Advancing Systematic and Factor Investing Strategies using Alternative Data and Machine Learning

David Happersberger

Department of Accounting and Finance
Lancaster University

A thesis submitted in fulfillment of the requirements
for the degree of
Doctor of Philosophy in Finance

September 2021

Supervisors: Prof. Ingmar Nolte
Dr. Harald Lohre

To my family Kristina, Jakob and Theo and
my parents Roswitha and Reiner

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

David Happersberger
September 2021

Acknowledgements

My PhD has been a challenging yet rewarding experience. The successful completion of this work would not have been possible without the help and support of many people to whom I am deeply thankful.

First of all, I would like to thank my supervisors Prof. Ingmar Nolte and Dr. Harald Lohre for their invaluable advice, continuous support and guidance throughout the whole journey of this PhD. I feel extremely lucky to have both as my supervisors. Their passion and enthusiasm for research are admirable and have encouraged me during all times of my PhD studies. Their levels of knowledge, patience and ingenuity are something I will always keep aspiring to. Our discussions and their insightful feedback helped sharpening my thinking and thus pushed my work to a higher level.

I would also like to express my appreciation and special thanks to my co-authors Carsten Rother and Maximilian Stroh. It was a great pleasure to work with such inspiring, skilled and likeable researchers, significantly broadening my research scope.

In the same vein, I would like to thank my colleagues from Invesco Quantitative Strategies (IQS) for many fruitful discussions and suggestions. Special thanks go to Matthias Kerling, who shared his immense expertise and experience in investment management and coding with me, and to Timo Biesenbach, who helped resolving any technical problem that came across.

Furthermore, I am grateful to Lancaster University Management School (LUMS) and the Economic and Social Research Council (UK) for providing me with generous financial support to fund my PhD. I would also like to thank all the academic and administrative staffs in the Department of Accounting and Finance of LUMS for organizing the PhD programme. With the support of the department and that of IQS I was able to participate in various seminars and conferences, which proved extremely important experiences for enhancing my research skills and building a network.

Many friends and colleagues have accompanied me during my PhD journey. There are too many names to mention here, but the past years certainly have been a very joyful and memorable experience for me because of you all. Especially, I would like to convey my appreciation and acknowledgements to Ahmad Al Gafari, Giorgio Mercantini, Yifan Li, Ananthalakshmi Ranganathan, Alexander Swade, Frederik Templiner, Enrico De Monte, Johannes Gerling and Robin Jäger for their help, support and friendship.

Finally, my lovely family deserves endless gratitude: To my parents-in-law, Nina and Willi, for their continuous encouragements and for taking care of our two little boys on numerous weekends. Their help allowed me to entirely focus on my dissertation at these times. To my parents, Reiner and Roswitha, for their love, their help, their encouragements and their endless support. For as long as I can remember, my parents have encouraged me to pursue my interests and provided me with everything needed to do so. Thank you, Mom and Dad. Likewise, to my dear siblings, Sarah and Simon, for their unconditional support, always having an open ear and giving me new perspectives on my research. Last but not least, to my wife, Kristina, and my little boys, Jakob and Theo: It is hard, if not impossible to find the right words for my gratitude and love to you. I feel extremely blessed to have you in my life. Your love, patience and understanding as well as the kids' contagious cheerfulness helped me throughout this long journey, especially through the hard times. I am deeply grateful that you kept everything away from me at all times so that I could concentrate on writing this dissertation. Without you believing in me, I never would have made it. Now, it is time to celebrate; you earned this degree right along with me.

So eine Arbeit wird eigentlich nie fertig, man muss sie für fertig erklären, wenn man nach Zeit und Umständen das Möglichste getan hat.[†]

Johann Wolfgang von Goethe

[†]Translated into English: Properly speaking, such work is never finished; one must declare it so when, according to time and circumstances, one has done one's best.

Abstract

This thesis advances systematic and factor investing strategies using alternative data and machine learning techniques. The first chapter studies the relevance of high-frequency news data for low-frequency factor investing strategies. We build various news-based equity factors for an investable global equity universe to investigate the factors' ability to extend the information inherent in standard factor models. Specifically, we document that incorporating news-based equity factors benefits multi-factor equity investments, employing diversified multi-factor equity allocations but also more dynamic factor timing strategies. The second chapter examines dynamic asset allocation strategies that focus on explicit downside risk management. We investigate suitable risk models that best inform tail risk protection strategies. In addition to forecasting portfolio risk based on standalone models such as extreme value theory or copula-GARCH, we propose a novel expected shortfall (ES) and value-at-risk (VaR) forecast combination approach that utilizes a loss function that overcomes the lack of elicitability for ES. This forecast combination method dominates simple and sophisticated standalone models as well as a simple average combination approach in terms of statistical accuracy. While the associated dynamic risk targeting or portfolio insurance strategies provide effective downside protection, the latter strategies suffer less from inferior risk forecasts, given the defensive portfolio insurance mechanics. The third chapter extends the above ES and VaR forecast combination approach using machine learning techniques. Building on a rich predictor set of VaR and ES forecasts from an array of econometric models (including GARCH, CAViaR-EVT, dynamic GAS and realized range models), we leverage shrinkage and neural network models to form combination forecasts. Such machine-learned VaR and ES forecasts outperform a set of competing forecast combination approaches in terms of statistical accuracy as well as economical relevance in dynamic tail risk protection strategies.

Table of Contents

List of Tables

List of Figures

Introduction

This dissertation addresses two salient objectives of investment management: enhance returns and reduce risk. To this end, quantitative investment processes are often guided by classic asset pricing theory (Sharpe, 1964; Lintner, 1965; Mossin, 1966; Ross, 1976), which focuses on harvesting systematic factor premiums. Well-known examples of factor premiums in the equity market include the value, momentum, quality and low-volatility premiums. The prevalence of these factor premiums has been extensively documented in the literature, and they have been shown to be robust over time and across different markets. Still, it is not a given that the factor evidence will continue to hold similarly in the future, given that markets are adaptive. It is therefore of great importance to constantly validate and evolve the employed quantitative model that guides the factor investing strategy. To improve on a given factor's definition, one can either seek to find better techniques using the same underlying, traditional data or exploit new, alternative data sources that may help to better capture the targeted factor premium. As for the latter, the advent of big and alternative data opens up tremendous opportunities. In particular, the processing of news data seems highly relevant, given that market prices ultimately aggregate all available information from news data into one figure.

In the first chapter of this dissertation, we thus investigate whether high-frequency news data can help improve low-frequency equity factor investing strategies. We build various news-based equity factors for an investable global equity universe, relating to news volume, news sentiment and further concepts. Our empirical analysis indicates that a global news sentiment factor shows promise and is most pertinent in European and emerging markets. While news sentiment factors are fairly correlated to classic momentum strategies, they are not subsumed by common equity factors and thus help expanding factor investors' opportunity set. As a result, incorporating news-based equity factors benefits diversified multi-factor equity allocation but also dynamic factor timing strategies.

Although equity factor investing strategies can deliver convincing risk and return characteristics, the associated investment risk cannot be borne by all types of investors. For instance, risk-averse investors often request further investment constraints or objectives that ultimately call for explicit downside risk management. To achieve the desired risk-return profile, one can implement risk-controlled asset allocation strategies that seek to align long-term expectations and short-term risks by actively managing exposure to risky assets such as equities. The associated asset allocation framework is typically built around three pillars: strategic asset allocation, tactical asset allocation and risk management. Strategic asset allocation targets a certain level of expected portfolio return and relies on long-term expectations of the assets' risk and return, typically over a horizon of five to ten years. The strategic asset allocation is supplemented by a tactical allocation model seeking to add value over the medium-term horizon, generally three to six months. The tactical asset allocation dynamically deviates from the strategic asset allocation weights to navigate short-term market fluctuations and takes into account the expected outperformance of risky assets in different market environments. While the hope is that a diversified strategic asset allocation together with a predictive tactical asset allocation allows navigating challenging downside markets, forecasting the latter is challenging and one thus typically brings in a third line of defense. To align the portfolio strategy with the investor's risk objective, one adds a short-term risk overlay to significantly reduce the probability of suffering from severe tail events, albeit sacrificing some of the underlying strategy's upside return potential. An effective risk control may be achieved via dynamic allocation strategies, which aim to improve the strategy's downside risk profile by switching between the underlying's portfolio current asset allocation and a risk-free asset such as a money market investment.

The success of dynamic tail risk protection strategies strongly depends on the success of forecasting tail risk. In the second chapter of this dissertation, we therefore investigate suitable risk models for timely managing the investment exposure in dynamic tail risk protection strategies. Specifically, we forecast the associated portfolio risk based on extreme value theory, expectile regression, copula-GARCH and dynamic generalized autoregressive score models. Utilizing a loss function that overcomes the lack of elicitability for expected shortfall, we further propose a novel expected shortfall (ES) and value-at-risk (VaR) forecast combination approach, which dominates simple and sophisticated standalone models as well as a simple average combination approach in modeling the tail of the portfolio return distribution. While the associated dynamic risk targeting or portfolio insurance strategies provide effective downside protection, the latter strategies suffer less from inferior risk forecasts, given the defensive portfolio insurance mechanics.

Given the importance of risk forecasting for dynamic tail risk protection strategies, we extend the above ES and VaR forecast combination approach in the third chapter and examine the ability of machine learning techniques to increase prediction accuracy. Building on a rich predictor set that contains VaR and ES forecasts from an array of econometric models, including GARCH, CAViaR-EVT, dynamic GAS and realized range models, we leverage shrinkage and neural network models to form combination forecasts. Such machine-learned VaR and ES forecasts outperform a set of competing forecast combination approaches in terms of statistical accuracy as well as economical relevance in dynamic tail risk protection strategies. Specifically, egalitarian shrinkage models demonstrate good relative accuracy in addition to convincing VaR and ES calibration backtesting results. In addition, neural network combination models are deemed relevant, particularly in recessionary periods. Still, the performance of the shrinkage models questions whether the additional complexity of the neural network models is needed. When evaluating the combination forecasts during the recent COVID-19 period, we observe lower VaR violation rates than in the global financial crisis, suggesting that the combination models have learned from previous recessions.

The three chapters in this cumulative dissertation consist of individual research papers, which I have written during my doctoral studies at Lancaster University and Invesco Quantitative Strategies in Frankfurt. The first chapter, *The Relevance of High-Frequency News Analytics for Low-Frequency Investment Strategies*, is a joint project with Carsten Rother and my supervisors Harald Lohre and Ingmar Nolte. The second chapter, *Estimating Portfolio Risk for Tail Risk Protection Strategies*, is co-authored with my supervisors and published in the *European Financial Management* journal (Happersberger, Lohre, and Nolte, 2020). The third chapter, *Combining Value-at-Risk and Expected Shortfall Forecasts using Machine Learning Techniques*, is joint work with Maximilian Stroh and my supervisors.

References

Happersberger, David, Harald Lohre, and Ingmar Nolte (2020). "Estimating portfolio risk for tail risk protection strategies". *European Financial Management* 26 (4), 1107–1146.

Lintner, John (1965). "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets". *Review of Economics and Statistics* 47 (1), 13–37.

Mossin, Jan (1966). "Equilibrium in a capital asset market". *Econometrica*, 768–783.

Ross, Stephen A (1976). "The arbitrage theory of capital asset pricing". *Journal of Economic Theory* 13 (3), 341–360.

Sharpe, William F (1964). "Capital asset prices: A theory of market equilibrium under conditions of risk". *Journal of Finance* 19 (3), 425–442.

Chapter 1

The Relevance of High-Frequency News Analytics for Low-Frequency Investment Strategies

This project is joint work with Harald Lohre, Ingmar Nolte and Carsten Rother. We thank Stefan Mittnik, Christoph Frey and the participants of the 2018 CEQURA Conference on Advances in Financial and Insurance Risk Management in Munich and the Frontiers of Factor Investing Virtual Conference 2021 for fruitful discussions and suggestions. Special thanks go to Matthias Kerling for valuable guidance on data processing. This work was supported by funding from the Economic and Social Research Council (UK).

1.1. Introduction

The proliferation of new alternative data sources opens new research avenues for enhancing investment strategies through improved return or risk forecasts. A recent route seeks to leverage news analytics that quantify textual information from news wire articles and social media using natural language processing techniques and researching the link between news and asset prices has been on the rise among both, academic scholars and industry practitioners. Tetlock (2007), Fang and Peress (2009), Heston and Sinha (2017), Engelberg, McLean, and Pontiff (2018) and Ke, Kelly, and Xiu (2019) are examples of this growing literature. While most studies concentrate on the short-term relationship between news and the cross-section of stock returns, there is only little evidence if and how news analytics can inform the practice of investing.

In contributing to this strand of research, we are particularly interested in the relevance of high-frequency news analytics for low-frequency investment strategies. First, we investigate the use of news data to construct news-based equity factors, looking into univariate tests as well as the factors' ability to extend the information inherent in standard factor models. Second, we analyze whether risk-based factor allocation strategies profit from adding news-based equity factors to a set of benchmark factors. From a dynamic factor allocation perspective, news data is used to inform the timing of standard investment factors using cross-sectional information.

For our analyses we utilize a unique global news data set that covers firm-level business news from all leading news providers and web aggregators between 2000 and 2017, collected by RavenPack News Analytics.[1] RavenPack does not only provide the flow of news articles related to a firm but also quantifies the content-relevant information in each news article based on natural language processing algorithms. In particular, it is determined which companies are mentioned in an article, how relevant the article is to a company and what the nature of the article's tone is with respect to that company. For example, a news article regarding a lapse in a company's corporate governance, a corruption scandal involving a company's executive or a bad earnings report would be associated with a negative score, whereas a news article regarding the announcement of a company's new product, a successful corporate acquisition or a positive earnings report would be associated with a positive score.

While other studies on news analytics concentrate on a single news phenomenon, we analyze various news analytics indicators in a unified framework, covering most indicators mentioned in the literature. These indicators can be divided into four news concepts: news

[1] RavenPack is a leading news data provider and its database has been used in many studies, see e.g. Kolasinski, Reed, and Ringgenberg (2013), Dang, Moshirian, and Zhang (2015), Beschwitz, Keim, and Massa (2020), and Audrino, Sigrist, and Ballinari (2020a).

volume, news sentiment, news trend and alternative news concepts. *News volume*, also referred to as media coverage or media attention, analyzes a firm's media presence (e.g., Barber and Odean, 2008; Fang and Peress, 2009). *News sentiment* was first studied by Tetlock (2007) and examines a news event's tone relating to a particular firm. *News trend* is about detecting time-series patterns in news sentiment (e.g., Leinweber and Sisk, 2011; Uhl, Pedersen, and Malitius, 2015). Alternative news concepts encompass more complex ways as to how news analytics can be used to inform investment strategies. These include the concept of news beta (Hafez, 2010) that measures the responsiveness of a firm's stock price to an aggregate news sentiment or news significance that captures both mean and variance of news sentiment.

Given the popularity of factor-based investment strategies, we first examine the predictive content of news-based indicators in the cross-section of stock returns. To this end, we form equally weighted long-short factor portfolios according to the respective news indicator using a global universe of stocks. Given the vast amount of factors available to explain the cross-section of expected stock returns, several studies (e.g. Hsu, Kalesnik, and Viswanathan, 2015; Harvey, Liu, and Zhu, 2016; Harvey, Liu, and Saretto, 2020; Arnott, Harvey, and Markowitz, 2019) emphasize the importance of following a rigorous research protocol when testing new factors in order to ensure their robustness. Accordingly, we adopt a five-step procedure when assessing the cross-sectional relevance of news analytics.

First and foremost, there needs to be a clear economic rationale behind the existence and persistence of a factor premium. There is an extensive theoretical literature (Basu, 1977; De Bondt and Thaler, 1985; Cutler, Poterba, and Summers, 1989; La Porta et al., 1997; Barberis, Shleifer, and Vishny, 1998; Daniel, Hirshleifer, and Subrahmanyam, 1998; Daniel, Hirshleifer, and Subrahmanyam, 2001) arguing that the arrival of news affects a firm's stock price. Under the biased expectations hypothesis, investors are too optimistic about some stocks and too pessimistic about others. When new information arrives in the form of a news story, investors update their expectations, resulting in a correction to the stock price (cf. Engelberg, McLean, and Pontiff, 2018). Thus, information contained in news flow data can help to predict future stock price fluctuations. Second, the cross-sectional effect should be robust to reasonable perturbations in a factor's definition (cf. Hsu, Kalesnik, and Viswanathan, 2015). We therefore build on different definitions when constructing the news-based indicators for each concept, including various look-back windows, and consider a market-capitalization weighting scheme in addition to equal-weighting when forming the long-short factors. The latter ensures that our results are not solely driven by smaller capitalized and less investable companies. Third, as suggested by Harvey, Liu, and Zhu (2016) and Harvey, Liu, and Saretto (2020), we account for multiple testing by

calculating various t-statistic thresholds a factor need to pass to be considered as statistically significant (e.g. Bonferroni, 1936; Holm, 1979; Benjamini and Hochberg, 1995; White, 2000; Benjamini and Yekutieli, 2001; Romano, Shaikh, and Wolf, 2008). This allows us to reduce the type I error, i.e. coming up with false positives which are not truly significant. Fourth, we check whether news-based factors are subsumed by common equity factors such as value and momentum or do indeed expand factor investors' opportunity set, employing mean-variance spanning tests according to Gibbons, Ross, and Shanken (1989), Cochrane (2009) and Kan and Zhou (2012). Finally, we follow Hsu, Kalesnik, and Viswanathan (2015) and analyze the cross-sectional effects of news-based factors in different regions and over multiple return horizons, complementing the literature on news-based equity factors, which is usually restricted to the US equity market and concentrates on one-month cross-sectional effects (e.g. Tetlock, 2007; Fang and Peress, 2009).

Overall, the key findings of the cross-sectional analysis are as follows: First, we document that global long-short factors based on news sentiment consistently pass the research protocol, confirming its overall relevance for the cross-section of stock returns. Still, the results differ when breaking the global universe down into its sub-regions. While the findings for Europe and emerging markets are even more pronounced than for the global universe, we do not find consistent significant cross-sectional stock return patterns of news sentiment for the US and the Japanese market. The fact that average momentum returns have historically been low in the Japanese market (see Daniel, Titman, and Wei, 2001; Hanauer, 2014) together with the finding that the momentum factor is highly correlated with news-based factors may explain the findings for the Japanese equity market. As for the US market, our findings seem to be at odds with previous studies (cf. Tetlock, 2007). This discrepancy may be explained by different underlying news data sets. While Tetlock (2007) solely analyzes a few US newspapers, our study is based on a comprehensive data set, including all types of news sources and a much longer sample period. Given the scope of the data set, our findings may be rationalized by the market structure, with the US market being simply more efficient than other markets, so that news are promptly incorporated in stock prices (see McLean and Pontiff, 2016; Jacobs and Müller, 2020).

Second, we document that only some of the news trend factors survive the stringent testing procedure. In contrast, we do not find consistent evidence that news volume or news beta factors are profitable investments, contradicting existing studies (e.g. Barber and Odean, 2008; Fang and Peress, 2009). As before, we rationalize this discrepancy by different underlying data sets.

As news sentiment-based equity factors earn significant returns and expand the investment opportunity set of common equity factors, we further investigate whether news analytics

are beneficial for multi-factor investment strategies. We first analyze whether risk-based factor allocation strategies can be enhanced by adding news-based factors to a representative set of global equity factors. Specifically, we consider an equally weighted portfolio, a minimum-variance portfolio and a risk parity portfolio. Empirically, we document that all three risk-based allocation strategies benefit from augmenting the benchmark portfolio by news sentiment-related equity factors.

Given the time variation in equity factor returns a forecasting-based factor allocation may add value over and above a diversified passive factor allocation (see e.g., Asness, 2016; Arnott, Beck, et al., 2016; Bender et al., 2018; Dichtl, Drobetz, Lohre, et al., 2019). We explore the benefits of active factor allocation when incorporating information from news flow data. To this end, we consider parametric portfolio policies that allow for timing factors based on cross-sectional information. We argue that the information contained in news flow data may help explaining the cross-section of factor returns as, similar to a stock level rationale, news data may entail information on the attractiveness of a factor itself. For example, a factor may be attractive when companies in the long leg have more positive news and/or companies in the short leg have more negative surprises. Hence, positive (net) news sentiment indicates positive factor returns. Following this rationale, we distill the set of news-based indicators on the level of equity factors to generate innovative equity factor characteristics and exploit them in the cross-sectional parametric portfolio policy framework of Brandt, Santa-Clara, and Valkanov (2009).

Indeed, news sentiment-related factor characteristics exhibit positive coefficients in a univariate parametric portfolio policy, which means that factors with positive news sentiment are overweighted in the factor allocation (relative to an equally weighted benchmark), whereas factors with negative news sentiment are underweighted. Associated performance statistics are consistently better than those for an equally weighted benchmark. When considering multiple characteristics jointly in multivariate parametric portfolio policies, it shows that factor timing allocations profit from using information contained in news flow data. The performance figures are in favor of those factor allocations that incorporate news sentiment data. We document higher risk-adjusted returns and positive information ratios for the news-related factor allocation strategies compared to an equally weighted benchmark portfolio, even after accounting for transaction costs. In addition, the coefficient for the news sentiment-related factor characteristic remains statistically significant, indicating its relevance for the dynamic factor allocation strategy.

The outline of the paper is as follows: Section 1.2 introduces the news analytics data and discusses the underlying ideas and the construction of the news-based indicators. Section 1.3 examines cross-sectional patterns in the derived news indicators. In Section 1.4,

we investigate the use of news flow data for multi-factor investment strategies. Section 1.5 concludes.

1.2. Condensing high-frequency news data into predictive indicators

1.2.1. News data

As main data source we utilize the news and sentiment data from RavenPack News Analytics. RavenPack systematically tracks, collects and analyzes real-time, firm-level business news from leading real-time news providers (including *Dow Jones Newswires*, the *Wall Street Journal*, *Barron's* and other major publishers) and web aggregators (including industry and business publications), regional and local newspapers, government and regulatory updates and trustworthy financial websites. In total, RavenPack features around 28,000 companies in over 130 countries (representing 98% of the investable global equity market) and covers news from a wide range of facts, opinions and corporate disclosures. The data is available from the year 2000 onwards, enabling us to analyze almost two decades of news data.

To transform unstructured news data items into structured granular data and corresponding scores RavenPack Analytics performs the following two steps. First, it classifies news articles into news event categories according to the RavenPack taxonomy, and both the topic and a firm's role in the news article are tagged and categorized.[2] Second, RavenPack constructs a set of scores, rating different aspects of the relevant news items with respect to the respective firm using natural language processing algorithms that effectively combine traditional linguistic analyses, financial expert consensus and market response methodologies (see RavenPack Analytics, 2017). The following four major scores form the basis of the news indicators we will build:

- *Event Sentiment Score (ESS):* A granular score between −1.00 and +1.00 that represents the news sentiment for a given company, where a negative (positive) score indicates negative (positive) sentiment and 0 indicates neutral sentiment. The ESS leverages RavenPack's event detection technology and produces a sentiment score every time an event is matched. In particular, the ESS is determined based on training sets in which experts with extensive experience and backgrounds in linguistics, finance and economics classify company-specific events and agree that these events generally convey a positive, neutral or negative sentiment (Hafez and Xie, 2011).

- *Relevance (REL):* An integer score between 0 and 100, with higher values indicating greater relevance of the underlying news story for a given company.

[2] See RavenPack's taxonomy scheme in Figure 1.C.1 for further details.

- *Event Relevance (EVR):* An integer score between 0 and 100 that reflects the relevance of the event in the story, with higher values indicating greater relevance.[3]
- *Event Similarity Days (ESD):* An integer between 0 and 365 indicating the number of days since a similar event was detected over the last 365 days. The ESD thus allows to isolate the first news article in a chain of similar articles about a given news event.

1.2.2. Global equity data

To allow for a comprehensive investigation of the news analytics data, we assemble a representative and investable equity universe encompassing the constituents of global and regional equity indices from MSCI, FTSE, S&P and STOXX. Company-specific data such as financial statement and price data are sourced from the Worldscope database. Having matched news and firm-level data, we consider a broad universe of, on average, 5350 companies per month and 1,155,342 relevant news events in the sample period from January 2000 to December 2017. On average, this translates to 94 news events per firm and month (cf. Table 1.1).

Panel A of Table 1.1 gives further descriptive statistics of the number of news events per month and firm, reflecting a company's media presence which we call news volume in the following. We only consider relevant news events and therefore require a relevance score of at least 75. This allows us to avoid clickbaits or biases by news stories just mentioning a company as reference. To get an overview of the data, we initially do not restrict the event similarity days analytics since a repeated dissemination of the same or similar news events may be a useful indication of a company's media presence. As a consequence, there is a sample maximum of 57,528 relevant news events for one company within a month. The covered company is Facebook, which went public in May 2012, constituting the biggest initial public offering in the technology sector. In the subsequent analyses, we, however, restrict the event similarity days analytics and focus on the most novel news events, see Section 1.2.3.

The positive skewness and the huge maximum number of news indicate that news volume is largely driven by company size. Indeed, large companies account for the majority of news events: large companies have, on average, 208 news events per firm and month compared to 53 and 21 news events for medium-sized and small companies, respectively (see also Figure

[3]To clarify the difference of the relevance and the event relevance score: The relevance score measures how relevant a whole news story is for a company, so that there is only one score for a story and company. In contrast, the event relevance score measures how relevant a news event is within a given news story. As there may be multiple news events for a company within a story, there may be multiple event relevance scores for a news story and company.

Table 1.1: Descriptive statistics of news data

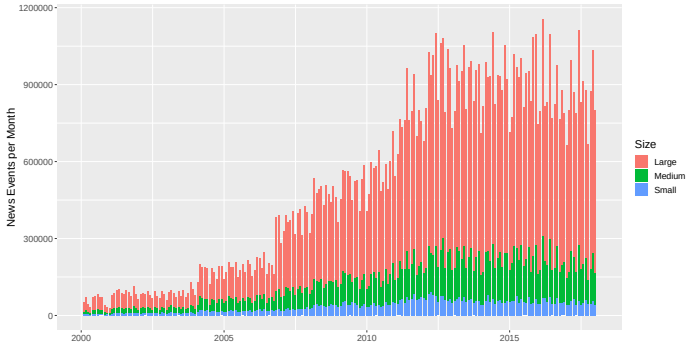
	Mean	Median	Min	Max	Sd	Skew	Kurt	Obs	Firms
Panel A: News Events									
Overall	93.95	19	1	57,528	523.17	33.22	1,739	1,155,342	5349
USA	223.41	75	1	57,528	949.97	20.51	617	272,781	1263
Japan	41.53	10	1	24,704	223.16	32.97	2,398	106,144	491
Europe	85.11	23	1	41,395	379.43	27.62	1,383	280,823	1300
RES	55.63	13	1	12,207	188.01	17.51	542	158,896	736
EM	31.03	9	1	26,325	190.88	53.95	4,169	336,698	1559
Large	208.35	57	1	57,528	859.26	20.99	685	385,191	1783
Medium	52.83	19	1	22,643	220.85	51.52	3,538	385,038	1783
Small	20.62	7	1	18,684	117.12	92.88	11,454	385,113	1783
Panel B: ESS									
Overall	0.17	0.23	-1.0	1.0	0.39	-0.52	-0.30	851,220	3941
USA	0.16	0.18	-1.0	1.0	0.34	-0.34	-0.13	250,088	1158
Japan	0.18	0.27	-1.0	1.0	0.40	-0.60	-0.44	74,719	346
Europe	0.19	0.27	-1.0	1.0	0.39	-0.63	-0.11	199,378	923
EMM	0.17	0.27	-1.0	1.0	0.41	-0.55	-0.51	222,669	1031
Large	0.19	0.22	-1.0	1.0	0.33	-0.56	0.24	283,806	1314
Medium	0.16	0.22	-1.0	1.0	0.39	-0.45	-0.38	283,667	1313
Small	0.16	0.27	-1.0	1.0	0.43	-0.50	-0.67	283,747	1314

This table shows the descriptive statistics of news volume (Panel A) and the average event sentiment score (Panel B) per month and firm. For news volume, i.e. the number of news events per month, we require a relevance score above 75. For the ESS we require an (according to the RavenPack taxonomy) assigned and non-neutral ESS score as well as a relevance, event relevance and event similarity score above 90. For each panel, we show the overall statistics as well as statistics for the regions USA, Japan, Europe and emerging markets (EM) and for large, medium-sized and small firms (in terms of market-capitalization). We show the following statistics: mean, median, minimum (Min), maximum (Max), standard deviation (Sd), skewness (Skew) and kurtosis (Kurt). Obs is the total number of observations and Firms gives the average number of firms per month. The time period spans from January 2000 to December 2017.

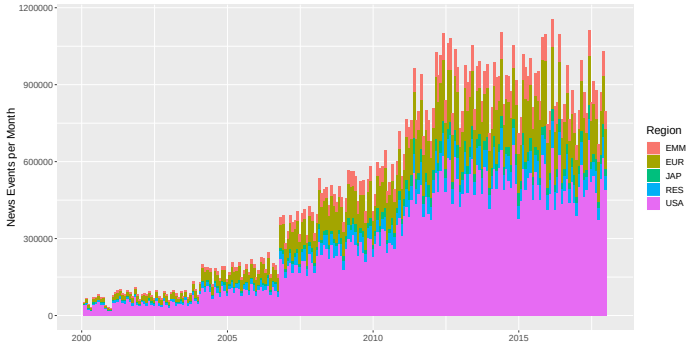
1.1a).[4] This fact is not only consistent with the literature on media and news indicating that large firms attract higher media attention but is also aligned with the intuition that large firms typically generate more news events (e.g. Ke, Kelly, and Xiu, 2019). To control for size effects, we will standardize the derived news indicators by market capitalization going forward (see details in Section 1.2.3).

Figure 1.1a shows the evolution of news volume over the sample period. The number of news articles increases substantially from the beginning of the sample in 2000 to the year

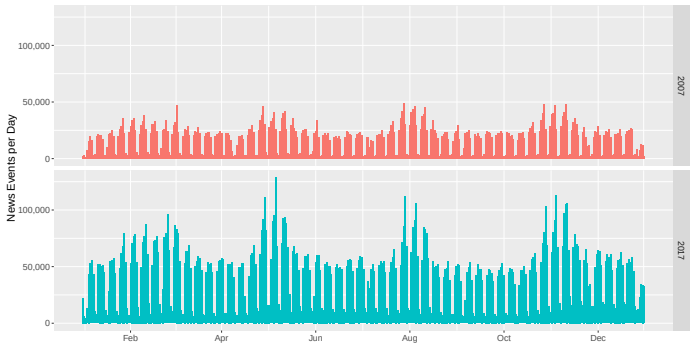
[4]We divide the universe of companies into three size buckets according to their market capitalization. That means, large companies refer to companies that are in the upper tercile, medium-sized and small companies in the middle and lower tercile, respectively.



(a) News volume per size



(b) News volume per region

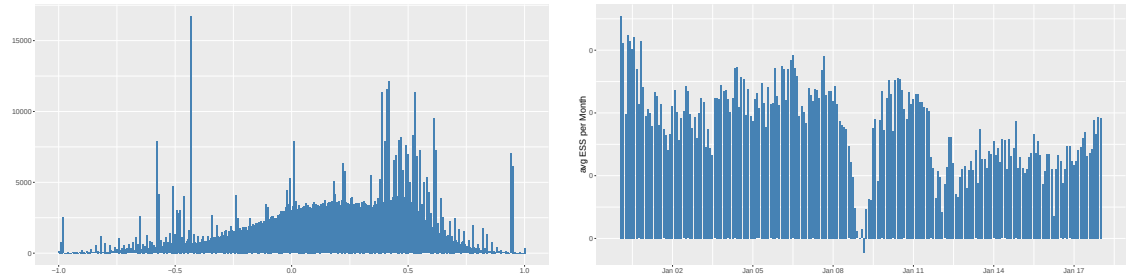


(c) Seasonality of news volume

Figure 1.1: Characteristics of news volume. This figure illustrates various characteristics of news volume over the sample period from January 2001 to December 2017. Panel (a) shows monthly news events allocated to the following regions: United States (USA), Japan (JAP), Europe (EUR), emerging markets (EM) and rest of the world (RES). Panel (b) shows news volume per market capitalization (large, medium-sized and small companies). Panel (c) illustrates the yearly pattern of daily news events for the years 2007 and 2017.

2012, but stabilizes afterwards. In addition to RavenPack's changing media coverage, this time-series pattern is driven by both an increasing intensity of media coverage and a growing amount of firm activities. Figure 1.1b shows the evolution of the number of monthly news events per region. We differentiate between United States (USA), Japan (JAP), Europe (EUR)

and emerging markets (EM).[5] It is not surprising that US stocks exhibit the largest fraction of news events, followed by European stocks (cf. Table 1.1). Figure 1.1c shows the number of daily news events over the years 2007 (upper part) and 2017 (lower part), conveying two distinct seasonal patterns: first, we observe a quarterly cycle that coincides with quarterly business reports (earnings announcements etc.).[6] Second, we observe a weekly cycle which shows a significantly reduced news dissemination on weekends. We control for both effects when constructing our indicators.



(a) Histogram of ESS

(b) ESS over the sample period

Figure 1.2: Characteristics of news sentiment. Panel (a) shows the histogram of the ESS, whereas Panel (b) shows the monthly average event sentiment score across all firms. The sample period spans from January 2000 to December 2017.

To explore the characteristics of the event sentiment score we examine Panel B of Table 1.1. The number of ESS scores and firms is lower than the number of news events for two reasons: first, an event sentiment score is only assigned to a news event when it can be classified according to the RavenPack taxonomy. Second, we exclude news events with a neutral ESS score of zero and require the ESS to pass filters of 90 for relevance, event relevance and novelty to further reduce noise (see Section 1.2.3 for more details on noise filtering). We document that sentiment is slightly positive on average: the ESS has a mean of 0.17 and a median of 0.23, respectively. Panel (a) of Figure 1.2 shows the histogram of all event sentiment scores, when applying the described filters. We observe a slightly negative skewed and fat-tailed distribution.[7] Panel 1.2b shows the evolution of the monthly

[5] Emerging markets include those countries that are classified as emerging market by MSCI, FTSE, S&P, and STOXX. This classification is time-dependent. Emerging market countries are, for example, Brazil, Russia and India.

[6] As a robustness check, we perform an analysis excluding news events corresponding to earnings announcements when constructing the set of news indicators. Unreported results do not show significant differences to the results including earnings announcements data, suggesting that the analysis of news-based indicators is not solely driven by events concerning quarterly business reports.

[7] According to the news data provider RavenPack News Analytics, the peaks in the histogram of the ESS correspond to the default sentiment scores that are assigned to events where no additional information was found that allowed for a more granular choice within the assigned sentiment range. These peaks occur more

ESS score averaged across firms, which is fairly stable with the exception of the time period of the global financial crisis in 2008.

1.2.3. News-based indicators

In this section, we develop a broad set of indicators that aim to explain and predict asset price fluctuations utilizing information extracted from news flow data. The general use of news data for this purpose can be rationalized via the efficient markets hypothesis of Malkiel and Fama (1970), which can be seen as the theoretical basis for any return prediction analysis. Therein, market efficiency predicts that the expected return of a stock is dominated by unforecastable news, as this news is rapidly (in its strongest form, immediately) and fully incorporated in its price. An alternative hypothesis is that information in news flow data is not fully absorbed by market prices instantaneously, for reasons such as limits-to-arbitrage and limited investor attention (e.g. Baker and Wurgler, 2006; Tetlock, 2007; Ke, Kelly, and Xiu, 2019). Under the biased expectations hypothesis (see e.g. Engelberg, McLean, and Pontiff, 2018), investors are generally too optimistic about some stocks and too pessimistic about others. When new information arrives in the form of a news story, investors update their beliefs, resulting in a correction to the stock price. Thus, information contained in news flow data can be predictive of future stock prices. While this alternative hypothesis is by now considered uncontroversial for very short horizons (e.g. daily or intradaily horizons), it is still not clear whether low-frequency investors can profit from information embedded in news flow data, facing investment horizons of one month or longer. Our analysis adds new evidence to the empirical literature investigating whether this alternative hypothesis also holds for longer horizons.

In computing news-based indicators, we first filter the news data to reduce the noise. In particular, we only include firms with at least one news story during our sample period. While it seems favorable to include as much information as possible (i.e. keep as many news events as possible), not all events are equally important. Therefore, we exclude news stories with $ESS = 0$ and filter the data based on relevance, event relevance and event similarity days according to Hafez (2010), Kolasinski, Reed, and Ringgenberg (2013), Dang, Moshirian, and Zhang (2015) and Beschwitz, Keim, and Massa (2020). Specifically, we only consider stories that are directly relevant to the mentioned company by only retaining data with a relevance score above 90. Therefore, we make sure to focus on the salient news items for each company. In a similar way, we only retain events with high relevance in a news story to avoid carrying unimportant news items, i.e. we require the event relevance score to be above

often for firms with small media coverage (and a small volume of news events) than for firms with large media coverage (and a large volume of events).

90. Furthermore, we only consider unique and novel news events for most of the indicators. We hypothesize that the first instance of an event is most impactful and any subsequent repetition thereof can be expected to have a lesser impact. By retaining only events that have an event similarity days analytic above 90, we filter our data set down to only the most novel events within the last 90 days. As such, any analysis of the news events is less likely to be driven by the repetitive dissemination of the same or similar news events. Still, we also investigate indicators that are less restrictive in terms of novelty, given that under the prospect theory and as suggested by behavioral studies, there is the possibility that repetitive news may change and force market participants to alter their attitude and trading strategies.[8]

In general, we proceed as follows when constructing a given news indicator: As our main analysis is conducted at a monthly frequency, we first aggregate the high-frequency news tick data to monthly indicators using indicator-specific functions. In this process, we calculate each indicator for each firm in our investment universe using various look-back windows. As the required information differs among indicators, not all indicators are based on the same number of firms. To alleviate robustness concerns we require a minimum number of 300 firms in each month when deriving the indicators. Furthermore, this allows us to test the predictive ability of the indicator itself and reduce the stock-specific impact when constructing the indicator. Second, as industries tend to perform differently across the business cycle and may also be at different stages in their life cycle, it seems reasonable to assume that the information extracted from news flow data is likely to reflect the broad industry context, potentially confounded with cues about firm-specific performance. For this reason, we settle for a standardization based on industry classifications by subtracting industry averages and dividing by the industry-specific standard deviation, limiting the impact from industry bets. Third, since a firm's news volume and news sentiment are likely driven by company size, we cross-sectionally neutralize the indicators by their market capitalization. Appendix 1.A gives further details on how we construct the individual news indicators.

The indicators that we derive from news flow data relate to various studies from the existing literature on news analytics and can be categorized into four broad concepts when building predictive indicators: News Volume, news sentiment, news trend and alternative news concepts. We describe each of these concepts in the following.

[8]We tested various filters around a value 90, but do not find significant differences in our results. Hence, we follow the studies from Hafez (2010), Kolasinski, Reed, and Ringgenberg (2013), Dang, Moshirian, and Zhang (2015) and Beschwitz, Keim, and Massa (2020) that also use RavenPack news flow data and perform analyses suggesting that RavenPack is good at identifying both relevance and sentiment of an article. Notably, for some indicators we deviate from REL, EVR and ESD filters of 90 for indicator-specific reasons. For further information see the detailed indicator description in Appendix 1.A.

News volume

News volume captures a firm's media presence measured by the number of news events within a specific time window. Following the literature (Chan, 2003; Barber and Odean, 2008; Da, Engelberg, and Gao, 2011; Hillert, Jacobs, and Müller, 2014) we investigate the "attention grabbing hypothesis", which states that investors are net buyers of stocks with high media presence.[9] Indeed, Barber and Odean (2008) find that returns of attention-grabbing stocks are (temporarily) higher than those of firms with low (or without) media presence. By using different time horizons (1, 3, 6 months), we analyze the persistence of the attention-grabbing effect, investigating whether an associated premium can be harvested by long-term investors. In addition to the standard filter settings (REL>90, EVR>90, ESD>90), we also examine a less restrictively filtered news volume indicator (REL>75) to check the effects related to the number of appearances of the same or similar news event across different news providers. The argument is that repetitive news may change and force market participants to alter their attitude and trading strategies, as proposed by the well-known known prospect theory and other behavioral rationales.

News sentiment

News sentiment synthesizes a news event's tone with respect to a particular firm. Positive sentiment corresponds to a news event that portrays positive surprises and opinions, resonating with generally good news or an outcome that is better than expected. Numerous studies (e.g., Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Heston and Sinha, 2017; Wang, Zhang, and Zhu, 2018; Ke, Kelly, and Xiu, 2019) demonstrate that firms' news sentiment contains information relevant for predicting the cross-section of stock returns. For instance, Tetlock (2007) shows that high media pessimism, i.e. negative sentiment, forecasts falling stock market prices.[10] In this vein, we construct four different firm-specific sentiment indicators. First, a straightforward indicator simply computes the monthly average of the event sentiment score over various look-back periods. Second, we construct a more robust version that does not depend on the magnitude of the event sentiment score emerging from the proprietary model of the news data provider.[11] Specifically, we divide the number of news events with a positive event sentiment score by the number of news events with a

[9] Conversely, other studies advocate the slogan "no news is good news". For instance, Fang and Peress (2009) find that stocks with no media presence earn higher returns than stocks with high media presence even after controlling for well-known risk factors.

[10] For a detailed literature review on news sentiment see Uhl, Pedersen, and Malitius (2015) or Coqueret (2020).

[11] To ensure the validity of the ESS provided by RavenPack Wang, Zhang, and Zhu (2018) compute a simple sentiment score using common text processing techniques as a robustness check. Their findings show that both sentiment scores provide similar results.

negative event sentiment score. The third news sentiment indicator seeks to exploit the sentiment score evolution over time by putting larger weight on more recent sentiment scores in the look-back window, because more recent news events might be more relevant than older news events. Finally, prospect-theory and behavioral finance literature typically argues that the market reaction to negative news is generally stronger than the reaction to positive news, which is empirically shown by Hafez, Guerrero-Colon, and Duprey (2015) and Heston and Sinha (2017). In this vein, we construct a firm-specific news sentiment indicator that gives higher weights to negative than to positive news, employing a weighting scheme that is based on the prospect theory of Tversky and Kahneman (1992).

News trend

News trend relates to changes in news sentiment rather than its average level. Analyzing associated time-series patterns, Leinweber and Sisk (2011) and Uhl, Pedersen, and Malitius (2015) argue that longer-term news sentiment cycles exist and can be exploited for return predictions and investment strategies, documenting that a positive trend in a firm's news sentiment has a positive impact on its future returns. To reduce noise and enable identifying longer-term trends in the news-sentiment indicator we follow Uhl, Pedersen, and Malitius (2015) and use a frequency filter to construct a corresponding news sentiment momentum indicator.[12] More simplistic approaches to pinpoint time trends are (1) to compare the distribution of the ESS between two different points in time (similar to a simple t-statistic of a change in ESS) or (2) to regress the cumulative ESS on the time index.

Alternative news concepts

Alternative news concepts covers the indicators news beta, news dispersion and news significance. *News beta* measures the sensitivity of a firm's stock return to changes in market sentiment. To this end, we calculate an overall market news sentiment by averaging the ESS across firms for each month. The idea is that positive news beta stocks, on average, outperform the market while negative news beta stocks tend to underperform (Hafez, 2010). *News dispersion* looks at the intraday variation of the ESS, while *news significance* captures both mean and variation of the ESS within a specific time horizon.

[12] We employ the cumulative sum (CUSUM) frequency filter to reduce the noise, see Appendix 1.A and Uhl, Pedersen, and Malitius (2015) for details.

1.3. News analytics and the cross-section of stock returns

If a certain firm characteristic is hypothesized to be relevant for the cross-section of stock returns, a corresponding long-short portfolio can be constructed to proxy for the underlying unknown factor. Given the biased expectation hypothesis (cf. Section 1.2.3), we therefore form long-short portfolios of stocks sorted on the proposed news indicators to examine the cross-sectional relevance of news analytics in a simple, non-parametric way. Specifically, we monthly divide the stock universe into quintile portfolios based on the prevailing scores of the selected news indicator and compute the equally weighted average return of each portfolio during the following month.[13] If the information embedded in the news indicator was already incorporated in stock prices, then the top quintile portfolio return should be similar to that of the bottom quintile portfolio. To test the pricing implications of news, we therefore form zero-investment trading strategies that are long in stocks with the highest news scores and short in stocks with the lowest news scores. Consequently, the ultimate long-short factor portfolio return emerges as the return difference between the top and bottom quintile portfolio returns.[14]

In this section, we test news-based equity factors for the global universe of stocks. Following Hsu, Kalesnik, and Viswanathan (2015), Harvey, Liu, and Zhu (2016), Arnott, Harvey, and Markowitz (2019) and Harvey, Liu, and Saretto (2020), we apply stringent criteria for qualifying factors, including multiple testing hurdles, mean-variance spanning tests and robustness across different regions and over multiple return horizons.

