GARCH part. With ARMA(p,q) we denote a model with p autoregressive and q moving average terms. More formally, we have the following specification

$$r_t = \mu + \sum_{j=1}^{p} \phi_j r_{t-j} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t,$$

where $r_\tau, \tau = t - p, \ldots, t$ are observations from the time series and $\phi_j, j = 1, \ldots, p$, $\theta_j, j = 1, \ldots, q$, and μ denote parameters. The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model by Bollerslev (1986) extends the ARCH model due to Engle (1982) by including lags of the conditional variances. More exactly, the variance equation according to the GARCH(r,s) model at time t is given by

$$\sigma_t^2 = \omega + \sum_{j=1}^{r} \beta_j \sigma_{t-j}^2 + \sum_{j=1}^{s} \alpha_j \epsilon_{t-j}^2,$$

where $\sigma_\tau^2, \tau = t - r, \ldots, t$ denotes the conditional variance, $\alpha_j, j = 1, \ldots, s$, $\beta_j, j = 1, \ldots, r$, and ω are parameters and all ϵ_t are of the form $\epsilon_t = z_t\sigma_t$ where z_t is an iid process with zero mean and unit variance. The parameters must fulfill some conditions in order to guarantee that the GARCH conditional variance estimates are always positive, see Nelson and Cao (1992) for details. In case of the GARCH(1,1) model, forecasts can be calculated as

$$\sigma_{t+h|t}^2 = \sigma^2 + (\alpha + \beta)^{h-1}(\sigma_{t+1}^2 - \sigma^2),$$

where $h > 2$ denotes the horizon of the forecasts and σ^2 denotes the unconditional variance given by $\sigma^2 = \frac{\omega}{1-\alpha-\beta}$ (see Bollerslev, 2010).[155]

B.1.2 Sklar's theorem

The popularity of copulas in multivariate dependence modeling is due to the theorem by Sklar (1959). Loosely speaking the theorem states that modeling of the marginals and of the multivariate dependence can be separated by means of copulas functions.

Theorem B.11 (Sklar's theorem) *Let* $\mathbf{X} = (X_1, \ldots, X_d) \sim F$ *be a d-dimensional random variable with marginal distributions* $F_j, j = 1, \ldots, d$. *We then have*

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)),$$

where C denotes some appropriate d-dimensional copula. If the multivariate distribution function F is absolutely continuous and the marginal distributions $F_1, \ldots, F_d$ *are strictly increasing continuous, we have*

$$f(x_1, \ldots, x_d) = \Big(\prod_{j=1}^{d} f_j(x_j) \Big) \cdot c(F_1(x_1), \ldots, F_d(x_d))$$

with the small letters denoting the corresponding probability density functions.

[155] Conditional variance estimates for the GARCH(1,1) model are positive almost surely given that $\omega > 0, \alpha \geq 0$ and $\beta \geq 0$. The model is covariance stationary provided that $\alpha + \beta < 1$.

The second part of the theorem highlights that the joint distribution of the random variable $\mathbf{X}$ can be modeled separately in terms of a "marginal term" $\prod_{j=1}^{d} f_j(x_j)$ and a "dependence term" $c(F_1(x_1), \ldots, F_d(x_d))$. The marginal term is based on information from the (univariate) marginals alone and does not contain any information about the multivariate dependence. On the other hand, the random variables $F_1(X_1), \ldots, F_d(X_d)$ are all uniformly distributed in the interval $[0, 1]$ and therefore do not contain any information on the marginals. For more details we refer to the books by Joe (2001) and Nelsen (2006).

B.1.3 Duration-based backtest

The duration-based VaR backtest by Christoffersen (2004), as the name implies, is based on the duration of days between VaR violations. The hit sequence of VaR_t violations is defined as

$$I_t = \begin{cases} 1, & \text{if } r_t < -VaR_t(p) \\ 0, & \text{else} \end{cases}$$

where r_t is a time series of daily ex-post portfolio returns and $VaR_t(p)$ a time series of ex-ante VaR forecasts with a coverage rate p. The time in days between two VaR violations is called the no-hit duration $D_i = t_i - t_{i-1}$ where t_i denotes the day of violation number i. The null hypothesis of the test is then: If the VaR model is correctly specified for coverage rate p, the no-hit duration or in other words the conditional expected duration between VaR violations should have no memory and a mean duration of $1/p$ days. Thus, under the null hypothesis the no-hit duration follows the exponential distribution

$$f_{exp}(D; p) = p\ exp(-pD)$$

whereas the alternative that allows for duration dependence follows a Weibull distribution

$$f_W(D; a, b) = a^b b D^{b-1} exp(-(aD)^b).$$

The tested null hypothesis of independence is then defined as

$$H_{0,ind} : b = 1.$$

For a detailed derivation of the no-memory property in terms of the discrete probability distribution and its hazard function, please see Christoffersen (2004).

B.1.4 Conditional calibration backtest

Nolde and Ziegel (2017) introduce a conditional calibration (CC) test and show that well-known traditional backtests can be unified within the concept of CC. The CC test comes in two versions: The simple version used in our analysis requires only risk forecasts (VaR and ES), whereas the general version additionally needs information on conditional volatility. Following Nolde and Ziegel (2017), $\mathcal{P}_0$ defines the class of Borel-probability distributions on $\mathbb{R}$. $\mathcal{P}_1 \subseteq \mathcal{P}_0$ denotes the class of all distributions with finite mean whereas $\mathcal{P}_V \subset \mathcal{P}_0$ describes distributions with unique quantiles. The chosen identification function for the pair (VaR_ν, ES_ν) for the level $\nu \in (0, 1)$ is

$$V(x_1, x_2, r) = \left(\frac{1 - \nu - \mathbb{I}_{(0,\infty)}(r - x_1)}{x_1 - x_2 - \frac{1}{1-\nu}\mathbb{I}_{(0,\infty)}(r - x_1)(x_1 - r)} \right)$$

with respect to $\mathcal{P}_1 \cap \mathcal{P}_V$, where $\mathbb{I}_{(0,\infty)}$ denotes the characteristic function of the open interval $(0, \infty)$. We further follow the notation of Nolde and Ziegel (2017) and define $\Theta = (\rho_1, ..., \rho_k)$ as the identifiable functional with identification function V with respect to $\mathcal{P}$. A series of negated log-returns defined as $\{r_t\}_{t\in\mathbb{N}}$ is adapted to the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$. Let $\{x_t\}_{t\in\mathbb{N}}$ be a sequence of predictions of Θ that are $\mathcal{F}_{t-1}$-measurable. All conditional distributions $\mathcal{L}(r_t|\mathcal{F}_{t-1})$ and all unconditional distributions $\mathcal{L}(r_t)$ are assumed to belong to $\mathcal{P}$ almost surely. The sequence of predictions $\{x_t\}_{t\in\mathbb{N}}$ is conditionally calibrated for Θ if

$$\mathbb{E}(V(x_t, r_t)|\mathcal{F}_{t-1}) = 0 \quad \text{almost surely}, \forall t \in \mathbb{N}.$$

The null hypothesis of the traditional backtest for CC considers this requirement: The sequence of predictions $\{x_t\}_{t\in\mathbb{N}}$ is conditionally calibrated for Θ. The requirement for the expected value is equivalent to $\mathbb{E}(h_t' V(x_t, r_t)) = 0$ for all $\mathcal{F}_{t-1}$-measurable $\mathbb{R}^k$-valued functions h_t. Nolde and Ziegel (2017) consider a $\mathcal{F}$-predictable sequence $\{\mathbf{h}_t\}_{t\in\mathbb{N}}$ of $q \times k$-matrices $\mathbf{h}_t$ called test functions to construct a Wald-type test statistic. For the simple version of the CC test, $\mathbf{h}_t$ equals the identity matrix. For more information on the Wald-type test statistic as well as a complete derivation of the CC test in both versions, please refer to the original paper.

B.1.5 Model confidence set procedure

In this section we more formally introduce the MCS procedure outlined in Section 4.2.4. Let therefore $\mathcal{M}^0$ denote the set of candidate models, $\mathcal{M}^*$ the true set of best models, and $\hat{\mathcal{M}}^*_{1-\alpha}$ the model confidence set at confidence level α. We further assume that $\mathcal{M}^0$ consists of a finite number m_0 of models. Based on an evaluation criterion (the *loss function*) one can calculate the losses $L_{i,t}$ that are associated with model i at time t. In the case of a VaR forecast, e.g., one might compare the risk forecast $VaR^{i,t}_\alpha$ of model i at time t with the actual realized return r_t by setting $L_{i,t} := L(VaR^{i,t}_\alpha, r_t)$ where L denotes an appropriate loss function. For $i, j \in \mathcal{M}^0$ one then defines the relative performance variable

$$d_{ij,t} := L_{i,t} - L_{j,t}$$

as well as the expected value of the performance variable

$$\mu_{ij} := E(d_{ij,t}).$$

The alternatives in $\mathcal{M}^0$ are now ranked based on their expected loss. That is, model i is preferred over model j if $\mu_{ij} < 0$. The MCS is now constructed based on a sequence of significance tests

$$H_{0,\mathcal{M}} : \mu_{ij} = 0 \quad \text{for all } i, j \in \mathcal{M},$$

with $\mathcal{M} \subseteq \mathcal{M}^0$. If $H_{0,\mathcal{M}}$ can be rejected, the elimination rule is applied to remove a model from $\mathcal{M}$ that is inferior to the remaining ones. The model confidence set is then defined as any subset of $\mathcal{M}^0$ that contains all best models with a given probability $1-\alpha$. More shortly, the MCS procedure can be summarized in the following algorithmic form (see Hansen et al., 2011):

Step 0: Set $\mathcal{M} := \mathcal{M}^0$.

Step 1: Test the null hypothesis $H_{0,\mathcal{M}}$ based on the equivalence test $\delta_{\mathcal{M}}$ at the confidence level α.

Step 2: If $H_{0,\mathcal{M}}$ is not rejected, set $\hat{\mathcal{M}}^*_{1-\alpha} := \mathcal{M}$, otherwise use the elimination rule $e_{\mathcal{M}}$ to eliminate a model from $\mathcal{M}$ and repeat the procedure from step 1.

Hansen et al. (2011) provide two t-statistics for the hypothesis test in step 1. We opt for the statistic t_{ij} that is also used in the test for comparing two forecasts (see Diebold and Mariano, 1995, West, 1996). We therefore define the sample loss statistic $\bar{d}_{ij} := \frac{1}{n}\sum_{t=1}^{n} d_{ij,t}$ and set

$$t_{ij} := \frac{\bar{d}_{ij}}{\sqrt{\hat{var}(\bar{d}_{ij})}},$$

where $\hat{var}(\bar{d}_{ij})$ denotes an estimate of $var(\bar{d}_{ij})$. The final test statistic is then defined as

$$T_{R,\mathcal{M}} := \max_{i,j\in\mathcal{M}} |t_{ij}|.$$

The asymptotic distribution of $T_{R,\mathcal{M}}$ is non-standard and derived via a bootstrapping scheme, see Hansen et al. (2011) for details. The natural elimination rule corresponding to the test statistic $T_{R,\mathcal{M}}$ is $e_{R,\mathcal{M}} := \arg\max_{i\in\mathcal{M}} \sup_{j\in\mathcal{M}} t_{ij}$ because the corresponding model is such that $t_{e_{R,\mathcal{M}}j} = T_{R,\mathcal{M}}$ is fulfilled for some $j \in \mathcal{M}$. Removing model $e_{R,\mathcal{M}}$ will therefore reduce (or at least not increase) the test statistic $T_{R,\mathcal{M}}$.

B.2 Figures

Figure B.1: Potential portfolio value under financial distress according to the 97.5% ES

This figure illustrates the economic significance of model risk arising from the disparity between different ES models. Here, we focus on the 97.5% ES for a well diversified portfolio ($100,000) and a 10 day holding period between November 4, 2004 and December 31, 2018. We provide the portfolio value minus the 5th and the 95th percentile of ES forecasts from all multivariate models that passed the conditional calibration backtest by Nolde and Ziegel (2017) on a daily basis. This corresponds to the potential portfolio value under financial distress according to the more (95th percentile) or less (5th percentile) conservative ES models.

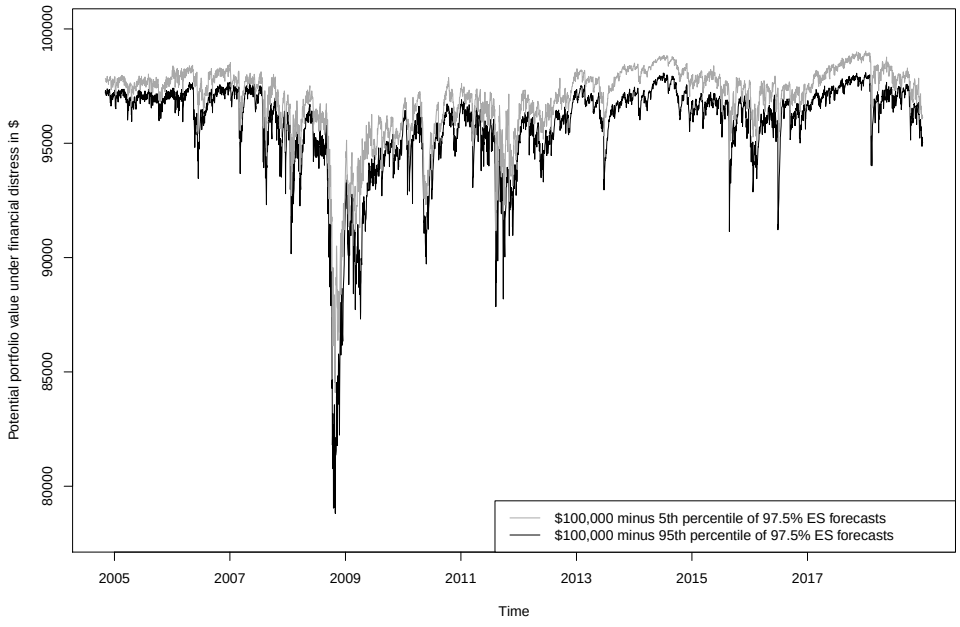
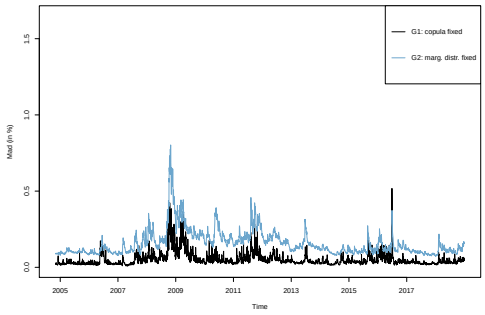
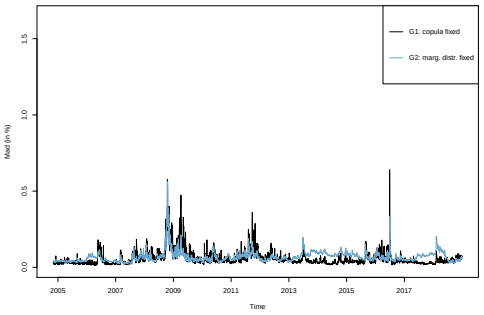


Figure B.2: Average model risk for alternative model risk measures (model sets with fixed and varying copula only)

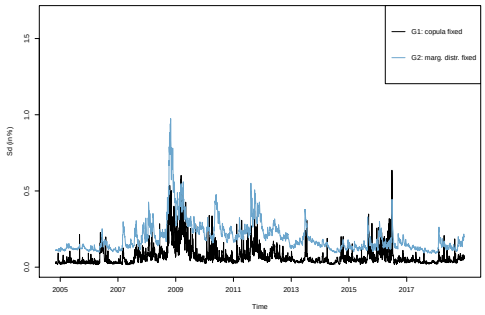
This figure shows the average model risk associated with one day ahead 99% VaR (Subfigures 1, 3, and 5) and 97.5% ES (Subfigures 2, 4, and 6) forecasts for a well diversified portfolio per group. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. Model risk is captured by different measures of one day ahead forecasts by various risk models within a model set. Our baseline measure is the mean absolute deviation (mad). We additionally include the standard deviation (sd) and interquartile range (iqr), see Section 4.2.3 for more details. Values are calculated on a daily basis between November 4, 2004 until December 31, 2018 in percent of the portfolio value based on all models that passed the respective backtest, see Section 4.2.2 for details.



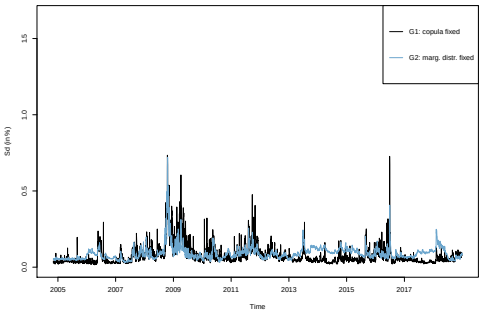
(a) Average model risk (99% VaR) - mad



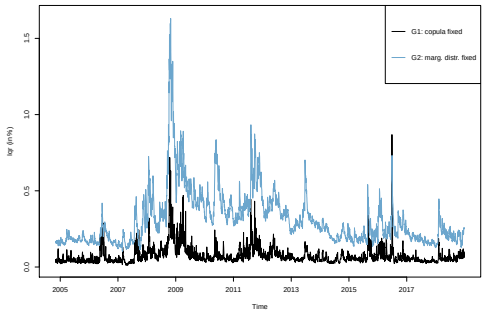
(b) Average model risk (97.5% ES) - mad



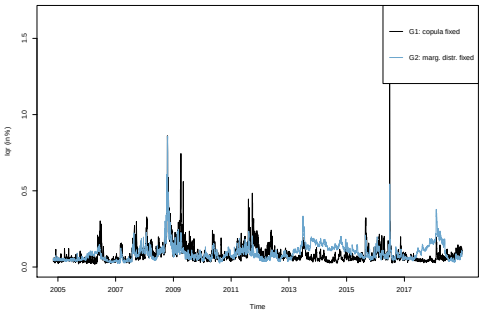
(c) Average model risk (99% VaR) - sd



(d) Average model risk (97.5% ES) - sd



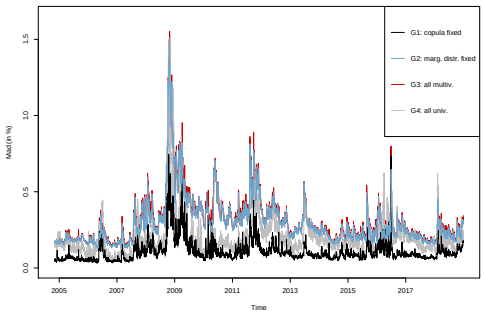
(e) Average model risk (99% VaR) - iqr



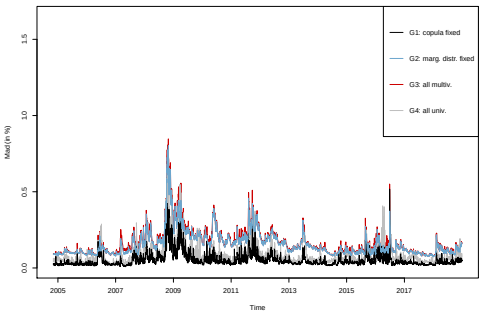
(f) Average model risk (97.5% ES) - iqr

Figure B.3: Average model risk for all groups under various VaR confidence levels