1.3.1. A robust framework to detect relevant news indicators

When testing 20 randomly selected factors, one factor will likely exceed the two-sigma threshold (t-statistic of 1.96 or above) by chance alone. Obviously, the standard t-statistic of 1.96 is not appropriate if more than one factor is tested (e.g. Arnott, Harvey, and Markowitz, 2019) and the early multiple testing literature in finance (e.g. Lo and MacKinlay, 1990b) already emphasized the importance to increase the t-statistic threshold to avoid false discoveries. Harvey, Liu, and Zhu (2016) and Harvey, Liu, and Saretto (2020) provide a

[13] We follow an equal-weighting scheme when forming long-short portfolios, because it is a simple and robust way of assessing the news indicators' predictive power across the firm size spectrum. Anecdotally, it is also close to the way that hedge funds use news text for portfolio construction (cf. Ke, Kelly, and Xiu, 2019). Notwithstanding, we will also use a market capitalization weighting scheme as robustness check.

[14] Note that we use the notion of long-short portfolios as well as factors interchangeable in the following, being aware that long-short portfolios are a tool to proxy the underlying unknown factor driving the cross-section of stock returns.

catalogue of different approaches to address the multiple testing problem.[15] These methods are designed to control false discoveries by either limiting the probability of a given number of false discoveries (i.e. the family-wise error rate) or controlling the proportion of false discoveries relative to the total number of discoveries (i.e. the false discovery rate).

We consider the following four approaches that strictly control the family-wise error rate (FWER) by allowing only one false discovery: (i) Bonferroni (1936), (ii) Holm (1979), (iii) the bootstrap reality check (BRC) of White (2000) and (iv) the StepM approach of Romano and Wolf (2005). The popular Bonferroni and Holm procedures asymptotically control the FWER under particular conditions and tend to do poorly in extreme situations, such as when tests exhibit negative correlation. In contrast, the BRC and the StepM procedures allow for any arbitrary dependence in the test statistics as they rely on resampling. However, when the number of factors being tested increases, these methods become more and more stringent (i.e., they lead to fairly high t-statistic thresholds). To alleviate this concern, the k-StepM method of Romano, Wolf, et al. (2007) extends the concept of FWER to allow for control of any arbitrary number of false discoveries, the so called k-FWER. A direct extension of the k-FWER is the idea of controlling the proportion of false discoveries (FDP). Romano, Shaikh, and Wolf (2008) introduce the FDP-StepM procedure, which is a sequence of k-StepM tests.

Instead of controlling the probability that the FDP is less than or equal to a threshold, one may also choose to control the average realized FDP, that is, the false discovery rate (FDR). The two main methods to control the FDR are from Benjamini and Hochberg (1995) (BH) and its extension by Benjamini and Yekutieli (2001) (BY). While the BH procedure controls the FDR under the assumption that test statistics are independent, the BY procedure allows for more general dependence. The former is therefore stricter (but less powerful) than the latter.

Given that some of the tested news factors are highly correlated (as we consider different factor definitions of the same news phenomena), we concentrate on multiple testing methods that allow for any dependence in the test statistics.[16]

[15] In statistics, multiple testing refers to simultaneous testing of more than one hypothesis. Biases arising from the multiple testing problem are known as data snooping, data fishing, data dredging or p-hacking (see Harvey, Liu, and Zhu, 2016; Harvey, Liu, and Saretto, 2020).

[16] In the tables, we provide the multiple testing hurdles of the BRC, the StepM, the FDP-StepM and the BH procedure. The methods not reported do deliver similar results. For more details on the described multiple testing procedures, see Harvey, Liu, and Saretto (2020).

1.3.2. News-based equity factor evidence

In Table 1.2 we report performance statistics for news-based equity factors constructed in a global investment universe. Given that the underlying news data ranges from 2000 to 2017 and the computation of indicators consumes up to the last twelve months of data, we report monthly scores and results from 2001 to 2017.

Table 1.2: News equity factors: Global universe

								Multiple testing					
Indicator	Return	Sd	Min	Max	SR	MDD	t-stat	Φ	Holm	BRC	FDP	BY	Firms
News Volume													
$VOL_{\text{REL}>75,1}$	-0.38	4.23	-4.70	9.12	-0.09	-14.46	-0.38						3421
VOL_1	0.53	2.90	-3.10	5.29	0.18	-7.81	0.76						2772
VOL_3	0.96	4.34	-3.56	10.46	0.22	-8.81	0.92						3576
VOL_6	0.80	5.00	-4.43	11.15	0.16	-16.20	0.66						3774
News Sentiment													
$SENT_1$	3.81	3.55	-6.32	3.38	1.07	-10.55	4.42	✓	✓	✓	✓	✓	2646
$SENT_3$	4.29	4.24	-7.34	4.25	1.01	-12.34	4.17	✓	✓	✓	✓	✓	3535
$SENT_6$	4.16	4.70	-8.09	4.55	0.88	-15.43	3.65	✓	✓	✓	✓	✓	3751
$rSENT_{l=u=0,1}$	3.30	3.35	-6.18	3.10	0.99	-8.45	4.06	✓	✓	✓	✓	✓	2646
$rSENT_{l=u=0,3}$	3.75	3.85	-5.62	3.93	0.97	-9.08	4.01	✓	✓	✓	✓	✓	3535
$rSENT_{l=u=0,6}$	3.70	4.42	-7.72	4.39	0.84	-14.74	3.45	✓	✓	✓	✓	✓	3751
$wSENT_{td,1}$	3.67	3.44	-5.29	3.42	1.07	-8.33	4.39	✓	✓	✓	✓	✓	2646
$wSENT_{td,3}$	4.74	3.77	-5.23	4.49	1.26	-8.94	5.18	✓	✓	✓	✓	✓	3535
$wSENT_{td,6}$	4.52	4.54	-7.08	4.45	1.00	-14.01	4.11	✓	✓	✓	✓	✓	3751
$wSENT_{as,1}$	3.37	3.84	-6.50	2.93	0.88	-11.25	3.62	✓	✓	✓	✓	✓	2646
$wSENT_{as,3}$	3.93	5.75	-10.62	4.80	0.68	-22.07	2.82	✓				✓	3535
$wSENT_{as,6}$	4.48	6.17	-11.33	5.51	0.73	-23.35	3.00	✓	✓			✓	3751
News Trend													
$SENTMOM$	2.71	2.79	-6.16	1.94	0.97	-9.59	4.00	✓	✓	✓	✓	✓	2676
$aSENTMOM_3$	1.38	3.15	-3.77	3.21	0.44	-10.98	1.80						2103
$aSENTMOM_6$	1.66	3.13	-4.04	2.56	0.53	-10.55	2.18	✓					2806
REG_6	0.77	4.93	-7.26	5.02	0.16	-17.61	0.65						847
REG_{12}	0.47	3.85	-5.84	2.29	0.12	-17.12	0.50						1928
Alternative News Concepts													
$NEWSBETA$	2.67	4.27	-5.52	4.83	0.62	-9.91	2.58	✓					2093
$DISP_1$	1.18	5.90	-4.22	12.32	0.20	-13.39	0.82						2080
SIG_1	2.38	3.81	-6.86	2.94	0.62	-14.00	2.57	✓					2034
SIG_3	3.89	4.44	-7.13	3.81	0.88	-13.06	3.62	✓	✓	✓	✓	✓	3287
SIG_6	3.78	4.82	-9.01	3.63	0.79	-17.00	3.24	✓	✓	✓	✓	✓	3629

This table shows performance statistics of equally weighted long-short portfolios for a set of news indicators using the global stock universe. Annualized mean returns are calculated using the arithmetic average of simple returns. Standard deviation (Sd) and Sharpe ratio (SR) are annualized through multiplication by $\sqrt{12}$. Min and Max denote the lowest and highest monthly excess return in the sample period. MDD is the maximum drawdown. Mean return, Sd, Min, Max and MDD are given in percentage points. The last column gives the average number of firms per month. t-stat is the t-statistic for testing against the Null of a zero mean return. To address the multiple testing problem, we show whether a factor passes common t-statistics thresholds (✓) such as: the usual value of 1.96 of the standard normal distribution (Φ), 2.97 based on Holm (1979), 3.13 using the bootstrap reality check of White (2000) (BRC), 3.02 using the FDP-StepM procedure of Romano, Shaikh, and Wolf (2008) (FDP) and 2.71 using the method of Benjamini and Yekutieli (2001) (BY) for a significance level of 5%. The time period spans from January 2001 to December 2017.

We find all long-short factors based on news volume to deliver statistically insignificant returns over the sample period, questioning the "attention grabbing hypothesis" put forward

by Barber and Odean (2008). This outcome may well be rationalized by important differences in the underlying data: While Barber and Odean (2008) (and others) rely on a few US newspapers, we analyze a significantly broader data set including all types of news sources and covering a longer sample period.

Next, we find significant long-short portfolio returns based on news sentiment indicators. For instance, the simple news sentiment factor at one month horizon, $SENT_1$, earns an annualized return of 3.81% at 3.55% annualized volatility. Irrespective of the factor definition, the ensuing return differentials between positive and negative news sentiment companies are statistically significant. Still, we document performance differences among the news sentiment-based global equity factors. Specifically, we find that a higher degree of sophistication in estimating news sentiment is rewarded. The ESS-based average sentiment factors earn higher monthly returns than the sentiment factors that merely build on the nature of a news event (positive versus negative). For instance, $SENT_1$ has a 51 basis points (bps) pick-up in monthly return relative to $rSENT_{l=u=0,1}$. Still, performance can be further enhanced by weighting the individual news events. For example, the news sentiment factor that gives higher weight to more recent news events ($wSENT_{td,3}$) earns a monthly return of 4.74% at a three month time horizon (compared to 4.29% for $SENT_3$). This finding is in line with the economic intuition that more recent news are more relevant in driving investor's decisions and ultimately stock prices (Beschwitz, Keim, and Massa, 2020). Notably, such return benefits do not result from higher risk. In terms of Sharpe ratio, risk-adjusted returns range from 0.68 ($wSENT_{as,3}$) to 1.26 ($wSENT_{td,3}$). The performance of the factors differs across look-back windows. While we observe higher monthly returns for three and six month horizons (e.g. 4.29% for $SENT_3$ and 4.16% for $SENT_6$ vs. 3.81% for $SENT_1$), the longer-horizon factor's return comes with an increase in risk (higher volatility and maximum drawdown). Risk-adjusted returns are highest for one and three month horizons. Except for $wSENT_{as,3}$ and $wSENT_{as,6}$, all news sentiment-based equity factors pass both the conventional two-sigma threshold of 1.96 and all multiple testing hurdles (e.g. 3.13 for the BRC or 3.02 for FDP-StepM approach). Overall, we document that stocks with higher news sentiment earn higher returns than stocks with lower news sentiment which is consistent with existing studies (e.g. Tetlock, 2007) but measured against a broader news data set and various factor definitions of news sentiment.

Concerning news trend factors, we only find the news sentiment momentum factor *SENTMOM* to exhibit statistically significant returns. Its long-short factor return is 2.71%, which corresponds to a Sharpe ratio of 0.97, and survives all multiple testing thresholds. All other factors that use simpler approaches to capture changes in news sentiment do not deliver convincing results. This finding suggests that a sophisticated technique that is able to

reduce the noise in the signal, such as the CUSUM frequency filter, is required to capitalize on news trend signals.

As for the alternative news concept indicators, both news beta and news significance factors exhibit statistically significant long-short returns when using the conventional two-sigma threshold of 1.96. Yet, when increasing the t-statistic threshold according to the multiple testing procedures, only the news significance factors using longer horizons (three and six months) do survive. This result is expected, as news sentiment is noisy and both indicators, news beta as well as news significance, get more stable with increasing time horizons. Having a clean market news sentiment for calculating beta helps in smoothing the overall signal, which therefore exhibits a higher predictive ability.

As a robustness check, we compare factors' performance based on equal weights with that of market capitalization weights, allowing to gauge the practical relevance of our findings. Table 1.3 reports the results of the capitalization-weighted long-short portfolios, showing similar patterns like their equally weighted counterparts. Overall, factor portfolios related to news sentiment have good performance, yet statistical significance is slightly reduced. While the conventional threshold of 1.96 is passed by all news sentiment factors, some are now failing the multiple testing hurdles. Consequently, news sentiment data is more predictive for future returns to small stocks' returns, all else being equal. Ke, Kelly, and Xiu (2019) provide a number of potential economic explanations for this observation: First, small stocks receive less investor attention and hence their prices respond more slowly to news. Second, small stocks' underlying fundamentals are more uncertain and opaque and hence it requires more effort to process associated news into information that can be used for assessing stock prices. Third, small stocks are less liquid and take longer time to trade and thus to incorporate information into prices.

1.3.3. Mean-variance spanning

Factor-based investment managers usually build on a comprehensive set of factors to enjoy the benefits of factor diversification. Hence, it is crucial to evaluate whether the proposed news factors expand the investor's investment opportunity set. Figure 1.C.2 in the Appendix shows the return correlation matrix of the news factors including a set of common equity factors, namely the Fama and French (1992, 2006) factors as well as the price momentum and short-term reversal factors of Jegadeesh and Titman (1993).[17] By construction, most news factors are highly correlated within their concept category. We further find the momentum factor to be highly correlated with some of the news sentiment factors, suggesting that news factors partly reflect the information embedded in momentum indicators.

[17] See Table 1.B.1 for a definition of the set of equity factors.

Table 1.3: News equity factors: Capitalization-weighting

								Multiple hypothesis testing					
Indicator	Return	Sd	Min	Max	SR	MDD	t-stat	Φ	Holm	BRC	FDP	BY	Firms
News volume													
$VOL_{\text{REL}>75,1}$	-0.79	2.83	-3.56	4.70	-0.28	-15.77	-1.15						3421
VOL_1	0.25	2.33	-2.45	2.75	0.11	-8.62	0.44						2772
VOL_3	0.62	3.21	-3.13	6.42	0.19	-8.91	0.79						3576
VOL_6	0.40	3.58	-3.94	6.60	0.11	-14.86	0.46						3774
News sentiment													
$SENT_1$	2.67	2.93	-4.93	3.00	0.91	-7.99	3.75	✓	✓	✓	✓	✓	2646
$SENT_3$	3.06	3.73	-6.35	4.17	0.82	-11.14	3.39	✓	✓	✓	✓	✓	3535
$SENT_6$	3.15	4.27	-7.32	4.54	0.74	-13.59	3.04	✓		✓	✓	✓	3751
$rSENT_{l=u=0,1}$	2.44	2.72	-4.54	2.52	0.90	-6.64	3.70	✓	✓	✓	✓	✓	2646
$rSENT_{l=u=0,3}$	2.54	3.35	-4.77	3.91	0.76	-9.14	3.12	✓	✓	✓	✓	✓	3535
$rSENT_{l=u=0,6}$	2.63	4.01	-6.89	4.39	0.66	-12.77	2.71	✓					3751
$wSENT_{td,1}$	2.58	2.81	-4.11	2.98	0.92	-6.06	3.78	✓	✓	✓	✓	✓	2646
$wSENT_{td,3}$	3.73	3.24	-4.64	4.50	1.15	-7.41	4.74	✓	✓	✓	✓	✓	3535
$wSENT_{td,6}$	3.46	4.02	-6.54	4.64	0.86	-12.50	3.55	✓	✓	✓	✓	✓	3751
$wSENT_{as,1}$	2.64	3.21	-5.86	2.65	0.82	-8.76	3.39	✓	✓	✓	✓	✓	2646
$wSENT_{as,3}$	2.67	5.35	-10.03	4.71	0.50	-21.00	2.06	✓					3535
$wSENT_{as,6}$	3.44	5.72	-10.25	5.33	0.60	-22.08	2.48	✓					3751
News trend													
$SENTMOM$	1.49	2.30	-4.30	1.87	0.65	-6.95	2.66	✓					2676
$aSENTMOM_3$	0.00	2.72	-3.09	3.06	0.00	-9.44	-0.01						2103
$aSENTMOM_6$	0.51	2.64	-3.09	2.18	0.19	-10.32	0.80						2806
REG_6	0.73	4.04	-5.84	5.44	0.18	-16.26	0.75						847
REG_{12}	0.13	3.16	-5.11	1.83	0.04	-14.10	0.17						1928
Alternative news concepts													
$NEWSBETA$	2.94	4.72	-5.55	4.99	0.62	-11.82	2.57	✓					2093
$DISP_1$	0.74	4.20	-3.10	6.92	0.18	-12.91	0.73						2080
SIG_1	1.39	3.03	-5.26	2.24	0.46	-10.64	1.90						2034
SIG_3	1.91	3.63	-5.85	2.98	0.53	-12.53	2.17	✓					3287
SIG_6	2.23	4.10	-8.11	2.95	0.55	-14.10	2.25	✓					3629

This table shows performance statistics of market capitalization-weighted long-short portfolios for a set of news indicators using the global stock universe. Annualized mean returns are calculated using the arithmetic average of simple returns. Standard deviation (Sd) and Sharpe ratio (SR) are annualized through multiplication by $\sqrt{12}$. Min and Max denote the lowest and highest monthly excess return in the sample period. MDD is the maximum drawdown. Mean return, Sd, Min, Max and MDD are given in percentage points. The last column gives the average number of firms per month. t-stat is the t-statistic for testing against the Null of a zero mean return. To address the multiple testing problem, we show whether a factor passes common t-statistics thresholds (✓) such as: the usual value of 1.96 of the standard normal distribution (Φ), 3.08 based on Holm (1979), 2.99 using the bootstrap reality check of White (2000) (BRC), 2.94 using the FDP-StepM procedure of Romano, Shaikh, and Wolf (2008) (FDP) and 2.87 using the method of Benjamini and Yekutieli (2001) (BY) for a significance level of 5%. The time period spans from January 2001 to December 2017.

To statistically examine whether news-based equity factors are subsumed by the set of common equity factors or help expanding an investor's opportunity set, we employ mean-variance spanning tests, see Gibbons, Ross, and Shanken (1989), Cochrane (2009) and Kan and Zhou (2012). These tests basically boil down to examining whether adding

factors to a set of benchmark factors improves the tangency portfolio. At their heart, the tests regress the returns of the news factors, $r_{N,t}$, on the returns of a set of benchmark factors, $r_{b,t}$:

$$r_{N,t} = \alpha + \sum_{b=1}^{B} \beta_b r_{b,t} + \varepsilon_t. \tag{1.1}$$

where α and β_b are the regression coefficients and ε_t is an independent and identically distributed innovation term with mean zero and unit variance. If a given news factor is fully explained by the set of benchmark factors, the estimated alpha $\hat{\alpha}$ should be insignificant. Gibbons, Ross, and Shanken (1989), Cochrane (2009) and Kan and Zhou (2012) propose different test statistics to test the null hypothesis of $H_0^1 : \alpha = 0$ (that are, the GRS, GMM and F1 step-down test).[18] Kan and Zhou (2012) add a second test that investigates whether the added factors benefit the global minimum-variance portfolio (F2 step-down test). The corresponding null hypothesis is $H_0^2 : \delta = 1 - \sum_{b=1}^{B} \beta_b = 0$. To this end, it imposes the restriction of $\alpha = 0$. Splitting up the hypotheses in this fashion allows to draw conclusions about the nature of the potential benefit of the news factors.

Table 1.4 provides the mean-variance spanning results for those news factors that pass multiple testing hurdles, where the set of common equity factors makes up the benchmark factors. We report regression statistics according to Equation (1.1) as well as the test statistics of the four spanning tests (GRS, GMM and the two F-tests). We find all factor alphas for news sentiment, news sentiment momentum and news significance to be significant at the 1% level, suggesting that this set of news factors may help explaining the cross-section of stock returns. Evaluating the R^2_{adj} we learn that the degree of added value decreases with the length of the respective factor's underlying time horizon: for instance, over 70% of the returns of the $SENT_6$ news factor can be explained by common equity factors (compared to only 43% for $SENT_1$). In line with the correlation analysis, we report significant loadings to the momentum factor for all news indicators, suggesting that price momentum effects are a crucial driver of most news factors.

These findings are reinforced by the mean-variance spanning tests. The GRS, the GMM as well as the F1 step-down test reject the null hypothesis of $\alpha = 0$ for all news factors at the 1% significance level, suggesting an improvement of the tangency portfolio. This finding is illustrated in Figure 1.3. We see that adding the news factors $SENT_1$, $wSENT_{td,3}$, $wSENT_{td,6}$ and $SENTMOM$ to the set of common factors helps shifting the efficient frontier to the Northwest. Taking the same level of risk, we can annually earn about 125 bps more when incorporating the news factors. Likewise, we document a significant improvement of the minimum-variance portfolio. The F2 step-down test rejects the null hypothesis of $\delta = 0$

[18] See the authors' papers for details on their test statistics.

Table 1.4: News equity factors: Mean-variance spanning

Indicator	Alpha	Market	Value	Quality	Size	MOM	STR	R^2_{adj}	GRS	GMM	F1-Test	F2-Test
$SENT_1$	**0.0027**	-0.029	**-0.132**	0.135	**-0.236**	**0.150**	-0.016	43.1%	**20.08**	**11.99**	**19.98**	**134.33**
	(3.61)	(-1.77)	(-2.62)	(1.14)	(-1.98)	(5.53)	(-0.60)		(0.00)	(0.00)	(0.00)	(0.00)
$SENT_3$	**0.0026**	0.008	**-0.087**	-0.007	-0.059	**0.268**	0.008	72.6%	**26.74**	**17.57**	**26.61**	**104.92**
	(4.48)	(0.70)	(-2.34)	(-0.08)	(-0.79)	(13.62)	(0.36)		(0.00)	(0.00)	(0.00)	(0.00)
$SENT_6$	**0.0027**	0.011	**-0.156**	-0.022	-0.092	**0.276**	0.005	70.5%	**22.65**	**18.43**	**22.53**	**106.25**
	(4.58)	(0.79)	(-3.27)	(-0.18)	(-1.04)	(13.17)	(0.18)		(0.00)	(0.00)	(0.00)	(0.00)
$wSENT_{td,1}$	**0.0027**	**-0.027**	**-0.166**	0.160	**-0.238**	**0.143**	-0.010	44.6%	**22.49**	**14.51**	**22.38**	**148.84**
	(4.05)	(-2.03)	(-3.04)	(1.60)	(-2.07)	(4.99)	(-0.52)		(0.00)	(0.00)	(0.00)	(0.00)
$wSENT_{td,3}$	**0.0031**	-0.005	**-0.082**	0.004	-0.104	**0.222**	-0.015	67.2%	**42.40**	**26.26**	**42.20**	**130.53**
	(5.57)	(-0.39)	(-1.93)	(0.04)	(-1.56)	(10.32)	(-0.67)		(0.00)	(0.00)	(0.00)	(0.00)
$wSENT_{td,6}$	**0.0027**	0.006	**-0.148**	0.087	-0.181	**0.273**	0.000	71.6%	**26.37**	**17.72**	**26.24**	**112.35**
	(4.32)	(0.47)	(-2.93)	(0.77)	(-1.72)	(11.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)
$wSENT_{as,1}$	**0.0024**	-0.019	**-0.124**	0.051	-0.128	**0.183**	-0.023	51.1%	**16.88**	**10.43**	**16.80**	**121.32**
	(3.17)	(-1.36)	(-2.30)	(0.40)	(-1.16)	(6.24)	(-0.84)		(0.00)	(0.00)	(0.00)	(0.00)
$rSENT_{l=u=0,1}$	**0.0022**	-0.026	**-0.122**	0.138	**-0.250**	**0.135**	-0.031	41.4%	**15.63**	**10.01**	**15.56**	**162.36**
	(3.12)	(-1.65)	(-2.57)	(1.22)	(-2.14)	(4.87)	(-1.15)		(0.00)	(0.00)	(0.00)	(0.00)
$SENTMOM$	**0.0018**	0.010	-0.075	0.010	-0.047	**0.133**	0.014	41.5%	**14.00**	**13.04**	**13.93**	**159.65**
	(3.99)	(0.74)	(-1.53)	(0.11)	(-0.59)	(5.03)	(0.49)		(0.00)	(0.00)	(0.00)	(0.00)
SIG_6	**0.0022**	0.014	**-0.091**	0.010	0.050	**0.309**	0.003	74.2%	**16.74**	**11.24**	**16.66**	**58.25**
	(3.69)	(1.00)	(-2.04)	(0.09)	(0.53)	(14.05)	(0.10)		(0.00)	(0.00)	(0.00)	(0.00)

This table shows results of spanning tests for the most promising equally weighted news analytics factors in the global stock universe. The regressors are the market return (represented by the MSCI World) and the common equity factors value, quality, size, momentum (MOM) and short-tem reversal (STR) that are known to affect the cross-section of stock returns. We report the estimated coefficients for the intercept (alpha) and the equity factors. t-statistics are computed from Newey-West adjusted standard errors and are given in parentheses. The last four columns report the test statistics and corresponding p-values (in parentheses) of the GRS test of Gibbons, Ross, and Shanken (1989), the GMM test according to Cochrane (2009) and the Kan and Zhou (2012) step-down test. The null hypothesis of all four tests is that news factors are spanned by the set of common equity factors. GRS, GMM and F1 test whether news factors improve the tangency portfolio, while F2 tests the ability of news factors to improve the minimum-variance portfolio. Coefficients and test statistics are in boldface if significant at a 5% level or better. The time period is from January 2001 to December 2017.

for all news factors at the 1% significance level. In a nutshell, the results of the spanning tests suggest that the most promising news factors significantly expand the investment opportunity set of investors by improving the tangency portfolio, representing minimum-variance hedging opportunities and offering diversification potential.

1.3.4. Robustness to different holding periods

Next, we investigate the persistence of the news indicators' predictive power, speaking to the ease with which these factors could be implemented in a portfolio. If the predictive power of the news indicators remains significant over various months, an investor may think of reducing the portfolio's rebalancing frequency and thus reduce implementation costs. To this end, we assess the performance of a strategy that represents an equally weighted average of the previous h monthly portfolios. The look-back period h is varied from one to twelve, meaning that a portfolio created up to twelve months ago would be considered for next month's factor portfolio. Figure 1.4 charts the associated cumulative returns for the news-based indicators. Table 1.5 reports the corresponding statistics.

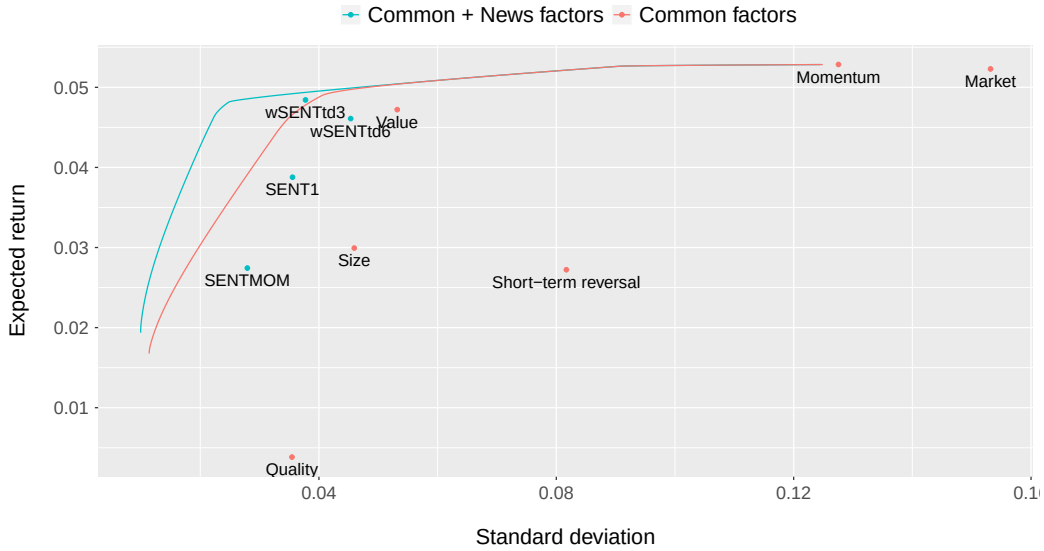
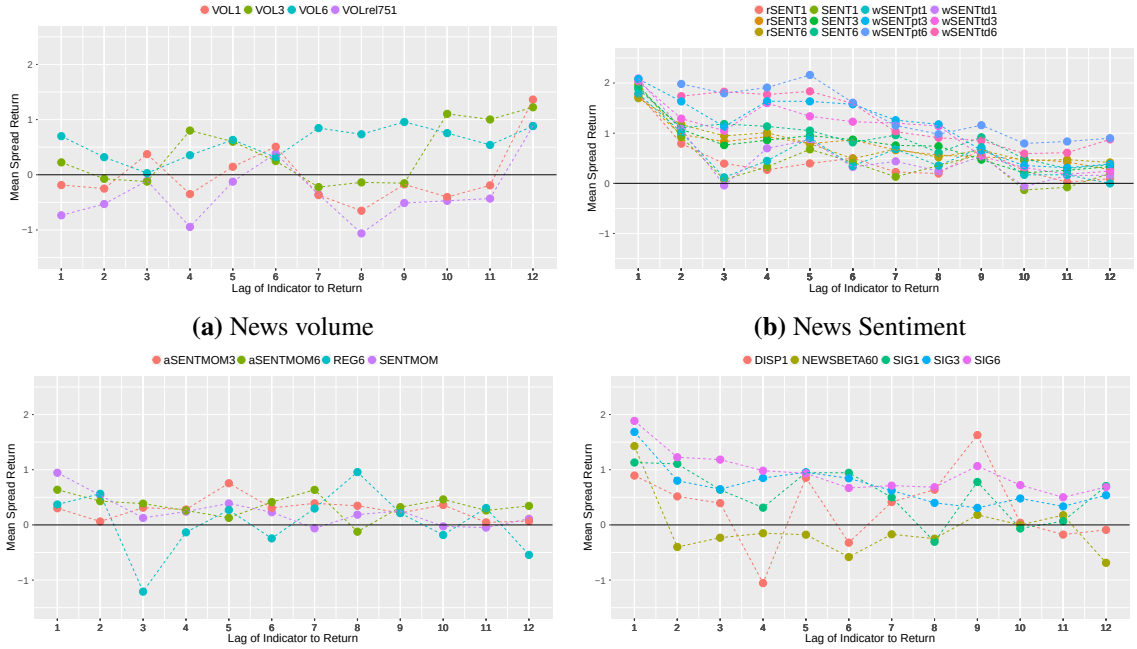


Figure 1.3: Mean-variance spanning of news equity factors. This figure illustrates the mean-variance characteristics of the news equity factors in relation to the set of common equity factors. Note that the underlying mean-variance optimizations include a full investment constraint and do not allow short selling. The red line shows the efficient frontier of the benchmark portfolio comprising of market, value, quality, size, momentum and short-term reversal. The green line shows the efficient frontier when adding $SENT_1$, $wSENT_{td,3}$, $wSENT_{td,6}$ and *SENTMOM* to the benchmark portfolio. Mean-variance inputs are derived from monthly return data over the sample period from January 2001 to December 2017.

The main findings are twofold: (1) Most factors with significant one-month long-short portfolio returns exhibit a fast signal decay in the following months. This finding is consistent with the biased expectation explanation, because news effects get incorporated into stock prices rather sooner than later. (2) Factors incorporating longer-term news sentiment (e.g. $SENT_6$ and $wSENT_{td,6}$) exhibit a quite stable return pattern, indicating that these factors may be useful for long-term investment management. Still, these factor returns are significant only at the conventional two-sigma threshold and do not pass any multiple testing hurdles.

1.3.5. Regional differences

Jacobs and Müller (2020) document regional differences when studying the pre- and post-publication return predictability of 241 cross-sectional anomalies in various international stock markets. In this vein, we divide the global stock universe into four regions—USA, Japan, Europe and emerging markets—and look for regional differences in the efficacy of the investigated news factors.



(a) News volume

(b) News Sentiment

(c) News Trend

(d) Alternative News Concepts

Figure 1.4: News equity factors: Long-horizon effects. This figure shows the returns of cross-sectional long-short portfolios based on news volume (Panel A), news sentiment (Panel B), news trend (Panel C) and alternative news concepts (Panel D) indicators for the global stock universe from January 2001 to December 2017.

Table 1.6 reports the performance statistics of the long-short portfolio returns for the four regions.[19] News volume factors do not seem to be relevant in any of the four regions, similar to the global universe evidence. The performance of news sentiment factors is mixed. While there is limited significance of few news sentiment factors in the US universe (e.g. $wSENT_{td,3}$; but only at the conventional t-statistic threshold of 1.96), we do not evidence any predictive power for the Japanese market. In contrast, the results are substantially better for European and emerging markets, with significant monthly returns of around 7% on average. Similar to the global universe the best performing news sentiment factors are the time-weighted average sentiment factors. For the news trend and alternative news concept, the news sentiment momentum and news significance factors show promising performance, however, again only in European and emerging markets.

In summary, we provide evidence of fairly weak results for the US and the Japanese market and strong results for the European universe and emerging markets. The fact that average momentum returns have been historically low in the Japanese market (see Daniel, Titman, and Wei, 2001; Hanauer, 2014) in conjunction with the finding that the momentum

[19] We exclude news factors with low coverage. From a regional perspective we require an average of at least 100 firms per month.

Table 1.5: News equity factors: Robustness to different holding periods

Indicator	Ret.1M	t-stat	Φ	FDP	Ret.3M	t-stat	Φ	FDP	Ret.6M	t-stat	Φ	FDP
News volume												
$VOL_{\text{REL}>75,1}$	-0.38	-0.38			0.51	0.51			1.94	2.04	✓	
VOL_1	0.53	0.76			1.16	1.61			2.28	3.09	✓	
VOL_3	0.96	0.92			-0.56	-0.63			0.64	0.79		
VOL_6	0.80	0.66			-0.13	-0.12			0.50	0.53		
News sentiment												
$SENT_1$	3.81	4.42	✓	✓	1.16	1.57			1.07	1.38		
$SENT_3$	4.29	4.17	✓	✓	1.54	1.64			2.00	2.34	✓	
$SENT_6$	4.16	3.65	✓	✓	2.40	2.21	✓		1.41	1.58		
$rSENT_{l=u=0,1}$	3.30	4.06	✓	✓	1.11	1.59			1.67	2.29	✓	
$rSENT_{l=u=0,3}$	3.75	4.01	✓	✓	1.35	1.58			2.05	2.58	✓	
$rSENT_{l=u=0,6}$	3.70	3.45	✓	✓	2.00	1.99	✓		1.35	1.62		
$wSENT_{td,1}$	3.67	4.39	✓	✓	1.25	1.55			1.11	1.43		
$wSENT_{td,3}$	4.74	5.18	✓	✓	1.62	1.82			1.44	1.74		
$wSENT_{td,6}$	4.52	4.11	✓	✓	2.00	1.90			1.86	1.99	✓	
$wSENT_{as,1}$	3.37	3.62	✓	✓	0.56	0.63			1.30	1.47		
$wSENT_{as,3}$	3.93	2.82	✓		1.60	1.35			2.33	2.03	✓	
$wSENT_{as,6}$	4.48	3.00	✓		2.70	1.99	✓		1.89	1.51		
News trend												
SENTMOM	2.71	4.00	✓	✓	1.00	1.53			0.74	1.17		
$aSENTMOM_3$	1.38	1.80			-0.05	-0.07			1.43	2.25	✓	
$aSENTMOM_6$	1.66	2.18	✓		0.86	1.39			0.46	0.76		
REG_6	0.77	0.65			1.15	0.99			-0.98	-0.81		
REG_{12}	0.47	0.50			-0.05	-0.06			-1.22	-1.37		
Alternative news concepts												
NEWSBETA	2.67	2.58	✓		1.51	1.47			0.74	0.69		
$DISP_1$	1.18	0.82			-0.19	-0.15			-0.12	-0.09		
SIG_1	2.38	2.57	✓		1.02	1.22			1.37	1.43		
SIG_3	3.89	3.62	✓	✓	1.11	1.05			1.63	1.72		
SIG_3	3.78	3.24	✓	✓	2.09	1.98	✓		0.70	0.73		

This table shows performance statistics of equally weighted long-short portfolios based on the news indicators for the global stock universe and longer return horizons. Annualized mean returns are calculated using the arithmetic average of simple returns and are given in percentage points. We use different lags of the news indicator to return: 1, 3 and 6 months. t-stat is the t-statistic for testing against the Null of a zero mean return. To address the multiple testing problem, we show whether a factor passes (✓) the standard value of 1.96 of the standard normal distribution (Φ) and 3.02, 3.07, 3.19 for the lags 1, 3, 6 using the FDP-StepM procedure of Romano, Shaikh, and Wolf (2008) (FDP) for a significance level of 5%. The time period spans from January 2001 to December 2017.

factor is highly correlated with news-based factors may explain the findings for the Japanese market. The US findings may be rationalized by the fact that it is generally difficult to explain the cross-section of stock returns in the US: the US stock market is likely more efficient than the other markets due to an extremely high analyst coverage, so that news are readily incorporated in stock prices (see McLean and Pontiff, 2016; Jacobs and Müller, 2020). In addition, we also check whether the findings for the US universe may be explained by pre- and post-publication effects. For this purpose, we divide the sample into the period before the seminal study of Tetlock (2007) regarding news flow data, spanning from January 2000

Table 1.6: News equity factors: Regional universes

	Global		USA		Japan		Europe		EM	
Indicator	Return	t-stat	Return	t-stat	Return	t-stat	Return	t-stat	Return	t-stat
News volume										
$VOL_{\text{REL>75},1}$	-0.38	-0.38	-0.14	-0.09	1.82	1.40	-0.62	-0.50	2.45	2.30
VOL_1	0.53	0.76	0.08	0.08	1.44	1.00	0.30	0.31	**2.96**	2.91
VOL_3	0.96	0.92	3.13	2.10	2.68	1.93	-0.94	-0.67	1.96	2.24
VOL_6	0.80	0.66	2.20	1.36	2.14	1.32	0.00	0.00	**2.40**	2.49
News sentiment										
$SENT_1$	**3.81**	4.42	1.27	1.11	0.73	0.50	**4.97**	4.91	**8.33**	6.19
$SENT_3$	**4.29**	4.17	1.83	1.31	-0.31	-0.22	**6.68**	5.10	**9.35**	8.37
$SENT_6$	**4.16**	3.65	1.35	0.85	0.41	0.30	**7.17**	4.95	**7.62**	7.26
$rSENT_{l=u=0,1}$	**3.30**	4.06	1.06	1.00	1.54	1.09	**4.62**	4.45	**6.51**	5.36
$rSENT_{l=u=0,3}$	**3.75**	4.01	1.77	1.44	-0.06	-0.05	**5.73**	4.56	**8.39**	8.40
$rSENT_{l=u=0,6}$	**3.70**	3.45	1.03	0.70	0.59	0.47	**6.21**	4.30	**7.36**	7.80
$wSENT_{td,1}$	**3.67**	4.39	1.07	0.98	0.97	0.65	**4.93**	4.81	**8.62**	6.50
$wSENT_{td,3}$	**4.74**	5.18	2.70	2.25	0.18	0.13	**6.73**	5.65	**10.02**	9.15
$wSENT_{td,6}$	**4.52**	4.11	1.93	1.26	-0.60	-0.44	**7.63**	5.45	**8.92**	8.51
$wSENT_{as,1}$	**3.37**	3.62	0.64	0.54	1.69	1.17	**5.03**	4.37	**7.80**	6.30
$wSENT_{as,3}$	3.93	2.82	1.06	0.51	-0.04	-0.03	**6.60**	4.38	**10.14**	8.98
$wSENT_{as,6}$	4.48	3.00	1.17	0.52	0.22	0.13	**7.29**	4.20	**8.75**	7.67
News trend										
$SENTMOM$	**2.71**	4.00	1.04	0.98	0.27	0.19	**3.16**	3.50	**7.58**	6.12
$aSENTMOM_3$	1.38	1.80	-0.54	-0.55	0.71	0.49	**2.87**	2.64	**4.80**	3.26
$aSENTMOM_6$	1.66	2.18	-0.51	-0.54	0.11	0.09	2.76	2.27	**3.26**	3.09
REG_6	0.77	0.65	-0.67	-0.49	2.50	0.48	5.06	2.20	6.23	1.40
REG_{12}	0.47	0.50	0.07	0.06	0.86	0.32	**3.35**	2.50	3.74	1.63
Alternative news concepts										
$NEWSBETA$	2.67	2.58	-1.12	-0.59	-1.35	-0.62	2.27	1.91	-2.56	-1.52
$DISP_1$	1.18	0.82	1.49	0.57	1.97	0.78	0.64	0.44	-0.86	-0.45
SIG_1	2.38	2.57	-0.64	-0.58	2.61	1.38	**4.87**	3.94	**9.24**	5.47
SIG_3	**3.89**	3.62	1.51	1.03	0.63	0.41	**5.77**	4.21	**9.31**	7.77
SIG_6	**3.78**	3.24	1.24	0.77	1.27	0.94	**6.19**	4.18	**8.13**	6.96

This table shows performance statistics of equally weighted long-short portfolios based on the news indicators for the regional universes USA, Japan, Europe and emerging markets (EM) in addition to the global stock universe. Annualized mean returns are calculated using the arithmetic average of simple returns and are given in percentage points. t-stat is the t-statistic for testing against the Null of a zero mean return. Mean returns are in boldface if their corresponding t-statistics exceed the FDP-StepM threshold of Romano, Shaikh, and Wolf (2008) at 5% significance, which is 3.02 for the global universe, 3.24 for USA, 3.07 for Japan, 2.29 for Europe, 2.21 for RES and 2.37 for EM. The time period spans from January 2001 to December 2017.

to December 2007, and thereafter. However, unreported results do not show significant performance differences between the pre- and post-publication period.

1.4. News analytics and multi-factor investment strategies

Following Section 1.3.3, news-based equity factors may expand an investor's equity factor opportunity set. In this section, we thus explore whether news analytics may be beneficial for constructing multi-factor investment strategies. Based on a benchmark set of factors we first examine whether simple multi-factor portfolios can be enhanced by adding news-

based factors. Second, we investigate the benefits of utilizing news flow data for dynamic factor allocation strategies. In particular, we use the parametric portfolio policy of Brandt, Santa-Clara, and Valkanov (2009) to arrive at meaningful factor timing allocations. To this end, we follow Dichtl, Drobetz, Lohre, et al. (2019) and construct a set of equity factors that expands the common factors used in Section 1.3.3 to further equity factors widely used and well documented in academic research. These factors can be roughly assigned to the following four categories:[20]

- $Value$: cash flow yield (CFY), dividend yield (DY), book-to-market ratio (BTM), earnings yield (EY), and profitability ($PROF$)
- $Momentum$: 12-month price momentum ($MOM12$), short-term reversal (STR), and long-term reversal (LTR)
- $Quality$: asset turnover (AT), change in long-term debt ($DLTD$), change in shares outstanding (DSO), asset growth (AG), cash productivity (CP), profit margin (PMA), leverage (LEV), return on assets (ROA), sales-to-cash (STC), sales-to-inventory (STI), and accruals (ACC)
- $Size$: Size ($SIZE$)

1.4.1. Diversified factor allocation

Taking an agnostic perspective regarding expected factor returns, risk-based factor allocations strategies are a common technique to construct diversified multi-factor portfolios. We examine how an equally weighted portfolio (1/N), a minimum-variance portfolio (MV) and a risk parity portfolio (RP) responds to the inclusion of news-based equity factors.[21]

Table 1.7 provides the performance statistics of the three strategies for the set of benchmark factors (Panel A) and the set of benchmark factors augmented by a news-based equity factor (Panel B). We compute the first optimal portfolio weights over a 36-month window, which expands over time. To enable a fair comparison to the subsequent dynamic factor allocation attempts, the out-of-sample period spans from January 2007 to December 2017. We enforce full investment and long-only constraints to make sure not to bet against a given factor premium.

[20] See Dichtl, Drobetz, Lohre, et al. (2019) for a concise definition of each factor.

[21] In brief, the 1/N strategy rebalances monthly to an equally weighted allocation scheme. The minimum-variance portfolio is the mean-variance efficient portfolio that is expected to have the lowest possible portfolio variance. The risk parity strategy allocates capital so that the factors' risk budgets contribute equally to overall portfolio risk.

Overall, we document that all three risk-based allocation strategies benefit from adding the news sentiment factor $wSENT_{td,3}$ to the benchmark portfolio. Moreover, the reported results are robust to the choice of the news-based factor to be added to the set of benchmark news factors, given that the used factor is among the factors tested in the mean-variance spanning tests in Table 1.4.