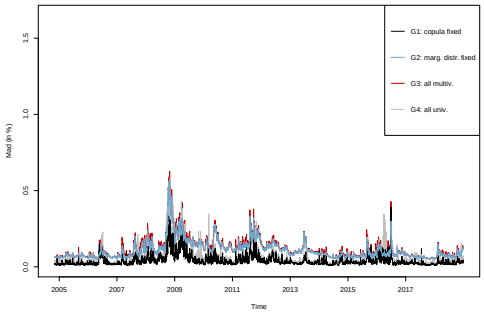
This figure shows the average model risk associated with one day ahead VaR forecasts for a well diversified portfolio and a confidence level of 99.9%, 99%, 97.5%, and 95% (Subfigures 1-4) per group. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. *Group 3* (G3) consists of all multivariate and *Group 4* (G4) of all univariate models. Model risk is measured in terms of the mean absolute deviation (mad) of one day ahead forecasts by various risk models within a model set. Values are calculated on a daily basis between November 4, 2004 until December 31, 2018 in percent of the portfolio value based on all models that passed the respective backtest, see Section 4.2.2 for details.



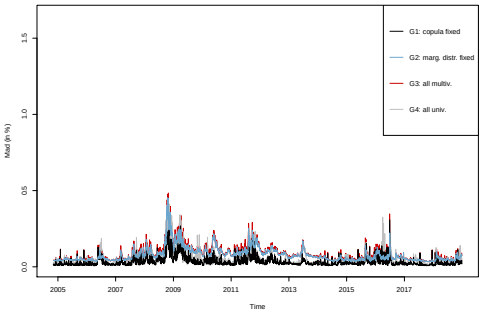
(a) Average model risk (99.9% VaR)



(b) Average model risk (99% VaR)



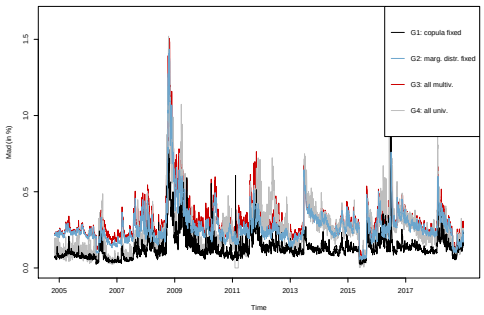
(c) Average model risk (97.5% VaR)



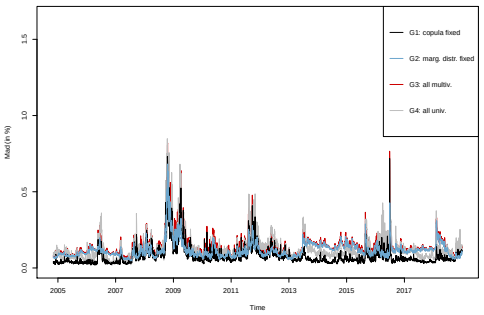
(d) Average model risk (95% VaR)

Figure B.4: Average model risk for all groups under various ES confidence levels

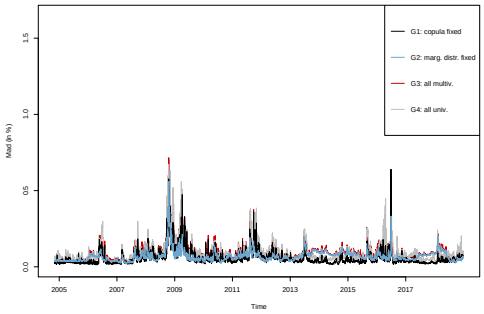
This figure shows the average model risk associated with one day ahead ES forecasts for a well diversified portfolio and a confidence level of 99.9%, 99%, 97.5%, and 95% (Subfigures 1-4) per group. *Group 1* (G1) includes all model sets in which a copula function is fixed while varying the marginal distribution. *Group 2* (G2) contains analogously the model sets with fixed marginal distribution and varying copula. *Group 3* (G3) consists of all multivariate and *Group 4* (G4) of all univariate models. Model risk is measured in terms of the mean absolute deviation (mad) of one day ahead forecasts by various risk models within a model set. Values are calculated on a daily basis between November 4, 2004 until December 31, 2018 in percent of the portfolio value based on all models that passed the respective backtest, see Section 4.2.2 for details.



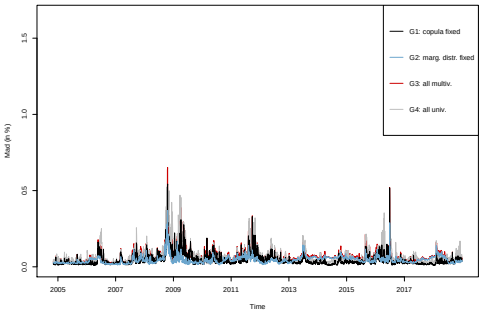
(a) Average model risk (99.9% ES)



(b) Average model risk (99% ES)



(c) Average model risk (97.5% ES)



(d) Average model risk (95% ES)

Appendix C

Supplementary Material for Chapter 5

Table C.1: Variable definitions and data sources.

The appendix presents definitions for the dependent and independent variables that are used in the empirical study and that have not been calculated using the LDA. Capital market data are retrieved from Thomson Reuters Datastream and accounting data are retrieved from Orbis Insurance Focus. All accounting data are collected in U.S. Dollar.

Variable	Description	Source
Capital market data		
Market value	Natural logarithm of the share price multiplied by the number of ordinary shares in issue.	Thomson Reuters Datastream
Market-to-book value	Market value of common equity divided by the balance sheet value of common equity in the company.	Thomson Reuters Datastream
Accounting data		
Total assets	Natural logarithm of an insurer's total assets at fiscal year end.	Orbis Insurance Focus
ROA	Return on Assets defined as net income over total assets (in %).	Orbis Insurance Focus
Total investment	Natural logarithm of an insurer's total amount of money invested into capital.	Orbis Insurance Focus
Solvency ratio	Net assets divided by net premiums written (in %).	Orbis Insurance Focus
Current ratio	Current assets divided by current liabilities (in %).	Orbis Insurance Focus
Foreign assets ratio	Foreign assets divided by total assets (in %).	Thomson Reuters Datastream

Appendix D

Publication Details

This cumulative dissertation comprises four independent research papers that were written with the co-authors Gregor Weiß, Felix Irresberger, Maike Timphus, and Philipp Scharner. This appendix provides publication details and a short description of the papers.

Paper 1 (Chapter 2):

Conditional Quantile Estimation via Leveraging Optimal Quantization

Author:

Simon Fritzsch

Abstract:

This paper proposes a new non-parametric estimator of conditional quantiles that is obtained by leveraging an ensemble of quantization-based estimators. The data-driven choice of the hyperparameters of the associated algorithm is discussed in detail and the added value of the new estimator is illustrated in an extensive simulation study. The estimator yields smooth quantile curves (in one dimension), extends well to multiple dimensions, and is competitive in terms of integrated squared errors. In an empirical application, the estimator is used to quantify the estimation risk of Value-at-Risk and Expected Shortfall forecasts by various GARCH-type models and to provide confidence bands. Among the considered models, the GARCH model exhibits the lowest estimation risk and the EGARCH model the highest, while in general estimation risk for Expected Shortfall is higher than for Value-at-Risk.

Publication details:

Working paper

Paper 2 (Chapter 3):

Cross-Section of Option Returns and the Volatility Risk Premium

Authors:

Simon Fritzsch, Felix Irresberger, Gregor Weiß

Abstract:

This paper presents a robust new finding that delta-hedged and raw equity option returns include a volatility risk premium. To separate volatility risk premia from confounding effects, we estimate conditional quantile curves of implied volatilities using machine learning. We find that a zero-cost trading strategy that is long (short) in the portfolio with low (high) implied volatility – conditional on the options' moneyness and realized volatility – produces an economically and statistically significant average monthly return. Using conditional quantile curves not only helps in distinguishing volatility risk premia from other effects, most notably realized volatility, it also leads to returns that are higher than those reported in previous work on similar volatility strategies.

Publication details:

Working paper

Paper 3 (Chapter 4):

Marginals Versus Copulas: Which Account for More Model Risk in Multivariate Risk Forecasting?

Authors:

Simon Fritzsch, Maike Timphus, Gregor Weiß

Abstract:

Copulas. We study the model risk of multivariate risk models in a comprehensive empirical study on copula GARCH models used for forecasting Value-at-Risk and Expected Shortfall. To determine whether model risk inherent in the forecasting of portfolio risk is caused by the candidate marginal or copula models, we analyze different groups of models in which we fix either the marginals, the copula, or neither. Model risk is economically significant, is especially high during periods of crisis, and is almost completely due to the choice of the copula. We then propose the use of the model confidence set procedure to narrow down the set of available models and reduce model risk for copula GARCH risk models. Our proposed approach leads to a significant improvement in the mean absolute deviation of one day ahead forecasts by our various candidate risk models.

Publication details:

Working paper

Paper 4 (Chapter 5):

Estimating the Relation Between Digitalization and the Market Value of Insurers

Authors:

Simon Fritzsch, Philipp Scharner, Gregor Weiß

Abstract:

We analyze the relation between digitalization and the market value of US insurance companies. To create a text-based measure that captures the extent to which insurers digitalize, we apply an unsupervised machine learning algorithm – Latent Dirichlet Allocation – to their annual reports. We show that an increase in digitalization is associated with an increase in market valuations in the insurance sector. In detail, capital market participants seem to reward digitalization efforts of an insurer in the form of higher absolute market capitalizations and market-to-book ratios. Additionally, we provide evidence that the positive relation between digitalization and market valuations is robust to sentiment in the annual reports and the choice of the reference document on digitalization, both being issues of particular importance in text-based analyses.

Publication details:

Fritzsch, S., P. Scharner, and G. Weiß (2021): "Estimating the relation between digitalization and the market value of insurers," *Journal of Risk and Insurance*, 88, 529–567.

Bibliography

Aas, K. and D. Berg (2009): "Models for construction of multivariate dependence – a comparison study," *European Journal of Finance*, 15, 639–659.

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009): "Pair-copula constructions of multiple dependence," *Insurance: Mathematics and Economics*, 44, 182–198.

Abaya, E. F. and G. L. Wise (1984): "Some remarks on the existence of optimal quantizers," *Statistics & Probability Letters*, 2, 349–351.

Acerbi, C. and B. Székely (2014): "Backtesting expected shortfall," *Risk*, 76–81.

Adrian, T. and M. K. Brunnermeier (2016): "CoVaR," *American Economic Review*, 106, 1705–1741.

Ahalt, S. C., A. K. Krishnamurthy, P. Chen, and D. E. Melton (1990): "Competitive learning algorithms for vector quantization," *Neural Networks*, 3, 277–290.

Alexander, C. and J. M. Sarabia (2012): "Quantile uncertainty and value-at-risk model risk," *Risk Analysis*, 32, 1293–1308.

Allen, F. (1993): "Strategic management and financial markets," *Strategic Management Journal*, 14, 11–22.

Andreou, E. and E. Ghysels (2020): "Predicting the VIX and the volatility risk premium: The role of short-run funding spreads Volatility Factors," *Journal of Econometrics*, 220, 366–398.

Antweiler, W. and M. Z. Frank (2004): "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *Journal of Finance*, 59, 1259–1294.

Arun, R., V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy (2010): "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations,"

in *Advances in knowledge discovery and data mining*, Berlin: Springer, Lecture notes in computer science, Lecture notes in artificial intelligence, 391–402.

Ausin, M. C. and H. F. Lopes (2010): "Time-varying joint distribution through copulas," *Computational Statistics & Data Analysis*, 54, 2383–2399.

Baele, L., J. Driessen, S. Ebert, J. M. Londono, and O. G. Spalt (2019): "Cumulative Prospect Theory, Option Returns, and the Variance Premium," *Review of Financial Studies*, 32, 3667–3723.

Bakshi, G. and N. Kapadia (2003): "Delta-Hedged Gains and the Negative Market Volatility Risk Premium," *Review of Financial Studies*, 16, 527–566.

Bali, T. G., H. Beckmeyer, M. Moerke, and F. Weigert (2021): "Option Return Predictability with Machine Learning and Big Data," *SSRN Electronic Journal*.

Bali, T. G., A. Goyal, D. Huang, F. Jiang, and Q. Wen (2020): "The Cross-Sectional Pricing of Corporate Bonds Using Big Data and Machine Learning," *Swiss Finance Institute Research Paper Series*.

Bally, V., G. Pages, and J. Printems (2005): "A QUANTIZATION TREE METHOD FOR PRICING AND HEDGING MULTIDIMENSIONAL AMERICAN OPTIONS," *Mathematical Finance*, 15, 119–168.

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

Bao, Y. and A. Datta (2014): "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures," *Management Science*, 60, 1371–1391.

Barendse, S., E. Kole, and D. van Dijk (2021): "Backtesting Value-at-Risk and Expected Shortfall in the Presence of Estimation Error," *Journal of Financial Econometrics*.

Barkur, G., K. Varambally, and L. L. Rodrigues (2007): "Insurance sector dynamics: towards transformation into learning organization," *The Learning Organization*, 14, 510–523.

Barone-Adesi, G., F. Bourgoin, and K. Giannopoulos (1998): "Don't look back," *Risk*, 11, 100–103.

Barone-Adesi, G., K. Giannopoulos, and L. Vosper (1999): "VaR without correlations for portfolios of derivative securities," *Journal of Futures Markets*, 19, 583–602.

Barraclough, K. and R. E. Whaley (2012): "Early Exercise of Put Options on Stocks," *Journal of Finance*, 67, 1423–1456.

Basel Committee on Banking Supervision (2013): "Consultative Document: Fundamental review of the trading book: a revised market risk framework," .

——— (2014): "Consultative Document: Fundamental review of the trading book: outstanding issues," .

——— (2019): "Minimum capital requirements for market risk," .

Bassett, G. and R. Koenker (1982): "An Empirical Quantile Function for Linear Models with iid Errors," *Journal of the American Statistical Association*, 77, 407–415.

Bassetti, F., M. E. D. Giuli, E. Nicolino, and C. Tarantola (2018): "Multivariate dependence analysis via tree copula models: An application to one-year forward energy contracts," *European Journal of Operational Research*, 269, 1107–1121.

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

Bayer, S. and T. Dimitriadis (2020a): "esback: Expected Shortfall Backtesting," .

——— (2020b): "Regression-Based Expected Shortfall Backtesting," *Journal of Financial Econometrics*.

Bedford, T. and R. M. Cooke (2001): "Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines," *Annals of Mathematics and Artificial Intelligence*, 32, 245–268.

——— (2002): "Vines–a new graphical model for dependent random variables," *Annals of Statistics*, 30, 1031–1068.

Bellstam, G., S. Bhagat, and J. A. Cookson (2020): "A Text-Based Analysis of Corporate Innovation," *Management Science*, 67, 1–28.

Bernard, C., R. Kazzi, and S. Vanduffel (2020): *A Practical Approach to Quantitative Model Risk Assessment*.

Bernard, C. and S. Vanduffel (2015): "A new approach to assessing model risk in high dimensions," *Journal of Banking & Finance*, 58, 166–178.

Beygelzimer, A., S. Kakadet, J. Langford, S. Arya, D. Mount, and S. Li (2019): "FNN: Fast Nearest Neighbor Search Algorithms and Applications," .

Bhattacharya, P. K. and A. K. Gangopadhyay (1990): "Kernel and nearest-neighbor estimation of a conditional quantile," *Annals of Statistics*, 18, 1400–1415.

Bianchi, D., M. Büchner, and A. Tamoni (2021): "Bond Risk Premiums with Machine Learning," *Review of Financial Studies*, 34, 1046–1089.

Bilgrau, A. E., P. S. Eriksen, J. G. Rasmussen, H. E. Johnsen, K. Dybkaer, and M. Boegsted (2016): "GMCM: Unsupervised Clustering and Meta-Analysis Using Gaussian Mixture Copula Models," *Journal of Statistical Software*, 70, 1–23.

Blei, D. M. (2012): "Probabilistic Topic Models," *Communications of the ACM*, 55, 77–84.

Blei, D. M., T. L. Griffiths, and M. I. Jordan (2010): "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, 57, 1–30.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Bodnaruk, A., T. Loughran, and B. McDonald (2015): "Using 10-K Text to Gauge Financial Constraints," *Journal of Financial and Quantitative Analysis*, 50, 623–646.

Bohnert, A., A. Fritzsche, and S. Gregor (2019): "Digital agendas in the insurance industry: the importance of comprehensive approaches," *Geneva Papers on Risk and Insurance - Issues and Practice*, 44, 1–19.

Bollerslev, T. (1986): "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

——— (2010): "Glossary to ARCH (GARCH *)," in *Volatility and time series econometrics*, ed. by R. F. Engle, M. W. Watson, T. Bollerslev, and J. R. Russell, Oxford: Oxford University Press, Advanced texts in econometrics.