Table 1.7: Diversified multi-factor allocation

Strategy	Excess Return	Sd	Min	Max	SR	MDD	t-stat
Panel A: Benchmark factors							
1/N	2.91	2.44	-1.58	2.31	1.20	-2.74	3.80
MVP	2.15	1.25	-0.59	1.35	1.72	-0.87	5.45
RP	2.39	1.41	-0.85	1.73	1.69	-1.63	5.36
Panel B: Benchmark + news factors							
1/N	2.97	2.27	-1.43	2.18	1.31	-2.44	4.15
MVP	2.41	1.13	-0.53	1.42	2.14	-1.00	6.79
RP	2.55	1.30	-0.71	1.69	1.97	-1.32	6.24

This table shows performance statistics of risk-based factor allocation strategies for the set of benchmark factors (Panel A) and the set of benchmark factors augmented by the news-based equity factor $wSENT_{td,3}$ (Panel B). Specifically, we examine an equally weighted portfolio (1/N), a minimum-variance portfolio (MVP) and a risk parity portfolio (RP). Annualized excess returns are calculated using the arithmetic average of simple returns. Standard deviation (Sd) and Sharpe ratio (SR) are annualized through multiplication by $\sqrt{12}$. Min and Max denote the lowest and highest monthly excess return in the sample period. MDD is the maximum drawdown. Excess return, Sd, Min, Max and MDD are given as percentages. t-stat is the t-statistic for testing against the Null of a zero return effect. The performance statistics are based on the out-of-sample period from January 2007 to December 2017.

The simple 1/N strategy earns an annualized excess return of 2.91% at 2.44% volatility. As commonly documented in the literature, the minimum-variance and risk parity portfolios exhibit lower excess returns than the 1/N portfolio (2.15% for the MV portfolio and 2.39% for the RP portfolio). The strengths of the MV and RP strategies are in (downside) risk hedging, translating to a significant reduction in volatility and maximum drawdown compared to the 1/N strategy. Specifically, we document volatility and maximum drawdown figures of 1.25% and -0.87% for the MV portfolio and of 2.39% and -1.63% for the RP portfolio, compared to 2.44% and -2.74% for the 1/N portfolio. The associated Sharpe ratios are in favor of the MV and RP strategies (1.72 and 1.69 versus 1.20 for 1/N).

When including the news-based factor, the strategies' Sharpe ratio increases by 0.32 (MV), 0.18 (RP) and 0.12 (1/N), respectively. This improvement is due to the favorable risk-return characteristics of the news equity factor, which are reflected by high weights in the minimum-variance and risk parity strategies. The news sentiment factor exhibits the second highest average weight (around 15% for the MV and 11% for the RP strategy), only exceeded

by the accrual factor (with average weights of 28% and 20%, respectively). The attractiveness of the news sentiment factor also shows in increased return and reduced volatility figures for all three risk-based factor allocation strategies. For instance, the minimum-variance strategy earns a 26 bps higher excess return at a decrease of 12 bps in volatility when including the news-based factor. Similar improvements can be documented for the risk parity and 1/N strategy. While the latter strategies also profit from the information contained in news data in terms of downside risk, we observe a slightly more severe maximum drawdown for the MV portfolio.

Due to the robustness and simplicity of the 1/N strategy (see DeMiguel, Garlappi, and Uppal, 2009), we benchmark the subsequent dynamic factor allocation strategies using the 1/N strategy.

1.4.2. Dynamic factor allocation

A popular way of dynamic factor allocation exploits cross-sectional differences in factor characteristics by tilting the factor allocation according to those characteristics. Utilizing the cross-sectional parametric policy framework developed by Brandt, Santa-Clara, and Valkanov (2009), we exploit factor characteristics based on the derived news indicators in addition to benchmark characteristics from Dichtl, Drobetz, Lohre, et al. (2019) to assess the relevance of the news analytics indicators.

Cross-sectional factor characteristics

To calculate news-based equity factor characteristics we follow Lee (2017) and look for "factors within factors". Therefore, we first build quintile portfolios for a given equity factor, such as value or momentum. We then compute the average news indicator score across all stocks in each quintile portfolio. A factor's news characteristic is finally computed as the spread between the news score of the top and that of the bottom quintile portfolio. For instance, one can thus back out whether a given factor has implicit positive news sentiment that might lend itself naturally to time factors. Similar to a stock level rationale, news sentiment on a factor level may entail information on the attractiveness of a factor itself. A positive net news sentiment for a factor is driven by more positive news for those companies on the long leg and/or more negative surprises for the short leg. Following the rationale that positive news sentiment indicates positive factor returns, we use the characteristic to compare factors and tilt towards those with positive sentiment and underweight those with negative sentiment.

In addition to a representative set of news-based characteristics we include the following factor characteristics that are well documented in the literature and used by Dichtl, Drobetz,

Lohre, et al. (2019): factor valuation, factor spread, factor momentum, factor volatility and factor crowding. *Factor valuation* applies the same rationale of value investing at a factor level, overweighting factors that, on aggregate, experience attractive valuation levels while underweighting those that look expensive. *Factor spread* measures the difference in a characteristic between the long and short leg. As a large factor spread might proxy for the factor's potential future return dispersion, we utilize this information to tilt towards factors with a high spread as this corresponds with a higher factor return opportunity as shown by Huang et al. (2010). Avramov et al. (2017) show that *factor momentum* is helpful in predicting the next month's factor return. Given that low-volatility stocks outperform high-volatility stocks on a risk-adjusted basis (e.g. Jensen, Black, and Scholes, 1972; Haugen and Baker, 1991), we calculate *factor volatility* to test for a volatility effect among equity factors. *Crowding* measures the risk and sensitivities to shocks investors are exposed to as they hold the same securities. We follow Cahan and Luo (2013) and capture crowding by the mean pairwise correlation within a given factor.[22]

Factor timing using cross-sectional characteristics

We incorporate the standardized cross-sectional characteristics into the parametric portfolio policy (PPP) of Brandt, Santa-Clara, and Valkanov (2009), which allows us to exploit the information content in a utility-based portfolio optimization. In contrast to other applications, the PPP allows to directly use characteristics for a dynamic factor allocation, avoiding to estimate the joint distribution of factor returns. While a mean-variance optimization needs to transform the expected returns to incorporate the timing component, the PPP proposes to model the portfolio weight as a linear function of asset characteristics $x_{i,t}$:

$$w_{i,t} = f(x_{i,t}; \phi) = w_{b,i,t} + \frac{1}{N_t} \phi' \hat{x}_{i,t}, \tag{1.2}$$

where $w_{i,t}$ denotes the portfolio weight for asset i, $w_{b,i,t}$ is the benchmark weight, N_t denotes the number of assets at time t, ϕ is the vector of coefficients to be estimated through utility maximization and $\hat{x}_{i,t}$ denotes the estimated standardized factor characteristics.

For a mean-variance utility function, the original problem can be restated as

$$\max_{\phi} \phi' \hat{\mu}_c - \left(\frac{\gamma}{2} \phi' \hat{\Sigma}_c \phi + \gamma \phi' \hat{\sigma}_{bc} \right), \tag{1.3}$$

where $\hat{\Sigma}_c$ is the sample covariance matrix, $\hat{\mu}_c$ is the mean of the characteristic return vector and $\hat{\sigma}_{bc}$ is the sample vector of covariances between the benchmark portfolio return and the

[22] See Dichtl, Drobetz, Lohre, et al. (2019) for a detailed description of these factor characteristics.

characteristic-return vector.[23] Thus, we can measure the information content embedded in our cross-sectional characteristics statistically via the coefficients ϕ and economically by analyzing the resulting information ratios.

Hence, the PPP directly incorporates cross-sectional characteristics into a dynamic factor allocation strategy and thus enables a comparison to the risk-based allocations discussed above.

Empirical results

Table 1.8 reports the estimation results and performance statistics for news-related factor timing allocations based on univariate and multivariate parametric portfolio policies. Panel B presents the results of six news-related allocations based on a univariate PPP. Across the univariate models, we obtain the only significant coefficients for the tilting characteristics $wSENT_{td,1}$ and $wSENT_{td,3}$, suggesting a short-term sentiment effect among equity factors. Hence, positive sentiment factors are overweighted relative to the equally weighted benchmark while negative sentiment factors are underweighted. The annualized returns of the corresponding PPP using $wSENT_{td,1}$ and $wSENT_{td,3}$ are 88 and 97 bps higher than the one for the equally weighted benchmark, whereas the volatility is increased by 23 bps for the $wSENT_{td,1}$ characteristics and decreased by 7 bps for $wSENT_{td,3}$. These figures correspond to an information ratio of 0.52 and 0.59, respectively.

While statistically weak, longer-horizon news sentiment-related characteristics have positive information ratios as well: $SENT_3$, $wSENT_{td,6}$, $SENTMOM$, and SIG_6 with information ratios of 0.55, 0.53, 0.26 and 0.54, respectively. Moreover, capturing news sentiment over a longer horizon seems to be more profitable: The $wSENT_{td,3}$ timing portfolio has a higher Sharpe ratio than the $wSENT_{td,1}$ timing portfolio and than the equally weighted benchmark (1.64 vs. 1.42 vs. 1.20). After accounting for transaction costs the $wSENT_{td,3}$ strategy's return and Sharpe ratio are reduced to 2.53% and 1.07, which is equivalent to an information ratio of 0.39 net of transaction costs. Notably, news sentiment-related timing allocations show similar performance statistics to allocations using common tilting characteristics such as factor crowding, factor valuation and factor volatility and seem to be more profitable than those for factor momentum and factor spread allocations.

Instead of relying on one factor characteristic only, we consider multiple characteristics jointly in different multivariate parametric portfolio policies. Panel C of Table 1.8 shows the results of a multivariate PPP based on the common tilting characteristics factor crowding,

[23] As all characteristics are standardized cross-sectionally across all factors at time t, deviations from the benchmark are equivalent to a zero-investment portfolio (DeMiguel, Martin-Utrera, et al., 2020; Dichtl, Drobetz, Lohre, et al., 2019). For a detailed description see also DeMiguel, Martin-Utrera, et al. (2020).

Table 1.8: Dynamic factor allocation

Characteristic	$\hat{\phi}$	Return p.a.		SD p.a.	Sharpe ratio		Maximum drawdown		t-statistic		Tracking error	Information ratio		Turnover p.a.
		gross	*net*		*gross*	*net*	*gross*	*net*	*gross*	*net*		*gross*	*net*	*two-way*
Panel A: Benchmark model														
1/N	–	2.91	1.89	2.44	1.20	0.78	2.74	2.99	3.80	2.47	–	–	–	–
Panel B: Univariate models														
$SENT_3$	3.22	3.83	2.51	2.33	1.65	1.08	4.07	4.65	5.22	3.42	1.67	0.55	0.37	1.83
$wSENT_{td,1}$	**4.63**	3.79	1.93	2.67	1.42	0.72	3.32	4.00	4.52	2.30	1.69	0.52	0.02	4.52
$wSENT_{td,3}$	**3.49**	3.88	2.53	2.37	1.64	1.07	3.93	4.51	5.19	3.39	1.64	0.59	0.39	1.95
$wSENT_{td,6}$	2.80	3.79	2.55	2.28	1.66	1.12	3.95	4.52	5.27	3.54	1.66	0.53	0.40	1.40
SENTMOM	1.64	3.38	1.74	3.16	1.07	0.55	4.16	6.51	3.39	1.74	1.78	0.26	-0.08	3.40
SIG_6	2.67	3.85	2.62	2.34	1.65	1.12	3.69	4.26	5.22	3.55	1.74	0.54	0.42	1.36
Volatility	5.41	3.89	2.53	2.87	1.35	0.88	3.36	3.92	4.3	2.79	1.76	0.55	0.36	1.98
Crowding	4.36	4.25	2.61	3.02	1.41	0.87	3.53	4.19	4.48	2.76	1.95	0.69	0.37	3.41
Momentum	1.32	3.57	1.70	2.91	1.22	0.58	3.09	3.57	3.89	1.83	1.45	0.45	-0.13	4.55
Spread	**16.82**	3.55	2.23	3.23	1.10	0.68	3.80	5.51	3.49	2.16	1.89	0.34	0.18	1.79
Valuation	-1.61	3.36	2.17	2.19	1.53	0.99	3.20	3.74	4.87	3.14	1.77	0.25	0.16	1.16
Panel C: Multivariate model based on benchmark characteristics														
Volatility	14.69	3.39	1.64	3.22	1.05	0.51	4.02	6.67	3.34	1.62	2.11	0.23	-0.12	3.94
Crowding	-11.12													
Momentum	0.84													
Spread	**17.90**													
Valuation	-0.22													
Panel D: Multivariate model based on benchmark + news characteristics														
Volatility	0.68	3.90	1.97	3.05	1.28	0.65	2.57	4.33	4.05	2.06	1.84	0.54	0.04	4.83
Crowding	-3.75													
Momentum	-0.80													
Spread	**27.34**													
Valuation	**7.29**													
$wSENT_{td,3}$	**12.87**													
Panel E: Multivariate cherry-picking model														
Spread	**24.48**	4.04	2.36	3.05	1.33	0.78	3.04	5.30	4.22	2.46	1.92	0.60	0.25	3.58
Valuation	**6.47**													
$wSENT_{td,3}$	**11.08**													

This table gives estimation results and performance statistics of univariate and multivariate parametric portfolio policies based on news-related and benchmark factor characteristics. The performance statistics are based on the out-of-sample period from January 2007 to December 2017. The second column gives the estimated coefficients of the PPP, highlighted in bold if significant at a 5% level or better. Annualized returns are calculated using the arithmetic average of simple returns. Standard deviation and Sharpe ratio are annualized through multiplication by $\sqrt{12}$. The information ratio uses arithmetic active returns of factor timing over the 1/N benchmark. Annualized turnover is stated as two-way turnover. All performance statistics are given as percentages, except for Sharpe ratio and t-statistic.

factor valuation, factor volatility, factor momentum and factor spread. Similar to the univariate results, we only document a significant coefficient for the spread characteristic, accompanied by a decline in performance. With a return of 3.39% at 3.22% volatility this multivariate timing allocation exhibits worse performance statistics than the associated univariate strategies, resulting in a negative information ratio net of transaction costs (-0.12).

A second multivariate PPP adds the news sentiment characteristic $wSENT_{td,3}$ to the set of tilting characteristics (cf. Panel D of Table 1.8). Similar to the univariate results, we find a significant coefficient for $wSENT_{td,3}$. The coefficient for spread remains statistically significant, whereas the one for valuation now becomes significant. This is due to interaction effects of the characteristics when utilizing the PPP. In the presence of the other characteristics, valuation gains importance as it adds to the overall allocation framework. The utility of news flow data for dynamic factor allocations is corroborated by improved performance statistics, documented by a higher return (3.90%) at lower volatility (3.05%) and slightly positive information ratio of 0.04 for the multivariate PPP that includes a news characteristic.

As transaction costs almost completely consume the benefits of the previous set of tilting characteristics, one may wonder whether a cherry-picked set of characteristics would lead to a more favorable outcome (cf. Dichtl, Drobetz, Lohre, et al., 2019). When focusing on the three characteristics that proved statistically significant over the whole sample period (spread, valuation and news), the dynamic factor allocation shows further increased performance statistics (cf. Panel E of Table 1.8). We find a slightly increased excess return of 4.04% at the same level of volatility, but document a substantially higher information ratio of 0.25 for the cherry-picking timing model.

Overall, utilizing news sentiment to dynamically tilt factor allocations shows promising results. It turns out that the news-related dynamic factor allocations are more profitable than using comparable predictors tested in the literature. For instance, we document that information extracted from news flow data may add value over and above a simple factor momentum timing strategy. Similar to other timing strategies, turnover is a key differentiator. Similar to one-month factor momentum, short-term news sentiment exhibits elevated turnover figures and subsequently lower net information ratios. In contrast to factor momentum, our results show that news sentiment entails information beyond the short-term window and deems meaningful in timing factors even at longer horizons, like three and six months.

In a nutshell, our empirical evidence suggests that news sentiment information is valuable for constructing multi-factor allocation strategies. Thus, our findings are in line with Uhl, Pedersen, and Malitius (2015) and Tetlock (2007) who document that news sentiment is useful for predicting future return movements and still hold in the context of predicting future factor returns.

1.5. Conclusion

This paper contributes to the literature on news analytics by investigating its relevance for the cross-section of stock returns and its ability to enhance multi-factor investment strategies.

Studying the cross-sectional characteristics of a broad set of indicators generated from news flow data suggests that the insights gathered from firm-specific news sentiment analysis can find their way into implementable trading strategies in a manner that adds over and above common drivers of equity returns. Passing a rigorous research protocol that includes multiple testing hurdles, long-short portfolios based on news sentiment indicators seem to be particularly profitable in a global and European stock universes, while results for US and Japanese equity markets are rather moderate.

Assessing the information embedded in news flow data in simple and dynamic factor allocation strategies reveals the relevance for practical equity factor investing. An equally weighted portfolio as well as minimum-variance and risk parity strategies benefit from adding news sentiment-related equity factors to a portfolio of common global equity factors. Building on these insights, we explore the benefits of active factor allocation when incorporating information stemming from news flow data. Utilizing parametric portfolio policies, we document that news sentiment-related factor characteristics help explaining the cross-section of factor returns, given that the news data entail information on a factor's attractiveness itself. Associated factor timing strategies generate statistically significant and economically relevant results, stressing the relevance of news analytics for dynamic factor allocation strategies.

Appendix 1.A The set of news indicators

This section describes how we construct indicators based on news flow data from RavenPack News Analytics. All indicators are filtered using the relevance score (REL), the event relevance score (EVR) and the event similarity days score (ESD). Unless otherwise indicated, we require all scores to be above the conventional level of 90 (cf. Hafez, 2010; Kolasinski, Reed, and Ringgenberg, 2013; Dang, Moshirian, and Zhang, 2015; Beschwitz, Keim, and Massa, 2020).

Let E_i be the ith news event for a specific firm in a given time horizon, as classified by the RavenPack taxonomy. The publication date of a news event is denoted as $\tau(\cdot)$. Then, the news volume indicator at time t, VOL_t, is computed as the number of news events within time horizon h, i.e.

$$VOL_{t,h} = \sum_{i \in I} \mathbb{1}_{\{\tau(E_i) \in [t-h,t]\}}, \tag{1.4}$$

where $I \subset \mathbb{N}$ captures all news events for a specific firm and $\mathbb{1}(\cdot)$ denotes the indicator function. In the empirical study, we calculate *VOL* using two filter settings: A less restrictive setting (REL $>$ 75) to cover a firm's overall media presence and the standard setting (REL $>$ 90, EVR $>$ 90, ESD $>$ 90) to focus on major events and thus only analyze a firm's meaningful media presence.

Further, let $ESS(\cdot)$ be the event sentiment score of a news event. Then, the average firm-specific news sentiment indicator *SENT* is given by

$$SENT_{t,h} = \frac{\sum_{i \in I} \mathbb{1}_{\{\tau(E_i) \in [t-h,t]\}} ESS(E_i)}{\sum_{i \in I} \mathbb{1}_{\{\tau(E_i) \in [t-h,t]\}}}. \tag{1.5}$$

The robust version of the news sentiment indicator, *rSENT*, is calculated as follows

$$rSENT_{t,h} = \frac{\sum_{i \in I} \mathbb{1}_{\{\tau(E_i) \in [t-h,t] \,|\, ESS(E_i) > u\}} - \sum_{i \in I} \mathbb{1}_{\{\tau(E_i) \in [t-h,t] \,|\, ESS(E_i) < l\}}}{\sum_{i \in I} \mathbb{1}_{\{\tau(E_i) \in [t-h,t] \,|\, ESS(E_i) > u, ESS(E_i) < l\}}}, \tag{1.6}$$

where l and u are lower and upper thresholds defining the range for the ESS. We use two threshold settings: first, we differentiate between positive and negative news by setting $u = l = 0$. Second, we further exclude sentiment scores that are close to zero, i.e. $u = 0.1$ and $l = -0.1$.

When constructing the weighted sentiment indicators *wSENT*, we resort to two different weighting schemes. The first one puts larger weight on more recent sentiment scores (i.e. those that are closer to the end of time horizon h). Specifically, we first compute the average ESS for each day according to the *SENT* indicator. Subsequently, we weight these daily

average sentiment scores within the look-back window using a linear decay function. Thus, the temporal decay (*td*) weighted sentiment factor is constructed as follows:

$$wSENT_{td,(t,h)} = \frac{\sum_{i=1}^{N_D} SENT_{i,\text{one-day}} \cdot i}{\sum_{j=1}^{N_D} j} \tag{1.7}$$

where N_D denotes the number of days in the look-back window $[t-h, t]$ and h is measured in months.

The second weighting scheme is based on the empirical observation that the market reaction to negative news is generally stronger than the reaction to positive news (Hafez, Guerrero-Colon, and Duprey, 2015). For this purpose, we utilize a weighting function from prospect theory (cf. Tversky and Kahneman, 1992) that is able to account for the asymmetric reaction of the stock market to the nature of news events. The asymmetrically weighted (*as*) news sentiment indicator is therefore given by:

$$wSENT_{as,(t,h)} = \frac{\sum_{i\in I}\left(\mathbb{1}_{\{ESS(E_i)>0|\tau(E_i)\in[t-h,t]\}} ESS(E_i)^{\alpha} - \mathbb{1}_{\{ESS(E_i)<0|\tau(E_i)\in[t-h,t]\}}\lambda(-ESS(E_i))^{\beta}\right)}{\sum_{i\in I}\left(\mathbb{1}_{\{ESS(E_i)>0|\tau(E_i)\in[t-h,t]\}} + \mathbb{1}_{\{ESS(E_i)<0|\tau(E_i)\in[t-h,t]\}}\lambda\right)}. \tag{1.8}$$

with $\alpha \approx \beta \approx 0.88$ and $\lambda \approx 2.25$, chosen according to Tversky and Kahneman (1992). The parameter λ captures the extend to which negative news receive a higher weight compared to positive news.

The news sentiment momentum indicator *SENTMOM* is constructed similar to the methodology of Uhl, Pedersen, and Malitius (2015). Based on the SENT indicator, we first calculate two moving average time series of different time horizons using a rolling window approach and then calculate the difference of these two time series (that is, for $h = 1$ and $h = 12$ we get $SENT_{t,1} - SENT_{t,12}$). Subsequently, we apply the cumulative sum (CUSUM) filter to this time series (see Uhl, Pedersen, and Malitius, 2015, for details). Finally, the indicator series is normalized to be bound by +1 and -1.

Another way to calculate a trend indicator for news sentiment is to standardize a crossing moving average time series (e.g. for $h = 1$ and $h = 3$, see previous paragraph) by its sample standard error instead of applying the CUSUM filter. Specifically, the *aSENTMOM* indicator is computed as follows

$$aSENTMOM_{t,h} = \frac{SENT_{t,1} - SENT_{t-h}}{\sqrt{\sigma_{t,1}^2/VOL_{t,1} - \sigma_{t,h}^2/VOL_{t,h}}}, \tag{1.9}$$

where the sample variance $\sigma^2_{t,h}$ is given by

$$\sigma^2_{t,h} = \frac{\sum_{i \in I} \mathbb{1}_{\{\tau(E_i) \in [t-h,t]\}} \left(ESS(E_i) - SENT_{t,h}\right)^2}{\left(\sum_{i \in I} \mathbb{1}_{\{\tau(E_i) \in [t-h,t]\}}\right) - 1}. \tag{1.10}$$

The third news trend indicator, *REG*, is simply based on the t-statistic from regressing the cumulative sum of the ESS on the time index within time horizon h.

Among the alternative news concept indicators, *NEWSBETA* measures the responsiveness of a firm's stock return to an aggregate market news sentiment within a specific horizon. Specifically, the indicator value is calculated as the t-statistic from regressing a firm's stock return on a market capitalization-weighted average of the ESS across all firms in the universe.

The news significance indicator *SIG* measures the significance of the ESS (similar to a t-statistic) and thus captures mean and variation in the ESS. Specifically, it is given by

$$SIG_{t,h} = \frac{SENT_{t,h}}{\sqrt{\sigma^2_{t,h}/VOL_{t,h}}}. \tag{1.11}$$

The news dispersion indicator measures the variation in the ESS and is computed as

$$DISP_{t,h} = \frac{\sqrt{\sigma^2_{t,h}}}{SENT_{t,h}}. \tag{1.12}$$

All indicators except *SENTMOM* and the regression-based indicators are computed for $h = 1, 3, 6$, where h is measured in months. While *SENTMOM* uses multiple time horizons by definition, *REG* is calculated for $h = 6,\ 12$ due to sample size requirements for time-series regressions. For the *NEWSBETA* indicator, we employ an expanding window estimation, with an initial window of $h = 12$. In a final step, we standardize all indicators by company size and industry classification.

Appendix 1.B Tables

Table 1.B.1: Equity Factor Description

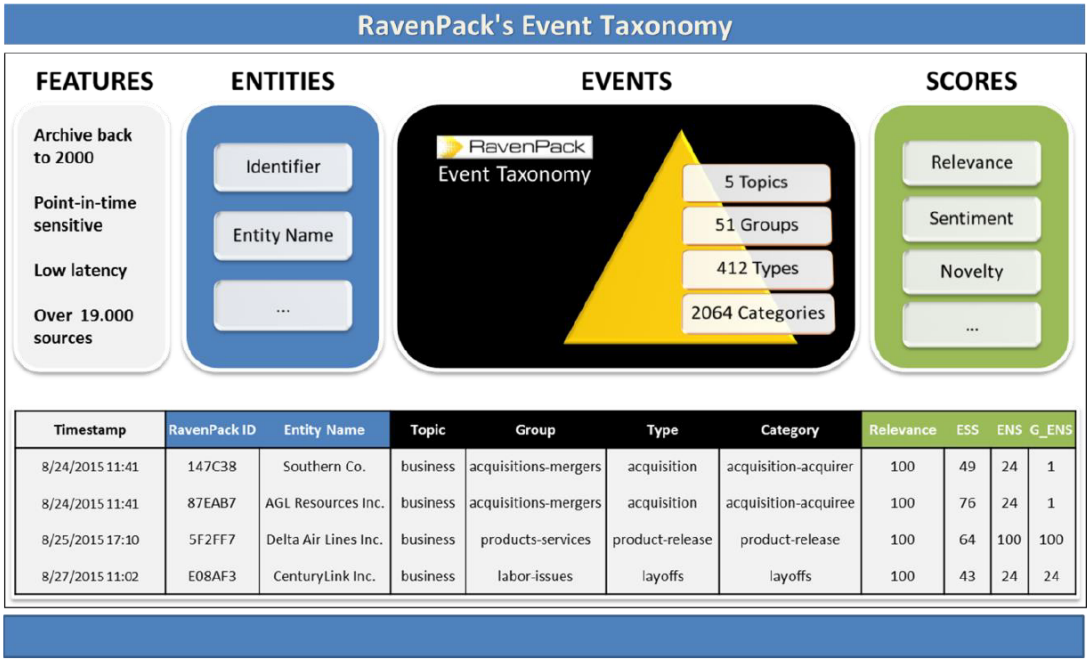
This table describes how we define common equity factors. In particular, we adopt the definitions of Dichtl, Drobetz, Lohre, et al. (2019). The necessary data are sourced from the Worldscope database.

Factor	Description	Related studies
Value	Cash flow yield is used as value factor. It is constructed as a zero-investment trading strategy that is long in stocks with high cash flow-to-price ratio and short in stocks with low cash flow-to-price ratio. Cash flows are measured as the sum of funds from operations, extraordinary items and funds from other operating activities	Sloan (1996), Da and Warachka (2009), Hou, Karolyi, and Kho (2011)
Quality	Profitability is employed as quality factor. This factor is constructed as a zero-investment trading strategy that is long in stocks with robust operating profitability and short in stocks with weak profitability. Profitability is measured by annual revenues less cost of goods sold and interest and other expenses, divided by the book value for the last fiscal year-end.	Haugen and Baker (1996), Cohen, Gompers, and Vuolteenaho (2002), Fama and French (2006), Novy-Marx (2013), Fama and French (2016)
Momentum	We use 12-month price momentum that captures a medium-term continuation effect in returns by buying recent winners and selling recent losers. We control for the short-term reversal effect by excluding the most recent month $(t-1)$ at time t.	Jegadeesh (1990), Jegadeesh and Titman (1993)
Size	The size factor builds on the observation that stocks with larger market capitalization tend to underperform stocks with smaller market capitalization. It is constructed as a zero-investment trading strategy that is long in stocks with small market capitalization and short in stocks with high market capitalization.	Banz (1981), Fama and French (1992), Sloan (1996), Da and Warachka (2009), Hou, Karolyi, and Kho (2011)
Short-term reversal	This factor captures the short-term reversal effect in the cross-section of stock returns. It is constructed as a zero-investment trading strategy that is long in stocks with weak previous month performance and short in stocks with high previous month performance.	Jegadeesh (1990), Lehmann (1990)

Appendix 1.C Figures

Figure 1.C.1: A schematic view of RavenPack's News Analytics

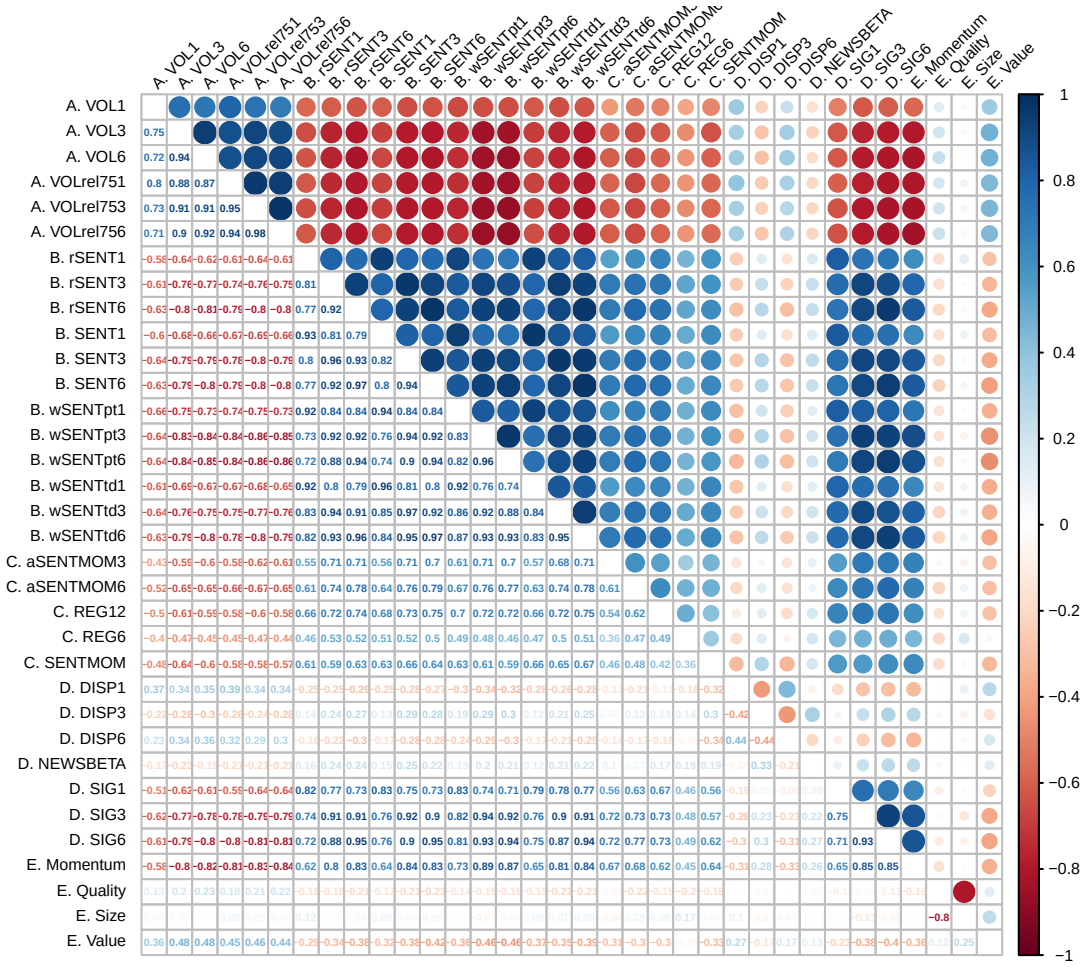
This figure shows a schematic view of RavenPack's news analytics data, including RavenPack's event taxonomy. Source: RavenPack (2016).



Timestamp	RavenPack ID	Entity Name	Topic	Group	Type	Category	Relevance	ESS	ENS	G_ENS
8/24/2015 11:41	147C38	Southern Co.	business	acquisitions-mergers	acquisition	acquisition-acquirer	100	49	24	1
8/24/2015 11:41	87EAB7	AGL Resources Inc.	business	acquisitions-mergers	acquisition	acquisition-acquiree	100	76	24	1
8/25/2015 17:10	5F2FF7	Delta Air Lines Inc.	business	products-services	product-release	product-release	100	64	100	100
8/27/2015 11:02	E08AF3	CenturyLink Inc.	business	labor-issues	layoffs	layoffs	100	43	24	24

Figure 1.C.2: Return correlation of news equity factors

This figure shows the correlation among news equity factors and common equity factors. Equity factors are derived from monthly return data for the global stock universe over the sample period from January 2001 to December 2017 and are grouped according to their concept category: news volume (A), news sentiment (B), news trend (c), alternative news concepts (D) and common equity factors (E).



References

Arnott, Rob, Noah Beck, Vitali Kalesnik, and John West (2016). "How can 'smart beta' go horribly wrong?" *Research Affiliates, February*.

Arnott, Rob, Campbell R Harvey, and Harry Markowitz (2019). "A backtesting protocol in the era of machine learning". *Journal of Financial Data Science* 1 (1), 64–74.

Asness, Clifford S (2016). "The siren song of factor timing aka "smart beta timing" aka "style timing"". *Journal of Portfolio Management* 42 (5), 1–6.

Audrino, Francesco, Fabio Sigrist, and Daniele Ballinari (2020a). "The impact of sentiment and attention measures on stock market volatility". *International Journal of Forecasting* 36 (2), 334–357.

Avramov, Doron, Si Cheng, Amnon Schreiber, and Koby Shemer (2017). "Scaling up market anomalies". *Journal of Investing* 26 (3), 89–105.

Baker, Malcolm and Jeffrey Wurgler (2006). "Investor sentiment and the cross-section of stock returns". *Journal of Finance* 61 (4), 1645–1680.

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

Barber, Brad M and Terrance Odean (2008). "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors". *Review of Financial Studies* 21 (2), 785–818.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny (1998). "A model of investor sentiment". *Journal of Financial Economics* 49 (3), 307–343.

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

Bender, Jennifer, Xiaole Sun, Ric Thomas, and Volodymyr Zdorovtsov (2018). "The promises and pitfalls of factor timing". *Journal of Portfolio Management* 44 (4), 79–92.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1), 289–300.

Benjamini, Yoav and Daniel Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency". *Annals of Statistics*, 1165–1188.

Beschwitz, Bastian von, Donald B Keim, and Massimo Massa (2020). "First to "read" the news: News analytics and algorithmic trading". *Review of Asset Pricing Studies* 10 (1), 122–178.

Bonferroni, Carlo E (1936). "Teoria statistica delle classi e calcolo delle probabilita'." *Liberia Internazionale Seeber*.

Brandt, Michael W, Pedro Santa-Clara, and Rossen Valkanov (2009). "Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns". *Review of Financial Studies* 22 (9), 3411–3447.

Cahan, Rochester and Yin Luo (2013). "Standing Out From the Crowd: Measuring Crowding in Quantitative Strategies". *Journal of Portfolio Management* 39, 14–23.

Chan, Wesley S (2003). "Stock price reaction to news and no-news: drift and reversal after headlines". *Journal of Financial Economics* 70 (2), 223–260.

Cochrane, John H (2009). *Asset pricing*. Princeton University Press.

Cohen, Randolph B, Paul A Gompers, and Tuomo Vuolteenaho (2002). "Who underreacts to cash-flow news? Evidence from trading between individuals and institutions". *Journal of Financial Economics* 66 (2), 409–462.

Coqueret, Guillaume (2020). "Stock-specific sentiment and return predictability". *Quantitative Finance* 20 (9), 1531–1551.

Cutler, David M, James M Poterba, and Lawrence H Summers (1989). "What moves stock prices?" *Journal of Portfolio Management* 15 (3), 4–12.

Da, Zhi, Joseph Engelberg, and Pengjie Gao (2011). "In search of attention". *Journal of Finance* 66 (5), 1461–1499.

Da, Zhi and Mitchell Craig Warachka (2009). "Cashflow risk, systematic earnings revisions, and the cross-section of stock returns". *Journal of Financial Economics* 94 (3), 448–468.

Dang, Tung Lam, Fariborz Moshirian, and Bohui Zhang (2015). "Commonality in news around the world". *Journal of Financial Economics* 116 (1), 82–110.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam (1998). "Investor psychology and security market under- and overreactions". *Journal of Finance* 53 (6), 1839–1885.

Daniel, Kent, Sheridan Titman, and KC John Wei (2001). "Explaining the cross-section of stock returns in Japan: factors or characteristics?" *Journal of Finance* 56 (2), 743–766.

Daniel, Kent D, David Hirshleifer, and Avanidhar Subrahmanyam (2001). "Overconfidence, arbitrage, and equilibrium asset pricing". *Journal of Finance* 56 (3), 921–965.

De Bondt, Werner FM and Richard Thaler (1985). "Does the stock market overreact?" *Journal of Finance* 40 (3), 793–805.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal (2009). "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?" *Review of Financial Studies* 22 (5), 1915–1953.

DeMiguel, Victor, Alberto Martin-Utrera, Francisco J Nogales, and Raman Uppal (2020). "A transaction-cost perspective on the multitude of firm characteristics". *Review of Financial Studies* 33 (5), 2180–2222.

Dichtl, Hubert, Wolfgang Drobetz, Harald Lohre, Carsten Rother, and Patrick Vosskamp (2019). "Optimal timing and tilting of equity factors". *Financial Analysts Journal* 75 (4), 84–102.

Engelberg, Joseph, R David McLean, and Jeffrey Pontiff (2018). "Anomalies and news". *Journal of Finance* 73 (5), 1971–2001.

Fama, Eugene F and Kenneth R French (1992). "The cross-section of expected stock returns". *Journal of Finance* 47 (2), 427–465.

— (2006). "Profitability, investment and average returns". *Journal of Financial Economics* 82 (3), 491–518.

— (2016). "Dissecting anomalies with a five-factor model". *Review of Financial Studies* 29 (1), 69–103.

Fang, Lily and Joel Peress (2009). "Media coverage and the cross-section of stock returns". *Journal of Finance* 64 (5), 2023–2052.

Gibbons, Michael R, Stephen A Ross, and Jay Shanken (1989). "A test of the efficiency of a given portfolio". *Econometrica*, 1121–1152.

Hafez, P. A., J. A. Guerrero-Colon, and S. Duprey (2015). "Thematic alpha streams improve equity portfolio performance". *RavenPack Research Paper*.

Hafez, P. A. and Junqiang Xie (2011). *Introducing the RavenPack sentiment index*. Tech. rep. RavenPack Analytics.

Hafez, Peter Agder (2010). "News beta - A new measure for risk & stock analysis". *RavenPack Research Paper*.

Hanauer, Matthias (2014). "Is Japan different? Evidence on momentum and market dynamics". *International Review of Finance* 14 (1), 141–160.

Harvey, Campbell R, Yan Liu, and Alessio Saretto (2020). "An evaluation of alternative multiple testing methods for finance applications". *The Review of Asset Pricing Studies* 10 (2), 199–248.

Harvey, Campbell R, Yan Liu, and Heqing Zhu (2016). "... and the cross-section of expected returns". *Review of Financial Studies* 29 (1), 5–68.

Haugen, Robert A and Nardin L Baker (1991). "The efficient market inefficiency of capitalization-weighted stock portfolios". *Journal of Portfolio Management* 17 (3), 35–40.

— (1996). "Commonality in the determinants of expected stock returns". *Journal of Financial Economics* 41 (3), 401–439.

Heston, Steven L and Nitish Ranjan Sinha (2017). "News versus sentiment: Predicting stock returns from news stories". *Financial Analysts Journal* 73 (3), 67–83.

Hillert, Alexander, Heiko Jacobs, and Sebastian Müller (2014). "Media makes momentum". *Review of Financial Studies* 27 (12), 3467–3501.

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure". *Scandinavian Journal of Statistics*, 65–70.

Hou, Kewei, G Andrew Karolyi, and Bong-Chan Kho (2011). "What factors drive global stock returns?" *Review of Financial Studies* 24 (8), 2527–2574.

Hsu, Jason, Vitali Kalesnik, and Vivek Viswanathan (2015). "A framework for assessing factors and implementing smart beta strategies". *Journal of Index Investing* 6 (1), 89–97.

Huang, Ethan, Victor Liu, Li Ma, and James Osiol (2010). "Methods in dynamic weighting". *Capital IQ Working Paper*.

Jacobs, Heiko and Sebastian Müller (2020). "Anomalies across the globe: Once public, no longer existent?" *Journal of Financial Economics* 135 (1), 213–230.

Jegadeesh, Narasimhan (1990). "Evidence of predictable behavior of security returns". *Journal of Finance* 45 (3), 881–898.

Jegadeesh, Narasimhan and Sheridan Titman (1993). "Returns to buying winners and selling losers: Implications for stock market efficiency". *Journal of Finance* 48 (1), 65–91.

Jensen, Michael C, Fischer Black, and Myron S Scholes (1972). "The capital asset pricing model: Some empirical tests". *Studies in the Theory of Capital Markets. Praeger Publishers*.

Kan, Raymond and GuoFu Zhou (2012). "Tests of mean-variance spanning". *Annals of Economics and Finance* 13 (1), 139–187.

Ke, Zheng Tracy, Bryan T Kelly, and Dacheng Xiu (2019). "Predicting Returns with Text Data". *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2019-69).

Kolasinski, Adam C, Adam V Reed, and Matthew C Ringgenberg (2013). "A multiple lender approach to understanding supply and search in the equity lending market". *Journal of Finance* 68 (2), 559–595.

La Porta, Rafael, Josef Lakonishok, Andrei Shleifer, and Robert Vishny (1997). "Good news for value stocks: Further evidence on market efficiency". *Journal of Finance* 52 (2), 859–874.

Lee, Wai (2017). "Factors timing factors". *Journal of Portfolio Management* 43 (5), 66–71.

Lehmann, Bruce N (1990). "Fads, Martingales, and Market Efficiency". *Quarterly Journal of Economics* 105 (1), 1–28.

Leinweber, David and Jacob Sisk (2011). "Event driven trading and the 'new news'". *Journal of Portfolio Management* 38 (1).

Lo, Andrew W and A Craig MacKinlay (1990b). "Data-snooping biases in tests of financial asset pricing models". *Review of Financial Studies* 3 (3), 431–467.

Malkiel, Burton G and Eugene F Fama (1970). "Efficient capital markets: A review of theory and empirical work". *Journal of Finance* 25 (2), 383–417.

McLean, R David and Jeffrey Pontiff (2016). "Does academic research destroy stock return predictability?" *Journal of Finance* 71 (1), 5–32.

Novy-Marx, Robert (2013). "The other side of value: The gross profitability premium". *Journal of Financial Economics* 108 (1), 1–28.

RavenPack Analytics (2017). *User Guide and Service Overview*. Tech. rep. RavenPack Analytics.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf (2008). "Formalized data snooping based on generalized error rates". *Econometric Theory* 24 (2), 404–447.

Romano, Joseph P and Michael Wolf (2005). "Stepwise multiple testing as formalized data snooping". *Econometrica* 73 (4), 1237–1282.

Romano, Joseph P, Michael Wolf, et al. (2007). "Control of generalized error rates in multiple testing". *Annals of Statistics* 35 (4), 1378–1408.

Sloan, R (1996). "Do stock prices fully reflect information in accruals and cash flows about future earnings?" *Accounting Review* 71 (3), 289–315.

Tetlock, Paul C (2007). "Giving content to investor sentiment: The role of media in the stock market". *Journal of Finance* 62 (3), 1139–1168.

Tetlock, Paul C, Maytal Saar-Tsechansky, and Sofus Macskassy (2008). "More than words: Quantifying language to measure firms' fundamentals". *Journal of Finance* 63 (3), 1437–1467.

Tversky, Amos and Daniel Kahneman (1992). "Advances in prospect theory: Cumulative representation of uncertainty". *Journal of Risk and Uncertainty* 5 (4), 297–323.

Uhl, Matthias W, Mads Pedersen, and Oliver Malitius (2015). "What's in the news? Using news sentiment momentum for tactical asset allocation". *Journal of Portfolio Management* 41 (2), 100.

Wang, Ying, Bohui Zhang, and Xiaoneng Zhu (2018). "The momentum of news". *SSRN Working Paper*.

White, Halbert (2000). "A reality check for data snooping". *Econometrica* 68 (5), 1097–1126.

Chapter 2

Estimating Portfolio Risk for Tail Risk Protection Strategies

This project is joint work with my supervisors Harald Lohre and Ingmar Nolte and is published in the European Financial Management journal (*European Financial Management*, 26(4), 1107-1146). We thank Torben Andersen, Ole Christian Bech-Moen, John A. Doukas, Christian Groll, Manel Kammoun, Mark Kritzman, Yifan Li, Stefan Mittnik, James Taylor, Ralf Wilke as well as the participants in the 2017 Paris Financial Management Conference (PFMC), the 2017 CEQURA Conference on Advances in Financial and Insurance Risk Management in Munich, the 3rd KoLa Workshop on Finance and Econometrics at Lancaster University Management School in 2017, the 2017 Doctoral Workshop on Applied Econometrics at the University of Strasbourg, the 2017 Global Research Meeting of Invesco Quantitative Strategies in Boston, the 2018 Frontiers of Factor Investing Conference in Lancaster, the 2018 FMA European Conference in Kristiansand and the 2018 IFABS Porto Conference for fruitful discussions and suggestions. This work was supported by funding from the Economic and Social Research Council (UK).