Bollerslev, T., G. Tauchen, and H. Zhou (2009): "Expected Stock Returns and Variance Risk Premia," *Review of Financial Studies*, 22, 4463–4492.

Boucher, C. M., J. Danielsson, P. S. Kouontchou, and B. B. Maillet (2014): "Risk models-at-risk," *Journal of Banking & Finance*, 44, 72–92.

Boudoukh, J., M. Richardson, and R. F. Whitelaw (1998): "The Best of Both Worlds: A Hybrid Approach to Calculating Value at Risk," *Risk*, 64–67.

Bouton, C. and G. Pagès (1997): "About the multidimensional competitive learning vector quantization algorithm with constant gain," *The Annals of Applied Probability*, 7.

Bouyé, E. and M. Salmon (2009): "Dynamic Copula Quantile Regressions and Tail Area Dynamic Dependence in Forex Markets," *European Journal of Finance*, 15, 721–750.

Boyer, B. H. and K. Vorkink (2014): "Stock Options as Lotteries," *Journal of Finance*, 69, 1485–1527.

Brechmann, E. C. and C. Czado (2013): "Risk management with high-dimensional vine copulas: An analysis of the Euro Stoxx 50," *Statistics & Risk Modeling*, 30.

Breiman, L. (1996): "Bagging predictors," *Machine Learning*, 24, 123–140.

——— (1998): "Arcing the edge," *Annals of Probability*, 26, 1683–1702.

Breiman, L. and P. Spector (1992): "Submodel Selection and Evaluation in Regression. The X-Random Case," *International Statistical Review / Revue Internationale de Statistique*, 60, 291.

Calabrese, R., M. Degl'Innocenti, and S. Osmetti (2017): "The effectiveness of TARP-CPP on the US Banking Industry: a new copula-based approach," *European Journal of Operational Research*, 256, 1029–1037.

Calabrese, R. and S. A. Osmetti (2019): "A new approach to measure systemic risk: A bivariate copula model for dependent censored data," *European Journal of Operational Research*, 279, 1053–1064.

Campbell, S. (2007): "A review of backtesting and backtesting procedures," *The Journal of Risk*, 9, 1–17.

Cao, J. and B. Han (2013): "Cross section of option returns and idiosyncratic stock volatility," *Journal of Financial Economics*, 108, 231–249.

Cao, J., B. Han, X. Zhan, and Q. Tong (2021): "Option Return Predictability," *Review of Financial Studies*.

Cao, J., A. Vasquez, X. Xiao, and X. Zhan (2019): "Volatility Uncertainty and the Cross-Section of Option Returns," *SSRN Electronic Journal*.

Cao, J., T. Xia, J. Li, Y. Zhang, and S. Tang (2009): "A density-based method for adaptive LDA model selection," *Neurocomputing*, 72, 1775–1781.

Cappiello, A. (2020): "The Digital (R)evolution of Insurance Business Models," *American Journal of Economics and Business Administration*, 12, 1–13.

Cardona, E., A. Mora-Valencia, and D. Velásquez-Gaviria (2019): "Testing expected shortfall: an application to emerging market stock indices," *Risk Management*, 21, 153–182.

Carr, P. and L. Wu (2009): "Variance Risk Premiums," *Review of Financial Studies*, 22, 1311–1341.

Catania, L. and M. Bernardi (2017): "MCS: Model Confidence Set Procedure," .

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and E. Schaumburg (2020): "Characteristic-Sorted Portfolios: Estimation and Inference," *The Review of Economics and Statistics*, 102, 531–551.

Chabi-Yo, F., S. Ruenzi, and F. Weigert (2018): "Crash Sensitivity and the Cross Section of Expected Stock Returns," *Journal of Financial and Quantitative Analysis*, 53, 1059–1100.

Chan, N. H., S.-J. Deng, L. Peng, and Z. Xia (2007): "Interval estimation of value-at-risk based on GARCH models with heavy-tailed innovations," *Journal of Econometrics*, 137, 556–576.

Charlier, I., D. Paindaveine, and J. Saracco (2015a): "Conditional quantile estimation based on optimal quantization: From theory to practice," *Computational Statistics & Data Analysis*, 91, 20–39.

——— (2015b): "Conditional quantile estimation through optimal quantization," *Journal of Statistical Planning and Inference*, 156, 14–30.

——— (2015c): "QuantifQuantile: Estimation of Conditional Quantiles using Optimal Quantization," .

Chen, L., M. Pelger, and J. Zhu (2020a): "Deep Learning in Asset Pricing," *arXiv*.

Chen, T. and C. Guestrin (2016): "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, ed. by B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, and R. Rastogi, New York, USA: ACM Press, 785–794.

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li (2020b): "xgboost: Extreme Gradient Boosting," .

Chen, X., R. Koenker, and Z. Xiao (2009): "Copula-based nonlinear quantile autoregression," *The Econometrics Journal*, 12, 50–67.

Chernozhukov, V., I. Fernandez-Val, and A. Galichon (2010): "Quantile and Probability Curves Without Crossing," *Econometrica*, 78, 1093–1125.

Cherubini, U., E. Luciano, and W. Vecchiato (2004): *Copula methods in finance*, Wiley finance series, Hoboken, NJ: John Wiley & Sons.

Christoffersen, P. (2004): "Backtesting Value-at-Risk: A Duration-Based Approach," *Journal of Financial Econometrics*, 2, 84–108.

Christoffersen, P. and S. Gonçalves (2005): "Estimation risk in financial risk management," *The Journal of Risk*, 7, 1–28.

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

Connor, G., M. Hagmann, and O. Linton (2012): "Efficient Semiparametric Estimation of the Fama-French Model and Extensions," *Econometrica*, 80, 713–754.

Connor, G. and O. Linton (2007): "Semiparametric estimation of a characteristic-based factor model of common stock returns," *Journal of Empirical Finance*, 14, 694–717.

Cont, R. (2006): "MODEL UNCERTAINTY AND ITS IMPACT ON THE PRICING OF DERIVATIVE INSTRUMENTS," *Mathematical Finance*, 16, 519–547.

Cont, R. and J. da Fonseca (2002): "Dynamics of implied volatility surfaces," *Quantitative Finance*, 2, 45–60.

Coval, J. D. and T. Shumway (2001): "Expected Option Returns," *Journal of Finance*, 56, 983–1009.

Cox, J. C., S. A. Ross, and M. Rubinstein (1979): "Option pricing: A simplified approach," *Journal of Financial Economics*, 7, 229–263.

Danielsson, J., K. R. James, M. Valenzuela, and I. Zer (2016): "Model risk of risk models," *Journal of Financial Stability*, 23, 79–91.

Danielsson, J. and J.-P. Zigrand (2006): "On time-scaling of risk and the square-root-of-time rule," *Journal of Banking & Finance*, 30, 2701–2713.

David Ardia, Kris Boudt, and Leopoldo Catania (2019): "Generalized Autoregressive Score Models in R: The GAS Package," *Journal of Statistical Software*, 88, 1–28.

de Bondt, W. F. M. and R. Thaler (1985): "Does the stock market overreact?" *Journal of Finance*, 40, 793–805.

de Fontnouvelle, P., R. P. H. Fishe, and J. H. Harris (2003): "The Behavior of Bid-Ask Spreads and Volume in Options Markets during the Competition for Listings in 1999," *Journal of Finance*, 58, 2437–2463.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990): "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, 41, 391–407.

Deloitte (2016): "Insurers on the brink: Disrupt or be disrupted," *White paper*.

Desyllas, P. and M. Sako (2013): "Profiting from business model innovation: Evidence from Pay-As-You-Drive auto insurance," *Research Policy*, 42, 101–116.

Dette, H. and S. Volgushev (2008): "Non-Crossing Non-Parametric Estimates of Quantile Curves," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70, 609–627.

Deveaud, R., E. Sanjuan, and P. Bellot (2014): "Accurate and effective latent concept modeling for ad hoc information retrieval," *Document numérique*, 17, 61–84.

Diebold, F., A. Hickman, A. Inoue, and T. Schuermann (1997): "Converting 1-Day Volatility to h-Day Volatitlity: Scaling by Root-h is Worse Than You Think," *Center for Financial Institutions Working Papers*.

Diebold, F. and R. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Ding, Z., C. W. Granger, and R. F. Engle (1993): "A long memory property of stock market returns and a new model," *Journal of Empirical Finance*, 1, 83–106.

Dissmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013): "Selecting and estimating regular vine copulae and application to financial returns," *Computational Statistics & Data Analysis*, 59, 52–69.

Doherty, N. A. and A. Richter (2002): "Moral Hazard, Basis Risk, and Gap Insurance," *Journal of Risk and Insurance*, 69, 9–24.

Driessen, J., P. J. Maenhout, and G. Vilkov (2009): "The Price of Correlation Risk: Evidence from Equity Options," *Journal of Finance*, 64, 1377–1406.

Duffy, N. and D. Helmbold (1999): "A Geometric Approach to Leveraging Weak Learners," in *Computational learning theory: 4th European conference; proceedings*, ed. by P. Fischer, Berlin: Springer, vol. 1572 of *Lecture notes in computer science, Lecture notes in artificial intelligence*, 18–33.

——— (2002): "Boosting Methods for Regression," *Machine Learning*, 47, 153–200.

Eisdorfer, A., A. Goyal, and A. Zhdanov (2020): "Cheap Options Are Expensive," *SSRN Electronic Journal*.

Eling, M. and M. Lehmann (2018): "The Impact of Digitalization on the Insurance Value Chain and the Insurability of Risks," *Geneva Papers on Risk and Insurance - Issues and Practice*, 43, 359–396.

Engle, R. (2002): "Dynamic Conditional Correlation," *Journal of Business & Economic Statistics*, 20, 339–350.

Engle, R. and K. Sheppard (2001): "Theoretical and Empirical properties of Dynamic Conditional Correlation Multivariate GARCH," *NBER working paper series*.

Engle, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, 987.

Engle, R. F. and S. Manganelli (2004): "CAViaR," *Journal of Business & Economic Statistics*, 22, 367–381.

Escanciano, J. C. and J. Olmo (2010): "Backtesting Parametric Value-at-Risk With Estimation Risk," *Journal of Business & Economic Statistics*, 28, 36–51.

Fama, E. F. and K. R. French (1992): "The cross-section of expected stock returns," *Journal of Finance*, 47, 427–465.

Fan, J., T.-C. Hu, and Y. K. Truong (1994): "Robust Non-Parametric Function Estimation," *Scandinavian Journal of Statistics*, 21, 433–446.

Feinerer, I. and K. Hornik (2020): "tm: Text Mining Package," .

Fernandez, C. and M. F. J. Steel (1998): "On Bayesian Modeling of Fat Tails and Skewness," *Journal of the American Statistical Association*, 93, 359.

Figlewski, S. (2003): "Estimation Error in the Assessment of Financial Risk Exposure," *New York University Stern School of Business Research Paper Series*.

Fischer, M., C. Köck, S. Schlüter, and F. Weigert (2009): "An empirical analysis of multivariate copula models," *Quantitative Finance*, 9, 839–854.

Fissler, T. and J. F. Ziegel (2016): "HIGHER ORDER ELICITABILITY AND OSBAND'S PRINCIPLE," *Annals of Statistics*, 44, 1680–1707.

Fissler, T., J. F. Ziegel, and T. Gneiting (2016): "Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting," *Risk Management*, 58–61.

Frazier, K. B., R. W. Ingram, and B. M. Tennyson (1984): "A Methodology for the Analysis of Narrative Accounting Disclosures," *Journal of Accounting Research*, 22, 318–331.

Freund, Y. and R. E. Shapire (1996): "Experiments with a New Boosting Algorithm," *Machine Learning: Proceedings of the Thirteenth International Converence*.

Freyberger, J., A. Neuhierl, and M. Weber (2020): "Dissecting Characteristics Nonparametrically," *Review of Financial Studies*, 33, 2326–2377.

Friedman, J., T. Hastie, and R. Tibshirani (2000): "Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)," *Annals of Statistics*, 28, 337–407.

Friedman, J. H. (2001): "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, 29, 1189–1232.

Fritzsch, S., P. Scharner, and G. Weiss (2021): "Estimating the relation between digitalization and the market value of insurers," *Journal of Risk and Insurance*, 88, 529–567.

Frongillo, R. and Ian A. Kash (2015): "Vector-Valued Property Elicitation," *Conference on Learning Theory*, 710–727.

Ganglmair, B. and M. Wardlaw (2017): "Complexity, Standardization, and the Design of Loan Agreements," *SSRN Electronic Journal*.

Gao, F. and F. Song (2008): "ESTIMATION RISK IN GARCH VaR AND ES ESTIMATES," *Econometric Theory*, 24, 1404–1424.

Gatzert, N. and D. Heidinger (2020): "An Empirical Analysis of Market Reactions to the First Solvency and Financial Condition Reports in the European Insurance Industry," *Journal of Risk and Insurance*, 87, 407–436.

Gelfand, A. E. and A. F. M. Smith (1990): "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Genest, C., B. Rémillard, and D. Beaudoin (2009): "Goodness-of-fit tests for copulas: A review and a power study," *Insurance: Mathematics and Economics*, 44, 199–213.

Ghalanos, A. (2019): "rmgarch: Multivariate GARCH models," .

——— (2020): "rugarch: Univariate GARCH models," .

Glasserman, P. and X. Xu (2014): "Robust risk measurement and model risk," *Quantitative Finance*, 14, 29–58.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993): "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks," *Journal of Finance*, 48, 1779–1801.

Gneiting, T. (2011): "Making and Evaluating Point Forecasts," *Journal of the American Statistical Association*, 106, 746–762.

Goldsmith-Pinkham, P., B. Hirtle, and D. O. Lucca (2016): "Parsing the Content of Bank Supervision," *Staff Reports*.

Goyal, A. (2012): "Empirical cross-sectional asset pricing: a survey," *Financial Markets and Portfolio Management*, 26, 3–38.

Goyal, A. and A. Saretto (2009): "Cross-section of option returns and volatility," *Journal of Financial Economics*, 94, 310–326.

Goyenko, R. and C. Zhang (2020): "The Joint Cross Section of Option and Stock Returns Predictability with Big Data and Machine Learning," *SSRN Electronic Journal*.

Grace, M. (2019): "Risk Identification: What is in the 10-K?" *Presented at: 2019 ARIA Annual Meeting*.

Graf, S. and H. Luschgy (2002): "Rates of convergence for the empirical quantization error," *Annals of Probability*, 30, 874–897.

Green, T. C. and S. Figlewski (1999): "Market Risk and Model Risk for a Financial Institution Writing Options," *Journal of Finance*, 54, 1465–1499.

Gries, S. T. and J. Newman (2013): "Creating and using corpora," in *Research methods in linguistics*, ed. by R. J. Podesva and D. Sharma, Cambridge: Cambridge Univ. Press, 257–287.

Griffiths, T. L. and M. Steyvers (2004): "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1, 5228–5235.

Griffiths, T. L., M. Steyvers, D. M. Blei, and J. B. Tenenbaum (2005): "Integrating topics and syntax," *Advances in Neural Information Processing Systems 17 - Proceedings of the 2004 Conference, NIPS 2004*.

Grosen, A. and P. L. Jørgensen (2002): "Life Insurance Liabilities at Market Value: An Analysis of Insolvency Risk, Bonus Policy, and Regulatory Intervention Rules in a Barrier Option Framework," *Journal of Risk and Insurance*, 69, 63–91.

Grossberg, S. (1976): "Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors," *Biological cybernetics*, 23, 121–134.

——— (1987): "Competitive learning: From interactive activation to adaptive resonance," *Cognitive Science*, 11, 23–63.

Grundke, P. and S. Polle (2012): "Crisis and risk dependencies," *European Journal of Operational Research*, 223, 518–528.

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

Hanelt, A., S. Firk, B. Hildebrandt, and L. M. Kolbe (2020): "Digital M&A, digital innovation, and firm performance: an empirical investigation," *European Journal of Information Systems*, 24, 1–24.

Hanley, K. W. and G. Hoberg (2010): "The Information Content of IPO Prospectuses," *Review of Financial Studies*, 23, 2821–2864.

Hansen, P. R., A. Lunde, and J. M. Nason (2011): "The Model Confidence Set," *Econometrica*, 79, 453–497.

Hansen, S., M. McMahon, and A. Prat (2018): "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach," *Quarterly Journal of Economics*, 133, 801–870.

Hastie, T., R. Tibshirani, and J. H. Friedman (2017): *The elements of statistical learning: Data mining, inference, and prediction*, Springer series in statistics, New York, NY: Springer, second edition, corrected at 12th printing 2017 ed.

Heagerty, P. J. and M. S. Pepe (1999): "Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48, 533–551.

Hendricks, D. (1996): "Evaluation of Value-at-Risk Models Using Historical Data," *SSRN Electronic Journal*.

Henry, E. and A. J. Leone (2016): "Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone," *Accounting Review*, 91, 153–178.

Hentschel, L. (1995): "All in the family Nesting symmetric and asymmetric GARCH models," *Journal of Financial Economics*, 39, 71–104.

Hoberg, G. and C. Lewis (2017): "Do fraudulent firms produce abnormal disclosure?" *Journal of Corporate Finance*, 43, 58–85.

Hoberg, G. and V. Maksimovic (2015): "Redefining Financial Constraints: A Text-Based Analysis," *Review of Financial Studies*, 28, 1312–1352.

Hoberg, G., G. Phillips, and N. Prabhala (2014): "Product Market Threats, Payouts, and Financial Flexibility," *Journal of Finance*, 69, 293–324.

Hofert, M., I. Kojadinovic, M. Maechler, and J. Yan (2020): "copula: Multivariate Dependence with Copulas," .

Hofmann, T. (1999): "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ed. by F. Gey, New York, NY: ACM, 50–57.

Hu, G. and K. Jacobs (2020): "Volatility and Expected Option Returns," *Journal of Financial and Quantitative Analysis*, 55, 1025–1060.

Huang, A. H., R. Lehavy, A. Y. Zang, and R. Zheng (2018): "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach," *Management Science*, 64, 2833–2855.