2.1. Introduction

Tail risk hedging strategies are of vital interest to many market participants to protect investment portfolios against extreme negative market moves. An obvious protection is the purchase of a put option. However, such a strategy can be expensive, since the option premium is payable each investment period, although the protection could prove unnecessary in the majority of periods. A possible alternative are dynamic asset allocation strategies, which aim to improve the downside risk profile of an investment strategy without jeopardizing its long-term return potential by dynamically shifting between a risky asset (or portfolio) and a risk-free asset.

Risk targeting strategies[1] are one such possibility (Hocquard, Ng, and Papageorgiou, 2013; Perchet et al., 2015; Bollerslev, Hood, et al., 2018a). A risk targeting strategy controls portfolio risk over time by taking advantage of the negative relationship between risk and return. Specifically, the investment exposure of the portfolio is adjusted according to updated risk forecasts in order to keep the *ex-ante* risk at a constant target level. A stricter way to limit downside risk is to apply portfolio insurance strategies, such as the constant proportion portfolio insurance (CPPI) strategy (Perold, 1986; Black and Jones, 1987, 1988; Perold and Sharpe, 1988), where the investor defines a minimum capital level to be preserved at the end of the investment period. The key element in determining the investment exposure to the risky asset is the so-called multiplier. This represents the inverse of the maximum sudden loss of the risky asset, so that a given risk budget will not be fully consumed and the portfolio value will not fall below the protection level. Initially, the multiplier was assumed to be static and unconditional (e.g. Bertrand and Prigent, 2002; Balder, Brandl, and Mahayni, 2009; Cont and Tankov, 2009). However, given the empirical characteristics of financial assets, such as time-varying volatility or volatility clustering (cf. Longin and Solnik, 1995; Andersen et al., 2006), other studies (e.g. Hamidi, Maillet, and Prigent, 2014) propose to model the multiplier as time-varying and conditional. The corresponding strategy is known as dynamic proportion portfolio insurance (DPPI).[2]

The success of both dynamic tail risk protection strategies strongly depends on the success of forecasting tail risk (Perchet et al., 2015). Given a plethora of available risk models, we contribute to the existing literature on tail risk protection strategies by investigating suitable risk models for timely management of the investment exposure in dynamic tail risk protection strategies. At the same time, we contribute to the literature on risk model evaluation by

[1] Risk targeting strategies are also known as constant risk, target risk or inverse risk weighting strategies.

[2] For a comprehensive literature review on portfolio insurance and CPPI/DPPI multipliers, see Benninga (1990), Black and Perold (1992), Basak (2002), Dichtl and Drobetz (2011), and Hamidi, Maillet, and Prigent (2014), among others.

assessing not only its statistical performance but also its economical relevance when testing the risk forecasts in a relevant portfolio application.

Risk targeting strategies have been extensively backtested using historical data, and are known to show superior performance compared to a simple buy-and-hold strategy (Cooper, 2010; Kirby and Ostdiek, 2012; Ilmanen and Kizer, 2012; Giese, 2012). Hallerbach (2012, 2015) demonstrates that the Sharpe ratio increases, even if the portfolio mean return is constant over time. Constant volatility portfolios not only deliver higher Sharpe ratios than their passive counterpart but also reduce drawdowns (Hocquard, Ng, and Papageorgiou, 2013). Similar to the risk targeting strategy that we apply, the dynamic value-at-risk (VaR) portfolio insurance strategy of Jiang, Ma, and An (2009) aims to control the exposure of a risky asset such that a specified VaR is not violated. However, their strategy can only be applied to parametric location-scale models, while the one we apply is compatible with any type of risk model. In a similar vein, Bollerslev, Hood, et al. (2018a) use a risk targeting strategy to compare realized volatility models to more conventional procedures that do not incorporate the information in high-frequency intraday data.

The literature on DPPI puts forward various ways to model the conditional time-varying multiplier. While Ben Ameur and Prigent (2007, 2014) employ generalized autoregressive conditional heteroskedasticity (GARCH) type models, Hamidi, Jurczenko, and Maillet (2009) and Hamidi, Maillet, and Prigent (2009) define the multiplier as a function of a dynamic autoregressive quantile model of the VaR according to Engle and Manganelli (2004). In contrast, Chen et al. (2008) propose a multiplier framework based on genetic programming. More recently, Hamidi, Maillet, and Prigent (2014) employ a dynamic autoregressive expectile (DARE) model to estimate the conditional multiplier.[3] All these papers provide evidence that the DPPI strategy, based on a time-varying conditional risk estimate, outperforms the traditional CPPI strategy.

We are particularly interested in comparing different ways to determine the risky investment exposure of dynamic tail risk protection strategies, assessing various models to estimate a portfolio's downside risk measured by VaR and expected shortfall (ES). While the literature suggests a myriad of VaR and ES models—Andersen et al. (2006, 2013), Kuester, Mittnik, and Paolella (2006), and Righi and Ceretta (2015) provide thorough summaries on market risk modeling—practitioners still only use a limited number of them. This discrepancy might be due to complexity, (computational) efficiency or the perception that the incremental benefit of implementing a highly sophisticated model is minor. Therefore, we examine

[3] Hamidi, Maillet, and Prigent (2014) model the multiplier as a function of the expected shortfall determined by a combination of quantile functions in order to reduce the potential model error. Specifically, they combine the historical simulation approach, three methods based on distributional assumptions, and four methods based on expectiles and conditional autoregressive specifications into the DARE approach.

simple methods that are common among practitioners as well as more involved methods to predict VaR and ES. Specifically, we consider: historical simulation (HS), Cornish-Fisher approximation (CFA), RiskMetrics, quantile and expectile regressions, extreme value theory, copula-GARCH and recent generalized autoregressive score (GAS) models (including one and two-factor GAS models as well as the hybrid GAS/GARCH model).

The primary issue of these (standalone) risk models is that their performance and reliability in accurately predicting risk often depend heavily on the data. While a parsimonious model can perform well in stable markets, it might fail during a volatile period. Likewise, highly parameterized models can be adequate during periods of high volatility, but might be easily outperformed by simpler approaches in less turbulent times (cf. Bayer, 2018). Hence, it is often beneficial to combine predictions originating from various approaches. Reviewing forecasting combinations, Timmermann (2006) provides three arguments in favor of combining forecasts to enhance the predictive performance relative to standalone models. First, there are diversification gains arising from the combination of forecasts computed from different assumptions, specifications or information sets. Second, combination forecasts tend to be robust against structural breaks. Third, the influence of potential misspecification biases and measurement errors of the individual models is reduced due to averaging over a set of forecasts derived from several models.

While there exist various approaches to combine VaR predictions (see Bayer, 2018, for a summary), the literature is lacking methods that combine ES predictions. This shortage relates to the fact that ES is not "elicitable", that is, there does not exist a loss function such that the correct ES forecast is the solution to minimizing the expected loss (cf. Gneiting, 2011). This lack of elicitability makes the estimation and backtesting of ES challenging (see Acerbi and Szekely, 2014; Embrechts and Hofert, 2014; Emmer, Kratz, and Tasche, 2015). As a remedy, Fissler and Ziegel (2016) recently introduced a class of loss functions that overcome the lack of elicitability for ES by jointly modeling ES and VaR. Drawing on their results, we propose a novel ES (and VaR) forecast combination approach.[4]

Based on a global multi-asset data set consisting of stock, bond, commodity and currency indices, our empirical study documents a clear superiority of the proposed forecast combination approach over both sophisticated and more naive standalone models using a state-of-the-art ES and VaR backtesting framework (Kupiec, 1995; Christoffersen, 1998; McNeil and Frey, 2000; Christoffersen and Pelletier, 2004; Engle and Manganelli, 2004; Berkowitz, Christoffersen, and Pelletier, 2011; Nolde and Ziegel, 2017; Bayer and Dimitriadis, 2020; Patton, Ziegel, and Chen, 2019). In the combination of ES (and VaR) forecasts, complexity

[4]In recent independent work, Taylor (2020) uses the same class of loss functions to combine VaR and ES predictions. While his methods for combining forecasts are in the spirit of Bates and Granger (1969) and Shan and Yang (2009), our methodology follows the approach of Halbleib and Pohlmeier (2012) and Bayer (2018).

seems to actually pay off as the proposed forecast combination approach outperforms a simple average forecast. Among the standalone models, sophisticated risk models such as the extreme value theory or the copula-GARCH approach outperform simple approaches in terms of historical accuracy and statistical fit. When subsequently feeding the risk forecasts in the tail risk protection framework, our findings are twofold. For the risk targeting strategy, we observe a clear outperformance of the more intricate methods, confirming the results from the statistical analysis. For the DPPI strategy, we likewise show that the use of more sophisticated risk models helps to protect investors from downside risk. Yet, more naive approaches do not fall short of providing downside protection. Given that the portfolio insurance strategy automatically reduces investment exposure when approaching the protection level, the less sophisticated methods' weaknesses seem to be compensated by this second line of defense.

The remainder of the paper is structured as follows. Section 2.2 discusses the tail risk protection strategies employed. Section 2.3 briefly presents the different models to estimate portfolio risk, including the novel forecast combination technique based on Fissler-Ziegel loss functions. In Section 2.4 we carry out the empirical study using a global multi-asset data set to compare the performance of dynamic tail risk protection strategies based on the different risk models. Section 2.5 concludes.

2.2. Tail risk protection strategies

We consider a risk-averse investor who aims to limit the downside risk of his investment over an investment horizon of H time steps. Let $t = 1, 2, ..., T$ be the time index of portfolio rebalancing and $I(t) = t - (\lceil t/H \rceil - 1)\, H$ a subindex for each investment period $\lceil t/H \rceil$, so that the latter runs from 1 to H in each investment period. At the beginning of each investment period $\lceil t/H \rceil$, that is, at $I(t) = 1$, the investor determines a risk target that should be achieved at the end of the period, that is, at $I(t) = H$. The management of the protected portfolio follows a dynamic portfolio allocation. In particular, the value of the protected portfolio, denoted by V_t, is invested in a risky asset (or portfolio) and a non-risky asset in such a way that the given risk target will not be violated. The price of the risky asset at time t is denoted as P_t, so that the logarithmic return from t to $t + 1$ is $r_{t+1} = \log(P_{t+1}/P_t)$. Accordingly, the price and return of the non-risky asset are denoted by B_t and $r_{f,t}$. To explicitly determine the exposure to the risky asset e_t, we need to forecast the downside risk of the risky asset over the next day, quantified by a risk measure $\rho(\cdot)$.

2.2.1. Risk targeting strategies

A risk targeting strategy systematically adjusts exposure to a given asset (or portfolio) conditional on its current risk (forecast) in order to maintain a pre-specified target risk level. Specifically, if a portfolio's current risk is higher than the target level, one would lower the investment exposure by shifting towards the risk-free asset, and vice versa if the current risk is lower than the target level. The rationale for maintaining a constant risk level is twofold (see Hocquard, Ng, and Papageorgiou, 2013). First, most significant market corrections have been preceded by an increase in risk. By conditioning their exposure on market risk, investors can dampen the impact of a market correction. Second, empirical evidence suggests that asset returns tend to be greater during periods of low risk. Consequently, investors should maximize asset exposure during these periods, taking advantage of a favorable risk-reward tradeoff. Conversely, they should decrease asset exposure when risk increases to maintain the desired risk level.

Given the level of *ex ante* risk of the underlying risky asset $\rho_t(r_{t+1})$ and the predefined target risk $\overline{\rho}$, the allocation to the risky asset e_t is simply $\overline{\rho}/\rho_t(r_{t+1})$. As $\rho_t(r_{t+1})$ is unknown, we utilize a forecast based on the information available at time t, $\mathcal{F}_t$:

$$e_t \equiv \frac{\overline{\rho}}{\hat{\rho}_t(r_{t+1}|\mathcal{F}_t)}. \tag{2.1}$$

Correspondingly, the weight of the risk-free asset is given by $1 - e_t$.

2.2.2. Constant and Dynamic Proportion Portfolio Insurance

The constant proportion portfolio insurance (CPPI) strategy (see Perold, 1986; Black and Jones, 1987, 1988; Perold and Sharpe, 1988) dynamically shifts between the risky and non-risky asset to guarantee that the investor at least recovers a given proportion of her initial capital. At the beginning of each investment period $\lceil t/H \rceil$, that is, at $I(t) = 1$, the investor determines this minimum portfolio value, or floor $F_{\lceil t/H \rceil}$, that should be preserved at the end of the period, that is, at $I(t) = H$. The corresponding risk capital, called the cushion, is derived as the difference of portfolio value, V_t, and the discounted floor (i.e. the net present value, NPV$(\cdot)$, of the floor):

$$C_t = V_t - \text{NPV}_t\left(F_{\lceil t/H \rceil}\right). \tag{2.2}$$

The cushion represents a certain amount of the portfolio value to absorb potential market shocks before the portfolio manager can rebalance the portfolio. In order to avoid a breach of the discounted floor, the investment exposure to the risky asset, defined as $E_t = e_t V_t$,

should be set such that the cushion at t is at least as high as the maximum sudden drop in the portfolio value between the rebalancing dates t and $t + 1$, that is

$$C_t \geq V_t \left| \inf \left(\log \left(\frac{V_{t+1}}{V_t} \right) \right) \right|. \tag{2.3}$$

As the portfolio consists of the risky and the non-risky asset, Equation (2.3) can be simplified to

$$C_t \geq e_t V_t \left| \inf \left(\log \left(\frac{P_{t+1}}{P_t} \right) \right) \right|. \tag{2.4}$$

Rearranging Equation (2.4) then yields the (total) exposure to the risky asset as

$$E_t \leq C_t |\inf(r_{t+1})|^{-1} = C_t m, \tag{2.5}$$

where $m \equiv |\inf(r_{t+1})|^{-1}$ is the multiplier.[5] The multiplier indicates how often a given cushion can be invested in the risky asset without breaching the floor. Thus, it reflects the investor's risk tolerance. The higher the multiplier, the more the investor will participate in upward market movements of the underlying. But the higher the multiplier, the faster the portfolio will reach the floor when there is a sustained decrease in the price of the underlying. In order to allow for the highest possible participation in the underlying risky asset, it is common to set E_t such that equality holds in Equation (2.5). The remainder is invested in the risk-free asset.

If rebalancing were continuous and price movements sufficiently smooth, the CPPI allocation rule would ensure that the portfolio does not fall below the floor (Cont and Tankov, 2009; Balder, Brandl, and Mahayni, 2009; Hamidi, Hurlin, et al., 2015; Ardia, Boudt, and Wauters, 2016). However, with discrete rebalancing and jumps in prices, there is a non-negligible probability that the floor is breached. This risk of losing more than the cushion between two rebalancing dates and thus failing to ensure the protection at the end of the investment period is called gap risk. A common way to minimize gap risk is to employ the minimum return of the risky asset over the sample history, that is, $\inf(r_{t+1}) = \min(r_1, \ldots, r_{t+1})$. Then, the CPPI strategy is based on a static unconditional multiplier, often reflecting a constant worst-case scenario. Although such a conservative stance would have meaningfully addressed catastrophic drawdowns during extreme market turmoil, it would also have unduly capped upside potential over the long term. Dynamic proportion portfolio insurance (DPPI)

[5] We follow Benninga (1990) and restrict the participation ratio to vary between 0% and 100% of the risky asset in order to rule out short positions. This approach leads to a slightly different exposure definition: $E_t = \max[\min(C_t m, V_t), 0]$.

is designed to introduce more flexibility. Instead of using a static multiplier, the risk budget adapts dynamically to changes in a risk forecast, measured by $\hat{\rho}(\cdot)$. Thus, the exposure changes to

$$E_t = C_t|\hat{\rho}_t(r_{t+1}|\mathcal{F}_t)|^{-1} = C_t m_t, \tag{2.6}$$

where the risk forecast $\hat{\rho}_{t+1}$ is based on information $\mathcal{F}_t$ and measures the risk when the risky asset price P evolves from t to $t+1$. The dynamic multiplier is therefore given by

$$m_t = |\hat{\rho}_t(r_{t+1}|\mathcal{F}_t)|^{-1}. \tag{2.7}$$

In this way, the portfolio's exposure to the risky asset reacts to changes in the risk forecast, ensuring that it does not remain artificially low as a result of a constant conservative risk assumption. For this to work in practice, the risk model must be capable of quickly homing in on volatility spikes, and just as quickly readjusting to a normalization of market volatility.

The advantage of the CPPI and DPPI strategy, respectively, is the simple practical implementation that does not call for forecasting the returns of the risky asset. Major drawbacks are the strategies' path dependencies as well as the lock-in effect. Depending on the underlying portfolio return path, the CPPI/DPPI strategy can deliver a wide range of outcomes. In general, the more volatile the risky asset, the lower the average participation ratio. While the CPPI strategy is fully exposed to the problem of path dependency, the DPPI strategy can mitigate this problem at least to some extent by quickly reacting to market changes. The lock-in effect occurs when the cushion is entirely consumed by portfolio losses. The CPPI/DPPI strategy is then fully invested in the risk-free asset until the end of the investment period and no participation in subsequent upward movements is possible.

2.3. Portfolio risk modeling

Given the vast amount of available risk models (see Kuester, Mittnik, and Paolella, 2006; Nadarajah, Zhang, and Chan, 2014), we focus on a few distinct approaches, ranging from rather simple models that are widely used by practitioners to more intricate, state-of-the-art models in the academic literature. We consider both portfolio-level (aggregated, "top-down") and asset-level (disaggregated, "bottom-up") risk modeling. In addition, we propose a new VaR and ES forecast combination approach based on a loss function that overcomes the

lack of elicitability for ES by jointly modeling ES and VaR. Following the description of downside risk measurement, we summarize the various methods in this section.[6]

2.3.1. Conditional risk measurement

The literature suggests various ways to quantify market risk of financial assets. As we are particularly interested in protecting risky portfolios against extreme market losses, we resort to the most common downside risk measures, value-at-risk (VaR) and expected shortfall (ES). VaR measures the maximum potential portfolio loss at a given confidence level.[7] Therefore, the VaR forecast from t to $t+1$ is simply the negative p-quantile of the conditional return distribution at $t+1$, that is,

$$\text{VaR}^{p}_{t+1|t} = Q_p(r_{t+1}|\mathcal{F}_t) = \inf_x\{x \in \mathbb{R} : P\left(r_{t+1} \leq x|\mathcal{F}_t\right) \geq p\}, \tag{2.8}$$

where $p \in (0,1)$ is the probability level, $Q_p(\cdot)$ denotes the quantile function, r_t reflects the return of the asset (portfolio) in period t and $\mathcal{F}_t$ represents the information available at time t.

Although VaR is still the risk measure of choice in the financial industry, it has been criticized for disregarding outcomes beyond the specified VaR-quantile. Moreover, VaR is not a subadditive risk measure. This property posits that the total portfolio risk should not be greater than the sum of the risks of its constituents (see Artzner et al., 1999; Acerbi and Tasche, 2002; Taylor, 2008). Expected shortfall, also known as conditional VaR or expected tail loss, is a risk measure that overcomes these weaknesses by aggregating information about the tail of the portfolio return distribution. It is defined as the conditional expectation of the return given that VaR is exceeded (see Yamai and Yoshiba, 2002), specifically

$$\text{ES}^{p}_{t+1|t} = p^{-1}\int_0^p \text{VaR}^{s}_{t+1|t}\, ds. \tag{2.9}$$

Throughout this paper, we focus on the probability level p = 1%, taking a conservative stance for the tail risk protection strategies.

[6] For a rigorous discussion of the risk models analyzed, see Kuester, Mittnik, and Paolella (2006), Andersen et al. (2013) and Righi and Ceretta (2015). Note that we do not impose the same estimate of the standard deviation in all location-scale risk models. We intentionally consider simple methods (such as historical simulation, Cornish-Fisher approximation and RiskMetrics) in a way that practitioners would often use them to compute VaR and ES forecasts. These models are then contrasted with more sophisticated methods popular in the academic literature, which we construct consistent with the original studies.

[7] The literature commonly uses low-probability terminology, hence we are speaking of a 1% VaR rather than a 99% VaR.

2.3.2. Conditional portfolio-level risk models

Generally, there are two classes of risk modeling, depending on the aggregation level. Portfolio-level analysis, as discussed in this section, requires only a univariate model based on aggregated portfolio returns. The latter can easily be constructed using portfolio holdings $\mathbf{w}_t = \left[w_{1,t}, w_{2,t}, ..., w_{N,t}\right]'$ and the individual asset returns $\mathbf{r}_t = \left[r_{1,t}, r_{2,t}, ..., r_{N,t}\right]'$:

$$r_{\text{PF},t} = \sum_{i=1}^{N} w_{i,t} r_{i,t} = \mathbf{w}_t' \mathbf{r}_t, \quad t = 1, 2, ..., T. \tag{2.10}$$

While aggregation generally entails the loss of information, Andersen et al. (2013) argue that there is no reason why a parsimonious dynamic model should not be estimated for portfolio-level returns. If one is interested in the portfolio return distribution, one may model it directly rather than via aggregation based on a larger, and almost inevitably less well-specified, multivariate model.[8]

Historical simulation

The simplest way to estimate VaR and ES is to use the sample quantile function based on historic return data, which is referred to as historical simulation (HS). Let $r_{\text{PF},(1)} \leq r_{\text{PF},(2)} \leq ... \leq r_{\text{PF},(t)}$ denote the order statistics in ascending order corresponding to the original historical pseudo portfolio returns $r_{\text{PF},1}, r_{\text{PF},2}, ..., r_{\text{PF},t}$. Then, the HS-VaR for $t + 1$ is simply the empirical $100p$th quantile or the tpth order statistic, that is,

$$\text{VaR}^p_{t+1|t} = r_{\text{PF},(\lceil tp \rceil)}. \tag{2.11}$$

Correspondingly, the ES estimate for $t + 1$ can be computed as

$$\text{ES}^p_{t+1|t} = \left(\sum_{i=\lceil tp \rceil}^{t} r_{\text{PF},(i)} \right) (t - \lceil tp \rceil)^{-1}. \tag{2.12}$$

The main advantage of the HS approach is its computational simplicity and non-parametric dimension, that is, VaR and ES do not rely on any distributional assumptions. In contrast, the HS approach cannot properly incorporate conditionality (see Andersen et al., 2006).[9] This deficiency of the conventional HS approach is oftentimes highlighted by a clustering of VaR violations in time, reflecting a failure to properly account for persistent changes in market

[8] In contrast to Andersen et al. (2013), some studies (e.g. Santos, Nogales, and Ruiz, 2012) take an opposite view regarding the validity of aggregation at portfolio level when estimating portfolio risk. They argue that univariate models are often misspecified if the true return generating process is multivariate.

[9] For a rigorous discussion of several serious issues of the HS approach, see Pritsker (2006).

volatility. The only source of dynamics in the HS risk estimates is the evolving window used to construct historical pseudo portfolio returns. Nevertheless, the choice of the window size is crucial: too few observations will lead to sampling error, whereas too many observations will slow down the reaction to changes in the true return distribution. Moreover, the risk estimates can jump when large negative returns either enter or exit the estimation window.

Cornish-Fisher approximation

Another simple approach is the Cornish-Fisher Approximation (CFA) method (Zangari, 1996), where the VaR is modeled as a Taylor series type expansion (cf. Cornish and Fisher, 1938) around the VaR of a normal distribution. Specifically, the CFA-VaR is an extension of the normal quantile function by accounting for skewness γ and kurtosis δ, and is calculated as

$$\text{VaR}^{p}_{t+1|t} = \mu_t + \sigma_t F^{-1}_{CF}(p), \tag{2.13}$$

where

$$F^{-1}_{CF}(p) \equiv \Phi^{-1}_p + \left([\Phi^{-1}_p]^2 - 1\right)\frac{\gamma}{6} + \left([\Phi^{-1}_p]^3 - 3\Phi^{-1}_p\right)\frac{\delta - 3}{24} - \left(2[\Phi^{-1}_p]^3 - 5\Phi^{-1}_p\right)\frac{\gamma^2}{36}$$

and $\Phi(\cdot)$ is the standard normal cumulative distribution function. Moreover, μ_t and σ_t are computed by the sample mean and sample standard deviation, respectively.

Boudt, Peterson, and Croux (2008) show how the Edgeworth and Cornish-Fisher expansions of the density and quantile functions can be used to obtain an estimator for ES that delivers accurate downside risk estimates even in the presence of non-normal returns. The modified or Cornish-Fisher ES is thus computed as

$$\text{ES}^{p}_{t+1|t} = \mu_t + \sigma_t \mathbb{E}_{F_{CF}}\left[z | z \leq F^{-1}_{CF}(p)\right] \tag{2.14}$$

where

$$\begin{aligned}\mathbb{E}_{F_{CF}}\left[z | z \leq F^{-1}_{CF}(p)\right] = -\frac{1}{p}\Bigg(&\phi(F^{-1}_{CF}(p)) + \frac{\delta}{24}\left[I^4 - 6I^2 + 3\phi(F^{-1}_{CF}(p))\right] + \frac{\gamma}{6}\left[I^3 - 3I^1\right] \\ &+ \frac{\gamma^2}{72}\left[I^6 - 15I^4 + 45I^2 - 15\phi(F^{-1}_{CF}(p))\right]\Bigg)\end{aligned}$$

with

$$I^q = \begin{cases} \sum_{i=1}^{q/2} \left(\frac{\prod_{j=1}^{q/2} 2j}{\prod_{j=1}^{i} 2j} \right) g_p^{2i} \phi(g_p) + \left(\prod_{j=1}^{q/2} 2j \right) \phi(g_p) & \text{for } q \text{ even} \\ \sum_{i=0}^{q^*} \left(\frac{\prod_{j=0}^{q^*} (2j+1)}{\prod_{j=0}^{i} (2j+1)} \right) g_p^{2i+1} \phi(g_p) - \left(\prod_{j=0}^{q^*} 2(j+1) \right) \phi(g_p) & \text{for } q \text{ odd} \end{cases}$$

and $q^* = (q-1)/2$, $g_p = F_{CF}^{-1}(p)$. $\phi(\cdot)$ denotes the standard normal probability density function.

The main advantage of the CFA method is its ability to account for fat tails. However, the CFA-VaR is not necessarily monotone, that is, the 1% VaR might be smaller than the 5% VaR. Martin and Arora (2017) also document that the CFA-VaR and CFA-ES suffer in terms of statistical efficiency.

Quantile/Expectile regression

As VaR and ES are directly linked to quantiles and expectiles, a natural approach to risk modeling employs quantile and expectile regressions. The basic idea of quantile regression is to directly model the conditional quantile rather than the whole distribution of portfolio returns. More precisely, the conditional p-quantile, $Q_p(r_{PF,t}|\mathcal{F}_{t-1}) = -\text{VaR}_{t|t-1}$, is modeled as a parametric function of the information $\mathcal{F}_{t-1}$:

$$\text{VaR}_{t|t-1}^p \equiv g_p\left(\mathcal{F}_{t-1}; \beta_p\right), \tag{2.15}$$

where $g(\cdot,\cdot)$ and the parameter vector β explicitly depend on p. Following Koenker and Bassett (1978), the conditional sample p-quantile can be found as the solution to

$$\min_{\beta_p} \left\{ \sum_{r_{\text{PF},t} \geq \text{VaR}_{t|t-1}^p} p|r_{\text{PF},t} - \text{VaR}_{t|t-1}^p| + \sum_{r_{\text{PF},t} < \text{VaR}_{t|t-1}^p} (1-p)|r_{\text{PF},t} - \text{VaR}_{t|t-1}^p| \right\}, \tag{2.16}$$

where we determine VaR_t^p by the conditional autoregressive value-at-risk (CAViaR) specification of Engle and Manganelli (2004). In particular, we adopt their asymmetric slope CAViaR model,[10] given by

$$\text{VaR}_{t|t-1}^p = \beta_0 + \beta_1 \text{VaR}_{t-1|t-2}^p + \beta_2 \max\left[r_{\text{PF},t-1}, 0\right] + \beta_3 \max\left[-r_{\text{PF},t-1}, 0\right]. \tag{2.17}$$

[10] For the sake of simplicity, we focus on one CAViaR model. Particularly, we choose the asymmetric slope specification because of its ability to accommodate the leverage effect.

In a similar fashion, we can use expectile regression to estimate ES. In particular, we employ the conditional autoregressive expectile (CARE) model of Taylor (2008). First, we suppose that the population τ_p expectile of $r_{\text{PF},t}$ is the parameter μ_{τ_p} that minimizes the function $\mathbb{E}\left[|\tau_p - \mathbb{1}(r_{\text{PF},t} - \mu_{\tau_p})|(r_{\text{PF},t} - \mu_{\tau_p})^2\right]$. Hence, we can represent the conditional expectile as a parametric function of past information, that is, $\mu_{\tau_p}(r_{\text{PF},t}) \equiv h_{\tau_p}\left(\mathcal{F}_{t-1}; \gamma_{\tau_p}\right)$. The parameters γ_{τ_p} can be estimated using asymmetric least squares (cf. Newey and Powell, 1987), that is,

$$\min_{\gamma_{\tau_p}} \left\{ \sum_{r_{\text{PF},t}} |\tau_p - \mathbb{1}(r_{\text{PF},t} < h_{\tau_p}\left(\mathcal{F}_{t-1}; \gamma_{\tau_p}\right))| \left(r_{\text{PF},t} - h_{\tau_p}\left(\mathcal{F}_{t-1}; \gamma_{\tau_p}\right)\right)^2 \right\}, \tag{2.18}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Similar to the asymmetric slope CAViaR model, we assume the conditional expectile to have an asymmetric slope specification. The ES can then be computed as the product of a correction term and the conditional expectile, that is,

$$\text{ES}^p_{t|t-1} = \left(1 + \frac{\tau_p}{(1-2\tau_p)\, p}\right)\left(\gamma_0 + \gamma_1 \mu_{\tau_p}(r_{\text{PF},t-1}) + \gamma_2 \max\left[r_{\text{PF},t-1}, 0\right] + \beta_3 \max\left[-r_{\text{PF},t-1}, 0\right]\right). \tag{2.19}$$

The quantile and expectile regression approach are appealing because no explicit distributional assumption for the time series behavior of returns is needed, thus reducing the risk of model misspecification. The main drawback of the CAViaR modeling strategy is that it might generate out-of-order quantiles similar to the CFA method. Also, estimation of model parameters is challenging.[11]

Extreme value theory

As we are primarily interested in the tails of the portfolio distribution, it seems natural to resort to extreme value theory (EVT) which estimates the tails based on extrapolating from available observations. McNeil and Frey (2000) propose a semi-parametric framework based on extreme value theory to describe the tail of the conditional distribution. The first step is to employ pseudo-maximum-likelihood fitting of GARCH(1,1) models to estimate conditional volatility forecasts $\hat{\sigma}_{t+1}$ (Engle, 1982; Bollerslev, 1986; Taylor, 1986). In a second step, we resort to EVT to estimate the tail of the innovation distribution of the GARCH(1,1) model. In particular, we use the peak-over-threshold method where a generalized Pareto distribution

[11] We thank James Taylor for providing the Gauss code for his CARE models.

(GPD) is fitted to the negative of portfolio returns over a specified threshold.[12] The quantile $\hat{z}_p$ can then be estimated as

$$\hat{z}_p = u + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{1-p}{n_u/n} \right)^{-\hat{\xi}} - 1 \right], \tag{2.20}$$

where $\hat{\beta}$ and $\hat{\xi}$ are the GPD estimates and n_u is the number of observations above threshold u. Consequently, the VaR and ES forecasts can be computed as

$$\mathrm{VaR}^p_{t+1|t} = \hat{\sigma}_{t+1} \hat{z}_p, \tag{2.21}$$

$$\mathrm{ES}^p_{t+1|t} = \hat{\sigma}_{t+1} \hat{z}_p \left(\frac{1}{1-\hat{\xi}} + \frac{\hat{\beta} - \hat{\xi} u}{(1-\hat{\xi}) \hat{z}_p} \right). \tag{2.22}$$

The GARCH fitting in the first step enables us to capture certain stylized facts such as time-varying volatility, fat tails and volatility clustering. Then, EVT is particularly suitable to estimate the tails of the distribution. The crucial assumption of EVT is, however, that one is in the tail of the distribution. Hence, the difficulty is the determination of the threshold. If the threshold is too low, then the approximation can hardly be justified and the associated risk estimates may be biased. If the threshold is too high, there are too few observations over the threshold, resulting in highly volatile estimates.

GAS Models

Patton, Ziegel, and Chen (2019) propose dynamic VaR and ES models drawing on recent results from statistical decision theory that overcome the lack of elicitability[13] for ES by jointly modeling ES and VaR (cf. Fissler and Ziegel, 2016).[14] These models are semi-parametric in that they impose parametric structures for the dynamics of ES and VaR according to the "generalized autoregressive score" (GAS) framework proposed by Creal, Koopman, and Lucas (2013) and Harvey (2013), but are completely agnostic about the conditional distribution of returns (aside from regularity conditions required for estimation and inference).

[12] We follow McNeil and Frey (2000) when determining the thresholds. See their paper for details.

[13] A statistical functional (e.g. a risk measure) is said to be "elicitable" if there exists a loss function such that the correct forecast of the functional is the solution to minimizing the expected loss (cf. Gneiting, 2011; Fissler and Ziegel, 2016; Patton, Ziegel, and Chen, 2019). For example, the mean is elicitable using the quadratic loss function, and VaR is elicitable using the piecewise-linear or "tick" loss function.

[14] Similar to the approach of Patton, Ziegel, and Chen (2019), Taylor (2019) proposes using the asymmetric Laplace distribution to jointly estimate dynamic models for VaR and ES.

Two-factor GAS model

The two-factor GAS(1,1) model allows ES and VaR to evolve as two separate, correlated processes:

$$\begin{bmatrix} \mathrm{VaR}^p_{t+1|t} \\ \mathrm{ES}^p_{t+1|t} \end{bmatrix} = \mathbf{v} + \mathbf{B} \begin{bmatrix} \mathrm{VaR}^p_{t|t-1} \\ \mathrm{ES}^p_{t|t-1} \end{bmatrix} + \mathbf{A}\mathbf{H}_t^{-1}\nabla_t \tag{2.23}$$

where $\mathbf{v}$ is a 2×1-vector and $\mathbf{B}$ and $\mathbf{A}$ are 2×2-matrices. The forcing variable in this model is a function of the derivative, ∇_t, and the Hessian, $\mathbf{H}_t$, of the "FZ loss function" (see Section 2.3.4 for details on this loss function):

$$\mathbf{H}_t^{-1}\nabla_t = \begin{bmatrix} \frac{-1}{k_p}\lambda_{\mathrm{VaR},t} \\ \frac{-1}{p}\left(\lambda_{\mathrm{VaR},t} + p\lambda_{\mathrm{ES},t}\right) \end{bmatrix}, \tag{2.24}$$

where k_p is a constant with the same sign as VaR_t and

$$\lambda_{\mathrm{VaR},t} \equiv -\mathrm{VaR}^p_{t|t-1}\left(\mathbb{1}\left(r_{\mathrm{PF},t} \leq \mathrm{VaR}^p_{t|t-1}\right) - p\right), \tag{2.25}$$

$$\lambda_{\mathrm{ES},t} \equiv \frac{1}{p}\mathbb{1}\left(r_{\mathrm{PF},t} \leq \mathrm{VaR}^p_{t|t-1}\right)\mathrm{VaR}^p_{t|t-1} - \mathrm{ES}^p_{t|t-1}. \tag{2.26}$$

As the second term in the model is a linear combination of the two elements of the forcing variable, and since the forcing variable is premultiplied by a coefficient matrix, say $\tilde{\mathbf{A}}$, we can equivalently use

$$\tilde{\mathbf{A}}\mathbf{H}_t^{-1}\nabla_t = \mathbf{A}\boldsymbol{\lambda}_t, \tag{2.27}$$

$$\text{where} \quad \boldsymbol{\lambda}_t \equiv \left[\lambda_{\mathrm{VaR},t}, \lambda_{\mathrm{ES},t}\right]'. \tag{2.28}$$

One-factor GAS model

As a simpler variant, Patton, Ziegel, and Chen (2019) introduce the one-factor GAS model, where both VaR and ES are driven only by a single variable, κ_t,

$$\mathrm{VaR}^p_{t+1|t} = a\exp(\kappa_{t+1}), \tag{2.29}$$

$$\mathrm{ES}^p_{t+1|t} = b\exp(\kappa_{t+1}), \tag{2.30}$$

where $b < a < 0$ and

$$\kappa_t = \omega + \beta\kappa_{t-1} + \gamma \frac{1}{b \exp\left(\mathrm{ES}^p_{t-1|t-2}\right)} \left(\frac{1}{p}\mathbb{1}\left(r_{\mathrm{PF},t-1} \leq a \exp\left(\mathrm{VaR}^p_{t-1|t-2}\right)\right) r_{\mathrm{PF},t-1} - b \exp\left(\mathrm{ES}^p_{t-1|t-2}\right)\right). \tag{2.31}$$

As ω, a and b are not separably identifiable we set $\omega = 0$.

Hybrid GAS/GARCH model

The hybrid GAS/GARCH model of Patton, Ziegel, and Chen (2019) is a direct combination of the forcing variable suggested by a GAS structure for a one-factor model of returns, described in Equation (2.31), with the successful GARCH model for volatility:

$$r_{t+1} = \exp(\kappa_{t+1})\,\eta_{t+1}, \quad \eta_t \sim \mathcal{F}_\eta(0,1) \tag{2.32}$$

where the log-volatility κ_t is specified as follows:

$$\kappa_t = \omega + \beta\kappa_{t-1} + \gamma \frac{1}{\mathrm{ES}^p_{t-1|t-2}} \left(\frac{1}{p}\mathbb{1}\left(r_{t-1} \leq \mathrm{VaR}^p_{t-1|t-2}\right) r_{t-1} - \mathrm{ES}^p_{t-1|t-2}\right) + \delta \log|r_{\mathrm{PF},t-1}|. \tag{2.33}$$

As the latent variable in this model is log-volatility, the authors use the lagged log absolute return rather than the lagged squared return, so that the units remain in line for the evolution equation for κ_t.

Similar to quantile and expectile regressions, the semi-parametric approach of the proposed GAS models eliminates the need to specify and estimate a conditional density. While removing the possibility of a model misspecification, there might be a loss of efficiency compared with a correctly specified density model. Unlike GARCH models, GAS models generate a smoother time series of VaR and ES estimates. While GARCH estimates are driven by lagged squared returns, and can thus be quite volatile, GAS model estimates only use information from returns when the VaR is violated, and revert deterministically to the long-run mean on other days.

2.3.3. Conditional asset-level risk models

The above models focus on dynamic risk modeling of univariate return time series. In contrast, conditional asset-level risk analysis is based on a multivariate model that additionally enables us to account for the dependence structure of the portfolio's assets.

The RiskMetrics approach

The RiskMetrics (RM) model is arguable the most simple and common approach among finance practitioners for estimating time-varying covariance matrices. It utilizes an exponentially weighted moving average filter that implicitly assumes a very tight parametric specification by incorporating conditionality via the exponential smoothing of individual squared returns and cross products. The estimate for the $N \times N$ covariance matrix at time $t+1$, $\hat{\Sigma}_{t+1}$, is then defined by

$$\hat{\Sigma}_{t+1} = \lambda \hat{\Sigma}_t + (1-\lambda)\mathbf{r}_t \mathbf{r}_t^{'}, \tag{2.34}$$

where $\lambda < 1$ is known as the decay factor.[15] The VaR and ES are then simply obtained as

$$\text{VaR}_{t+1|t}^{p} = \left(\mathbf{w}_t^{'} \hat{\Sigma}_{t+1} \mathbf{w}_t\right)^{1/2} \Phi_p^{-1}, \tag{2.35}$$

$$\text{ES}_{t+1|t}^{p} = \left(\mathbf{w}_t^{'} \hat{\Sigma}_{t+1} \mathbf{w}_t\right)^{1/2} \frac{\phi\left(\Phi_p^{-1}\right)}{p}. \tag{2.36}$$

The RM model is appealing because no parameters need to be estimated, thanks to the implicit assumption of zero mean returns, a fixed smoothing parameter and conditional normality. At the same time, the RM approach is very restrictive, imposing the same degree of smoothness on all elements of the covariance matrix. Moreover, the RM model tends to underestimate VaR and ES under the normality assumption. We therefore employ a Student's t-distribution instead.

The copula-GARCH approach

The copula-GARCH (CG) approach proposed by Jondeau and Rockinger (2006) and Patton (2006) is based on the concept of inference from margins, that is, dependencies between the marginal distributions are captured by a copula. In the first step, univariate GARCH(1,1) models are fitted to the underlying return series. Assuming a return process $(r_{i,t})_{i \in \mathbb{N}, t \in \mathbb{Z}}$, the mean and variance equations are given by

$$r_{i,t} = \mu_i + \varepsilon_{i,t}, \tag{2.37}$$

$$\varepsilon_{i,t} = z_{i,t}\sqrt{\sigma_{i,t}^2}, \tag{2.38}$$

$$z_{i,t} \sim \mathcal{D}_i(0, 1, \xi_i, \nu_i), \tag{2.39}$$

$$\sigma_{i,t}^2 = \omega_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i \sigma_{i,t-1}^2, \tag{2.40}$$

[15] In practice, λ is typically fixed at a preset value of 0.94 when using daily returns.

where $\omega_i > 0, \alpha_i \geq 0$ and $\beta_i \geq 0$, $i = 1, ..., N$. Moreover, $r_{i,t}$ are the returns of the ith portfolio asset at time t, and $\mathcal{D}_i$ reflects the skewed t-distribution with skewness parameter ξ_i and shape parameter ν_i according to Hansen (1994).

In the second step, we use a time-varying copula to estimate the marginal distributions of the asset returns together with the dependence structure. In particular, the joint distribution of the N GARCH return processes can be expressed depending on an N-dimensional copula C:

$$F_t\left(\mathbf{r}_t|\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t\right) = C_t\left(F_{1,t}(r_{1,t}|\mu_{1,t}, \sigma_{1,t}), ..., F_{N,t}(r_{N,t}|\mu_{N,t}, \sigma_{N,t})|\mathcal{F}_{t-1}\right), \tag{2.41}$$

where $F_1(\cdot), ..., F_N(\cdot)$ are the conditional marginal distributions of the return processes. The dependence structure of the margins is assumed to follow a Student's t-copula with conditional correlation $\mathbf{R}_t$ and constant shape parameter η. We opt for the Student's t-copula for modeling the dependence of financial assets since the normal copula cannot account for tail dependence. The conditional density of the Student's t-copula at time t is given by

$$c_t\left(u_{i,t}, ..., u_{N,t}|\mathbf{R}_t, \eta\right) = \frac{f_t\left(F_{i,t}^{-1}(u_{i,t}|\eta), ..., F_{i,t}^{-1}(u_{N,t}|\eta)|\mathbf{R}_t, \eta\right)}{\prod_{i=1}^{n} f_i\left(F_{i,t}^{-1}(u_{i,t}|\eta)|\eta\right)}, \tag{2.42}$$

where $u_{i,t} = F_{i,t}(r_{i,t}|\mu_{i,t}, \sigma_{i,t}, \xi_i, \nu_i)$ is the probability integral transformation of each series by its conditional distribution $F_{i,t}$ estimated via the first-stage GARCH process, $F_{i,t}^{-1}(u_{i,t}|\eta)$ represents the quantile transformation of the uniform margins subject to the common shape parameter of the multivariate density, $f_t(\cdot|\mathbf{R}_t, \eta)$ is the multivariate density of the Student's t-distribution with conditional correlation $\mathbf{R}_t$ and shape parameter η and $f_i(\cdot|\eta)$ is the univariate margins of the multivariate Student's t-distribution with common shape parameter η. Furthermore, we allow the parameters of the conditional copula to vary with time in a manner analogous to a GARCH model for conditional variance (e.g. Patton, 2006). Specifically, we assume the dynamics of $\mathbf{R}_t$ to follow an asymmetric generalized dynamic conditional correlation (AGDCC) model according to Cappiello, Engle, and Sheppard (2006).

Based on the copula estimates, we then generate N sets of random pseudo-uniform variables and transform these into corresponding realizations of the error processes by using the quantile function of the margins. These simulated numbers are then used together with the conditional volatility forecast of the GARCH models to derive a Monte Carlo set of returns for each asset. By means of the portfolio's weight vector we can then compute a distribution of portfolio returns for $t + 1$ which allows us to calculate VaR and ES forecasts.