Huang, D., C. Schlag, I. Shaliastovich, and J. Thimme (2019): "Volatility-of-Volatility Risk," *Journal of Financial and Quantitative Analysis*, 54, 2423–2452.

Hull, J. and W. Suo (2002): "A Methodology for Assessing Model Risk and Its Application to the Implied Volatility Function Model," *Journal of Financial and Quantitative Analysis*, 37, 297–318.

Hwang, C. and J. Shim (2005): "A Simple Quantile Regression via Support Vector Machine," in *Advances in Natural Computation*, ed. by K. Chen, Y. S. Ong, and L. Wang, Berlin Heidelberg: Springer-Verlag GmbH, 512–520.

Israelsen, R. D. (2014): "Tell It Like It Is: Disclosed Risks and Factor Portfolios," *SSRN Electronic Journal*.

Ivaşcu, C.-F. (2021): "Option pricing using Machine Learning," *Expert Systems with Applications*, 163.

Jayech, S. (2016): "The contagion channels of July–August-2011 stock market crash: A DAG-copula based approach," *European Journal of Operational Research*, 249, 631–646.

Jegadeesh, N. (1990): "Evidence of predictable behavior of security returns," *Journal of Finance*, 45, 881–898.

Jegadeesh, N. and S. Titman (1993): "Returns to buying winners and selling losers: Implications for stock market efficiency," *Journal of Finance*, 48, 65–91.

Jegadeesh, N. and D. Wu (2013): "Word power: A new approach for content analysis," *Journal of Financial Economics*, 110, 712–729.

——— (2017): "Deciphering Fedspeak: The Information Content of FOMC Meetings," *SSRN Electronic Journal*.

Jiang, G. J. and Y. S. Tian (2005): "The Model-Free Implied Volatility and Its Information Content," *Review of Financial Studies*, 18, 1305–1342.

Jiang, L., K. Wu, and G. Zhou (2018): "Asymmetry in Stock Comovements: An Entropy Approach," *Journal of Financial and Quantitative Analysis*, 53, 1479–1507.

Joe, H. (1996): "Families of m-Variate Distributions with Given Margins and m(m-1)/2 Bivariate Dependence Parameters," *Lecture Notes-Monograph Series*, 28, 120–141.

——— (2001): *Multivariate models and dependence concepts*, vol. 73 of *Monographs on statistics and applied probability*, Boca Raton, Fla.: Chapman & Hall/CRC, 1. crc reprint ed.

Jondeau, E. and M. Rockinger (2006): "The Copula-GARCH model of conditional dependencies: An international stock market application," *Journal of International Money and Finance*, 25, 827–853.

Jorion, P. (1996): "Risk^2: Measuring the Risk in Value at Risk," *Financial Analysts Journal*, 52, 47–56.

Kabaila, P. and R. Mainzer (2018): "Estimation risk for value-at-risk and expected shortfall," *The Journal of Risk*, 20, 29–47.

Ke, Z. T., B. T. Kelly, and D. Xiu (2019): "Predicting Returns With Text Data," *NBER working paper series*.

Kearns, M. J., R. E. Schapire, and L. M. Sellie (1994): "Toward efficient agnostic learning," *Machine Learning*, 17, 115–141.

Kelly, B. T., S. Pruitt, and Y. Su (2019): "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, 134, 501–524.

Koenker, R. (2005): *Quantile regression*, vol. 38 of *Econometric Society monographs*, Cambridge: Cambridge University Press.

——— (2011): "Additive models for quantile regression: Model selection and confidence bandaids," *Brazilian Journal of Probability and Statistics*, 25.

——— (2020): "quantreg: Quantile Regression," .

Koenker, R. and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R. and I. Mizera (2004): "Penalized triograms: total variation regularization for bivariate smoothing," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 145–163.

Koenker, R., P. I. N. Ng, and S. Portnoy (1994): "Quantile smoothing splines," *Biometrika*, 81, 673–680.

Kohonen, T. (1982): "Analysis of a simple self-organizing process," *Biological cybernetics*, 44, 135–140.

——— (1989): *Self-Organization and Associative Memory*, vol. 8 of *Springer Series in Information Sciences*, Berlin and Heidelberg: Springer, 3 ed.

——— (1992): "New developments of learning vector quantization and the self-organizing map," in *Symposium on Neural Networks in Senri, Osaka, Japan.*

——— (2001): *Self-Organizing Maps*, vol. 30 of *Springer Series in Information Sciences*, Berlin and Heidelberg: Springer, 3rd ed.

Kole, E., K. Koedijk, and M. Verbeek (2007): "Selecting copulas for risk management," *Journal of Banking & Finance*, 31, 2405–2423.

Kole, E., T. Markwat, A. Opschoor, and D. van Dijk (2017): "Forecasting Value-at-Risk under Temporal and Portfolio Aggregation," *Journal of Financial Econometrics*, 15, 649–677.

Kozak, S., S. Nagel, and S. Santosh (2020): "Shrinking the cross-section," *Journal of Financial Economics*, 135, 271–292.

Kraus, D. and C. Czado (2017): "D-vine copula based quantile regression," *Computational Statistics & Data Analysis*, 110, 1–18.

Kullback, S. and R. A. Leibler (1951): "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

Kurowicka, D. and R. Cooke (2006): *Uncertainty analysis with high dimensional dependence modelling*, Wiley series in probability and statistics, Chichester: Wiley.

Lan, H., B. L. Nelson, and J. Staum (2010): "A Confidence Interval Procedure for Expected Shortfall Risk Measurement via Two-Level Simulation," *Operations Research*, 58, 1481–1490.

Ledoit, O. and M. Wolf (2008): "Robust performance hypothesis testing with the Sharpe ratio," *Journal of Empirical Finance*, 15, 850–859.

Lee, J.-P. and M.-T. Yu (2002): "Pricing Default-Risky CAT Bonds With Moral Hazard and Basis Risk," *Journal of Risk and Insurance*, 69, 25–44.

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, 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, 423–434.

Lönnbark, C. (2010): "A corrected Value-at-Risk predictor," *Applied Economics Letters*, 17, 1193–1196.

——— (2013): "On the role of the estimation error in prediction of expected shortfall," *Journal of Banking & Finance*, 37, 847–853.

López de Prado, M. and M. J. Lewis (2019): "Detection of false investment strategies using unsupervised learning methods," *Quantitative Finance*, 19, 1555–1565.

Lopez-Lira, A. (2019): "Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns," *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

Loughran, T. and B. McDonald (2011): "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, 66, 35–65.

——— (2016): "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, 54, 1187–1230.

Lowry, M., R. Michaely, and E. Volkova (2020): "Information Revealed through the Regulatory Process: Interactions between the SEC and Companies ahead of Their IPO," *Review of Financial Studies*, 33, 5510–5554.

Lubatkin, M. and R. E. Shrieves (1986): "Towards Reconciliation of Market Performance Measures to Strategic Management Research," *Academy of Management Review*, 11, 497–512.

Manning, C. D., P. Raghavan, and H. Schütze (2009): *Introduction to information retrieval*, Cambridge: Cambridge Univ. Press, reprinted. ed.

Mayhew, S. (2002): "Competition, Market Structure, and Bid-Ask Spreads in Stock Option Markets," *Journal of Finance*, 57, 931–958.

McKinsey (2017): "Digital disruption in insurance: Cutting through the noise," *White paper*.

McNeil, A. J. and R. Frey (2000): "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach," *Journal of Empirical Finance*.

McNeil, A. J., R. Frey, and P. Embrechts (2005): *Quantitative risk management: Concepts, techniques and tools*, Princeton series in finance, Princeton and Oxford: Princeton University Press, 2015 revised ed.

Meier, A. and H. Stormer (2012): *eBusiness & eCommerce: Management der digitalen Wertschöpfungskette*, Berlin, Heidelberg: Springer, 3rd ed.

Meinshausen, N. (2006): "Quantile Regression Forests," *Journal of Machine Learning Research*, 7, 983–999.

Meir, R. and G. Rätsch (2003): "An Introduction to Boosting and Leveraging," in *Advanced Lectures on Machine Learning: Machine Learning Summer School 2002 Canberra, Australia, February 1122, 2002 Revised Lectures*, ed. by S. Mendelson, Berlin, Heidelberg: Springer, Lecture Notes in Computer Science, 118–183.

Moritz, B. and T. Zimmermann (2016): "Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns," *SSRN Electronic Journal*.

Nadarajah, S. (2005): "A generalized normal distribution," *Journal of Applied Statistics*, 32, 685–694.

Nagler, T. (2020): "vinereg: D-Vine Quantile Regression," .

Nagler, T., U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, and T. Erhardt (2019): "VineCopula: Statistical Inference of Vine Copulas," .

Nelsen, R. B. (2006): *An introduction to Copulas*, Springer series in statistics, New York, NY: Springer New York, 2. (2010) ed.

Nelson, D. B. (1991): "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 59, 347.

Nelson, D. B. and C. Q. Cao (1992): "Inequality Constraints in the Univariate GARCH Model," *Journal of Business & Economic Statistics*, 10, 229.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703.

——— (1994): "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies*, 61, 631–653.

Nicoletti, B. (2016): *Digital Insurance: Business Innovation in the Post-Crisis Era*, Palgrave Studies in Financial Services Technology, Basingstoke: Palgrave Macmillan.

Nikoloulopoulos, A. K., H. Joe, and H. Li (2012): "Vine copulas with asymmetric tail dependence and applications to financial return data," *Computational Statistics & Data Analysis*, 56, 3659–3673.

Noh, H., A. E. Ghouch, and I. van Keilegom (2015): "Semiparametric Conditional Quantile Estimation Through Copula-Based Multivariate Models," *Journal of Business & Economic Statistics*, 33, 167–178.

Nolde, N. and J. F. Ziegel (2017): "Elicitability and backtesting: Perspectives for banking regulation," *The Annals of Applied Statistics*, 11, 1833–1874.

Oh, D. H. and A. J. Patton (2017): "Modeling Dependence in High Dimensions With Factor Copulas," *Journal of Business & Economic Statistics*, 35, 139–154.

——— (2018): "Time-Varying Systemic Risk: Evidence From a Dynamic Copula Model of CDS Spreads," *Journal of Business & Economic Statistics*, 36, 181–195.

Pagès, G. (1998): "A space quantization method for numerical integration," *Journal of Computational and Applied Mathematics*, 89, 1–38.

Pagès, G. and J. Printems (2003): "Optimal quadratic quantization for numerics: the Gaussian case," *Monte Carlo Methods and Applications*, 9, 135–165.

Patton, A. and A. Timmermann (2010): "Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts," *Journal of Financial Economics*, 98, 605–625.

Patton, A. J. (2006): "Estimation of multivariate models for time series of possibly different lengths," *Journal of Applied Econometrics*, 21, 147–173.

Patton, A. J., J. F. Ziegel, and R. Chen (2019): "Dynamic semiparametric models for expected shortfall (and Value-at-Risk)," *Journal of Econometrics*, 211, 388–413.

Picard, R. R. and R. D. Cook (1984): "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79, 575–583.

Porter, M. F. (1980): "An algorithm for suffix stripping," *Program*, 14, 130–137.

Rapach, D. E., J. K. Strauss, and G. Zhou (2013): "International Stock Return Predictability: What Is the Role of the United States?" *Journal of Finance*, 68, 1633–1662.

Rayport, J. F. and J. F. Sviokla (1995): "Exploiting the virtual value chain," *Harvard Business Review*, 73, 75–85.

Roberts, M. E., B. M. Stewart, and E. M. Airoldi (2016): "A Model of Text for Experimentation in the Social Sciences," *Journal of the American Statistical Association*, 111, 988–1003.

Rossi, A. G. (2018): "Predicting Stock Market Returns with Machine Learning," *Working paper*.

Rothfuss, J., F. Ferreira, S. Walther, and M. Ulrich (2019): "Conditional Density Estimation with Neural Networks: Best Practices and Benchmarks," *arXiv*.

Santos, A. A. P., F. J. Nogales, and E. Ruiz (2013): "Comparing Univariate and Multivariate Models to Forecast Portfolio Value-at-Risk," *Journal of Financial Econometrics*, 11, 400–441.

Savu, C. and M. Trede (2008): "Goodness-of-fit tests for parametric families of Archimedean copulas," *Quantitative Finance*, 8, 109–116.

Schmidt, C. (2018): *Insurance in the digital age: A view on key implications for the economy and society*, Zurich: The Geneva Association - International Association for the Study of Insurance Economics.

Scott, S. V., J. van Reenen, and M. Zachariadis (2017): "The long-term effect of digital innovation on bank performance: An empirical study of SWIFT adoption in financial services," *Research Policy*, 46, 984–1004.

Seitz, M. (2017): "Online Insurance Management among German Farmers," *Proceedings of the 17th International Joint Conference Central and Eastern Europe in the Changing Business Environment*, 212–220.

Sirignano, J. and R. Cont (2019): "Universal features of price formation in financial markets: perspectives from deep learning," *Quantitative Finance*, 19, 1449–1459.

Sklar, A. (1959): "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Spokoiny, V., W. Wang, and W. Härdle (2013): "Local quantile regression," *Journal of Statistical Planning and Inference*, 143, 1109–1129.

Stone, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *Annals of Statistics*, 8, 1348–1360.

Subramanyam, K. R. (2014): *Financial statement analysis*, New York: McGraw Hill Education, 11th ed.

Takeuchi, I., Q. V. Le, T. D. Sears, and A. J. Smola (2006): "Nonparametric Quantile Estimation," *Journal of Machine Learning Research*, 7, 1231–1264.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006): "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581.

Tetlock, P. C. (2007): "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *Journal of Finance*, 62, 1139–1168.

Tewari, A., M. J. Giering, and A. Raghunathan (2011): "Parametric Characterization of Multimodal Distributions with Non-gaussian Modes," in *2011 IEEE 11th International Conference on Data Mining workshops (ICDMW 2011)*, ed. by M. Spiliopoulou, Piscataway, NJ: IEEE, 286–292.

Tidd, J. and J. R. Bessant (2018): *Managing innovation: Integrating technological, market and organizational change*, Hoboken: Wiley, 6th ed.

Toft, K. B. and B. Prucyk (1997): "Options on Leveraged Equity: Theory and Empirical Tests," *Journal of Finance*, 52, 1151–1180.

Valiant, L. G. (1984): "A theory of the learnable," *Communications of the ACM*, 27, 1134–1142.

van Rossum, A., H. de Castries, and R. Mendelsohn (2002): "The Debate on the Insurance Value Chain," *Geneva Papers on Risk and Insurance - Issues and Practice*, 27, 89–101.

Vedavathi, K., K. Srinivasa Rao, and K. Nirupama Devi (2014): "Unsupervised learning algorithm for time series using bivariate AR(1) model," *Expert Systems with Applications*, 41, 3402–3408.

Venables, W. N. and B. D. Ripley (2002): *Modern Applied Statistics with S*, New York: Springer, fourth ed.

Wainwright, M. J. and M. I. Jordan (2008): "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, 1, 1–305.

Wallach, H. M. (2006): "Topic modeling," in *Proceedings of the 23rd international conference on Machine learning*, ed. by W. Cohen, New York, NY: ACM, 977–984.

Wang, J.-N., J.-H. Yeh, and N. Y.-P. Cheng (2011): "How accurate is the square-root-of-time rule in scaling tail risk: A global study," *Journal of Banking & Finance*, 35, 1158–1169.

Wang, T. and J. S. Dyer (2012): "A Copulas-Based Approach to Modeling Dependence in Decision Trees," *Operations Research*, 60, 225–242.

Wang, X. and E. Grimson (2007): "Spatial Latent Dirichlet Allocation," *Advances in Neural Information Processing Systems*, 20, 1577–1584.

Weiss, G. (2013): "Copula-GARCH versus dynamic conditional correlation: an empirical study on VaR and ES forecasting accuracy," *Review of Quantitative Finance and Accounting*, 41, 179–202.

Weiss Hanley, K. and G. Hoberg (2019): "Dynamic Interpretation of Emerging Risks in the Financial Sector," *Review of Financial Studies*, 32, 4543–4603.

West, K. D. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067.

Wu, S. (2014): "Construction of asymmetric copulas and its application in two-dimensional reliability modelling," *European Journal of Operational Research*, 238, 476–485.

Wu, T. Z., K. Yu, and Y. Yu (2010): "Single-index quantile regression," *Journal of Multivariate Analysis*, 101, 1607–1621.

Xiao, Z. and R. Koenker (2009): "Conditional Quantile Estimation for Generalized Autoregressive Conditional Heteroscedasticity Models," *Journal of the American Statistical Association*, 104, 1696–1712.

Yang, Y. (2007): "Consistency of cross validation for comparing regression procedures," *Annals of Statistics*, 35, 2450–2473.

Yekini, L. S., T. P. Wisniewski, and Y. Millo (2016): "Market Reaction to the Positiveness of Annual Report Narratives," *British Accounting Review*, 48, 415–430.

Yu, K. and M. C. Jones (1998): "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228.

Zador, P. (1964): "Development and Evaluation of Procedures for Quantizing Multivariate Distributions," *ProQuest LLC, Ann Arbor (MI. Thesis Ph.D.), Stanford University*.

Zakoian, J.-M. (1994): "Threshold heteroskedastic models," *Journal of Economic Dynamics and Control*, 18, 931–955.

Zhang, P. (1993): "Model Selection Via Multifold Cross Validation," *Annals of Statistics*, 21, 299–313.

Zheng, S. (2012): "QBoost: Predicting quantiles with boosting for regression and binary classification," *Expert Systems with Applications*, 39, 1687–1697.

Ziegel, J. F. (2016): "Coherence and elicitability," *Mathematical Finance*, 26, 901–918.

Declaration of academic integrity

I hereby declare that I have composed this dissertation myself and without inadmissible outside help, in particular without the help of a doctoral consultant. I have used no other sources and aids than those stated. I have indicated all text passages that are incorporated, verbatim or in substance, from published or unpublished writings. I have indicated all data or information that is based on oral communication. All material or services provided by other persons are indicated as such.

Leipzig, September 29, 2021 Simon Fritzsch