The copula-GARCH model has several advantages over more simplistic approaches. The GARCH models with skewed t-distribution applied in the first stage capture the main empirical characteristics of financial asset returns. Moreover, the use of copulas in the second stage helps overcome the deficiency of the Pearson correlation that merely captures linear relationships. In particular, copulas allow dependencies of portfolio assets to be modeled in a more flexible way. Given the associated computational effort and complexity, however, most practitioners resort to simpler methods.

2.3.4. Risk forecast combination

The forecasting literature (e.g. Timmermann, 2006) generally argues that combining forecasts may enhance the predictive performance relative to standalone models because of diversification benefits, robustness against structural breaks and a reduction of the dangers of model misspecifications. While there exist various approaches to combine VaR predictions (see Bayer, 2018, for a summary), the literature is lacking a method that combines ES predictions. In this section, we propose a new technique for the combination of ES (and VaR) forecasts based on a loss function of Fissler and Ziegel (2016).

Let $\text{VaR}^p_{m,t+1|t}$ and $\text{ES}^p_{m,t+1|t}$ be the VaR and ES forecast for day $t+1$ of model $m = 1, ..., M$ based on the information available at t and $\mathbf{VaR}^p_{t+1|t} = \left[\text{VaR}^p_{1,t+1|t}, ..., \text{VaR}^p_{M,t+1|t}\right]'$ and $\mathbf{ES}^p_{t+1|t} = \left[\text{ES}^p_{1,t+1|t}, ..., \text{ES}^p_{M,t+1|t}\right]'$ be the vectors of all forecasts. The linear combination of the M forecasts is then given by

$$\text{VaR}^p_{\text{comb},t+1|t} = \sum_{m=1}^{M} \beta^{\text{VaR}}_{m,t} \text{VaR}^p_{m,t+1|t} = \left(\mathbf{VaR}^p_{t+1|t}\right)' \boldsymbol{\beta}^{\text{VaR}}_t, \tag{2.43}$$

$$\text{ES}^p_{\text{comb},t+1|t} = \sum_{m=1}^{M} \beta^{\text{ES}}_{m,t} \text{ES}^p_{m,t+1|t} = \left(\mathbf{ES}^p_{t+1|t}\right)' \boldsymbol{\beta}^{\text{ES}}_t, \tag{2.44}$$

where $\boldsymbol{\beta}^{\text{VaR}}_t$ and $\boldsymbol{\beta}^{\text{ES}}_t$ are the combination weight vectors.

Simple average

The most naive combination approach would simply average the forecasts of all standalone models. The corresponding combination weights are given by

$$\widehat{\boldsymbol{\beta}}_{m,t} = \frac{1}{M}, \quad \forall\, m = 1, ..., M. \tag{2.45}$$

According to the mean forecasting literature (see Timmermann, 2006), this approach is empirically successful and hard to beat by more sophisticated combination methods. Therefore, we consider the simple average as a benchmark approach.

FZ loss forecast combination

Given that parsimonious models often perform well in stable markets, whereas highly parameterized models show their strengths during periods of high volatility we are concerned that the simple average combination approach fails to leverage the time-dependent benefits satisfactorily. Therefore, we propose a new technique for the combination of ES (and VaR) forecasts that incorporates the most recent information into the model parameters.

In order to determine the optimal combination weights, we resort to the class of loss functions proposed by Fissler and Ziegel (2016). Similar to Patton, Ziegel, and Chen (2019), we choose the parameters of the function class in such a way that the loss differences of two forecasts are homogeneous of degree zero, given that VaR and ES are strictly negative. While Patton, Ziegel, and Chen (2019) use the FZ loss function for creating standalone risk models, we use it for the purpose of risk forecast combination. The FZ loss function is given by

$$L_{FZ}(r, \text{VaR}, \text{ES}, p) = -\frac{1}{p\,\text{ES}}\mathbb{1}\left(r \leq \text{VaR}\right)\left(\text{VaR} - r\right) + \frac{\text{VaR}}{\text{ES}} + \log(-\text{ES}) - 1. \tag{2.46}$$

Fissler and Ziegel (2016) show consistency of this loss function, implying that the true VaR and ES predictions minimize the expected loss. Equipped with a consistent loss function, the optimal forecast combination weights consequently minimize the expected loss of the FZ loss function,

$$\left(\left(\boldsymbol{\beta}_t^{\text{VaR}}\right)^*, \left(\boldsymbol{\beta}_t^{\text{ES}}\right)^*\right) = \underset{\boldsymbol{\beta}_t^{VaR}, \boldsymbol{\beta}_t^{ES}}{\arg\min}\ \mathbb{E}\left[L_{FZ}\left(r_{\text{PF},t+1}, \left(\mathbf{VaR}_{t+1|t}^{p}\right)'\boldsymbol{\beta}_t^{\text{VaR}}, \left(\mathbf{ES}_{t+1|t}^{p}\right)'\boldsymbol{\beta}_t^{\text{ES}}\right)|\mathcal{F}_t\right]. \tag{2.47}$$

Consistent and asymptotically normal estimators of the combination weights can be obtained by minimizing the average FZ loss,[16]

$$\left(\widehat{\boldsymbol{\beta}}_t^{\text{VaR}}, \widehat{\boldsymbol{\beta}}_t^{\text{ES}}\right) = \underset{\boldsymbol{\beta}_t^{VaR}, \boldsymbol{\beta}_t^{ES}}{\arg\min}\ \frac{1}{t}\sum_{\tau=0}^{t-1} L_{FZ}\left(r_{\text{PF},\tau+1}, \left(\mathbf{VaR}_{\tau+1|\tau}^{p}\right)'\boldsymbol{\beta}_t^{\text{VaR}}, \left(\mathbf{ES}_{\tau+1|\tau}^{p}\right)'\boldsymbol{\beta}_t^{\text{ES}}\right), \tag{2.48}$$

[16] A proof of consistency and asymptotic normality of the presented estimators is similar to that in Patton, Ziegel, and Chen (2019).

which can then be used to form the combined forecast

$$\widehat{\mathrm{VaR}}^{p}_{\mathrm{comb},t+1|t} = \left(\mathbf{VaR}^{p}_{t+1|t}\right)' \widehat{\boldsymbol{\beta}}^{\mathrm{VaR}}_{t}, \tag{2.49}$$

$$\widehat{\mathrm{ES}}^{p}_{\mathrm{comb},t+1|t} = \left(\mathbf{ES}^{p}_{t+1|t}\right)' \widehat{\boldsymbol{\beta}}^{\mathrm{ES}}_{t}. \tag{2.50}$$

Hence, our approach allows the combination weights to differ for VaR and ES predictions, which is favorable as the quality of a method's VaR and ES may differ. Following the forecast combination literature (Timmermann, 2006; Hansen, 2008), we impose convexity on the combination weights as this restriction typically improves upon the unconstrained estimator in terms of predictive performance. Convex weights are non-negative and sum to unity, that is, $0 \leq \beta_m \leq 1$ for $m = 1, \ldots, M$ and $\sum_{m=1}^{M} \beta_m = 1$.[17]

2.4. Empirically validating risk models for tail risk protection

In this section we describe the design and the results of the empirical study to compare the various methods for portfolio risk modeling using tail risk protection strategies.

2.4.1. Data and return synchronization

We use a global multi-asset data set, encompassing four major market risk factors: equity, fixed income, commodities and exchange rates. In particular, we utilize the following representative assets: Nikkei 225, EURO STOXX 50, FTSE 100, S&P 500 and MSCI EM equity futures; JGB 10Y, Euro Bund, UK Gilt and US 10Y bond futures; total return indices for the commodities oil, gold and copper; and JPY/USD, EUR/USD, GBP/USD spot market foreign exchange rates. The money market investment is based on the 3-month US Treasury Bill. We retrieve all data from Bloomberg. All asset prices are in local currency. Portfolio returns (and associated portfolio risk figures) are computed from the perspective of a US investor who is hedging any currency exposure. The sample spans the period from January 2, 1991 to March 31, 2017, giving rise to 6847 daily return observations for each series.

To calculate portfolio risk figures, we assume a few static strategic allocations of portfolio weights. Alternatively, we could consider a dynamic weight structure driven by a tactical asset allocation component, complicating the determination of whether an increase in performance

[17] We use an optimization procedure similar to that of Engle and Manganelli (2004). We first generate 10^5 vectors of parameters from a uniform random number generator such that the convex weight restriction is fulfilled. For each of these vectors, we compute the average loss from the FZ loss function and select the 10 vectors that produce the lowest average loss as initial values for the optimization routine. Using the augmented Lagrange multiplier method with a sequential quadratic programming interior algorithm according to Ye (1987), we minimize the average loss for each of the 10 resulting vectors and select the vector producing the lowest average loss as the final parameter vector.

is due to superior risk forecasts or due to predictability of the tactical component. As a base case we use a broadly diversified, conservative multi-asset portfolio, but we also investigate four alternative allocations: a pure equity portfolio, a pure bond portfolio, a 30/70 equity/bond portfolio and 60/40 equity/bond portfolio. Table 2.1 reports the corresponding allocation of portfolio weights as well as the descriptive statistics of the log returns of each asset and portfolio: all time series exhibit the typical features of financial assets such as fat tails and non-normality.

Table 2.1: Descriptive statistics and test portfolio allocations

								Portfolio weights				
	Mean	Med	Min	Max	Sd	Skew	Kurt	MA	EQ	BO	30/70	60/40
Individual assets												
Stocks												
Nikkei 225	-0.00	0.00	-14.0	18.82	1.51	-0.20	119.32	5	9.8	0	2.9	5.9
Euro STOXX 50	0.03	0.05	-9.44	11.38	1.39	-0.12	88.46	5	12.5	0	3.8	7.5
MSCI EM	0.02	0.08	-9.99	10.07	1.14	-0.52	108.25	5	15.2	0	4.6	9.1
FTSE 100	0.02	0.00	-9.70	9.58	1.13	-0.15	86.28	5	9.9	0	3.0	5.9
S&P 500	0.02	0.03	-10.4	13.20	1.13	-0.15	142.71	15	52.6	0	15.8	31.6
Bonds												
JGB 10Y	0.01	0.00	-1.55	2.18	0.25	-0.28	82.85	10	0	10	7	4
Euro Bund	0.02	0.01	-1.73	1.96	0.33	-0.19	48.7	10	0	20	14	8
UK Gilt	0.01	0.00	-2.34	3.65	0.41	0.06	62.65	10	0	10	7	4
US 10Y	0.01	0.00	-2.63	3.53	0.37	-0.10	62.63	10	0	40	28	16
Commodities												
Oil	0.00	0.00	-38.4	13.34	2.16	-0.95	206.1	5	0	0	0	0
Gold	0.02	0.00	-9.81	8.84	1.01	-0.17	112.25	5	0	0	0	0
Copper	0.02	0.00	-11.7	11.65	1.61	-0.19	76.55	5	0	0	0	0
Foreign exchange rates												
EUR/USD	-0.00	0.00	-3.38	3.93	0.62	0.04	52.18	15	0	20	14	8
GBP/USD	-0.01	0.00	-7.94	5.24	0.60	-0.49	113.7	15	0	10	7	4
JPY/USD	0.00	0.00	-4.07	7.06	0.68	0.46	83.9	15	0	10	7	4
Asset portfolios												
Multi-asset	0.02	0.03	-3.83	3.67	0.46	-0.27	93.07	-	-	-	-	-
Equity	0.02	0.06	-8.42	10.24	0.93	-0.32	130.9	-	-	-	-	-
Bond	0.02	0.02	-1.93	1.98	0.36	-0.12	47.99	-	-	-	-	-
30/70	0.02	0.03	-2.72	2.76	0.34	-0.15	79.18	-	-	-	-	-
60/40	0.02	0.04	-5.12	6.04	0.55	-0.24	124.77	-	-	-	-	-
3-M US T-Bill	0.01	0.01	-0.00	0.02	0.01	0.06	13.86	-	-	-	-	-

This table reports the descriptive statistics of the daily log returns of the individual assets and test portfolios over the period from January 2, 1991 to March 31, 2017 (including 6847 observations). The following statistics are reported: mean, median (Med), minimum (Min), maximum (Max), standard deviation (Sd), skewness (Skew) and kurtosis (Kurt). All statistics are given as percentages, except skewness and kurtosis. In addition, we provide the static weights of the test portfolio allocations (multi-asset (MA), equity (EQ), bond (BO), 30/70 equity/bond (30/70), 60/40 equity/bond (60/40)) as percentages in the last five columns.

When modeling risk using international daily return data, one has to properly account for different market closing times.[18] Even worse, for some markets trading times do not overlap at all, as is the case for the USA and Japan. Obviously, these differences will make equity markets appear less (cor)related than they actually are. As a result, portfolio risk estimates will overstate the diversification benefit attached to investing across these assets (see Scholes and Williams, 1977; Lo and MacKinlay, 1990a; Burns, Engle, and Mezrich, 1998; Scherer, 2013). Ideally, daily returns can be computed for all series using the same time-stamp. This approach, however, is hardly feasible, even when using high-frequency data. Instead, the literature suggests synchronizing daily returns by extrapolating asset prices for those markets that close earlier, based on information from markets that close latest. While Burns, Engle, and Mezrich (1998) use a first-order vector moving average model with a multivariate GARCH covariance matrix to estimate synchronized returns, Audrino and Bühlmann (2004) employ a simple first-order vector autoregressive model (see Appendix 2.A for details on the return synchronization methodology). We follow the latter approach due to its computational efficiency.

Based on our sample, we compare the synchronized daily returns to the original ones. Table 2.2 shows the descriptive statistics of the original and synchronized daily returns. We observe that differences in the mean are only marginal, whereas volatilities are slightly higher when synchronizing. Thus, the return characteristics of the original data are maintained.

Table 2.2: Synchronized vs. original daily returns

	Nikkei 225	JGB10Y	Euro Bund	UK Gilt	EURO STOXX 50	FTSE 100
Original returns						
Mean	-0.0035	0.0136	0.0165	0.0149	0.0305	0.0178
Standard deviation	1.5109	0.2486	0.3341	0.4098	1.3875	1.1306
First-order autocorrelation	-5.3639	-3.3298	0.2096	0.9714	-2.6876	-2.1746
Synchronized returns						
Mean	-0.0034	0.0136	0.0165	0.0149	0.0307	0.0179
Standard deviation	1.6338	0.2561	0.3581	0.4386	1.5829	1.2916
First-order autocorrelation	-13.7891	-6.314	-6.3808	-5.5118	-12.9119	-12.8367

This table reports descriptive statistics for the synchronized and original daily returns. As we anchor the synchronization of daily returns in US markets, the US time series remain unchanged and are thus not reported. Non-US data are forecasted to the closing time of the US market by the VAR(1). All figures are given as percentages.

[18] The opening times of the markets in our sample are as follows: Japanese markets are open from 19:00(-1) to 1:00 ET, EU/UK markets from 3:00 to 11:30 ET, and US markets open from 09:30 to 16:15 ET.

To check the effectiveness of synchronization, we have identified the correlation matrices of both return types (not shown). For the synchronized daily returns, the chosen VAR(1) model is successfully re-correlating the within-asset class correlations. While equity correlations are no longer underestimated, equity-bond correlations tend to be more negative when using synchronized returns. Hence, the improved equity-bond diversification could mitigate the pick-up in equity risk. However, we learn that the latter effect dominates, and unreported results evidence an average increase of 15% in portfolio risk figures for the conservative multi-asset portfolio.[19] These findings are in line with Scholes and Williams (1977) and Lo and MacKinlay (1990a).

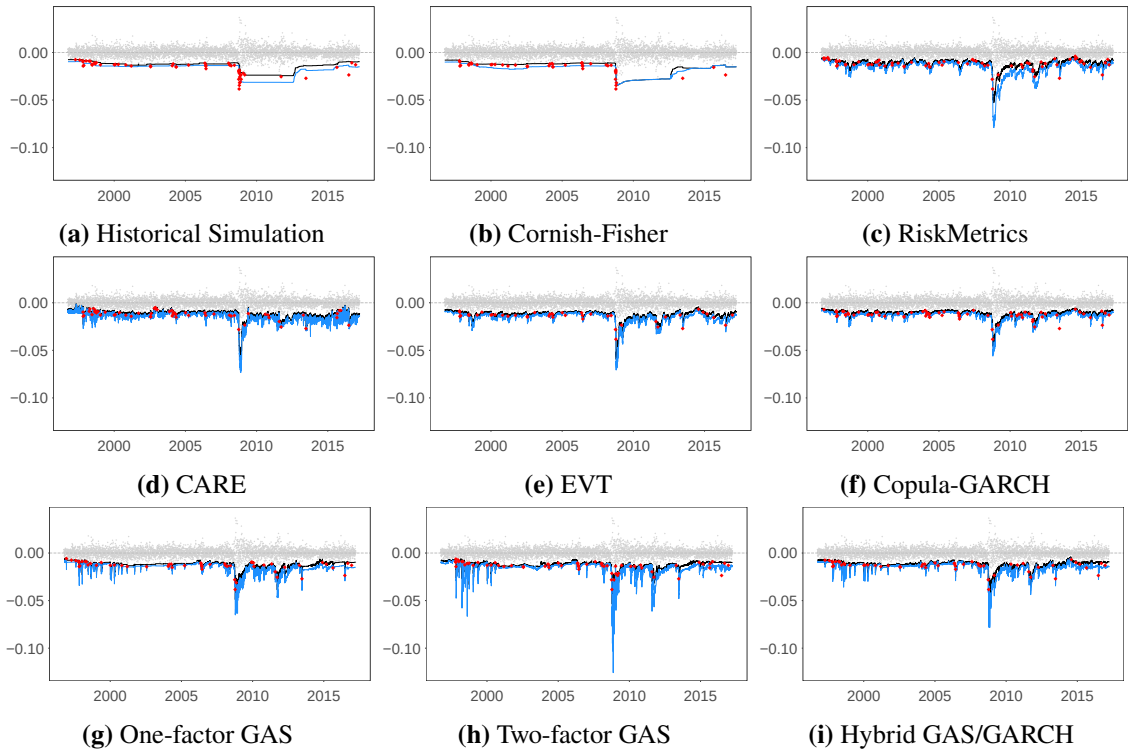
2.4.2. Estimating portfolio risk

The empirical study considers one day-ahead estimation of the conditional VaR and ES at a 1% confidence level, consistent with the portfolio rebalancing frequency of the tail risk protection strategies considered. Like Kuester, Mittnik, and Paolella (2006) and Taylor (2008), we use a moving window of 1000 observations to re-estimate parameters for the various standalone risk methods on a daily basis. Similarly, forecast combination weights are re-estimated daily using a moving window of 500 observations. Unreported results show that the combination weights are robust to the window length. Given this decision, the out-of-sample estimation period ranges from October 2, 1996 to March 31, 2017, consisting of 5348 daily VaR and ES forecasts for each method (standalone and combination) and portfolio.

Standalone risk forecasts

Figure 2.1 presents the predicted 1% VaR and associated ES figures of the standalone risk models for the multi-asset portfolio over the whole out-of-sample period. The figures show that the average ES was estimated at around -2%, rising to some -0.7% in the mid 2000s, and attaining extreme values around -12% during the financial crisis in late 2008. Corresponding VaR figures are greater by construction. Like volatility, ES and VaR fluctuate substantially over time. We further observe that simple methods like historical simulation and Cornish-Fisher approximation produce forecasts that take some time to adjust to current market conditions, whereas more flexible risk models are more sensitive and quicker to react to the prevailing risk environment.

[19] Note that the synchronized returns are used for estimating and forecasting VaR and ES, but not for out-of-sample evaluation. Instead, we use the original returns for assessing the statistical validity of the risk forecasts as well as their performance in the tail risk protection strategies.



(a) Historical Simulation **(b)** Cornish-Fisher **(c)** RiskMetrics

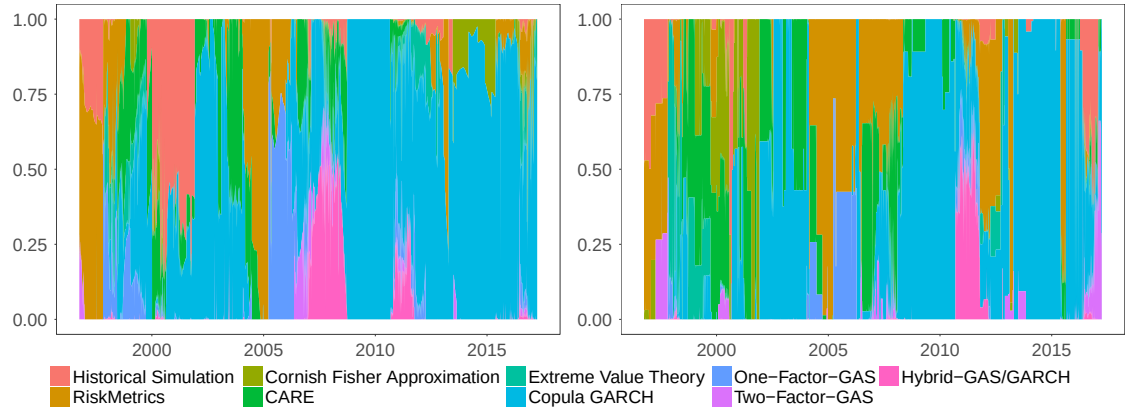
(d) CARE **(e)** EVT **(f)** Copula-GARCH

(g) One-factor GAS **(h)** Two-factor GAS **(i)** Hybrid GAS/GARCH

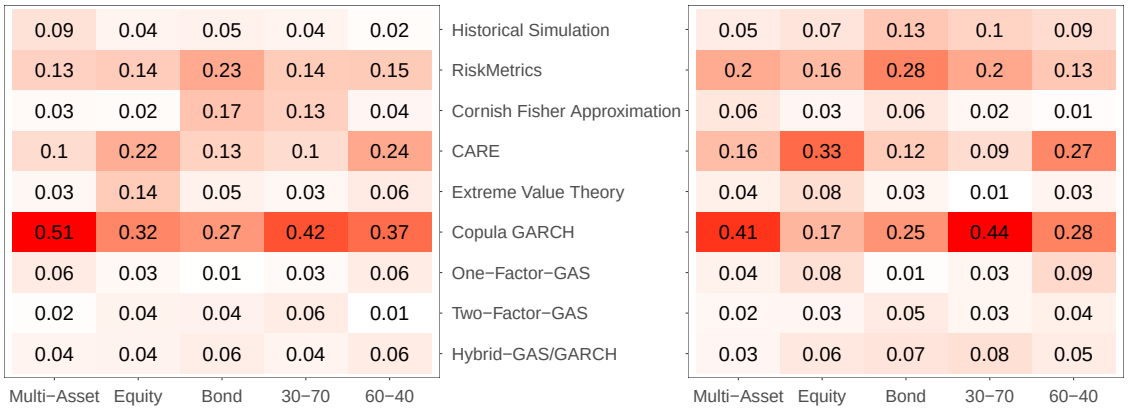
Figure 2.1: Standalone VaR and ES forecasts over time. This figure shows the daily 1% VaR forecasts (in black) and associated ES forecasts (in blue) of the different standalone risk models as well as the realized returns of the multi-asset portfolio (grey dots) over the period from October 2, 1996 to March 31, 2017. VaR violations are marked in red. At a confidence level of 1%, a total of 53 violations are expected over the model period.

Combination weights and combination risk forecasts

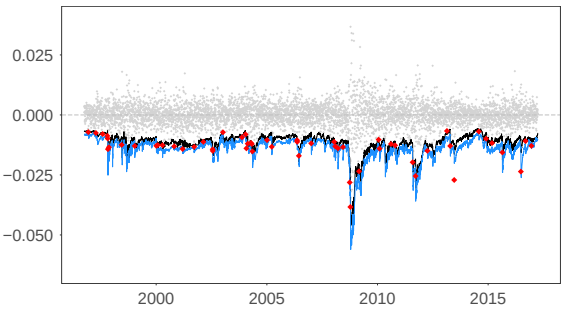
To foster intuition with regard to how the FZ loss combination approach estimates the combination weights and selects the standalone models, Figure 2.2 shows the combination weights for ES and VaR forecasts. Figure 2.2a exhibits the time evolution of the out-of-sample weights for the multi-asset portfolio. ES and VaR weights are not restricted to be identical, and we indeed observe different weight patterns over the sample period. While we document an average weight overlap of 64.1% when comparing ES and VaR combination weights through time, we note that we still can observe periods with zero overlap. Although the composition of the estimated weights is sensitive to the current market conditions for both risk measures, VaR weights are slightly less volatile than ES weights. On average, this difference translates to a daily weight change of 8.8% for the ES weights and 7.4% for the VaR weights. In terms of weight composition, we find the copula-GARCH model to be the most important component in the FZ loss combination approach, see Figure 2.2b giving the average weights for all portfolios over the out-of-sample period. The average ES weight



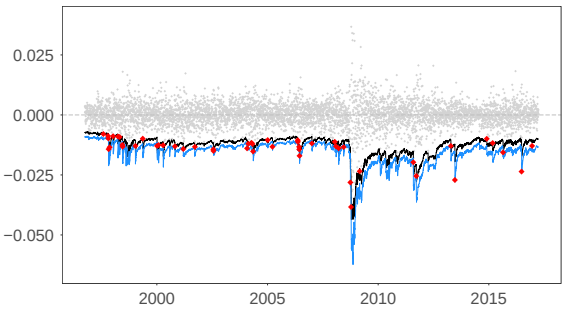
(a) ES and VaR combination weights over time



(b) ES and VaR mean combination weights



(c) FZ loss combination forecasts



(d) Average combination forecasts

Figure 2.2: Forecast combination. This figure illustrates the weights and forecasts of the FZ loss combination approach for the multi-asset portfolio. Panel (a) shows the estimated ES and VaR combination weights over the sample period from 1996 to 2017. Panel (b) illustrates the mean of ES and VaR combination weights for all portfolios. Panel (c) and (d) show the daily 1% VaR forecasts (in black) and associated ES forecasts (in blue) from the FZ loss and the simple average combination approaches. Realized returns of the multi-asset portfolio and VaR violations are marked in grey and red, respectively.

for the copula-GARCH approach ranges from 27% for the bond portfolio to 51% for the multi-asset portfolio. Another important component is the RiskMetrics forecast, even though it does not perform well individually (see Section 2.4.3). Given an estimation error of zero,

the RiskMetrics model serves as a stabilizing component. On average, the ES estimation weights for all other standalone models lie between 1% and 20%. We also see that the model weights differ across portfolios, suggesting that a data-driven selection of the standalone models may offer advantages compared to a simple average forecast.

In addition to the characteristics of the combination weights, Figure 2.2 also gives the forecasts of the FZ loss (Figure 2.2c) and the simple average combination approach (Figure 2.2d) over time. As expected, the FZ loss combination forecasts are slightly more sensitive to the prevailing risk environment compared to the average forecast as less weight is, on average, given to simple methods such as historical simulation and Cornish-Fisher approximation.

2.4.3. Statistical validity of risk forecasts

To assess forecasting performance from an econometric perspective we perform various VaR and ES tests proposed in the literature. The objective of such statistical tests is to consider the *ex ante* portfolio risk forecasts from a specific model and compare them with the *ex post* realized portfolio returns.

Value-at-Risk tests

Testing VaR forecasts boils down to evaluating the distribution of VaR violations. That is, one needs to count and investigate those realized return observations that fall below the predicted VaR level for a given estimation period. For instance, there should be 2.5 violations in a set of 250 forecasts of daily 1% VaRs per year. The test for unconditional coverage (UC) of Kupiec (1995) assesses whether the frequency of violations is consistent with the quantile of loss that the VaR measure is intended to reflect. However, this test does not account for serial independence of the number of violations. In this vein, the conditional coverage (CC) test of Christoffersen (1998) offers a remedy by jointly testing the frequency as well as the independence of violations, assuming that VaR violations are modeled with a first-order Markov chain. This test could reject a VaR model that generates too many clustered violations. As the original likelihood ratio test of Christoffersen (1998) has inferior size and power properties compared to more recent alternatives (see Berkowitz, Christoffersen, and Pelletier, 2011), we also consider the dynamic quantile (DQ) test of Engle and Manganelli (2004). Specifically, the authors propose a regression-based test that checks whether VaR estimates satisfy the criteria of unbiasedeness, independent violations and independence of the quantile estimates. To account for clustering of extremes we further consider the duration (DU) test of Christoffersen and Pelletier (2004), which examines the duration between violations by testing the null hypothesis that the duration

between violations is exponentially distributed against a Weibull alternative. More recently, Patton, Ziegel, and Chen (2019) proposed the generalized residual (GR) test. It is based on simple regressions of standardized versions of the "generalized residuals" (as given in Equation (2.25)), on elements of the information set available at the time the forecast was made. As these standardized generalized residuals are conditionally mean zero under the correct specification, forecast optimality can be assessed by testing wether all parameters in these regressions are zero, against a two-sided alternative.

Expected Shortfall tests

The GR test is also applicable to test ES predictions, because the generalized residuals are derived from the FZ loss function, which incorporates both VaR and ES (see Equation (2.26)). A close cousin of the GR test is the ES regression test (ESR) of Bayer and Dimitriadis (2020) which, in contrast, only needs ES forecasts as input parameters. It is based on a regression framework modeling the conditional ES as a linear function, where returns are used as the response variable and ES forecasts as the explanatory variable including an intercept term. For correct ES forecasts, the intercept and slope parameters should be equal to zero and one, respectively. A Wald statistic is then employed to test for these parameter values.

One of the first and most frequently used ES tests is the exceedance residual (ER) test of McNeil and Frey (2000). This testing procedure is based on the ES residuals that exceed VaR, $er_t = (r_{\text{PF},t} - \widehat{\text{ES}}_t)\mathbb{1}(r_{\text{PF},t} \leq \widehat{\text{VaR}}_t)$, which should have zero mean under the null hypothesis of a correctly specified risk model. Using a bootstrap hypothesis test, it is tested whether the expected value of the exceedance residuals, $\mathbb{E}[er_t]$, is zero. In addition, we consider the conditional conditional calibration (CAL) test of Nolde and Ziegel (2017) for testing ES. This approach is based on a Wald-type test statistic that uses moment functions of VaR and ES.[20]

Empirical evidence

Table 2.3 presents the p-values from the above VaR and ES tests. Entries greater than 0.10 indicate no evidence against optimality at the 10% significance level. Our main findings are as follows. First, we find the simple methods, including HS, CFA and RiskMetrics to struggle in most of the tests. Although the HS and CFA methods show a conclusive number of violations over whole the sample period (close to the expected number of 53 violations) and therefore pass the UC test, they fail the remainder of VaR tests because the violations are not occurring independently, but rather in clusters. Given that a correctly specified VaR

[20] See Bayer and Dimitriadis (2020) for a rigorous discussion of most of the ES tests in use.

model is the basis of estimating ES, the subsequent ES tests may be considered useless. The RiskMetrics approach fails most tests due to the large deviation from the expected number of violations (72 realized violations).

Table 2.3: VaR and ES backtesting

				VaR tests					ES tests			
	Viol	FZ	Tick	UC	CC	DQ	DU	GR	ER	CAL	ESR	GR
Standalone models												
Historical Simulation	57	60.11	1.91	**0.63**	0.00	0.00	0.00	0.00	**0.40**	**0.44**	*0.10*	0.02
Cornish-Fisher-Approximation	44	58.15	1.87	**0.18**	0.00	0.00	0.00	0.00	*0.07*	0.01	**0.21**	0.00
RiskMetrics	72	38.84	1.51	0.02	0.01	0.00	**0.83**	0.01	**0.12**	0.00	0.01	0.01
CARE	58	46.67	1.58	**0.54**	**0.75**	**0.81**	0.03	0.00	0.03	**0.67**	0.02	0.00
Extreme Value Theory	44	36.25	1.51	**0.18**	**0.28**	*0.08*	**0.24**	**0.27**	**0.46**	**0.35**	**0.49**	**0.25**
Copula-GARCH	71	*34.79*	*1.48*	0.02	*0.05*	0.00	**0.15**	**0.16**	**0.38**	**0.11**	**0.13**	**0.14**
One-Factor-GAS	54	41.83	1.57	**0.94**	**0.57**	0.03	**0.30**	0.00	**0.88**	**0.99**	**0.29**	0.00
Two-Factor-GAS	64	46.66	1.62	**0.16**	*0.05*	0.00	0.02	**0.25**	**0.45**	**0.28**	*0.06*	**0.18**
Hybrid-GAS/GARCH	55	40.90	1.55	**0.83**	**0.85**	**0.40**	**0.79**	0.03	**0.36**	**0.93**	**0.15**	*0.10*
Combination models												
Average	49	38.96	1.54	**0.53**	**0.64**	0.00	0.01	*0.10*	**0.89**	**0.68**	**0.32**	**0.13**
FZ loss	56	**28.11**	**1.41**	**0.73**	**0.83**	**0.42**	**0.83**	**0.94**	**0.83**	**0.80**	**0.90**	**0.98**

This table reports the results of VaR and ES tests for evaluating 1% VaR and 1% ES predictions based on the 11 forecasting models applied to the multi-asset portfolio over the out-of-sample period from October 2, 1996 to March 31, 2017. For testing VaR we include the unconditional coverage (UC), the conditional coverage (CC), the dynamic quantile (DQ) and the duration (DU) test. For testing ES we resort to the exceedance residual (ER), the conditional calibration (CAL) and the ES regression (ESR) test. The generalized residual (GR) test allows us to test both VaR and ES. We report p-values in bold if greater than 0.10, indicating no evidence against optimality at the 10% significance level. Values between 0.05 and 0.10 are in italics. We further report the number of realized VaR violations (second column) and the average loss using the FZ loss function (third column) and the tick loss function (fourth column), the latter two being scaled by 100. The lowest average loss in each column is highlighted in bold, the second-lowest in italics. The expected number of violations is 53 over the whole out-of-sample period.

Second, we find the more sophisticated standalone models (except for the CARE approach) to pass most of the VaR and ES tests. Notably, none of the models passes all tests at the 10% significance level. Among the combination models, we provide evidence that the simple average approach delivers decent results, passing most of the tests. The newly proposed FZ loss approach is, however, even more convincing: it is the only model clearly passing all VaR and ES tests.

In addition to the number of violations and the p-values from the various tests, we show the average out-of-sample losses, based on the FZ loss function from Equation (2.46) and the piecewise-linear or "tick" loss function (only appropriate for VaR forecast evaluation), see the third and fourth column of Table 2.3. The FZ loss forecast combination approach is the

preferred model, exhibiting the lowest value for both loss function. As expected, the HS and CFA models are the worst models. While average losses are useful to eyeball out-of-sample forecast performance, we still need to investigate whether the gains are statistically significant. Table 2.4 presents t-statistics of modified Diebold-Mariano (DM) tests on the loss differences using the FZ loss function, according to Diebold and Mariano (1995), Harvey, Leybourne, and Newbold (1997), and Patton, Ziegel, and Chen (2019).[21] The tests are conducted and repeated as "row model minus column model", such that a positive number indicates that the column model outperforms the row model. All entries for the FZ loss forecast combination approach are positive, showing that this model outperforms all competing models. Also, this outperformance is highly significant for all comparisons, with DM t-statistics between 2.91 and 6.65. The second and third best model according to the DM tests are the copula-GARCH model and the EVT approach. The two are not statistically different from each other and are dominated by the FZ loss combination approach only. In a nutshell, we thus find that the FZ loss forecast combination approach dominates both the more sophisticated standalone risk models and the simple mean combination approach.[22]

Table 2.4: Diebold-Mariano tests

	HS	CFA	RM	CARE	EVT	CG	1F-GAS	2F-GAS	Hyb-GAS	Average	FZ
HS		0.33	1.91	1.32	2.19	2.26	1.83	1.37	1.98	2.12	2.91
CFA	-0.33		2.94	1.97	3.62	3.69	3.21	2.14	3.47	3.95	5.01
RM	-1.91	-2.94		-1.72	1.25	1.40	-0.95	-2.00	-0.69	-0.04	3.74
CARE	-1.32	-1.97	1.72		2.64	2.95	1.39	0.00	1.68	2.32	5.08
EVT	-2.19	-3.62	-1.25	-2.64		0.61	-2.37	-3.00	-1.93	-1.49	5.09
CG	-2.26	-3.69	-1.40	-2.95	-0.61		-2.65	-3.44	-2.34	-1.70	3.20
1F-GAS	-1.83	-3.21	0.95	-1.39	2.37	2.65		-1.86	0.63	1.94	6.65
2F-GAS	-1.37	-2.14	2.00	0.00	3.00	3.44	1.86		2.48	3.14	5.84
Hyb-GAS	-1.98	-3.47	0.69	-1.68	1.93	2.34	-0.63	-2.48		1.19	5.51
Average	-2.12	-3.95	0.04	-2.32	1.49	1.70	-1.94	-3.14	-1.19		6.48
FZ	-2.91	-5.01	-3.74	-5.08	-5.09	-3.20	-6.65	-5.84	-5.51	-6.48	

This table reports t-statistics from modified Diebold–Mariano tests according to Harvey, Leybourne, and Newbold (1997) comparing the average losses using the FZ loss function over the out-of-sample period from October 2, 1996 to March 31, 2017 for the 11 risk models based on the multi-asset portfolio. The first nine rows correspond to the standalone models: historical simulation (HS), RiskMetrics (RM), Cornish-Fisher approximation (CFA), conditional autoregressive expectile model (CARE), extreme value theory (EVT), copula-GARCH (CG), one-factor GAS (1F-GAS), two-factor GAS (2F-GAS) and hybrid-GAS/GARCH (Hyb-GAS) model. The last two rows correspond to the combination models: the simple average forecast (Average) and the proposed FZ loss approach (FZ). A positive value indicates that the row model's average loss is higher than that for the column model. Values greater than 1.96 in absolute value indicate that the average loss difference is significantly different from zero at the 95% confidence level. Values along the main diagonal are all identically zero and are omitted for interpretability.

[21] The outcomes of the different VaR and ES tests represent good examples of a problem highlighted in Nolde and Ziegel (2017). All of the more sophisticated models pass most of the goodness-of-fit tests, complicating the discussion of their relative performance. The comparative Diebold-Mariano tests are therefore an important element of the overall testing framework.

[22] Note that the testing results and rankings for the other portfolios are qualitatively similar to those for the multi-asset portfolio discussed here.

2.4.4. The economic relevance of risk forecasting for tail risk protection

We consider two steps when evaluating the various risk models in the tail risk protection framework. First, we analyze the historical path of each strategy. That is, assessing how each strategy would have performed when implemented over the whole out-of-sample period. For this, we assume an investment horizon of one calendar year—a typical choice of institutional and private investors alike (see Benartzi and Thaler, 1995). For the DPPI strategy, the floor is then adjusted to the current portfolio value at the start of each year to initialize the cushion. This procedure helps to mitigate the lock-in effect.

Analyzing the historical path suffers from path dependency; therefore, we additionally conduct a historical block-bootstrap analysis[23] in the second step. Following Annaert, Van Osselaer, and Verstraete (2009), Bertrand and Prigent (2011), Dichtl and Drobetz (2011) and Dichtl, Drobetz, and Wambach (2017), we draw blocks of 250 subsequent daily portfolio and risk-free returns on a rolling window basis and implement the tail risk protection strategies in each draw. Thus, we obtain 5597 overlapping yearly returns as a basis for the comparison of our methods. Intuitively, this historical block-bootstrap approach enables us to assess a strategy's robustness with respect to alternative entry dates. Moreover, the available data is used in the most efficient way while preserving all dependency effects in the series, such as autocorrelation and conditional heteroskedasticity (see Dichtl and Drobetz, 2011).

As the objective of tail risk protection strategies is twofold—providing downside protection while still enjoying the upside potential of the risky portfolio—the performance should be evaluated accordingly. Alongside standard measures like the Sharpe ratio and maximum drawdown we therefore employ specific downside risk measures commonly used in the portfolio insurance literature such as the Calmar, Sortino and Omega ratios (see Bertrand and Prigent, 2011).[24]

We implement the tail risk protection strategies without short sales or leverage and assume round-trip transaction costs of 10 basis points. To avoid portfolio shifts triggered by rather small market movements, we also apply a trading filter of 2%, acting only on exposure changes in excess of 2% (cf. Dichtl, Drobetz, and Wambach, 2017).

[23] This method is sometimes referred to as historical simulation, see Dichtl and Drobetz (2011).

[24] While the Calmar ratio is defined as the ratio of annualized return over the absolute value of the maximum drawdown, the Sortino ratio is the difference between mean return and minimum acceptable return (here, zero) divided by downside deviation (which measures the variability of underperformance below a minimum target rate). The Omega ratio is calculated by dividing the upper partial moment of degree one by the lower partial moment of degree one. Lower (upper) partial moments indicate the return potential below (above) a predefined threshold return. See Bertrand and Prigent (2011) for details on these performance risk measures.

Tail risk protection via risk targeting

Figure 2.3 illustrates the performance of the ES targeting strategy for the historical path and the historical block-bootstrap, based on the 1% ES of the FZ loss combination approach.[25] The underlying is the multi-asset portfolio and we target an ES level of 1.5%, which is a reasonable assumption given the conservative underlying. Figure 2.3a shows the evolution of the protected portfolio, the underlying multi-asset strategy and a money market investment over the out-of-sample period from 1994 to 2017. We observe a decrease in exposure of the ES targeting strategy during the financial market crisis in 2008, thus avoiding the huge drawdowns of the underlying but also reducing upside participation at the end of the sample period.

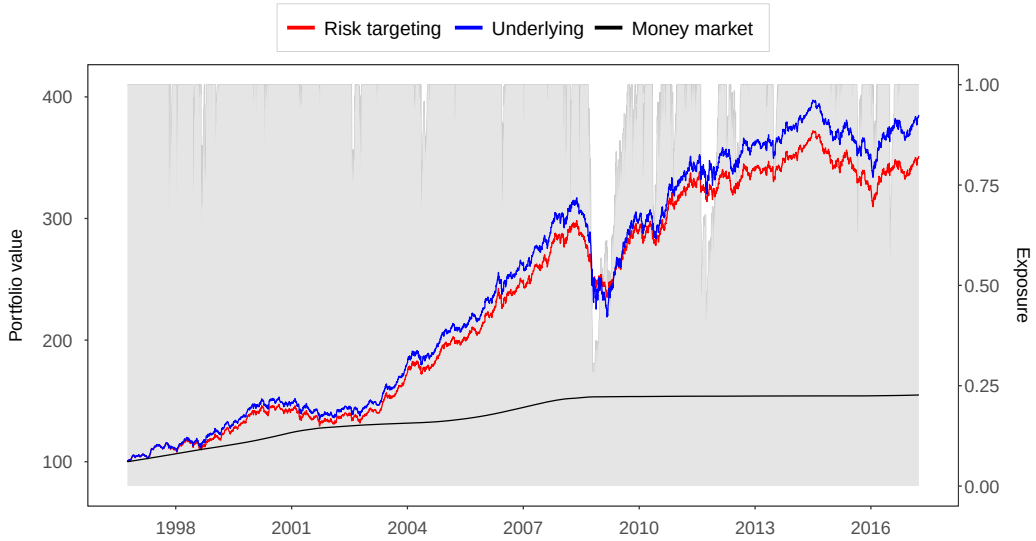
Figure 2.3b shows the distribution of simulated yearly returns of the protected portfolio in comparison with a pure buy-and-hold portfolio investment strategy. We see that the distribution of the ES targeting strategy is shifted to the right, thus reducing the mass in the left tail. However, this reduction comes at the cost of some return potential in the upper right tail.

Table 2.5 complements Figure 2.3 with the estimation results of the ES targeting strategy based on all different 1% ES forecasts for the historical path and the historical block-bootstrap. Panel A reports the results for the historical path. We find a similar size of risk-adjusted returns (cf. Sharpe ratio), but lower maximum drawdowns and thus higher Calmar ratios for all risk methods compared to the underlying. These figures confirm the ability of the ES targeting strategy to reduce downside risk. Comparing across risk models, we observe the best risk-adjusted performance (measured in Sharpe ratio) and downside risk measures (measured in Calmar ratio) for the copula-GARCH (CG), the hybrid GAS/GARCH (Hyb-GAS) and the FZ loss approach (FZ). Thus, our results indicate that the ES targeting strategy is more profitable when using these more flexible methods. This finding is confirmed by the historical block-bootstrap analysis shown in Panel B. We observe higher Omega ratios for CG, Hyb-GAS and FZ (5.22, 5.07 and 5.25); comparing these figures to 4.45 for the HS method, for example.[26] Also in terms of Sharpe and Sortino ratios, these three approaches outperform all other risk models.

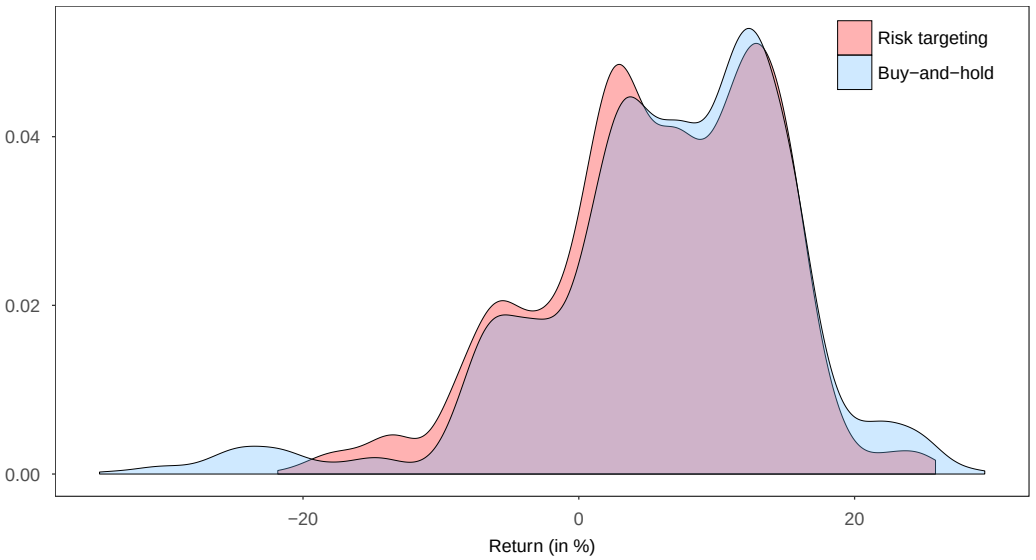
As the results of the ES targeting strategy may be sensitive to the choice of portfolio allocation and risk target, we also investigate the strategy using different underlying portfolios

[25] Note that volatility or VaR targeting strategies deliver similar results. To be consistent with the DPPI strategy that is based on ES we only report the results for the ES targeting strategy.

[26] Note that we rely on the Omega ratio rather than the mean of the yearly Calmar ratios in the historical block-bootstrap analysis. While the Calmar ratio is based on daily returns (and thus needs to be transformed via the mean), the Omega ratio is usually calculated for longer-horizon returns such as yearly returns and is therefore more appropriate in the historical block-bootstrap analysis (see Bertrand and Prigent, 2011; Dichtl, Drobetz, and Wambach, 2017).



(a) Historical path



(b) Historical block-bootstrap

Figure 2.3: Historical path and historical block-bootstrap of risk targeting. This chart illustrates the performance of the ES targeting strategy with multi-asset underlying portfolio (35% equities, 40% fixed income, 15% commodities, 45% currencies). Panel (a) shows the historical path of the protected portfolio (red line) over the sample period 1996–2017. Exposure is calculated based on the 1% ES of the FZ loss combination approach. The target level is a 1.5% ES. For comparison, we include the performance of the underlying multi-asset portfolio (blue line) and a money market investment (black line). Panel (b) shows the distribution of simulated yearly returns of the protected portfolio (red shading) and that of a buy-and-hold portfolio invested in the simulated multi-asset underlying (blue shading).

(a pure equity, a pure bond, a 30/70 equity/bond and a 60/40 equity/bond portfolio in addition to the multi-asset portfolio) at various ES target levels (1%, 1.25%, 1.5%, 1.75%, 2%). Table 2.6 reports the corresponding results. Assuming appropriate portfolio-specific ES

Table 2.5: Risk targeting for multi-asset portfolio

	Method	Return	SD	Sharpe	MDD	Calmar	Sortino	Omega	Part	TO	ES
Panel A: Historical path											
	Underlying portfolio	6.36	7.73	0.55	-31.80	0.20	0.07	1.16	100	0	-1.86
	Money market	2.13	0.17	0.00	0.00	-	-	-	0	0	0.00
Risk targeting	Historical Simulation	5.50	6.33	0.53	-25.82	0.21	0.08	1.16	87.41	0.04	-1.47
	Cornish-Fisher	5.59	6.24	0.55	-25.58	0.22	0.08	1.17	86.78	0.05	-1.43
	RiskMetrics	5.31	6.13	0.52	-19.68	0.27	0.08	1.15	90.34	0.81	-1.29
	CARE	5.61	6.46	0.54	-21.80	0.26	0.08	1.16	92.36	2.79	-1.34
	Extreme Value Theory	5.54	6.34	0.54	-21.87	0.25	0.08	1.16	92.05	1.01	-1.34
	Copula-GARCH	5.96	6.64	0.58	-21.63	0.28	0.08	1.16	95.67	0.75	-1.36
	One-Factor-GAS	5.67	6.44	0.55	-21.21	0.27	0.08	1.16	92.44	1.47	-1.34
	Two-Factor-GAS	5.30	6.41	0.49	-24.23	0.22	0.07	1.15	91.81	2.18	-1.38
	Hybrid-GAS/GARCH	5.79	6.43	0.57	-20.96	0.28	0.08	1.16	93.24	1.36	-1.34
	Average	5.64	6.32	0.55	-22.00	0.26	0.08	1.16	91.82	0.65	-1.33
	FZ loss	5.92	6.60	0.57	-21.39	0.28	0.08	1.16	95.30	0.85	-1.35
Panel B: Historical block-bootstrap											
	Underlying portfolio	6.15	9.41	0.43	-7.49	1.51	1.23	4.81	100	0	-28.56
	Money market	2.10	2.07	0.00	0.00	-	-	-	0	0	0.02
Risk targeting	Historical Simulation	5.27	8.63	0.37	-6.32	1.50	1.15	4.45	86.78	0.37	-24.04
	Cornish-Fisher	5.36	8.56	0.38	-6.24	1.52	1.18	4.62	86.15	0.37	-23.97
	RiskMetrics	5.08	8.02	0.37	-6.36	1.39	1.39	4.59	89.92	1.17	-17.36
	CARE	5.43	8.29	0.40	-6.58	1.38	1.47	4.86	92.24	3.12	-18.82
	Extreme Value Theory	5.32	8.23	0.39	-6.54	1.40	1.40	4.69	91.82	1.38	-19.28
	Copula-GARCH	5.75	8.23	0.44	-6.68	1.43	1.56	5.22	95.48	1.13	-18.71
	One-Factor-GAS	5.45	8.30	0.40	-6.49	1.44	1.43	4.89	92.13	1.86	-19.18
	Two-Factor-GAS	5.08	8.47	0.35	-6.73	1.40	1.21	4.33	91.54	2.54	-22.72
	Hybrid-GAS/GARCH	5.58	8.20	0.42	-6.46	1.47	1.51	5.07	92.94	1.75	-18.43
	Average	5.42	8.22	0.40	-6.45	1.44	1.41	4.87	91.47	1.01	-19.63
	FZ loss	5.71	8.17	0.44	-6.65	1.44	1.59	5.25	95.11	1.24	-18.41

This table reports the backtesting results of the risk targeting strategy based on different 1% ES forecasts for the historical path (Panel A) and the historical block-bootstrap (Panel B) over the sample period 1996–2017. For comparison, we include the performance of the underlying multi-asset portfolio and the money market investment. We target an ES of 1.5% over the whole out-of-sample period. We report the annualized mean return (Return), annualized standard deviation (SD), Sharpe ratio (SR), maximum drawdown (MDD), Calmar ratio, Sortino ratio, Omega ratio, participation in the risky multi-asset portfolio (Part), turnover (TO) and the 1% ES. Return, Sd, MDD, Part, TO and ES are given as percentages. For the historical path, the performance measures are calculated using the daily returns resulting from the strategy. For the historical block-bootstrap, the performance measures are based on the simulated yearly returns, except for MDD, Calmar ratio and participation. Those are based on the daily risky asset exposure of the corresponding draw and show the yearly mean of the specific measure.

target levels,[27] most portfolios benefit from the more flexible methods showing higher Calmar ratios (historical path; see Panel A) and Omega ratios (historical block-bootstrap; see Panel B), respectively. The same holds true for the robustness checks with respect to the risk target level. For most levels and for both analyses, historical path and historical block-bootstrap, we find superior performance of the sophisticated models (such as the FZ loss approach) in terms of Calmar and Omega ratios.

[27] We choose the following portfolio-specific ES target levels: 1.5% for the multi-asset portfolio, 4% for the pure equity portfolio, 1.5% for the pure bond portfolio, 1.5% for the 30/70 equity/bond portfolio and 2.5% for the 60/40 equity/bond portfolio.

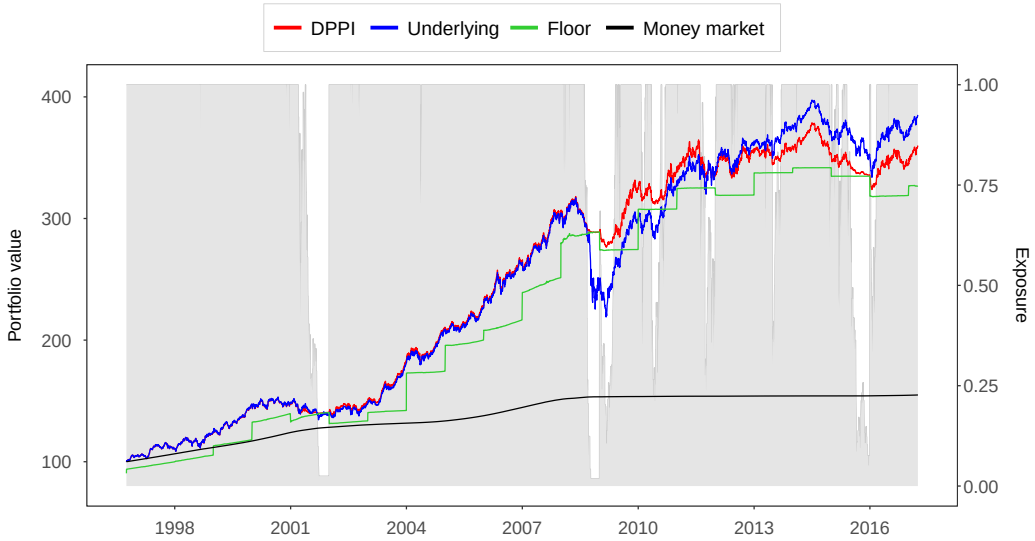
Table 2.6: Risk targeting: Various portfolios and target levels

	Portfolios					Target levels				
	Multi-Asset	Equity	Bond	30-70	60-40	1%	1.25%	1.5%	1.75%	2%
Panel A: Historical path (Calmar ratio)										
Historical Simulation	0.21	0.20	0.58	0.32	0.19	0.25	0.23	0.21	0.20	0.20
Cornish-Fisher	0.22	0.20	0.57	0.35	0.20	0.25	0.23	0.22	0.21	0.21
RiskMetrics	0.27	0.21	0.66	0.36	0.22	0.30	0.28	0.27	0.27	0.27
CARE	0.26	0.21	0.64	0.16	0.19	0.28	0.26	0.26	0.25	0.25
Extreme Value Theory	0.25	0.20	0.64	0.33	0.23	0.28	0.26	0.25	0.25	0.25
Copula-GARCH	0.28	0.20	0.64	0.37	0.23	0.31	0.29	0.28	0.27	0.25
One-Factor-GAS	0.27	0.20	0.64	0.35	0.27	0.30	0.27	0.27	0.26	0.26
Two-Factor-GAS	0.22	0.21	0.65	0.37	0.23	0.23	0.21	0.22	0.22	0.21
Hybrid-GAS/GARCH	0.28	0.21	0.60	0.42	0.24	0.30	0.27	0.28	0.28	0.27
Average	0.26	0.21	0.64	0.36	0.24	0.30	0.27	0.26	0.26	0.25
FZ loss	0.28	0.20	0.64	0.36	0.23	0.33	0.29	0.28	0.27	0.25
Panel B: Historical block-bootstrap (Omega ratio)										
Historical Simulation	4.45	2.13	25.29	12.04	3.86	5.62	4.88	4.45	4.29	4.33
Cornish-Fisher	4.62	2.21	24.65	12.05	4.00	5.80	5.05	4.62	4.42	4.56
RiskMetrics	4.59	2.47	24.43	12.17	4.74	4.77	4.33	4.59	5.15	5.32
CARE	4.86	2.44	20.28	3.79	3.82	5.28	4.75	4.86	5.10	5.38
Extreme Value Theory	4.69	2.27	23.70	11.57	4.31	5.15	4.70	4.69	4.98	5.28
Copula-GARCH	5.22	2.31	24.96	14.41	4.49	4.95	4.75	5.22	5.39	5.36
One-Factor-GAS	4.89	2.45	21.81	12.94	5.22	5.49	4.88	4.89	5.18	5.25
Two-Factor-GAS	4.33	2.34	21.24	12.60	4.65	4.38	4.04	4.33	4.58	4.67
Hybrid-GAS/GARCH	5.07	2.30	22.42	17.33	4.70	5.16	4.64	5.07	5.40	5.49
Average	4.87	2.32	25.53	12.95	4.76	5.84	4.94	4.87	5.18	5.28
FZ loss	5.25	2.39	23.86	13.06	4.36	5.74	5.02	5.25	5.49	5.46

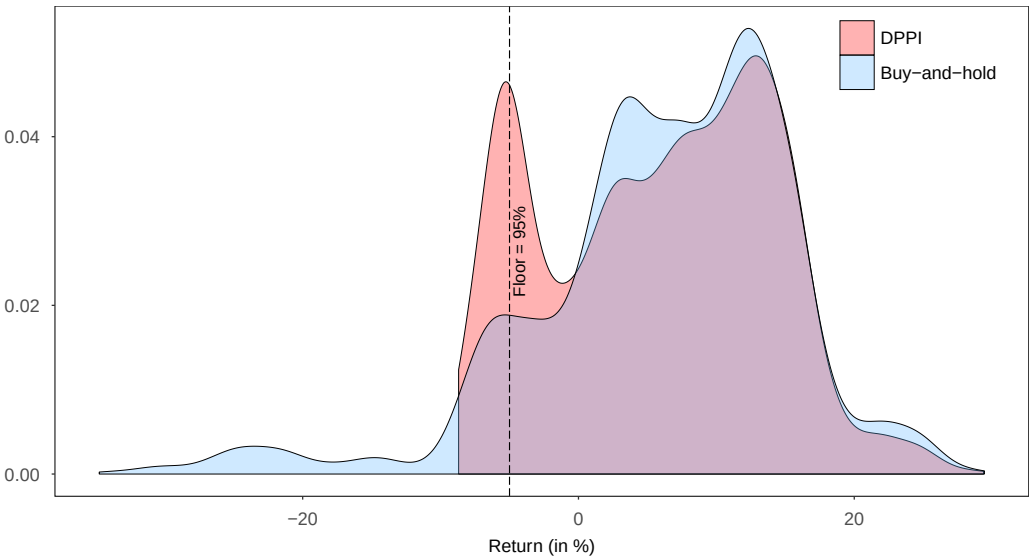
This table reports the backtesting results of the risk targeting strategy based on different risk forecasts for the historical path and the historical block-bootstrap over the sample period 1996–2017 using various portfolios and various target levels. To benchmark the results of the ES targeting strategy based on the multi-asset portfolio we estimate the ES targeting strategy also for four different test portfolio allocations. We choose a pure equity portfolio (4% ES target), a pure bond portfolio (1.5% ES target), a 30/70 equity/bond portfolio (1.5% ES target) and a 60/40 equity/bond portfolio (2.5% ES target) as underlyings for the ES targeting strategy. Moreover, we check the robustness of the multi-asset results with respect to the chosen ES target level (1%, 1.25%, 1.5%, 1.75% and 2%). We base our comparison on the Calmar ratio for the historical path and on the Omega ratio for the historical block-bootstrap.

Tail risk protection via DPPI

While the ES targeting strategy is able to mitigate downside risk to some extent, it falls short in clearly reducing maximum drawdown. A stricter way to limit downside risk is the DPPI strategy. Figure 2.4a illustrates how the mechanism of a DPPI strategy generally works. The chart shows the performance of the conservative multi-asset portfolio using the DPPI strategy in relation to the floor over time. The investment exposure is mainly driven by two components: the floor and the multiplier. If the portfolio value of the underlying risky investment approaches the floor from above, that is, the cushion shrinks, the investment exposure is reduced by shifting into the risk-free asset. Similarly, the exposure is reduced if risk estimates predict too high (overnight) risk, that is, the multiplier decreases, given that



(a) Historical path



(b) Historical block-bootstrap

Figure 2.4: Historical path and historical block-bootstrap of DPPI. This chart illustrates the performance of the DPPI strategy with multi-asset underlying portfolio (35% equities, 40% fixed income, 15% commodities, 45% currencies). Panel (a) shows the historical path of the protected portfolio (red line) in relation to the floor (green line) over the sample period from 1996–2017. Exposure is calculated based on the 1% ES of the FZ loss combination approach. The floor level of the DPPI strategy is 95%. For comparison, we include the performance of the underlying multi-asset portfolio (blue line) and a money market investment (black line). Panel (b) shows the distribution of simulated yearly returns of the protected portfolio (red shading) and that of a buy-and-hold portfolio invested in the simulated multi-asset underlying (blue shading).

the distance to the floor is not excessive. In this example, the conditional multiplier is based on the 1% ES of the FZ loss combination approach.[28]

[28] In order to reflect the preferences of risk-averse investors, we follow Soupé, Heckel, and De Carvalho (2014) and scale the risk forecast by a term consisting of an investor's risk aversion parameter and the expected

Examining the whole sample period, the DPPI strategy did indeed prevent severe drawdowns. With the onset of the global financial crisis, investment exposure drops to zero, so that the portfolio value at the end of 2008 is equal to the floor. Even in the subsequent V-shaped return pattern (sudden decline followed by a rapid recovery) in early 2009—a major impediment for portfolio insurance—the DPPI portfolio does not end up in a "cash lock". It partly participates in the subsequent recovery. On the whole, the DPPI portfolio has an average investment exposure of approximately 60% to 90%, depending on the chosen risk method, and delivers slightly lower returns compared to the pure multi-asset portfolio (cf. Table 2.7). However, the risk-adjusted results are clearly in favor of the DPPI portfolio. This relative advantage remains when considering downside risk measures. The lower maximum drawdown of the DPPI portfolio evidences that downside protection is effective, irrespective of the choice of risk method. Comparing the performance of the DPPI portfolio across risk models yields less clear-cut results. Panel A of Table 2.7 shows the corresponding results. In terms of returns, we observe a 76bp difference between the best-performing risk model, the copula-GARCH model, and the weakest model, the two-factor GAS model. However, in terms of risk-adjusted returns, this spread is diminished, resulting in marginal differences across models. In particular, the Sharpe ratios range from 0.53 to 0.63. The same conclusions can be drawn in terms of maximum drawdown. Evaluating the risk models on the basis of the Calmar, Sortino and Omega ratios shows only marginal differences as well. The respective ranges are from 0.34 to 0.42 (Calmar), around 0.09 (Sortino), and from 1.17 to 1.19 (Omega) and suggest that even rather naive approaches did not fail to provide downside protection in the context of DPPI. This finding can be rationalized as follows. In general, few allocation changes are necessary to protect from downside risks if the DPPI strategy is reasonably calibrated. In particular, the investment exposure is reduced when approaching the floor, irrespective of the underlying risk forecast. This embedded line of defense is most likely preventing less accurate risk forecasts from impeding overall performance. As a result, any DPPI strategy dominates the underlying risky portfolio when evaluating Calmar, Sortino and Omega ratios.

Similar to several studies (Bertrand and Prigent, 2002; Ben Ameur and Prigent, 2007; Hamidi, Jurczenko, and Maillet, 2009; Ben Ameur and Prigent, 2014; Hamidi, Maillet, and Prigent, 2014), we also benchmark the DPPI performance with multipliers based on the different risk models against the CPPI performance based on a static unconditional multiplier. In particular, the latter is calculated as the maximum loss of the underlying over the whole sample period, resulting in a multiplier of 8. In terms of downside measures, the CPPI

Sharpe ratio given a constant relative risk aversion utility function of the investor. Specifically, assuming a risk-averse investor, we set the risk aversion parameter to 0.15 and the expected Sharpe ratio to 0.6.

Table 2.7: DPPI for multi-asset portfolio

	Method	Return	SD	Sharpe	MDD	Calmar	Sortino	Omega	Part	TO	ES
Panel A: Historical path											
	Underlying portfolio	6.36	7.73	0.55	-31.80	0.20	0.07	1.16	100	0	-1.86
	Money market	2.13	0.17	0.00	0.00	-	-	-	0	0	0.00
	CPPI (m=8)	4.64	4.23	0.59	-9.92	0.47	0.10	1.21	58.91	0.94	-0.93
DPPI	Historical Simulation	5.80	5.83	0.63	-13.91	0.42	0.09	1.19	83.46	0.88	-1.25
DPPI	Cornish-Fisher	5.83	5.88	0.63	-14.27	0.41	0.09	1.19	84.49	0.96	-1.27
DPPI	RiskMetrics	5.94	6.09	0.62	-14.66	0.41	0.09	1.18	87.46	0.87	-1.33
DPPI	CARE	5.62	6.13	0.57	-16.48	0.34	0.08	1.17	86.68	1.96	-1.34
DPPI	Extreme Value Theory	5.86	6.14	0.61	-14.18	0.41	0.09	1.18	87.74	0.90	-1.33
DPPI	Copula-GARCH	6.06	6.25	0.63	-14.68	0.41	0.09	1.18	89.05	0.89	-1.36
DPPI	One-Factor-GAS	5.92	6.10	0.62	-14.25	0.42	0.09	1.18	86.79	1.23	-1.32
DPPI	Two-Factor-GAS	5.30	5.97	0.53	-14.17	0.37	0.08	1.17	85.16	1.43	-1.31
DPPI	Hybrid-GAS/GARCH	5.94	6.15	0.62	-14.53	0.41	0.09	1.18	87.70	1.15	-1.33
	Average	5.88	6.08	0.62	-14.25	0.41	0.09	1.18	86.89	1.02	-1.31
	FZ loss	6.04	6.22	0.63	-14.60	0.41	0.09	1.18	88.79	0.92	-1.35
Panel B: Historical block-bootstrap											
	Underlying portfolio	6.15	9.41	0.43	-7.49	1.51	1.23	4.81	100	0	-28.56
	Money market	2.10	2.07	0.00	0.00	-	-	-	0	0	0.02
	CPPI (m=8)	5.17	6.84	0.45	-4.54	1.30	2.86	8.16	70.57	1.16	-7.87
DPPI	Historical Simulation	5.87	8.05	0.47	-5.77	1.38	2.23	5.78	88.37	0.93	-9.03
DPPI	Cornish-Fisher	5.91	8.07	0.47	-5.80	1.39	2.25	5.79	88.58	0.93	-9.07
DPPI	RiskMetrics	5.78	8.06	0.46	-5.82	1.37	2.15	5.49	88.83	0.97	-8.30
DPPI	CARE	5.72	8.12	0.45	-5.87	1.36	2.10	5.33	89.01	1.56	-8.40
DPPI	Extreme Value Theory	5.79	8.04	0.46	-5.84	1.37	2.18	5.57	89.07	0.95	-8.34
DPPI	Copula-GARCH	5.87	8.12	0.46	-5.91	1.38	2.14	5.47	89.71	0.95	-8.43
DPPI	One-Factor-GAS	5.90	8.09	0.47	-5.80	1.39	2.23	5.71	89.54	1.11	-8.43
DPPI	Two-Factor-GAS	5.52	8.10	0.42	-5.87	1.33	2.05	5.14	88.26	1.35	-8.43
DPPI	Hybrid-GAS/GARCH	5.82	8.09	0.46	-5.85	1.38	2.18	5.53	89.34	1.11	-8.47
	Average	5.86	8.04	0.47	-5.82	1.38	2.22	5.68	89.25	1.00	-8.47
	FZ loss	5.85	8.11	0.46	-5.89	1.38	2.17	5.51	89.62	0.99	-8.40

This table reports the backtesting results of the multi-asset DPPI strategy with conditional multipliers based on different 1% ES forecasts for the historical path (Panel A) and the historical block-bootstrap (Panel B) over the sample period 1996–2017. For comparison, we include a static multiplier (CPPI) based on the maximum portfolio loss (resulting in $m = 8$) as well as the performance of the underlying multi-asset portfolio and the money market investment. In each calendar year, a floor of 95% of the initial portfolio value is installed. We report the annualized mean return (Return), annualized standard deviation (SD), Sharpe ratio (SR), maximum drawdown (MDD), Calmar ratio, Sortino ratio, Omega ratio, participation in the risky multi-asset portfolio (Part), turnover (TO) and the 1% ES. Return, Sd, MDD, Part, TO and ES are given as percentages. For the historical path, the performance measures are calculated using the daily returns resulting from the strategy. For the historical block-bootstrap, the performance measures are based on the simulated yearly returns, except for MDD, Calmar ratio and participation, which are based on the daily risky asset exposure of the corresponding draw and show the yearly mean of the specific measure.

shows slightly better results than the competing DPPI strategies (Calmar ratio of 0.47 versus approximately 0.41) owing to a rather defensive investment exposure (approximately 60%). As a result, there is a severe performance drag relative to the DPPI strategies: the static multiplier underperforms in terms of mean return (4.6% versus approximately 6.0%). In turn, the CPPI strategy embeds a severely higher insurance premium compared to the DPPI strategy which is unfavorable from an investor's perspective.

The analysis of the historical block-bootstrap confirms these findings. Figure 2.4b shows the distribution of the simulated yearly returns of the DPPI strategy. For comparison, we

also include the return distribution of a pure buy-and-hold portfolio investment strategy. The chart clearly highlights the effect of portfolio insurance. The left tail of the return distribution is shifted towards the floor level such that downside risk is reduced, albeit at the expense of some return potential in the right tail. Panel B in Table 2.7 reports the corresponding performance statistics. Compared to the historical path, we obtain slightly different results. Concerning the performance of the underlying, Sortino and Omega ratios increase substantially for all strategies. This finding can be explained by the fact that the massive drawdown year 2008 loses weight when performing the historical block-bootstrap. In other words, the crisis year 2008 is "averaged out" to some extent. Again, we detect only marginal differences across risk models. In essence, the results support the conclusion drawn from the historical path analysis. DPPI strategies building on sophisticated risk models do a good job in protecting investors from downside risk. Given that the mechanics of the portfolio insurance strategy automatically reduce investment exposure when approaching the protection level, a less sophisticated risk forecast is mainly benefiting from this second line of defense.

Again, we provide robustness checks with respect to the choice of portfolio allocation and floor level. We use the same portfolios as in the robustness check of the ES targeting strategy and employ the following floor levels: 93%, 94%, 95%, 96% and 97%. The corresponding results are shown in Table 2.8. Assuming appropriate portfolio-specifc floor levels,[29] we find no significant differences across the risk models for most portfolios when analyzing the historical path (Panel A). Only for the bond and the 30/70 equity/bond portfolio we do document an outperformance of the more flexible methods, such as the copula-GARCH and the FZ loss approach (in terms of Calmar ratio). Conversely, we cannot observe a clear pattern which risk method dominates for the historical block-bootstrap analysis (Panel B). Analyzing robustness with respect to the floor level delivers similar results. Notably, we observe that the less sophisticated HS and CFA risk models are superior for higher floor levels. This finding can be explained by the fact that the DPPI strategy hardly acts on the risk forecasts for tighter floor levels and thus favors more conservative methods. Overall, these robustness checks confirm our findings for the DPPI strategy for the multi-asset portfolio.

[29] Here, we choose the following portfolio-specific floor levels: 95% for the multi-asset portfolio, 80% for the pure equity portfolio, 95% for the pure bond portfolio, 95% for the 30/70 equity/bond portfolio and 90% for the 60/40 equity/bond portfolio.

Table 2.8: DPPI: Various portfolios and floors

	Portfolios					Floors				
	Multi-Asset	Equity	Bond	30-70	60-40	93%	94%	95%	96%	97%
Panel A: Historical path (Calmar ratio)										
Historical Simulation	0.42	0.17	0.53	0.55	0.25	0.37	0.38	0.42	0.48	0.57
Cornish-Fisher	0.41	0.17	0.51	0.57	0.26	0.37	0.38	0.41	0.47	0.54
RiskMetrics	0.41	0.14	0.60	0.65	0.20	0.37	0.39	0.41	0.39	0.34
CARE	0.34	0.19	0.51	0.49	0.24	0.36	0.36	0.34	0.37	0.39
Extreme Value Theory	0.41	0.17	0.60	0.60	0.23	0.37	0.39	0.41	0.42	0.42
Copula-GARCH	0.41	0.15	0.60	0.70	0.24	0.38	0.39	0.41	0.40	0.46
One-Factor-GAS	0.42	0.17	0.57	0.60	0.24	0.38	0.39	0.42	0.45	0.49
Two-Factor-GAS	0.37	0.16	0.57	0.54	0.23	0.34	0.36	0.37	0.37	0.35
Hybrid-GAS/GARCH	0.41	0.16	0.56	0.59	0.24	0.38	0.39	0.41	0.42	0.43
Average	0.41	0.17	0.59	0.61	0.25	0.37	0.39	0.41	0.44	0.43
FZ loss	0.41	0.18	0.59	0.65	0.24	0.37	0.39	0.41	0.44	0.47
Panel B: Historical block-bootstrap (Omega ratio)										
Historical Simulation	5.78	2.56	13.61	14.21	4.36	5.39	5.58	5.78	6.07	6.44
Cornish-Fisher	5.79	2.58	13.26	14.54	4.47	5.41	5.59	5.79	6.06	6.44
RiskMetrics	5.49	2.39	13.51	14.33	4.00	5.29	5.38	5.49	5.61	5.70
CARE	5.33	2.87	9.88	9.70	4.18	5.29	5.31	5.33	5.38	5.39
Extreme Value Theory	5.57	2.62	14.42	14.59	4.12	5.36	5.45	5.57	5.69	5.78
Copula-GARCH	5.47	2.51	12.74	16.37	4.10	5.31	5.38	5.47	5.59	5.70
One-Factor-GAS	5.71	2.85	12.43	15.19	4.71	5.52	5.64	5.71	5.80	5.93
Two-Factor-GAS	5.14	2.71	11.33	12.72	4.57	5.21	5.20	5.14	5.09	5.13
Hybrid-GAS/GARCH	5.53	2.69	12.00	15.69	4.72	5.37	5.45	5.53	5.68	5.86
Average	5.68	2.67	13.27	14.84	4.40	5.41	5.53	5.68	5.84	5.98
FZ loss	5.51	2.78	12.81	14.83	4.18	5.34	5.43	5.51	5.66	5.79

This table shows the backtesting results of the DPPI strategy with conditional multipliers based on different risk forecasts for the historical path and the historical block-bootstrap over the sample period 1996–2017 using various portfolios and various floor levels. To benchmark the results of the DPPI strategy based on the multi-asset portfolio we also backtest the DPPI strategy for four different test portfolio allocations. To this end, we choose a pure equity portfolio (80% floor), a pure bond portfolio (95% floor), a 30/70 equity/bond portfolio (95% floor) and a 60/40 equity/bond portfolio (90% floor) as underlying for the DPPI strategy. Moreover, we check the robustness of the results of the multi-asset DPPI strategy with respect to the chosen floor level (93%, 94%, 95%, 96% and 97%. We base our comparison on the Calmar ratio for the historical path and on the Omega ratio for the historical block-bootstrap.

2.5. Conclusion

Tail risk protection strategies are an effective way to limit downside risk of a given investment portfolio while maintaing most of its upside return potential. Given the limitations of option-based hedging strategies, dynamic asset allocations strategies such as risk targeting and dynamic proportion portfolio insurance are popular choices among practitioners. As the success of both dynamic strategies strongly depends on the success of forecasting (tail) risk, this paper investigates a number of forecasting models to generate portfolio risk estimates that are especially suitable in timely managing the investment exposure of these strategies. To this end, we analyze risk models both prominent in the academic literature and popular among practitioners, including simple historical simulation, the RiskMetrics approach, the Cornish-Fisher Approximation, quantile/expectile regressions, extreme value theory, the copula-GARCH approach and dynamic GAS models. In addition to standalone models, we

propose a novel ES (and VaR) forecast combination approach based on a loss function that overcomes the lack of elicitability for ES by jointly modeling ES and VaR.

Empirically, we build our analysis on a global multi-asset return data set including stocks, bonds, commodities and foreign exchange rates. To take account of different market closing times we apply a return synchronization technique by extrapolating prices of closed markets, based on information from markets which close later. It turns out that the forecasts of the proposed forecast combination approach dominates both sophisticated and more naive standalone models as well as a simple average combination approach in modeling the tail of the portfolio return distribution using a comprehensive VaR/ES testing framework. When feeding the forecasts of the different risk models into the risk targeting strategy, we show that the more flexible methods, such as the copula-GARCH, the hybrid GAS/GARCH and the FZ loss combination approach, outperform more naive methods. For the DPPI strategy, however, our results are less clear-cut. We provide evidence that dynamic portfolio insurance strategies building on sophisticated risk models are capable of protecting investors from downside risk. However, more naive approaches are also able to provide downside protection. Given that portfolio insurance only leads to a few allocation changes, simple risk models might simply have been lucky. Going forward, the more accurate FZ loss forecast combination approach appears to be more likely to help mitigate the next downturn.

Appendix 2.A Return synchronization

In this section we describe the return synchronization methodology that we apply to the global multi-asset data set (see Audrino and Bühlmann, 2004). Let $S_{t_i,i}$ denote the continuous time price of asset i ($i = 1, ..., N$), where time t_i is the closing time of market i measured in local time of the base market, that is, the market with which to synchronize. The corresponding synchronized price $S^s_{t,i}$ is then defined as

$$\log\left(S^s_{t,i}\right) = \mathbb{S}\left[\log\left(S_{t,i}\right)|\mathcal{F}_t\right] = \mathbb{E}\left[\log\left(S_{t_i+1,i}\right)|\mathcal{F}_t\right], \quad t_i \leq t \leq t_i + 1 \ (t \in \mathbb{N}), \tag{2.51}$$

where $t = t_1$ and $\mathcal{F}_t$ is the complete information of all recorded prices up to time t. Logarithms are used for consistency with continuously compounded returns. Clearly, if the closing price S is observed at time $t \in \mathbb{N}$, its conditional expectation given $\mathcal{F}_t$ is the observed price. This is the case for the assets from the base market. If the market closes before t, its past prices and all the other markets may be useful in predicting S at time t. As a simplifying approximation, the authors therefore assume that, given the information $\mathcal{F}_t$, the best predicted log-prices at t and at the nearest succeeding closing time $t_i + 1$ remain the same, meaning that future changes up to $t_i + 1$ are unpredictable.

Then we denote by r_t the vector of log-returns in different markets using the multi-index $t = (t_1, t_2, \cdots, t_N)$ and define the synchronized returns r^s_t as the change in the logarithms of the synchronized prices:

$$r_t = \begin{bmatrix} \log\left(\frac{S_{t_1,1}}{S_{t_1-1,1}}\right) \\ \vdots \\ \log\left(\frac{S_{t_N,N}}{S_{t_N-1,N}}\right) \end{bmatrix}, \qquad r^s_t = \begin{bmatrix} \log\left(\frac{S^s_{t,1}}{S^s_{t-1,1}}\right) \\ \vdots \\ \log\left(\frac{S^s_{t,N}}{S^s_{t-1,N}}\right) \end{bmatrix}. \tag{2.52}$$

In order to estimate the relationship between the individual asset markets, the authors employ a simple "auxiliary" VAR(1) model:

$$r_t = \mathbf{A} r_{t-1} + \varepsilon_t, \tag{2.53}$$

where the innovations ε_t are independent and identically normally distributed with mean zero and variance-covariance Σ, independent of $\{r_s; s < t\}$, and $\mathbf{A}$ is the matrix of VAR coefficients. We can then derive the synchronized returns as follows

$$\begin{aligned} r_t^s &= \log\left(\mathbf{S}_t^s\right) - \log\left(\mathbf{S}_{t-1}^s\right) \\ &= \mathbb{E}\left[\log\left(\mathbf{S}_{t+1}\right)|\mathcal{F}_t\right] - \mathbb{E}\left[\log\left(\mathbf{S}_t\right)|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\log\left(\mathbf{S}_{t+1}\right) - \log\left(\mathbf{S}_t\right)|\mathcal{F}_t\right] - \mathbb{E}\left[\log\left(\mathbf{S}_t\right) - \log\left(\mathbf{S}_{t-1}\right)|\mathcal{F}_{t-1}\right] + \log\left(\frac{\mathbf{S}_t}{\mathbf{S}_{t-1}}\right) \\ &= \mathbb{E}\left[r_{t+1}|\mathcal{F}_t\right] - \mathbb{E}\left[r_t|\mathcal{F}_{t-1}\right] + r_t \\ &= \mathbf{A}r_t - \mathbf{A}r_{t-1} + r_t \\ &= r_t + \mathbf{A}(r_t - r_{t-1}). \end{aligned} \tag{2.54}$$

That is, any synchronized return r_t^s is still anchored in the actual realized return r_t plus an anticipated innovation according to the estimated VAR relation as captured in matrix $\mathbf{A}$. The "missing" dynamics of markets closing early in the day are thus proxied according to the short-term relationship with respect to those markets closing later that day.

Sorting markets according to their closing times enables us to readily formulate a restriction matrix for the VAR model such that markets are explained only by those markets with a later closing time. Given that US markets are the last to close in our sample, we anchor our synchronization of daily returns in US markets. Thus, the US time series remain unchanged but are still included in the VAR model to serve as explanatory variables, that is, the final set of synchronized daily returns does not build on forecasted time series for the USA but uses their original daily returns. Non-US data are forecasted to the closing time of the US market by the VAR(1).

References

Acerbi, Carlo and Balazs Szekely (2014). "Backtesting expected shortfall". *Risk* 27 (11), 76–81.

Acerbi, Carlo and Dirk Tasche (2002). "On the coherence of expected shortfall". *Journal of Banking & Finance* 26 (7), 1487–1503.

Andersen, Torben G, Tim Bollerslev, Peter F Christoffersen, and Francis X Diebold (2006). "Volatility and correlation forecasting". In: *Handbook of Economic Forecasting*. Ed. by Elliott, G., Granger, Clive W.J., and Timmermann, A. Vol. 1. Elsevier. Chap. 15, 777–878.

— (2013). "Financial risk measurement for financial risk management". In: *Handbook of the Economics of Finance*. Ed. by Constantinides, George M, Harris, Milton, and Stulz, René M. Vol. 2. Elsevier. Chap. 17, 1127–1220.

Annaert, Jan, Sofieke Van Osselaer, and Bert Verstraete (2009). "Performance evaluation of portfolio insurance strategies using stochastic dominance criteria". *Journal of Banking & Finance* 33 (2), 272–280.

Ardia, David, Kris Boudt, and Marjan Wauters (2016). "Smart beta and CPPI performance". *Finance* 37 (3), 31–65.

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, and David Heath (1999). "Coherent measures of risk". *Mathematical Finance* 9 (3), 203–228.

Audrino, Francesco and Peter Bühlmann (2004). "Synchronizing multivariate financial time series". *Journal of Risk* 6 (2), 81–106.

Balder, Sven, Michael Brandl, and Antje Mahayni (2009). "Effectiveness of CPPI strategies under discrete-time trading". *Journal of Economic Dynamics and Control* 33 (1), 204–220.

Basak, Suleyman (2002). "A comparative study of portfolio insurance". *Journal of Economic Dynamics and Control* 26 (7), 1217–1241.

Bates, John M and Clive WJ Granger (1969). "The combination of forecasts". *Journal of the Operational Research Society* 20 (4), 451–468.

Bayer, Sebastian (2018). "Combining Value-at-Risk forecasts using penalized quantile regressions". *Econometrics and Statistics* 8, 56–77.

Bayer, Sebastian and Timo Dimitriadis (Sept. 2020). "Regression-based expected shortfall backtesting". *Journal of Financial Econometrics*. nbaa013.

Ben Ameur, Hachmi and Jean-Luc Prigent (2007). "Portfolio insurance: Determination of a dynamic CPPI multiple as function of state variables". *Working paper* THEMA (University of Cergy) and ISC (Paris).

— (2014). "Portfolio insurance: Gap risk under conditional multiples". *European Journal of Operational Research* 236 (1), 238–253.

Benartzi, Shlomo and Richard H Thaler (1995). "Myopic loss aversion and the equity premium puzzle". *Quarterly Journal of Economics* 110 (1), 73–92.

Benninga, Simon (1990). "Comparing portfolio insurance strategies". *Financial Markets and Portfolio Management* 4 (1), 20–30.

Berkowitz, Jeremy, Peter Christoffersen, and Denis Pelletier (2011). "Evaluating value-at-risk models with desk-level data". *Management Science* 57 (12), 2213–2227.

Bertrand, Philippe and Jean-Luc Prigent (2002). "Portfolio insurance: The extreme value approach to the CPPI method". *Finance* 23 (2), 69–86.

Bertrand, Philippe and Jean-Luc Prigent (2011). "Omega performance measure and portfolio insurance". *Journal of Banking & Finance* 35 (7), 1811–1823.

Black, Fischer and Robert W Jones (1987). "Simplifying portfolio insurance". *Journal of Portfolio Management* 14 (1), 48–51.

— (1988). "Simplifying portfolio insurance for corporate pension plans". *Journal of Portfolio Management* 14 (4), 33–37.

Black, Fischer and AndréF Perold (1992). "Theory of constant proportion portfolio insurance". *Journal of Economic Dynamics and Control* 16 (3-4), 403–426.

Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". *Journal of Econometrics* 31 (3), 307–327.

Bollerslev, Tim, Benjamin Hood, John Huss, and Lasse Heje Pedersen (2018a). "Risk everywhere: Modeling and managing volatility". *Review of Financial Studies* 31 (7), 2729–2773.

Boudt, Kris, B Peterson, and Christophe Croux (2008). "Estimation and decomposition of downside risk for portfolios with non-normal returns". *Journal of Risk* 11 (2), 79–103.

Burns, Patrick, Robert F Engle, and Joseph J Mezrich (1998). "Correlations and volatilities of asynchronous data". *Journal of Derivatives* 5 (4), 7–18.

Cappiello, Lorenzo, Robert F Engle, and Kevin Sheppard (2006). "Asymmetric dynamics in the correlations of global equity and bond returns". *Journal of Financial Econometrics* 4 (4), 537–572.

Chen, Jiah-Shing, Chia-Lan Chang, Jia-Li Hou, and Yao-Tang Lin (2008). "Dynamic proportion portfolio insurance using genetic programming with principal component analysis". *Expert Systems with Applications* 35 (1), 273–278.

Christoffersen, Peter and Denis Pelletier (2004). "Backtesting value-at-risk: A duration-based approach". *Journal of Financial Econometrics* 2 (1), 84–108.

Christoffersen, Peter F (1998). "Evaluating interval forecasts". *International Economic Review* 39 (4), 841–862.

Cont, Rama and Peter Tankov (2009). "Constant proportion portfolio insurance in the presence of jumps in asset prices". *Mathematical Finance* 19 (3), 379–401.

Cooper, Tony (2010). "Alpha generation and risk smoothing using managed volatility". *Working Paper*.

Cornish, Edmund A and Ronald A Fisher (1938). "Moments and cumulants in the specification of distributions". *Revue de l'Institut International de Statistique* 5 (4), 307–320.

Creal, Drew, Siem Jan Koopman, and André Lucas (2013). "Generalized autoregressive score models with applications". *Journal of Applied Econometrics* 28 (5), 777–795.

Dichtl, Hubert and Wolfgang Drobetz (2011). "Portfolio insurance and prospect theory investors: Popularity and optimal design of capital protected financial products". *Journal of Banking & Finance* 35 (7), 1683–1697.

Dichtl, Hubert, Wolfgang Drobetz, and Martin Wambach (2017). "A bootstrap-based comparison of portfolio insurance strategies". *European Journal of Finance* 23 (1), 31–59.

Diebold, Francis X and Robert S Mariano (1995). "Comparing predictive accuracy". *Journal of Business & Economic Statistics* 13 (3), 253–263.

Embrechts, Paul and Marius Hofert (2014). "Statistics and quantitative risk management for banking and insurance". *Annual Review of Statistics and Its Application* 1, 493–514.

Emmer, Susanne, Marie Kratz, and Dirk Tasche (2015). "What is the best risk measure in practice? A comparison of standard measures". *Journal of Risk* 18 (2), 31–60.

Engle, Robert F (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". *Econometrica* 50 (4), 987–1007.

Engle, Robert F and Simone Manganelli (2004). "CAViaR: Conditional autoregressive value at risk by regression quantiles". *Journal of Business & Economic Statistics* 22 (4), 367–381.

Fissler, Tobias and Johanna F Ziegel (2016). "Higher order elicitability and Osband's principle". *Annals of Statistics* 44 (4), 1680–1707.

Giese, Guido (2012). "Optimal design of volatility-driven algo-alpha trading strategies". *Risk* 25 (6), 68–73.

Gneiting, Tilmann (2011). "Making and evaluating point forecasts". *Journal of the American Statistical Association* 106 (494), 746–762.

Halbleib, Roxana and Winfried Pohlmeier (2012). "Improving the value at risk forecasts: Theory and evidence from the financial crisis". *Journal of Economic Dynamics and Control* 36 (8), 1212–1228.

Hallerbach, Winfried G (2012). "A proof of the optimality of volatility weighting over time". *Journal of Investment Strategies* 1 (4), 87–99.

— (2015). "Advances in portfolio risk control". In: *Risk-Based and Factor Investing*. Ed. by Jurczenko, Emmanuel. Vol. 1. Elsevier. Chap. 1, 1–30.

Hamidi, B, E Jurczenko, and B Maillet (2009). "A CAViaR modelling for a simple time-varying proportion portfolio insurance strategy". *Bankers, Markets & Investors* 102, 4–21.

Hamidi, Benjamin, Christophe Hurlin, Patrick Kouontchou, and Bertrand Maillet (2015). "A DARE for VaR". *Finance* 36 (1), 7–38.

Hamidi, Benjamin, Bertrand Maillet, and Jean-Luc Prigent (2014). "A dynamic autoregressive expectile for time-invariant portfolio protection strategies". *Journal of Economic Dynamics and Control* 46, 1–29.

Hamidi, Benjamin, Bertrand B Maillet, and Jean-Luc Prigent (2009). "A risk management approach for portfolio insurance strategies". In: *Proceedings of the 1st EIF International Financial Research Forum, Economica*.

Hansen, Bruce E (1994). "Autoregressive conditional density estimation". *International Economic Review* 35, 705–730.

— (2008). "Least-squares forecast averaging". *Journal of Econometrics* 146 (2), 342–350.

Harvey, A C (2013). "Dynamic models for volatility and heavy tails: with applications to financial and economic time series". In: *Econometric Society Monographs*. Vol. 52. Cambridge University Press.

Harvey, David, Stephen Leybourne, and Paul Newbold (1997). "Testing the equality of prediction mean squared errors". *International Journal of Forecasting* 13 (2), 281–291.

Hocquard, Alexandre, Sunny Ng, and Nicolas Papageorgiou (2013). "A constant-volatility framework for managing tail risk". *Journal of Portfolio Management* 39 (2), 28–40.

Ilmanen, Antti and Jared Kizer (2012). "The death of diversification has been greatly exaggerated". *Journal of Portfolio Management* 38 (3), 15–27.

Jiang, Chonghui, Yongkai Ma, and Yunbi An (2009). "The effectiveness of the VaR-based portfolio insurance strategy: An empirical analysis". *International Review of Financial Analysis* 18 (4), 185–197.

Jondeau, Eric and Michael Rockinger (2006). "The Copula-GARCH model of conditional dependencies: An international stock market application". *Journal of International Money and Finance* 25 (5), 827–853.

Kirby, Chris and Barbara Ostdiek (2012). "It's all in the timing: Simple active portfolio strategies that outperform naive diversification". *Journal of Financial and Quantitative Analysis* 47 (2), 437–467.

Koenker, Roger and Gilbert Bassett (1978). "Regression quantiles". *Econometrica* 46 (1), 33–50.

Kuester, Keith, Stefan Mittnik, and Marc S Paolella (2006). "Value-at-risk prediction: A comparison of alternative strategies". *Journal of Financial Econometrics* 4 (1), 53–89.

Kupiec, Paul H (1995). "Techniques for verifying the accuracy of risk measurement models". *Journal of Derivatives* 3 (2), 73–84.

Lo, Andrew W and A Craig MacKinlay (1990a). "An econometric analysis of nonsynchronous trading". *Journal of Econometrics* 45 (1-2), 181–211.

Longin, Francois and Bruno Solnik (1995). "Is the correlation in international equity returns constant: 1960–1990?" *Journal of International Money and Finance* 14 (1), 3–26.

Martin, R Douglas and Rohit Arora (2017). "Inefficiency and bias of modified value-at-risk and expected shortfall". *Journal of Risk* 19 (6), 59–84.

McNeil, Alexander J and Rüdiger Frey (2000). "Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach". *Journal of Empirical Finance* 7 (3), 271–300.

Nadarajah, Saralees, Bo Zhang, and Stephen Chan (2014). "Estimation methods for expected shortfall". *Quantitative Finance* 14 (2), 271–291.

Newey, Whitney K and James L Powell (1987). "Asymmetric least squares estimation and testing". *Econometrica* 55, 819–847.

Nolde, Natalia and Johanna F Ziegel (2017). "Elicitability and backtesting: Perspectives for banking regulation". *Annals of Applied Statistics* 11 (4), 1833–1874.

Patton, Andrew J (2006). "Modelling asymmetric exchange rate dependence". *International Economic Review* 47 (2), 527–556.

Patton, Andrew J, Johanna F Ziegel, and Rui Chen (2019). "Dynamic semiparametric models for expected shortfall (and value-at-risk)". *Journal of Econometrics* 211 (2), 388–413.

Perchet, Romain, Raul Leote De Carvalho, Thomas Heckel, and Pierre Moulin (2015). "Predicting the success of volatility targeting strategies: Application to equities and other asset classes". *Journal of Alternative Investments* 18 (3), 21–38.

Perold, Andre (1986). "Constant proportion portfolio insurance". *Harvard Business School*.

Perold, Andre F and William F Sharpe (1988). "Dynamic strategies for asset allocation". *Financial Analysts Journal* 44 (1), 16–27.

Pritsker, Matthew (2006). "The hidden dangers of historical simulation". *Journal of Banking & Finance* 30 (2), 561–582.

Righi, Marcelo Brutti and Paulo Sergio Ceretta (2015). "A comparison of expected shortfall estimation models". *Journal of Economics and Business* 78, 14–47.

Santos, André AP, Francisco J Nogales, and Esther Ruiz (2012). "Comparing univariate and multivariate models to forecast portfolio value-at-risk". *Journal of Financial Econometrics* 11 (2), 400–441.

Scherer, B (2013). "Synchronize your data or get out of step with your risks". *Journal of Derivatives* 20 (3), 75–84.

Scholes, Myron and Joseph Williams (1977). "Estimating betas from nonsynchronous data". *Journal of Financial Economics* 5 (3), 309–327.

Shan, Kejia and Yuhong Yang (2009). "Combining regression quantile estimators". *Statistica Sinica* 19 (3), 1171–1191.

Soupé, François, Thomas Heckel, and Raul Leote De Carvalho (2014). "Portfolio insurance with adaptive protection (PIWAP)". *Journal of Investment Strategies* 5 (3), 1–15.

Taylor, James W (2008). "Estimating value at risk and expected shortfall using expectiles". *Journal of Financial Econometrics* 6 (2), 231–252.

— (2019). "Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution". *Journal of Business & Economic Statistics* 37 (1), 121–133.

— (2020). "Forecast combinations for value at risk and expected shortfall". *International Journal of Forecasting* 36 (2), 428–441.

Taylor, Stephen J (1986). *Modelling Financial Time Series*. Chichester: Wiley.

Timmermann, Allan (2006). "Forecast combinations". In: *Handbook of Economic Forecasting*. Ed. by Elliott, G., Granger, C.W., and Timmermann, A. Vol. 1. Elsevier. Chap. 4, 135–196.

Yamai, Yasuhiro and Toshinao Yoshiba (2002). "On the validity of value-at-risk: Comparative analyses with expected shortfall". *Monetary and Economic Studies* 20 (1), 57–85.

Ye, Yinyu (1987). "Interior algorithms for linear, quadratic, and linearly constrained nonlinear programming". PhD thesis. Department of ESS, Stanford University.

Zangari, Peter (1996). "A VaR methodology for portfolios that include options". *RiskMetrics Monitor* 1, 4–12.

Chapter 3

Combining Value-at-Risk and Expected Shortfall Forecasts using Machine Learning Techniques

This project is joint work with Harald Lohre, Ingmar Nolte and Maximilian Stroh. We thank Roxana Halbleib, Stefan Mittnik, Winfried Pohlmeier, Artem Prokhorov, Michael Rockinger and the participants of the 2019 CEQURA Conference on Advances in Financial and Insurance Risk Management in Munich and the 6th KoLaMaFr Workshop on Financial Econometrics for fruitful discussions and suggestions. This work was supported by funding from the Economic and Social Research Council (UK).

3.1. Introduction

Extreme events such as the recent COVID-19 pandemic and its economic consequences or the global financial crisis of 2007-2008 stress the importance of measuring and forecasting market risk for financial applications such as asset allocation, hedging or risk management. To assess financial market risks, expected shortfall (ES) is a highly relevant metric due to its ability to mitigate problems of the popular value-at-risk (VaR) measure, such as ignoring the tail of the distribution and the lack of sub-additivity (see Artzner et al., 1999; Acerbi and Tasche, 2002). Specifically, ES gives the average return of a risky asset below a given quantile of its return distribution, thus summarizing the tail risk information beyond the VaR.

Given their prominence, there exists an extensive literature on estimating and predicting VaR and ES (Andersen et al., 2006, 2013; Kuester, Mittnik, and Paolella, 2006; Louzis, Xanthopoulos-Sisinis, and Refenes, 2014; Righi and Ceretta, 2015; Nieto and Ruiz, 2016; Happersberger, Lohre, and Nolte, 2020). The primary challenge with VaR and ES forecasting is, however, that the models' performance and reliability in accurately predicting the risk often heavily depends on the data. While a parsimonious model can perform well in stable markets, it might fail during a volatile period. Likewise, highly parameterized models can be adequate during periods of high volatility, but might be easily outperformed by simpler approaches in less turbulent times (Bayer, 2018). If the best model is unknown or likely to change over time, a promising alternative is to combine the predictions originating from various models (Bates and Granger, 1969). Timmermann (2006) puts forward three main arguments in favor of combining forecasts to enhance the predictive performance relative to standalone models. First, there are diversification gains arising from the combination of forecasts computed from different assumptions, specifications or information sets. Second, combination forecasts tend to be robust against structural breaks. Third, the influence of potential misspecification biases and measurement errors of the individual models is reduced due to averaging over a set of forecasts derived from various models (Bayer, 2018).

While the literature offers a number of methods for combining VaR forecasts (see Bayer, 2018, for a summary), there is a lack of ES forecast combination methods. This relates to the fact that ES is not "elicitable", that is, there does not exist a loss function such that the correct ES forecast is actually minimizing the expected loss (cf. Gneiting, 2011). This lack of elicitability renders the estimation and backtesting of ES challenging (see Acerbi and Szekely, 2014; Embrechts and Hofert, 2014; Emmer, Kratz, and Tasche, 2015). As a remedy, Fissler and Ziegel (2016) introduced a class of loss functions that overcome the lack of elicitability for ES by jointly modeling ES and VaR. There exist two approaches that draw on these results to form linear ES (and VaR) combination forecasts. Happersberger, Lohre,

and Nolte (2020) simply estimate the optimal forecast combination weights by minimizing the average loss of the FZ loss function using a linear combination of the individual forecasts for both VaR and ES. Taylor (2020)'s approach slightly differs in the sense that it does not combine ES forecasts, but instead combines forecasts of the difference between ES and VaR.

In this paper, we extend these simple combination approaches utilizing machine learning[1] techniques to increase prediction accuracy and allowing for non-linear forecast combinations. In particular, we examine the merits of using shrinkage and neural network models to form VaR and ES combination forecasts.

Simple linear ES (and VaR) combination schemes may suffer from multicollinearity, if we combine a large number of forecasts that are based on the same data or similar mathematical approaches. Indeed, in our empirical application we observe high pairwise correlations among ES and VaR forecasts of the individual models, indicating the presence of pronounced multicollinearity. Also, highly correlated forecasts may lead to overfitting. That is, we may find a model that fits the in-sample data well, but fails to properly generalize to new data (Hastie, Tibshirani, and Friedman, 2011; Bayer, 2018). In this situation, the approaches of Happersberger, Lohre, and Nolte (2020) or Taylor (2020) may produce unstable estimates of the combination weights. An obvious solution is to focus on combining forecasts that exhibit small to moderate cross-correlations. However, we aim to avoid manually selecting models; instead, we utilize shrinkage models that are able to handle high correlations among the input variables. Specifically, we consider the popular least shrinkage and selection operator (LASSO) of Tibshirani (1996), the ridge penalty of Hoerl and Kennard (1970a,b) and the elastic net penalty proposed by Zou and Hastie (2005) that linearly combines the penalties of the LASSO and ridge methods. According to Bayer (2018), the latter shrinkage model is particularly appealing for forecast combination, as it produces stable weight estimates, reduces overfitting and automatically selects the individual forecasts.

In addition, we consider the egalitarian LASSO model and related variants proposed by Diebold and Shin (2019). The authors suggest that simple averages are a natural shrinkage direction when combining forecasts, given that simple average combinations are frequently found to perform well (despite being theoretically sub-optimal). In this vein, they propose LASSO-based procedures that shrink combination weights towards equal weights instead of

[1]The definition of "machine learning" is inchoate and often context-specific (cf. Gu, Kelly, and Xiu, 2020). We use the term to describe a diverse collection of high-dimensional models that computers use for making and improving statistical predictions from a given data set. Among the collection of machine learning models, we focus on supervised machine learning in this paper. This type of machine learning concentrates on prediction problems, where we have a data set for which we already know the outcome of interest and want to learn to predict the outcome for new data. More precisely, the machine learning algorithm learns a model by estimating parameters (like weights, as in our case) or learning structures and is guided by a loss function that is minimized (cf. Molnar, 2019).

shrinking towards zero. They put forward the partially-egalitarian LASSO as the optimal solution, which uses the standard LASSO to select the appropriate forecasts in the first step and then shrink these towards equal weights by applying the egalitarian ridge, the egalitarian LASSO or the simple average in the second step.

Although shrinkage combination models are able to handle a large number of (potentially) highly correlated individual forecasts, they still assume linear relationships between the individual forecasts and the target combination forecast. However, linear combinations may not be optimal in terms of prediction accuracy if the models that generate the individual forecasts are non-linear or if the target forecast's true underlying expectation is a non-linear function of the information sets on which the individual forecasts are based (Donaldson and Kamstra, 1996). As neural networks have the ability to meaningfully approximate whatever functional form best characterizes the data, they could be well-suited for forecast combination when the optimal combination of individual forecasts is potentially non-linear. Donaldson and Kamstra (1996) successfully apply the class of neural networks to volatility forecasting. They demonstrate that a neural network-based combination model dominates traditional linear combining procedures. However, the quality of ES and VaR predictions can only be assessed with the help of relative rare events, and therefore the amount of relevant data is limited. Thus, introducing too much flexibility into the forecast combination model might result in poor out-of-sample results. In this spirit, we investigate whether the complexity of applying simple feed-forward neural networks to ES and VaR forecast combination is actually beneficial.

In the empirical part of this paper, we assess the performance of the proposed shrinkage and neural network combination models applied to a data set that encompasses 12 major equity indices over a period of 30 years. Combining is most promising when the individual methods use different information or use information in different ways. Hence, we resort to a diverse set of individual VaR and ES models, including non-parametric, parametric and semi-parametric techniques as well as methods capturing intraday volatility: historical simulation, weighted historical simulation, CAViaR, dynamic GAS and various location-scale models with GARCH, RiskMetrics, realized GARCH and HAR volatility processes and innovation processes based on filtered historical simulation and extreme value theory. For forecast evaluation we employ a comprehensive VaR and ES backtesting framework comprising methods based on calibration tests (Kupiec, 1995; Christoffersen, 1998; McNeil and Frey, 2000; Christoffersen and Pelletier, 2004; Nolde and Ziegel, 2017; Bayer and Dimitriadis, 2020) and loss functions (Diebold and Mariano, 1995; Hansen, Lunde, and Nason, 2011; Patton, Ziegel, and Chen, 2019). Our empirical results indicate that the machine-learned VaR and ES forecasts outperform a set of existing combination approaches in terms of

statistical accuracy. While all combination approaches exhibit high passing rates of the calibration tests (across equity indices and multiple probability levels), machine learning methods show a better forecasting accuracy than the (majority of) competing combination approaches when comparing the models with relative tests such as the model confidence set or average loss ranking. Specifically, egalitarian shrinkage models such as the egalitarian ridge or partially-egalitarian LASSO models are well-suited for combining VaR and ES predictions. When splitting the evaluation sample into calm and recession periods, the egalitarian models emerge as particularly well-performing in calm periods, whereas the neural network combination model dominates in recession periods. When evaluating the combination forecasts during the recent COVID-19 period, we observe lower VaR violation rates than in the global financial crisis, suggesting that the combination models have learned from previous recessions.

In addition to assessing the statistical accuracy of the combination forecasts, we investigate their relevance in a portfolio management application. In particular, we implement a risk targeting strategy that controls portfolio risk over time by systematically adjusting the investment exposure according to its current risk (forecast) in order to keep the *ex ante* risk at a constant target level. In terms of downside risk measures we find the best performance for the egalitarian shrinkage models as well as for the simple average combination approach.

Our work extends the empirical literature on ES and VaR modeling in two ways: To the best of our knowledge, this study is the first to apply machine learning techniques to ES forecast combination. While Bayer (2018) already uses standard shrinkage methods in the context of combining VaR forecasts, egalitarian shrinkage and neural network models are new to the VaR and ES forecast combination literature. Also, the variety of non-parametric combination approaches has not been used for combining ES forecasts before, allowing for a thorough comparison of available combination methods. In addition, we are the first to investigate the performance of VaR and ES (combination) forecasts in the prevailing COVID-19 period.

The remainder of the paper is structured as follows. In Section 3.2 we present the shrinkage and neural network models applied to ES (and VaR) forecast combination. Section 3.3 outlines the data, the models to be combined, the set of competing combination techniques and the forecast evaluation methodology. Section 3.4 presents the results of the empirical application. Section 3.5 summarizes and concludes the paper.

3.2. Forecast combination based on machine learning

This section describes the set of machine learning methodologies implemented for combining VaR and ES forecasts. We start with the statistical framework that includes the definition of VaR and ES, corresponding combination forecasts and a consistent loss function for VaR and ES that is the basis for the estimation procedure of all proposed combination methods. What follows is a detailed description of the individual machine learning-based combination techniques. As most of these methods require tuning of hyperparameters, we conclude the section with details on corresponding calibration procedures.

3.2.1. Risk measures and loss function

Let r_t be the daily log return of a single financial asset at time t, with conditional (on information set $\mathcal{F}_{t-1}$) distribution F_t, which we assume to be strictly increasing with finite mean. Then, the VaR forecast for period $t+1$ is simply the α-quantile of the conditional return distribution at $t+1$, that is,

$$\text{VaR}_{t+1|t}(\alpha) \equiv Q_\alpha\left(r_{t+1}|\mathcal{F}_t\right) = \inf\{x \in \mathbb{R} : P\left(r_{t+1} \leq x|\mathcal{F}_t\right) \geq \alpha\}, \tag{3.1}$$

where $\alpha \in (0,1)$ is the probability level, $Q_\alpha(\cdot)$ denotes the quantile function and $\mathcal{F}_t$ represents the information available at time t. The corresponding ES forecast for period $t+1$ is defined as the expected return conditional on the return being below its VaR, specifically,

$$\text{ES}_{t+1|t}(\alpha) \equiv \mathbb{E}\left[r_{t+1}|r_{t+1} \leq \text{VaR}_{t+1|t}, \mathcal{F}_t\right] = \frac{1}{\alpha}\int_0^\alpha \text{VaR}_{t+1|t}(s)ds. \tag{3.2}$$

Throughout the paper, we suppress the probability level α to keep the notation simple.

In the following, we assume that the forecaster has a set of individual methods at hand that each produces a forecast for VaR and ES. Specifically, we define $\text{VaR}_{m,t+1|t}$ and $\text{ES}_{m,t+1|t}$ to be the VaR and ES forecast of model $m = 1, ..., M$ for day $t+1$ based on the information available at t. When combining the forecasts from the individual methods we generally allow the combination forecasts for VaR to depend also on the individual ES forecasts and the combination forecasts for ES to depend also on the individual VaR forecasts. In addition, we allow the combination weights for VaR and ES, denoted by $\beta_{m,t}^{\text{VaR}}$ and $\beta_{m,t}^{\text{ES}}$, to differ as the

quality of a method's VaR and ES forecasts may differ. Thus, the combination forecasts for VaR and ES, denoted by $\text{VaR}_{c,t+1|t}$ and $\text{ES}_{c,t+1|t}$, are given by

$$\text{VaR}_{c,t+1|t} = \tilde{g}_1\left(\mathbf{VaR}_{t+1|t}, \mathbf{ES}_{t+1|t}; \tilde{\boldsymbol{\beta}}_{t,1}^{\text{VaR}}, \tilde{\boldsymbol{\beta}}_{t,1}^{\text{ES}}\right), \tag{3.3}$$

$$\text{ES}_{c,t+1|t} = \tilde{g}_2\left(\mathbf{VaR}_{t+1|t}, \mathbf{ES}_{t+1|t}; \tilde{\boldsymbol{\beta}}_{t,2}^{\text{VaR}}, \tilde{\boldsymbol{\beta}}_{t,2}^{\text{ES}}\right), \tag{3.4}$$

where $\tilde{g}_1(\cdot)$ and $\tilde{g}_2(\cdot)$ are some functions that may be linear or non-linear in the weight vectors, $\boldsymbol{\beta}_{t,1}^{\text{VaR}}$, $\boldsymbol{\beta}_{t,1}^{\text{ES}}$ and $\boldsymbol{\beta}_{t,2}^{\text{VaR}}$, $\boldsymbol{\beta}_{t,2}^{\text{ES}}$, in the vectors of the M individual forecasts, $\mathbf{VaR}_{t+1|t}$ and $\mathbf{ES}_{t+1|t}$, or both.[2]

Despite different combination schemes, i.e. different specifications of the function $g(\cdot)$, all methods share the basic objective of minimizing the same class of loss functions (or "scoring rules"). While there exists a consistent loss function for VaR, ES is only elicitable jointly with VaR.[3] Fissler and Ziegel (2016) show that the following class of loss functions is consistent for VaR and ES, if both VaR and ES are strictly negative and ES is smaller than VaR (which follows naturally from the definition of VaR and ES). That is, minimizing the expected loss using any of these loss functions returns the true VaR and ES:

$$\begin{aligned} L(r, \text{VaR}, \text{ES}, \alpha, G_1, G_2) = (\mathbb{1}\{r \leq \text{VaR}\} - \alpha)\left(G_1(\text{VaR}) - G_1(r) + \frac{1}{\alpha} G_2(\text{ES})\,\text{VaR}\right) \\ - G_2(\text{ES})\left(\frac{1}{\alpha}\mathbb{1}\{r \leq \text{VaR}\}\, r - \text{ES}\right) - \mathcal{G}_2(\text{ES}), \end{aligned} \tag{3.5}$$

where the function G_1 is weakly increasing, the function G_2 is strictly increasing and strictly positive, and $\mathcal{G}_2' = G_2$. Similar to Patton, Ziegel, and Chen (2019), we choose the parameters of the function class in such a way that the loss differences of two forecasts are homogeneous of degree zero, given that VaR and ES are strictly negative, i.e. $G_1(x) = 0$ and $G_2(x) = -1/x$. The resulting loss function, which we call "FZ loss function" in the following, is then given by

$$L_{FZ}(r, \text{VaR}, \text{ES}, \alpha) = -\frac{1}{\alpha\,\text{ES}}\mathbb{1}\{r \leq \text{VaR}\}\,(\text{VaR} - r) + \frac{\text{VaR}}{\text{ES}} + \log(-\text{ES}) - 1. \tag{3.6}$$

Given its strict consistency, we can use the FZ loss function as objective function for estimating the optimal combination weights (see Gneiting and Raftery, 2007).

[2] Note that we omit the corresponding index of $\boldsymbol{\beta}_{t,1}^{\text{VaR}}$ and $\boldsymbol{\beta}_{t,2}^{\text{ES}}$ if $\text{VaR}_{c,t+1|t}$ only depends on $\mathbf{VaR}_{t+1|t}$ and $\text{ES}_{c,t+1|t}$ only depends on $\mathbf{ES}_{t+1|t}$.

[3] See Taylor (2020) for more details on loss functions for VaR and ES.

3.2.2. Minimum loss

A simple way to estimate the optimal forecast combination weights is to minimize the average loss of the FZ loss function using a linear combination of the M individual forecasts (see Happersberger, Lohre, and Nolte, 2020):

$$\left(\widehat{\boldsymbol{\beta}}_t^{\text{VaR}}, \widehat{\boldsymbol{\beta}}_t^{\text{ES}}\right) = \underset{\boldsymbol{\beta}_t^{VaR}, \boldsymbol{\beta}_t^{ES}}{\arg\min} \frac{1}{t} \sum_{\tau=0}^{t-1} L_{FZ}\left(r_{\tau+1}, \left(\mathbf{VaR}_{\tau+1|\tau}\right)' \boldsymbol{\beta}_t^{\text{VaR}}, \left(\mathbf{ES}_{\tau+1|\tau}\right)' \boldsymbol{\beta}_t^{\text{ES}}\right). \tag{3.7}$$

Following the forecast combination literature (Timmermann, 2006; Hansen, 2008; Happersberger, Lohre, and Nolte, 2020) we impose convexity on the combination weights as this restriction typically improves upon the non-constrained estimator in terms of predictive performance. Convex weights are non-negative and sum to one, i.e. $0 \leq \beta_{m,t}^x \leq 1$ for $m = 1, ..., M$ and $\sum_{m=1}^M \beta_{m,t}^x = 1, x \in \{\text{VaR}, \text{ES}\}$.

To estimate the combination weights we use an optimization procedure similar to that of Engle and Manganelli (2004) and Happersberger, Lohre, and Nolte (2020). First, we generate vectors of parameters from a uniform random number generator such that the convex weight restriction is fulfilled. Subsequently, we compute the average loss from the FZ loss function for each of these vectors and select the ten vectors that produce the lowest average loss as initial values for the optimization routine. Finally, we minimize the average loss for each of the ten resulting vectors utilizing the augmented Lagrange multiplier method with a sequential quadratic programming interior algorithm according to Ye (1987) and select the vector producing the lowest average loss as the final parameter vector (cf. Happersberger, Lohre, and Nolte, 2020). The optimization procedure includes the restriction that both VaR and ES are negative and ES is always below VaR.

3.2.3. Shrinkage methods

When combining a large number of ES (or VaR) forecasts, there is a high chance that some forecasts will be largely redundant, given that they might be based on the same information and possibly resort to similar mathematical approaches. This phenomenon is called multicollinearity. If there is multicollinearity among the set of individual forecasts, the simple minimum loss approach will deliver unstable weight estimates: small variations in the data can lead to large changes in the estimated combination weights. Likewise, the minimum loss approach may suffer from the problem of overfitting in the presence of highly correlated ES (or VaR) forecasts. This is not problematic from an in-sample perspective, because the estimated coefficients still minimize the loss function. However, such imprecisely

estimated parameters can be harmful for out-of-sample purposes, because the model may fail to properly generalize to new data (cf. Hastie, Tibshirani, and Friedman, 2011; Bayer, 2018).

A simple and common solution is to append a penalty to the objective function (i.e. the average loss function) in order to favor more robust and parsimonious specifications:

$$\mathcal{L}_t\left(\boldsymbol{\beta}_t^{\text{VaR}}, \boldsymbol{\beta}_t^{\text{ES}}; \cdot\right) = \left(\frac{1}{t}\sum_{\tau=0}^{t-1} L_{FZ}\left(r_{\tau+1}, \left(\mathbf{VaR}_{\tau+1|\tau}\right)'\boldsymbol{\beta}_t^{\text{VaR}}, \left(\mathbf{ES}_{\tau+1|\tau}\right)'\boldsymbol{\beta}_t^{\text{ES}}\right)\right) + \phi\left(\boldsymbol{\beta}_t^{\text{VaR}}, \boldsymbol{\beta}_t^{\text{ES}}; \cdot\right), \tag{3.8}$$

where $\phi(\cdot)$ is the penalty function. The model's in-sample performance suffers from this regularization of the estimation problem, which, however, usually improves the out-of-sample performance in terms of parameter stability. This improvement will be achieved if the penalization manages to increase the model's signal-to-noise ratio by reducing the noise (Gu, Kelly, and Xiu, 2020).

Elastic net, ridge and LASSO

We first consider the popular elastic net penalty of Zou and Hastie (2005), which is a convex combination of the ridge penalty of Hoerl and Kennard (1970a,b) and the LASSO of Tibshirani (1996). Given the parameter vector β, its most general form presents as follows:

$$\phi(\beta; \lambda, \delta) = \lambda \sum_{m=1}^{M}\left(\delta|\beta_m| + \frac{1}{2}(1-\delta)\beta_m^2\right), \tag{3.9}$$

where λ is the non-negative regularization parameter, which controls the amount of regularization, and $\delta \in [0, 1]$ balances the ridge and the LASSO term. The case $\delta = 0$ corresponds to the ridge penalty, which shrinks the parameter coefficients by imposing a penalty on their size. In particular, coefficients are shrunk towards zero (and each other), without being set exactly equal to zero. The case $\delta = 1$ corresponds to the LASSO, which also shrinks the coefficients but additionally selects variables by setting sufficiently small coefficients exactly to zero. This variable selection feature leads to simpler and more interpretable combination models (compared to the ridge), given that only a subset of the predictors are involved (see Hastie, Tibshirani, and Friedman, 2011). For intermediate values of δ, the elastic net combines the strengths of both approaches so that Zou and Hastie (2005) interpret the elastic net as a stabilized version of the LASSO penalization.

Given the two-dimensional parameter space, i.e. β^{VaR} and β^{ES}, we need to adjust the standard elastic net penalty to our framework. For this purpose, we append the objective

function by a separate elastic penalty term for both VaR and ES. Thus, the FZ loss estimator with elastic net penalization is given by

$$\begin{aligned}\left(\widehat{\boldsymbol{\beta}}_t^{\text{VaR}}, \widehat{\boldsymbol{\beta}}_t^{\text{ES}}\right) = \underset{\boldsymbol{\beta}_t^{VaR}, \boldsymbol{\beta}_t^{ES}}{\arg\min} \; & \frac{1}{t}\sum_{\tau=0}^{t-1} L_{FZ}\left(r_{\tau+1}, \left(\mathbf{VaR}_{\tau+1|\tau}\right)'\boldsymbol{\beta}_t^{\text{VaR}}, \left(\mathbf{ES}_{\tau+1|\tau}\right)'\boldsymbol{\beta}_t^{\text{ES}}\right) \\ & + \lambda_t^{\text{VaR}} \sum_{m=1}^{M}\left(\delta_t^{\text{VaR}}|\beta_{m,t}^{\text{VaR}}| + \frac{1}{2}(1-\delta_t^{\text{VaR}})\left(\beta_{m,t}^{\text{VaR}}\right)^2\right) \\ & + \lambda_t^{\text{ES}} \sum_{m=1}^{M}\left(\delta_t^{\text{ES}}|\beta_{m,t}^{\text{ES}}| + \frac{1}{2}(1-\delta_t^{\text{ES}})\left(\beta_{m,t}^{\text{ES}}\right)^2\right).\end{aligned} \tag{3.10}$$

As suggested by Hastie, Tibshirani, and Friedman (2011) and Bayer (2018), we only estimate λ_t^{VaR} and λ_t^{ES} and consider pre-selected values of and δ_t^{VaR} and δ_t^{ES}. In particular, we consider the three cases of $\delta_t^{\text{VaR}} = \delta_t^{\text{ES}} = 0$ (ridge), $\delta_t^{\text{VaR}} = \delta_t^{\text{ES}} = 1$ (LASSO) and $\delta_t^{\text{VaR}} = \delta_t^{\text{ES}} = 0.5$ (elastic net) in our empirical application.

Egalitarian LASSO and its relatives

Although theoretically sub-optimal, simple-average combinations are frequently found to perform well and even outperform more sophisticated combination methods—a finding often referred to as the forecast combination "equal weights puzzle" (see Clemen, 1989; Diebold, 1989; Stock and Watson, 2004).[4] Therefore, Diebold and Shin (2019) suggest that simple averages (equal weights) are a natural shrinkage direction when combining forecasts. In this vein, they propose LASSO-based procedures that shrink combination weights towards equal weights instead of shrinking towards zero.

The egalitarian ridge ("eRidge") penalty is a modified ridge penalty that centers around $1/M$:

$$\phi(\beta;\lambda) = \frac{1}{2}\lambda\sum_{m=1}^{M}\left(\beta_m - \frac{1}{M}\right)^2. \tag{3.11}$$

Although it shrinks towards equality, the eRidge does not select variables, similar to the standard ridge penalty.

The egalitarian LASSO ("eLASSO") penalty changes the standard LASSO to

$$\phi(\beta;\lambda) = \lambda\sum_{m=1}^{M}\left|\beta_m - \frac{1}{M}\right|. \tag{3.12}$$

[4]The theoretical sub-optimality of equal weights is for example shown by Diebold and Shin (2019).

That is, instead of shrinking the weights towards zero, it shrinks the deviations from equal weights towards zero. Like the standard LASSO, the eLASSO shrinks and selects, but whereas the standard LASSO shrinks towards zero and selects to zero, the eLASSO shrinks towards equality and selects to equality.

In a similar vein, we can construct the egalitarian elastic net penalty ("eElasticNet"):

$$\phi(\beta;\lambda,\delta) = \lambda \sum_{m=1}^{M} \left(\delta|\beta_m - \frac{1}{M}| + \frac{1}{2}(1-\delta)\left(\beta_m - \frac{1}{M}\right)^2 \right), \tag{3.13}$$

where $\delta \in [0,1]$ balances the eRidge and the eLASSO term.

Still, according to Diebold and Shin (2019) the optimal solution is to set some combination weights to zero (i.e. select to zero) and shrink the remaining weights towards equality. For this purpose, they put forward the partially-egalitarian LASSO ("peLASSO"), which fulfills both requirements. As difficult to optimize in one step, the authors propose a two-step implementation. In a first step, one selects k forecasts from among the full set of M forecasts using the standard LASSO. In a second step, one shrinks the combination weights of the k forecasts that survive step 1 towards $1/k$. For the second step, one can use both the eRidge or the eLASSO. The only difference is that the eLASSO sets some of the surviving weights to exactly $1/k$ and shrinks the rest towards $1/k$, whereas the eRidge shrinks all surviving weights towards $1/k$. A simple alternative for the second step is to average the surviving k forecasts, thus avoiding a second optimization step.

3.2.4. Neural network combination model

Inspired by the organization and functioning of biological neurons, neural networks represent a class of flexible non-linear models that have been developed in different fields, such as biostatistics, image processing, neuroscience and artificial intelligence. The distinguishing feature of neural networks is that they are universal function approximators for any smooth predictive association (Hornik, Stinchcombe, White, et al., 1989; Cybenko, 1989). That means, they do not restrict the shape of the distribution or the relationship between the distribution's shape and the inputs that it is conditioned on. Instead, they are able to meaningfully approximate whatever functional form best characterizes the data. In particular, they learn the intrinsic relationship in the data through a number of interconnected processing elements, called neurons, spread in different layers (Friedman, Hastie, and Tibshirani, 2001). Thus, neural networks are ideally suited to the problem of forecast combination when the optimal combination of individual forecasts is potentially non-linear.

We focus our analysis on traditional feed-forward networks, also called multi-layer perceptrons. Given that we already model temporal dependencies within the individual

VaR and ES forecasting methods, we refrain from employing more complex, sequential data compatible network structures such as recurrent neural networks, convolutional neural networks or long-short term memory networks for the combination of forecasts. Simple feed-forward neural networks generally consist of an input layer of raw predictors, one or more hidden layers that interact and non-linearly transform the predictors, and an output layer that aggregates hidden layers into an ultimate outcome prediction. Analogous to axons in a biological brain, layers of the networks represent groups of "neurons" with each layer connected by "synapses" that transmit signals (i.e. the sample information) among neurons of different layers (see Gu, Kelly, and Xiu, 2020). Thus, in this type of architecture information is passed "forward" from the input layer through the hidden layers to the output layer without feedback.

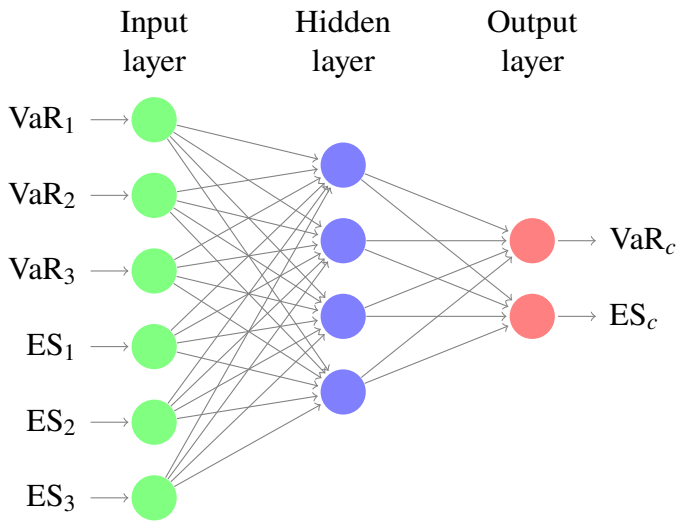


Figure 3.1: Neural network schematic. This figure shows the schematic of a single hidden layer, feed-forward neural network. The green circles represent the input variables, which are the individual VaR and ES models in our case. The blue circles represent the fully connected neurons in the hidden layer. The red circles represent the output variables consisting of the combined VaR and ES predictions.

Figure 3.1 illustrates the architecture of a single hidden layer, feed-forward neural network as implemented in our empirical analysis. The green circles represent the input variables, which are the individual VaR and ES models in our case. The blue circles represent the fully connected neurons in the hidden layer. The red circles represent the output variables consisting of the combined VaR and ES predictions, given the two dimensions of the FZ loss function that we use for combining VaR and ES forecasts. This architecture is in contrast to most applications of neural networks in finance, which usually use a standard mean-squared error loss function and therefore require only a single output variable.

Formally, we define the neural network for our application of combining VaR and ES forecasts in the following. Note that we estimate such a neural network at every time step, but suppress the time index t for simplicity. Let K_l be the number of neurons (or units) in each hidden layer $l = 1, ..., L$. Moreover, we denote the output of neuron k in layer l as $z_{k,l}$ and the vector of outputs for this layer, augmented to include a constant,[5] as $\mathbf{z}_l = \left(1, z_{l,1}, ..., z_{l,K_l}\right)'$. To initialize the network, we similarly define the input layer, which contains the raw predictors, in our case the individual VaR and ES forecasts, as $\mathbf{z}_0 = (1, \text{VaR}_1, ..., \text{VaR}_M, \text{ES}_1, ..., \text{ES}_M)'$. Each neuron k in layer l linearly aggregates information from all of the units in the layer below and then applies the univariate non-linear "activation function" ϕ_l before sending its output $z_{l,k}$ to the next layer.[6] Hence, the recursive output formula for the neural network at each neuron k in layer $l > 0$ is given by

$$z_{l,k} = \phi_l \left(\mathbf{z}'_{l-1}\boldsymbol{\theta}_{l-1,k}\right), \tag{3.14}$$

where $\boldsymbol{\theta}_{l,k} = \left(\theta^{(0)}_{l,k}, \theta^{(1)}_{l,k}, ..., \theta^{(K_l)}_{l,k}\right)'$ denotes the vector of connection weights between neuron k in layer l and all K_{l-1} neurons in the layer below. The final network output then takes the form

$$\text{VaR}_c = g_1\left(\mathbf{VaR}, \mathbf{ES}; \boldsymbol{\theta}\right) = \mathbf{z}'_L\boldsymbol{\theta}_{L,1}, \tag{3.15}$$

$$\text{ES}_c = g_2\left(\mathbf{VaR}, \mathbf{ES}; \boldsymbol{\theta}\right) = \mathbf{z}'_L\boldsymbol{\theta}_{L,2}, \tag{3.16}$$

where $\boldsymbol{\theta}$ summarizes the complete set of model parameters to be estimated. The activation function in the output layer is set as the identity function, so that the two final outputs enjoy the freedom of assuming any real value.

[5] The constant term, also known as bias term, adds flexibility to the hidden nodes and the output-node responses (activations) in a way similar to the constant term in linear regression models (cf. Kuan and White, 1994).

[6] Activation plays a key role within neural networks. Similar to biologic neurons in the human brain, neurons (or nodes) receive inputs from adjacent neurons. When the information in these inputs accumulate beyond a certain threshold the neuron is "activated" suggesting that there is a signal.

To put it differently, a neural network is just a hierarchical model of the form

$$
\begin{aligned}
g_j\left(\mathbf{VaR}, \mathbf{ES}; \boldsymbol{\theta}\right) &= \mathbf{z}_L' \boldsymbol{\theta}_{L,j}, \ \forall\ j = 1, 2 && \text{(Output layer)} \\
z_{L,k} &= \phi_L\left(\mathbf{z}_{L-1}' \boldsymbol{\theta}_{L-1,k}\right), \ \forall\ k = 1, ..., K_L && \text{(Hidden layer } L\text{)} \\
z_{L-1,k} &= \phi_{L-1}\left(\mathbf{z}_{L-2}' \boldsymbol{\theta}_{L-2,k}\right), \ \forall\ k = 1, ..., K_{L-1} && \text{(Hidden layer } L-1\text{)} \\
&\vdots \\
z_{1,k} &= \phi_1\left(\mathbf{z}_0' \boldsymbol{\theta}_{0,k}\right), \ \forall\ k = 1, ..., K_1 && \text{(Hidden layer 1)} \\
\mathbf{z}_0 &= (1, \mathrm{VaR}_1, ..., \mathrm{VaR}_M, \mathrm{ES}_1, ..., \mathrm{ES}_M)'. && \text{(Input layer)}
\end{aligned}
$$

The number of weight parameters in each hidden layer l is $K_l(1+K_{l-1})$, plus another $3(1+K_L)$ weights for the output layer.

Model implementation

The network architecture, comprising the number of layers ("model depth") and neurons ("model width") as well as the activation function, is crucial for the predictive performance of neural networks. The more layers and neurons we add the more opportunities for new features and patterns to be learned (commonly referred to as the model's capacity). However, higher complexity comes with higher computational burden and greater risk of overfitting the data, resulting in poor generalization. For this reason, we focus on a simple network architecture with a single hidden layer, which is usually sufficient for many applications. As common in the neural network literature (see Goodfellow et al., 2016), we choose the number of neurons in accordance with the number of input units. All nodes are fully connected, so that each neuron receives an input from all units in the input layer.

We use the popular rectified linear unit (ReLU) as activation function at all nodes within the hidden layer. The ReLu is defined as $\phi(x) = \max(x, 0)$ and has some advantageous characteristics compared to other common activation functions such as the hyperbolic tangent or sigmoid functions. Specifically, the ReLU encourages sparsity in the number of active neurons, is computationally attractive and avoids vanishing gradient problems (cf. Gu, Kelly, and Xiu, 2020; Bianchi, Büchner, and Tamoni, 2021).

Each neural network specification is trained by minimizing the FZ loss function, in its simplest form given by:[7]

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{t} \sum_{\tau=0}^{t-1} L_{FZ}\left(r_{\tau+1}, \mathrm{VaR}_c\left(\mathbf{VaR}_{\tau+1|\tau}, \mathbf{ES}_{\tau+1|\tau}; \boldsymbol{\theta}\right), \mathrm{ES}_c\left(\mathbf{VaR}_{\tau+1|\tau}, \mathbf{ES}_{\tau+1|\tau}; \boldsymbol{\theta}\right)\right). \tag{3.17}$$

The estimates of the model parameters $\widehat{\boldsymbol{\theta}}$ are solutions of a non-convex optimization problem, making brute force optimization highly computationally intensive. Therefore, conventional estimation procedures of neural networks rely on stochastic optimization routines such as the stochastic gradient descent (SGD) algorithm. In contrast to the standard descent algorithm that utilizes the whole training sample to evaluate the loss function at each iteration of the optimization process, SGD evaluates the loss function based on a small randomly selected subset of the data and iteratively approaches the (local) minimum through back propagation. This stochastic approximation usually results in smoother and much faster convergence, with the cost of sacrificing some accuracy (see Goodfellow et al., 2016; Gu, Kelly, and Xiu, 2020; Bianchi, Büchner, and Tamoni, 2021). We implement two common extensions of the standard SGD. First, the "learning rate shrinkage" algorithm of Kingma and Ba (2015), called Adam, uses an adaptive learning rate[8], which shrinks the learning rate towards zero as the gradient approaches zero. Thus, it is more accurate than the standard SGD, while also providing faster convergence. Second, the RMSprop alogrithm developed by Hinton (2012) is also an adaptive learning rate method that divides the learning rate by an exponentially decaying average of squared gradients.

To reduce the risk of overfitting the data, we place constraints on the model's complexity using a variety of regularization techniques. Similar to the ridge combination model, we use an L2 penalty to add a cost to the size of the node weights, called weight decay in the context of neural networks. Regularizing the weights will force small signals (noise) to have weights nearly equal to zero and only allow consistently strong signals to have larger weights. In addition to the L2 penalization of the weight parameters, we use three other regularization techniques in our estimation procedure: dropout, early stopping and ensembling.

Dropout regularizes the choice of the number of nodes in a hidden layer (see Hinton and Salakhutdinov, 2006; Srivastava et al., 2014). This is achieved by randomly dropping out nodes in each layer at each iteration. That is, temporarily removing a fraction of neurons

[7] The optimization procedure is subject to a selected set of hyperparameters. See Section 3.2.5.

[8] The learning rate controls the step size of the descent.

from the network, along with all its incoming and outgoing connections with a certain probability.[9]

The "early stopping" procedure reduces the risk of overfitting by terminating the training algorithm when the loss on the validation sample has not improved for a pre-specified number of consecutive iterations. This approach is also useful to improve the computational efficiency as the training process may be stopped far before the maximum number of iterations is reached (see Goodfellow et al., 2016; Bianchi, Büchner, and Tamoni, 2021).

Finally, we use ensemble averaging when training our neural networks (see Hansen and Salamon, 1990; Dietterich, 2000). In particular, we train the same neural network model using different starting values (generated by multiple random seeds) and then construct a final prediction by averaging forecasts from all separately trained models. This approach reduces the prediction variance, as different starting values may produce different forecasts, given the stochastic nature of the optimization process. In addition, ensembling increases the overall robustness of the model since the impact of a local fit that is only sub-optimal is reduced (cf. Gu, Kelly, and Xiu, 2020).

All algorithms of the neural network combination model are implemented in R using the Keras interface to the TensorFlow library (see Chollet et al., 2015).

3.2.5. Hyperparameter tuning

Both the shrinkage and the neural network combination models rely on a choice of hyperparameters. These may have a crucial impact on a model's prediction accuracy as they define its structure and behavior.[10] While the shrinkage models restrict to a maximum of two hyperparameters—the balancing parameter for the elastic net (which we pre-select) and the regularization parameter—neural networks rely on a variety of hyperparameters by construction. These include, for example, the dropout rate, the regularization rate, the learning rate, the number of epochs or the batch size.

Generally, there is little theoretical guidance for how to tune hyperparameters, that is, finding the best set of hyperparameters for our data. Hence, we determine the hyperparameters adaptively from the data via cross-validation (CV), which is the standard approach in the machine learning literature. The general idea of cross-validation is to minimize the out-of-sample loss by evaluating the model on unseen data, i.e. data that were not used to train the model. Given the dependence in financial return data (and derived VaR and ES predicitions)

[9] Given that dropout and batch normalization, another regularization technique, are similar in spirit, we omit the usually weaker batch normalization.

[10] Parameters controlling the learning rate and a model's complexity are called hyperparameters (or, synonymously, "tuning parameters") to distinguish them from model parameters that are estimated on the training data set.

we cannot apply common cross-validation methods such as leave-ν-out or K-fold CV (see Arlot, Celisse, et al., 2010).[11] We therefore resort to the time-series cross-validation method of Hart (1994) that accounts for the dependence in the data.[12]

As common in cross-validation, the estimation sample is divided into two disjoint time periods (at each estimation step t) that maintain the temporal ordering of the data. The first, or "training", sub-sample is used to estimate (i.e. "train") the model parameters, e.g. the connection weights in a neural network, and is subject to a specific set of hyperparameter values. The second, or "validation", sub-sample is used for tuning the hyperparameters. In particular, we construct pseudo out-of-sample forecasts for data points in the validation sample based on the estimated model from the training sample.[13] We then calculate the average FZ loss associated to these forecasts. The "cross-validation loss" of the times-series CV method, denoted by CV_t, is therefore given by

$$\mathrm{CV}_t\left(\boldsymbol{\Omega}\right) = \frac{1}{t - n_{\min}} \sum_{\tau=n_{\min}}^{t-1} L_{FZ}\left(r_{\tau+1}, \mathrm{VaR}_{c,\tau+1|\tau}\left(\boldsymbol{\Omega}\right), \mathrm{ES}_{c,\tau+1|\tau}\left(\boldsymbol{\Omega}\right)\right), \tag{3.18}$$

where $\boldsymbol{\Omega}$ is a given vector of hyperparameter values. $\mathrm{VaR}_{c,\tau+1|\tau}$ and $\mathrm{ES}_{c,\tau+1|\tau}$ denote the combined VaR and ES prediction for $\tau + 1$ estimated from the training sample (i.e. based on information available up to τ), which follows an expanding window scheme with $n_{\min}$ as the minimum number of observations. Similarly, the validation sample is also expanding, which means that it gradually includes more recent observations but also retains the entire history in the validation sample.[14] The entire approach ensures that only past information is used to generate forecasts, thus entailing robustness with respect to autocorrelation in the data (see Hart and Lee, 2005).

From a pre-specified multi-dimensional hyperparameter space, we choose the subset of hyperparameter values that minimizes the CV loss. For the shrinkage combination models, which only require the selection of one tuning parameter, we create a simple grid of hyperparameter values and evaluate every position in the grid. However, this procedure, known as grid search, is computationally inefficient for methods with many hyperparameters and a huge search space. For the neural network combination models, we therefore resort

[11] Applying standard cross-validation methods to financial return data results in a violation of the fundamental assumption of cross-validation, which states that estimation and evaluation samples need to be independent (Arlot, Celisse, et al., 2010).

[12] The time-series cross-validation method of Hart (1994) is also known as temporal order approach.

[13] The predictions in the validation set cannot be considered as truly out-of-sample as they are employed to tune the hyperparameters, which are then used to estimate the final model (cf. Bianchi, Büchner, and Tamoni, 2021).

[14] We tested different specifications of the size of training and validation samples and set $n_{\min}$ to 1000 days in our application.

to the random search approach of Bergstra and Bengio (2012), which is a technique that randomly samples points from the hyperparameter grid to find the best solution for the built model. The authors show that random search is empirically and theoretically more efficient for hyperparameter optimization than grid search, given a large set of parameters to be tuned.

As the time-series cross-validation method of Hart (1994) is computationally very expensive (especially for more complex models such as neural networks) we slightly adjust the methodology and refrain from cross-validating at each estimation step. Instead, we periodically tune the set of hyperparameters every few years (depending on the model complexity) and use the optimized vector of hyperparameters for fitting and predicting over the subsequent year(s). In particular, we find a tuning horizon of one year to be a suitable choice for the shrinkage combination models, balancing performance and computational burden. By contrast, we tune the neural network models every four years, given the advanced model complexity. Unreported results show that the corresponding hyperparameters are stable over the full sample period for this tuning horizon. Additionally, we randomly tested other tuning horizons without any significant enhancements.

3.3. Research design

In the empirical application, we compare the predictions of the proposed machine learning based combination approaches with forecasts of a large range of competing combination approaches. This section outlines the corresponding research design, including the data, the models to be combined, the set of competing combination techniques and the forecast evaluation methodology. The empirical results are presented in Section 3.4.

3.3.1. Data description and estimation setup

We use daily price data of 12 major developed equity markets. In particular, we retrieve the daily closing, high and low prices for the following equity indices from Bloomberg: AEX (Netherlands), CAC 40 (France), DAX (Germany), FTSE 100 (UK), Hang Seng (Hongkong), IBEX 35 (Spain), KOSPI (Korea), Nikkei 225 (Japan), OMX 30 (Sweden), SMI (Switzerland), S&P 500 (US) and S&P/ASX 200 (Australia). The daily closing prices are used to compute daily log returns, whereas high and low prices are employed for deriving the realized range estimator (see Section 3.3.2). The sample spans the period from May 11, 1992 to May 7, 2021, giving rise to a total of 7706 trading days for each stock index.

Table 3.1 presents full-sample summary statistics on the daily return series of all stock indices. Average annualized returns range from 1.57% for the Nikkei 225 to 8.38% for the OMX 30, and annualized standard deviations range from 16.35% (S&P/ASX 200) to 24.41%

(Hang Seng). We observe mild negative skewness (around -0.30) for most of the equity indices. All return series exhibit substantial kurtosis (between 7.77 and 21.40).

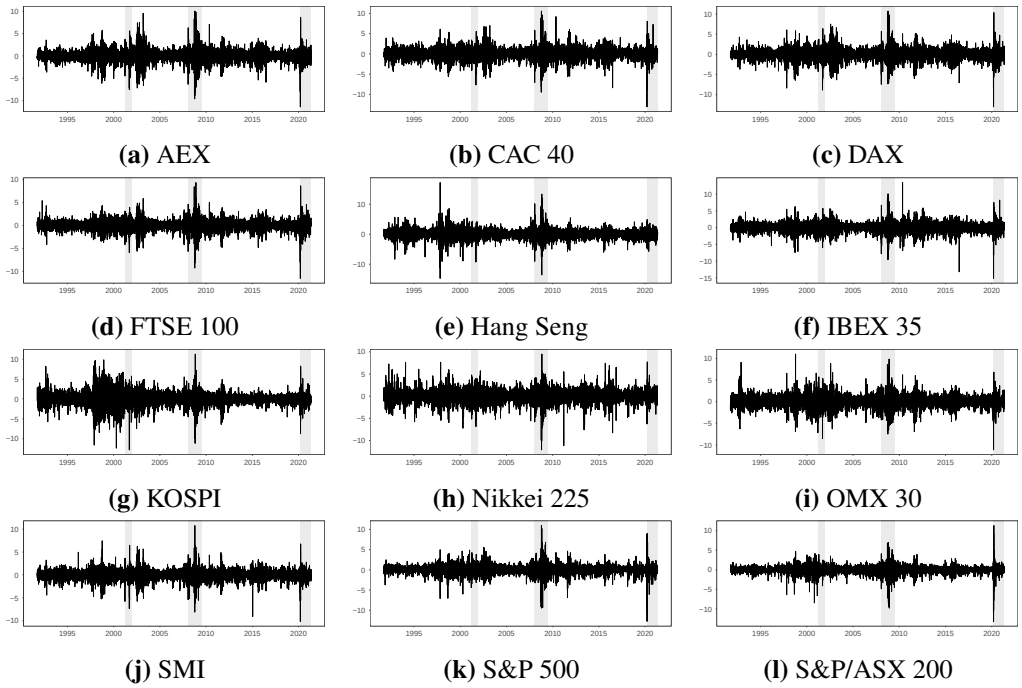
Table 3.1: Descriptive statistics of the daily return data

	Mean	Sd	Min	Max	Skewness	Kurtosis
AEX	5.38	20.85	-11.38	10.03	-0.26	10.44
CAC 40	3.77	21.72	-13.10	10.59	-0.20	9.02
DAX	7.27	22.26	-13.05	10.80	-0.24	8.93
FTSE 100	3.21	17.68	-11.51	9.38	-0.31	10.90
Hang Seng	5.44	24.41	-14.73	17.25	0.00	12.90
IBEX 35	3.99	22.28	-15.15	13.48	-0.31	10.62
KOSPI	5.59	25.45	-12.80	11.28	-0.18	9.46
Nikkei 225	1.57	22.71	-12.11	13.23	-0.24	8.90
OMX 30	8.38	22.39	-11.17	11.02	-0.00	7.77
SMI	5.83	17.81	-10.13	10.79	-0.30	10.30
S&P 500	7.75	18.18	-12.77	10.96	-0.43	14.98
S&P/ASX 200	5.86	16.35	-13.18	11.29	-1.00	21.40

This table reports the descriptive statistics of the daily log-returns spanning the period from May 11, 1992 to May 7, 2021. We report the annualized mean, annualized standard deviation (Sd), minimum (Min), maximum (Max), skewness and kurtosis in percentage points.

Figure 3.1 complements the summary statistics by depicting the daily log-return series of the 12 stock indices. All series show the typical stylized facts of financial return series, including periods of volatility clustering. Notably, we indicate periods of recession, as determined by the National Bureau of Economic Research (NBER). Our sample includes the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the still prevalent COVID-19 crisis (March 2020 to sample end). Strikingly, all equity indices did not experience their minimum return in the global financial crisis but in the recent COVID-19 crisis. From a tail risk modeling and management perspective it is therefore of particular interest to take a closer look at the commonalities and differences of the considered models in these two extreme recessions, see Section 3.4.5.

For predicting one-day-ahead VaR and ES we follow common convention and consider 1% and 5% probability levels (e.g. Kuester, Mittnik, and Paolella, 2006). Additionally, we include the 2.5% probability level because it is the new standard according to the third Basel Accord (Basel Committee on Banking Supervision, 2016). As we require data to estimate the individual models, to estimate the set of hyperparameters and to combine the forecasts, we split our data as follows. First, we use a rolling window of 1000 days, which we move forward by one day at a time, to re-estimate the parameters of the individual forecasting methods. Then, we are left with 6706 daily VaR and ES forecasts per stock index that serve as input to the estimation of the combination weights. Following Bayer (2018), who shows



(a) AEX **(b)** CAC 40 **(c)** DAX

(d) FTSE 100 **(e)** Hang Seng **(f)** IBEX 35

(g) KOSPI **(h)** Nikkei 225 **(i)** OMX 30

(j) SMI **(k)** S&P 500 **(l)** S&P/ASX 200

Figure 3.1: Daily return series over time. This figure shows the daily log-returns series for the 12 equity indices. The sample spans the period from May 11, 1992 to May 7, 2021. The grey areas indicate recessions as determined by the National Bureau of Economic Research: the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the still prevalent COVID-19 crisis (March 2020 to sample end).

that it is reasonable to use all available information[15], we implement a recursively expanding window approach with 1000 days as starting size for this combination exercise.

For all combination methods that require hyperparameter tuning, we adopt the time-series cross-validation approach using 200 iterations of grid search (shrinkage models) or random search (neural networks), with the full search grid documented in Table 3.A.1. As described in Section 3.2.5, we tune the shrinkage models once per year and the neural networks every four years. In every tuning step, we divide the in-sample data into a training and a validation sample, which both recursively increase by one (four) year(s) with every tuning step.

In the non-tuning periods, we refit the models using the last tuned hyperparameters. As machine learning models generally require a large sample size to estimate the model parameters due to their rich parametrization, we pool the individual VaR and ES forecasts from all stock indices in the machine learning combination models. We argue that the index-specific information is already captured by the individual methods. Unreported results

[15]Bayer (2018) evaluates the out-of-sample predictive performance of VaR combination forecasts using penalized quantile regressions for a recursively expanding window and rolling windows of the sizes 250, 500, 1000 and 1500. The author documents that it is reasonable to use all available information for the estimation of the combination weights and therefore recommends to use the recursively expanding window approach.

show that the pooled estimation is even more effective, given that more data and thus more information can be used.

For a fair comparison, we apply the same expanding window of data to all combination methods that do not require hyperparameter tuning. Thus, we obtain an out-of-sample evaluation period that ranges from June 3, 1999 to May 7, 2021, consisting of 5706 daily VaR and ES forecasts for each combination method and index series.

3.3.2. Description of the individual methods

Combining is most promising when the individual methods use different pieces of information or use information in different ways. Hence, we consider a large collection of different methods. The selected methods cover frequently used parametric, semi-parametric and non-parametric techniques and include methods capturing intraday volatility. We follow the implementation of the authors of the original paper (if possible). Moreover, we omit detailed descriptions of each estimator in the interest of space, and instead refer the interested reader to the original papers.

Historical simulation

The historical simulation (HS) approach predicts the next day's VaR by the empirical α-quantile, $Q_\alpha(\cdot)$, of the past w returns and the next day's ES by the average of the returns beyond the VaR:

$$\widehat{\mathrm{VaR}}_{t+1|t} = Q_\alpha\left(\{r_\tau\}_{\tau=t-w+1}^{t}\right), \tag{3.19}$$

$$\widehat{\mathrm{ES}}_{t+1|t} = \frac{1}{\alpha w}\sum_{\tau=t-w+1}^{t} r_\tau \, \mathbb{1}\{r_\tau \leq \mathrm{VaR}_{t+1|t}\}. \tag{3.20}$$

Weighted historical simulation

While the standard HS gives the same weights to all past returns, the weighted historical simulation (WHS) technique of Boudoukh, Richardson, and Whitelaw (1998) employs a geometrically declining weighting scheme, giving higher importance to more recent returns. Specifically, each of the most recent w returns is associated with a weight, which is computed as $\eta_i = \eta^{i-1}(1-\eta)/(1-\eta^w)$ for return i. We set $\eta = 0.99$. After ordering the returns in ascending order, we sum the corresponding weights until α is reached, starting from the

lowest return. The VaR is then the return corresponding to the last weight used in the previous sum:

$$\widehat{\text{VaR}}_{t+1|t} = \sum_{\tau=t-w+1}^{t} r_\tau \, \mathbb{1}\left\{ \sum_{i=1}^{w} \eta_i \mathbb{1}\{r_{t+1-i} \leq r_\tau\} = \alpha \right\}. \tag{3.21}$$

As for historical simulation, the ES is computed as the average of the returns beyond the VaR (cf. Equation 3.20).

Location-scale models

The probably most prominent and flexible approach to estimate and predict VaR and ES is the class of location-scale models. The underlying assumption of this class of models is that returns can be decomposed into $r_t = \mu_t + \sigma_t z_t$, where μ_t is the mean of the conditional distribution of r_t, σ_t is the volatility process and z_t is an independent and identically distributed innovation term with mean zero and unit variance. Assuming that returns are not predictable we set the conditional mean to zero. Then, corresponding VaR and ES forecasts can be computed as

$$\widehat{\text{VaR}}_{t+1|t} = \hat{\sigma}_{t+1|t} Q_\alpha(z_t), \tag{3.22}$$

$$\widehat{\text{ES}}_{t+1|t} = \hat{\sigma}_{t+1|t} \widetilde{\text{ES}}_\alpha(z_t), \tag{3.23}$$

where $\sigma_{t+1|t}$ is the one-step-ahead volatility forecast, $Q_\alpha(z_t)$ is the unconditional α-quantile of the innovations and $\widetilde{\text{ES}}_\alpha(z_t)$ is the unconditional α-ES of the innovations.

Given the vast amount of volatility models, we concentrate on the most prominent models, which we divide into three broad categories similar to Louzis, Xanthopoulos-Sisinis, and Refenes (2014). First, among the classical GARCH-type models we implement the standard GARCH(1,1) of Bollerslev (1987) and the GJR-GARCH(1,1) of Glosten, Jagannathan, and Runkle (1993), which additionally accounts for the leverage effect by responding asymmetrically with respect to positive and negative returns. As widely used among practitioners (due to its simple implementation), we also include the exponential smoothing RiskMetrics method (RiskMetrics Group, 1996). These interday models take the form

$$\text{GARCH}(1,1) \qquad \sigma_t^2 = \omega + \gamma \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{3.24}$$

$$\text{GJR-GARCH}(1,1) \qquad \sigma_t^2 = \omega + \gamma \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 + \delta \varepsilon_{t-1}^2 \mathbb{1}\{\varepsilon_{t-1} < 0\}, \tag{3.25}$$

$$\text{RiskMetrics} \qquad \sigma_t^2 = (1-\lambda) r_{t-1}^2 + \lambda \sigma_{t-1}^2 \quad \text{with } \lambda = 0.94, \tag{3.26}$$

where $\varepsilon_t = \sigma_t z_t$.

The second and third category of volatility models were developed to exploit the information content of high-frequency intraday data and thus require us to estimate intraday volatility. The most common estimator in the literature is the realized variance estimator, calculated as the sum of squared intraday returns (see Liu, Patton, and Sheppard, 2015). However, as intraday data can be expensive and given the ready availability of daily high and low prices, an alternative way of capturing intraday volatility is to use the realized range estimator (Parkinson, 1980; Alizadeh, Brandt, and Diebold, 2002; Brownlees and Gallo, 2010; Gerlach and Chen, 2015; Taylor, 2020), which takes the form $\text{RV}_t = 1/(4\ln(2))\left(p_{\text{high},t} - p_{\text{low},t}\right)^2$, where $p_{\text{high},t}$ is the largest log-price and $p_{\text{low},t}$ is the lowest log-price between open- and close-of-day t.

The second category of volatility models consists of the realized GARCH model of Hansen, Huang, and Shek (2012). This model retains a GARCH-type structure but instead of utilizing squared interdaily returns, it employs realized volatility measures, denoted by RV_t:

$$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \gamma \text{LRV}_{t-1}, \tag{3.27}$$

$$\text{LRV}_t = \psi + \varphi \ln(\sigma_t^2) + \pi(z_t) + u_t, \tag{3.28}$$

where $\text{LRV}_t = \ln(\text{RV}_t)$. The term $\pi(z_t) = \pi_1 z_t + \pi_2(z_t^2 - 1)$ captures the asymmetric impact of negative shocks on the volatility process and $u_t \sim \text{i.i.d.}(0, \sigma_u^2)$ is mutually independent with z_t. For further details on the realized GARCH model, see Hansen, Huang, and Shek (2012) or Louzis, Xanthopoulos-Sisinis, and Refenes (2014).

The third category of volatility models comprises heterogeneous autoregressive (HAR) realized volatility models (Corsi, 2009; Corsi and Renò, 2012; Louzis, Xanthopoulos-Sisinis, and Refenes, 2012), which account for the long-memory property of the realized volatility measures and fat tails. In particular, we implement the leverage HAR model of Corsi and Renò (2012), which extends the basic HAR structure by accounting for leverage effects:

$$\begin{aligned}\text{LRV}_t = c &+ \gamma_{(d)}\text{LRV}_{t-1}^{(d)} + \gamma_{(w)}\text{LRV}_{t-1}^{(w)} + \gamma_{(m)}\text{LRV}_{t-1}^{(m)} \\ &+ \varphi_{(d)}\upsilon_{t-1}^{(d)} + \varphi_{(w)}\upsilon_{t-1}^{(w)} + \varphi_{(m)}\upsilon_{t-1}^{(m)} + u_t,\end{aligned} \tag{3.29}$$

where $u_t \sim \text{i.i.d.}N(0, \sigma_u^2)$. Furthermore, $\text{LRV}_{t-1}^{(h)}$ and the leverage component $\upsilon_{t-1}^{(h)}$ are averaged over the common frequencies (h=d=1 for daily, h=d=5 for weekly, h=d=22 for monthly). That means, $\text{LRV}_{t-1}^{(h)} = 1/h \sum_{j=1}^{h} \text{LRV}_{t-j}$ and $\upsilon_{t-1}^{(h)} = \min\left(1/h \sum_{j=1}^{h} \text{r}_{t-j}, 0\right)$. As realized volatility and returns are not jointly modeled within HAR models, we implement the two-step approach of Giot and Laurent (2004). The first step is to calculate the conditional expectation of the realized volatility using the transformation $\text{RV}_{t|t-1} = \exp\left(\text{LRV}_t - \hat{u}_t + \frac{1}{2}\hat{\sigma}_u^2\right)$, where u_t refers to the estimated residual and the residual variance $\hat{\sigma}_u^2$ is modeled as a GARCH(1,1)

process (Louzis, Xanthopoulos-Sisinis, and Refenes, 2014). The final step then models the daily conditional variance as a linear function of the estimated conditional realized volatility forecast: $\sigma_t^2 = \omega_1 + \omega_2 \, \mathrm{RV}_{t|t-1}$.

For estimating the quantile and the ES of the innovation process z_t (denoted by $Q(z_t)$ and $\tilde{\mathrm{ES}}(z_t)$), we assume three alternative approaches: First, we implement fully parametric methods given by the normal (N) and the skewed t-distribution (SSTD). Second, we employ the semi-parametric filtered historical simulation (FHS) method of Barone-Adesi, Giannopoulos, and Vosper (1999), which applies the historical simulation approach, described in Section 3.3.2, to the standardized residuals z_t. Third, we use extreme value theory (EVT) and apply the peak-over-threshold method to the tail of the innovation distribution (cf. McNeil and Frey, 2000). For details on these tail models see e.g. Louzis, Xanthopoulos-Sisinis, and Refenes (2012) or Happersberger, Lohre, and Nolte (2020).

Combining the five variance processes with the four assumptions on the innovations yields a total of 20 location-scale models. Given the (from a machine learning perspective) small sample we have at hand, we do not use all combinations in our empirical study and pre-select a few models. In particular, we restrict to models that are successfully applied in other studies in the VaR and ES literature or that are popular among practitioners in the finance industry.

CAViaR-EVT

The conditional autoregressive VaR (CAViaR) class of models introduced by Engle and Manganelli (2004) directly models the conditional quantile rather than the whole return distribution by means of quantile regression. Although modeling VaR directly has some advantages—no explicit distributional assumption for the time series behavior of returns is needed—it does not consider on how model ES. This limitation is addressed by Manganelli and Engle (2004). The first step of their approach is to estimate a CAViaR model for a tail quantile that is not as extreme as the VaR of interest (as Manganelli and Engle we choose the 7.5% quantile). Specifically, we adopt the asymmetric slope CAViaR model due to its ability to accommodate the leverage effect:

$$q_t^\theta = \beta_0 + \beta_1 q_{t-1}^\theta + \beta_2 \max\left[r_{t-1}, 0\right] + \beta_3 \max\left[-r_{t-1}, 0\right], \tag{3.30}$$

where q_t^θ denotes the less extreme θ quantile.

In the second step, peak-over-threshold EVT is applied to the exceedances beyond the θ-quantile after standardizing the exceedances by the corresponding quantile estimates. Based on the fitted extreme value distribution we can then compute the VaR and ES forecasts:

$$\widehat{\text{VaR}}_{t+1|t} = \hat{q}^{\theta}_{t+1|t}\,(1 + z_t)\,, \tag{3.31}$$

$$\widehat{\text{ES}}_{t+1|t} = \hat{q}^{\theta}_{t+1|t}\left(\frac{1+\gamma-\xi u}{1-\xi}\right), \tag{3.32}$$

where z_t denotes the EVT estimate of the quantile of the standardized residuals, γ and ξ are EVT parameters and u is the extreme value threshold.

GAS

Patton, Ziegel, and Chen (2019) leverage "generalized autoregressive score" (GAS) models and the FZ loss function to estimate semi-parametric VaR and ES forecasts. Their models are semi-parametric in that they impose parametric structures for the dynamics of ES and VaR according to the GAS framework proposed by Creal, Koopman, and Lucas (2013) and Harvey (2013), but are completely agnostic about the conditional distribution of returns (aside from regularity conditions required for estimation and inference). We adopt the one-factor GAS model, in which both VaR and ES are driven only by a single variable, κ_t,

$$\widehat{\text{VaR}}_{t+1|t} = a\exp(\kappa_{t+1}), \tag{3.33}$$

$$\widehat{\text{ES}}_{t+1|t} = b\exp(\kappa_{t+1}), \tag{3.34}$$

where $b < a < 0$ and

$$\kappa_t = \beta\kappa_{t-1} + \gamma\frac{1}{b\exp\left(\text{ES}_{t-1|t-2}\right)}\left(\frac{1}{\alpha}\mathbb{1}\{r_{t-1} \leq a\exp\left(\text{VaR}_{t-1|t-2}\right)\}r_{t-1} - b\exp\left(\text{ES}_{t-1|t-2}\right)\right).$$

3.3.3. Competing combination approaches

This section outlines a range of competing forecast combination approaches. Most approaches are easy to implement, only the difference spacing and relative score method require a proper dynamic optimization procedure. All approaches induce convex weights, i.e. their weights are non-negative and sum to one.

Average

The most naive combination approach simply averages all individual methods' forecasts. The corresponding combination weights are given by

$$\widehat{\beta}^{\text{VaR}}_{m,t} = \widehat{\beta}^{\text{ES}}_{m,t} = \frac{1}{M}, \quad \forall\ m = 1, ..., M. \tag{3.35}$$

Trimmed average

Timmermann (2006) proposes the trimmed average combination approach, which uses the individual model's relative rankings to discard the models with the worst performance by setting the corresponding weights to zero. This method is found to be more robust than the simple average, since only the best performing models are included in the combination forecast (Bayer, 2018). The corresponding weights take the form

$$\widehat{\beta}^{\text{VaR}}_{m,t} = \widehat{\beta}^{\text{ES}}_{m,t} = \begin{cases} \dfrac{1}{\lfloor \eta M \rfloor}, & \text{if } R_{m,t} \leq \eta M \\ 0, & \text{else} \end{cases} \qquad \forall\ m = 1, ..., M, \tag{3.36}$$

where $R_{m,t}$ is the rank of model m at time t with respect to the average FZ loss up to time t, given by $\sum_{\tau=0}^{t-1} L_{FZ}\left(r_{\tau+1}, \text{VaR}_{m,\tau+1|\tau}, \text{ES}_{m,\tau+1|\tau}\right)$. We set $\eta = 0.25$, which means that we average over the forecasts of the four best models of the history up to time t.

Trimmed best-average

The trimmed best-average method according to Diebold and Shin (2019) follows the first step of the trimmed average method by trimming the models with the worst performance. In a second step, it computes the average weights of each combination of the remaining models and selects the combination with the best historical performance.

Inverse loss

Timmermann (2006) suggests to weight the forecasts inversely proportional to the historical losses of the individual models,

$$\widehat{\beta}^{\text{VaR}}_{m,t} = \widehat{\beta}^{\text{ES}}_{m,t} = \frac{(L_{m,t})^{-1}}{\sum_{n=1}^{M} (L_{n,t})^{-1}}, \quad \forall\ m = 1, ..., M. \tag{3.37}$$

Inverse rank

A more robust alternative to the inverse loss weighting scheme is the inverse rank approach[16], which weights the forecasts inversely proportional to their rank instead of the losses directly, since ranks are expected to be less sensitive to outliers than losses (see Timmermann, 2006; Bayer, 2018),

$$\widehat{\beta}^{\text{VaR}}_{m,t} = \widehat{\beta}^{\text{ES}}_{m,t} = \frac{(R_{m,t})^{-1}}{\sum_{n=1}^{M}(R_{n,t})^{-1}}, \quad \forall\ m = 1, ..., M. \tag{3.38}$$

Difference spacing

Taylor (2020) proposes a similar combination approach to the minimum loss method of Happersberger, Lohre, and Nolte (2020), described in Section 3.2.2. The only difference is that Taylor (2020) does not combine ES forecasts, but instead combines forecasts of the difference between ES and VaR. This allows us to clearly distinguish the VaR accuracy from the ES accuracy, given that the ES depends on the VaR as being the mean of the exceedances beyond the VaR. The corresponding combination weights can be obtained as follows,

$$\left(\widehat{\boldsymbol{\beta}}^{\text{VaR}}_t, \widehat{\boldsymbol{\beta}}^{\text{ES}}_t\right) = \underset{\beta^{VaR}_t, \beta^{ES}_t}{\arg\min} \frac{1}{t}\sum_{\tau=0}^{t-1} L_{FZ}\left(r_{\tau+1}, \text{VaR}_{c,\tau+1|\tau}\left(\boldsymbol{\beta}^{\text{VaR}}_t\right), \text{ES}_{c,\tau+1|\tau}\left(\boldsymbol{\beta}^{\text{VaR}}_t, \boldsymbol{\beta}^{\text{ES}}_t\right)\right), \tag{3.39}$$

where

$$\text{VaR}_{c,\tau+1|\tau}\left(\boldsymbol{\beta}^{\text{VaR}}_t\right) = \left(\widehat{\mathbf{VaR}}_{\tau+1|\tau}\right)' \boldsymbol{\beta}^{\text{VaR}}_t, \tag{3.40}$$

$$\text{ES}_{c,\tau+1|\tau}\left(\boldsymbol{\beta}^{\text{VaR}}_t, \boldsymbol{\beta}^{\text{ES}}_t\right) = \left(\widehat{\mathbf{VaR}}_{\tau+1|\tau}\right)' \boldsymbol{\beta}^{\text{VaR}}_t + \left(\widehat{\mathbf{ES}}_{\tau+1|\tau} - \widehat{\mathbf{VaR}}_{\tau+1|\tau}\right)' \boldsymbol{\beta}^{\text{ES}}_t. \tag{3.41}$$

Relative score

Following Bates and Granger (1969) and Shan and Yang (2009), Taylor (2020) suggests to combine VaR and ES forecasts by setting the combination weights to be inversely proportional to the (individual) models' relative historical performance, that is,

$$\widehat{\beta}^{\text{VaR}}_{m,t} = \widehat{\beta}^{\text{ES}}_{m,t} = \widehat{\beta}_{m,t} = \frac{\exp\left(-\lambda \sum_{\tau=0}^{t-1} L_{FZ}\left(r_t, \widehat{\text{VaR}}_{m,\tau+1|\tau}, \widehat{\text{ES}}_{m,\tau+1|\tau}\right)\right)}{\sum_{n=1}^{M}\exp\left(-\lambda \sum_{\tau=0}^{t-1} L_{FZ}\left(r_t, \widehat{\text{VaR}}_{n,\tau+1|\tau}, \widehat{\text{ES}}_{n,\tau+1|\tau}\right)\right)}, \quad \forall\ m = 1, ..., M, \tag{3.42}$$

[16]The inverse rank approach is also called triangular weighting scheme (see Timmermann, 2006).

where λ is a tuning parameter that controls how much the weights rely on the loss performance. High values of λ entail using the best performing individual models, whereas values of λ close to zero reduce the method to the simple average. We optimize λ following the hyperparameter tuning approach as described in Section 3.2.5.

Shrinkage-to-equal

Stock and Watson (2004) propose shrinkage towards the average of forecasts. Let $\tilde{\beta}^{\text{VaR}}_{m,t}$ and $\tilde{\beta}^{\text{ES}}_{m,t}$ be the minimum loss estimates of the weight on the m-th model. Then, the combination weights considered by the authors take the form

$$\widehat{\beta}^{\text{VaR}}_{m,t} = \psi\tilde{\beta}^{\text{VaR}}_{m,t} + (1-\psi)/M, \tag{3.43}$$

$$\widehat{\beta}^{\text{ES}}_{m,t} = \psi\tilde{\beta}^{\text{ES}}_{m,t} + (1-\psi)/M, \tag{3.44}$$

$$\psi = \max\left(0, 1 - \kappa M/(T-M-2)\right), \tag{3.45}$$

where κ regulates the strength of shrinkage. Similar to Stock and Watson (2004) we focus on $\kappa = 0.5$. If the in-sample size T rises relative to M, the minimum loss estimate gets a larger weight.

Single best

This combination method simply uses the best model from the previous period, that means

$$\widehat{\beta}^{\text{VaR}}_{m,t} = \widehat{\beta}^{\text{ES}}_{m,t} = \begin{cases} 1 & \text{if } R_{m,t} = 1 \\ 0, & \text{else} \end{cases} \quad \forall\ m = 1, ..., M. \tag{3.46}$$

3.3.4. Forecast evaluation

To assess the forecasting performance of the proposed ES and VaR combination approaches we adopt the backtesting framework of Happersberger, Lohre, and Nolte (2020). That means, we first employ various VaR and ES tests that are popular in the literature to gauge the forecasts' statistical accuracy.[17] In a second step, we investigate the forecasts' performance in a simple risk targeting strategy to check their relevance from a portfolio management perspective.

[17] See Bayer and Dimitriadis (2020) and Happersberger, Lohre, and Nolte (2020) for a summary of the applied VaR and ES tests and the corresponding original papers for a detailed description.

Backtesting VaR and ES forecasts with calibration tests

Traditionally, VaR and ES forecasts are evaluated using backtests that Nolde and Ziegel (2017) classify as unconditional and conditional calibration tests. The objective of such tests is to consider the ex ante risk forecasts from a specific model and compare them with the ex post realized returns.

We consider four common VaR calibration tests that are based on evaluating the distribution of VaR violations. That means, counting and analyzing those realized return observations that fall below the predicted VaR level for a given estimation period. (1) The test for unconditional coverage of Kupiec (1995) examines the frequency of violations, which shall be consistent with the quantile of loss that the VaR measure is intended to reflect. However, Kupiec's likelihood ratio test does not account for serial independence of the number of violations. (2) The conditional coverage test of Christoffersen (1998) offers a remedy by jointly testing the frequency as well as the independence of violations, assuming that VaR violations are modeled with a first-order Markov chain. This test could reject a VaR model that generates too many clustered violations. To account for clustering of extremes we further consider (3) the duration test of Christoffersen and Pelletier (2004), which examines the duration between violations by testing the null hypothesis that the duration between violations is exponentially distributed against a Weibull alternative. A more recent alternative is (4) the generalized residual test proposed by Patton, Ziegel, and Chen (2019), which allows to test both VaR and ES because it is derived from the FZ loss function. Here, standardized versions of the generalized residuals, given by $\mathbb{1}\{r_\tau \leq \widehat{\text{VaR}}_t\} - \alpha$ and $\frac{1}{\alpha}\mathbb{1}\{r_\tau \leq \widehat{\text{VaR}}_t\}\frac{r_t}{\widehat{\text{ES}}_t} - 1$, are simply regressed on elements of the information set available at the time the forecast was made. As these standardized generalized residuals are conditionally mean zero under the correct specification, forecast optimality can be assessed by testing that all parameters in these regressions are zero, against a two-sided alternative.

In addition to the generalized residual test we consider three other ES calibration tests. First, the ES regression test of Bayer and Dimitriadis (2020) adopts a similar approach to that of Patton, Ziegel, and Chen (2019), with the difference of only needing ES forecasts as input parameters. The idea of this Wald-type test is to regress the realized returns on the conditional ES forecasts. Intercept and slope parameters are then evaluated, required to be zero and one for correct ES forecasts. Second, the exceedance residual test of McNeil and Frey (2000) relies on the ES residuals that exceed VaR, given by $(r_t - \widehat{\text{ES}}_t)\mathbb{1}\{r_t \leq \widehat{\text{VaR}}_t\}$, which should have zero mean under the null hypothesis of a correctly specified risk model. A bootstrap hypothesis test then checks whether the expected value of the exceedance residuals is zero. Finally, we perform the conditional calibration test of Nolde and Ziegel (2017), which uses a Wald-type test statistic based on the moment functions of VaR and ES.

Relative evaluation of VaR and ES forecasts

As sophisticated VaR and ES forecasting methods often pass the majority of calibration tests, it is important to consider tests that allow for relative comparisons between the different methods. To this end, we evaluate the relative precision of the forecasts by comparing the FZ losses by means of the model confidence set (MCS) testing framework of Hansen, Lunde, and Nason (2011), similar to Bernardi and Catania (2016), Bayer (2018) or Taylor (2020). The MCS procedure consists of a sequence of equivalence tests (cf. Diebold and Mariano, 1995), which allows to construct a set of "superior" models. In particular, at each iteration we evaluate the null hypothesis of equal predictive ability (EPA), given by $\mathbb{E}[d_{ij}] = 0$ for all $i, j = 1, ..., M$, where d_{ij} is the loss differential between the forecasts of model i and model j. Whenever the hypothesis of EPA among all forecasts can be rejected, the worst performing model is discarded and the MCS algorithm starts again. The iterative procedure stops if the null hypothesis is accepted for each model in the set. Then, we are left with a set of models that statistically cannot be further distinguished at a pre-specified significance level (Bernardi and Catania, 2016; Bayer, 2018). We follow Hansen, Lunde, and Nason (2011) and implement the MCS at the 75% confidence level.

Backtesting VaR and ES forecasts with risk targeting strategies

Risk targeting, also known as constant risk, target risk, or inverse risk weighting, is a widely applied asset allocation strategy in the investment management industry if the aim is to protect an investment against extreme negative market moves. Taking advantage of the negative relationship between risk and return, this strategy controls portfolio risk over time by dynamically shifting between a risky and a risk-free asset (Hocquard, Ng, and Papageorgiou, 2013; Perchet et al., 2015; Bollerslev, Hood, et al., 2018b). The exposure to the risky asset is systematically adjusted conditional on its current risk (forecast) in order to maintain a predefined target risk level. Specifically, the exposure simply calculates as $\overline{\rho}/\rho_t(r_{t+1})$, where $\rho(\cdot)$ is a risk measure, typically VaR or ES, $\overline{\rho}$ is the predefined target risk level and $\rho_t(r_{t+1})$ is the level of ex-ante risk of the risky asset. As the latter is unknown at time t, we utilize a forecast based on the information available at time t. This makes the risk targeting strategy perfectly suited for evaluating the forecasting performance of VaR and ES methods from a portfolio management perspective.[18]

[18] See Happersberger, Lohre, and Nolte (2020) for further details on the risk targeting strategy.

3.4. Empirical analysis

This section presents the forecasting results of the empirical study outlined in the previous section, including the evaluation of the predictors' relative importance, statistical backtesting and a portfolio management application.[19]

3.4.1. Relative importance of the individual methods

In order to obtain some intuition on how the combination methods select their predictors and estimate the corresponding weights, Figure 3.1 shows importance scores of the individual methods for all forecast combination methods at the 1% probability level. For the shrinkage

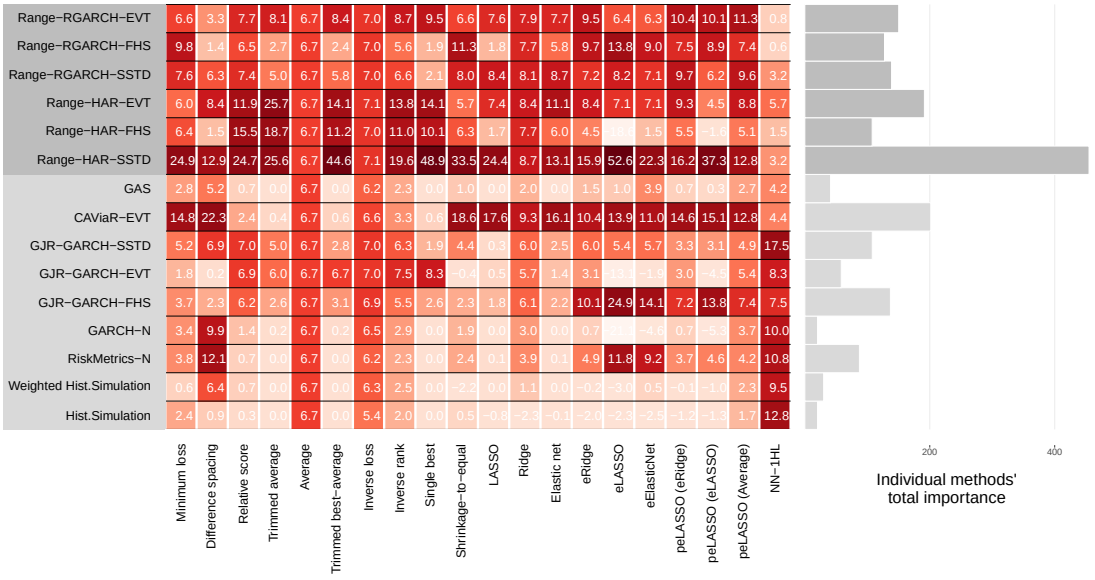


Figure 3.1: Individual methods' importance. This figure shows the importance of the individual methods for all combination approaches at the 1% probability level. The figures are the average ES combination weights for the shrinkage and competing models across time and all 12 equity indices over the period from June 3, 1999 to May 7, 2021. For the neural network combination models, the figures are the mean of the permutation feature importance scores calculated every four years. The darker the red tone, the higher the importance. The right part of the figure shows the total importance of each individual method; the bars represent the sum of scores across all combination approaches.

and competing combination methods we report the average ES combination weights across time and equity indices over the out-of-sample period covering 5706 daily observations. For the neural network combination models we compute permutation feature importance

[19] We provide details on both the individual and combination risk forecasts in the appendix. Table 3.A.2 and Table 3.A.3 report the summary statistics of and correlations between the different individual 5% ES forecasts for the S&P500. Figure 3.B.1, Figure 3.B.2, Figure 3.B.3 and Figure 3.B.4, Figure 3.B.5, Figure 3.B.6 show the VaR and ES forecasts over time for both the individual and combination methods at all three probability levels, illustrated for the S&P500.

scores[20] every four years and report the corresponding average. The right part of the chart summarizes the total importance of the individual methods by summing up the average scores across the combination methods. We observe a dominance of the realized range methods, which may be explained by their informational advantage: they process slightly more price information than the standard forecasting methods (daily high and low prices versus closing prices only). Among the realized range estimators, the Range-HAR-SSTD turns out to be the most important predictor. In contrast, the methods with the lowest importance scores are the GARCH-N, HS and GAS method, which even negatively contribute to VaR and ES forecasts for some combination approaches. These findings hold true for most combination approaches and across all probability levels (see Figure 3.B.7 in the appendix for the 2.5% and 5% probability level).

As the typical features of shrinkage models are not completely visible in Figure 3.1 (due to netting effects that occur because of averaging), we examine Figure 3.2, showing the temporal course of the 1% ES combination weights for the S&P 500 as an illustration.[21] When evaluating the estimated weights of the ridge for that particular stock (Figure 3.2a) we find very similar weights for the different individual models (except for the GAS, HS and WHS). This finding reveals the grouping effect of the ridge penalty: the coefficients of highly correlated variables are shrunk towards each other. In contrast, the LASSO model (Figure 3.2c) sets the weights of many individual methods to zero: from the common methods only the CAViaR-EVT and the GJR-GARCH model have significant positive contributions over time. The relevance of the CAViaR-EVT method is particularly interesting, as this method does not perform well individually in the statistical backtests (see Table 3.A.4). This finding shows that an individually weak performing method may still positively contribute to combination forecasting and thus stresses the advantage of data-driven model selection instead of manually deciding on the model components. The elastic net penalty equally combines the shrinkage feature from the ridge and the selection characteristic from the LASSO, as can be seen in Figure 3.2c. The egalitarian versions show similar features as their standard counterparts with the exception that they do not select and/or shrink towards zero but towards equal weights.

[20] Permutation feature importance is a model-agnostic ML interpretation method that measures a feature's (or predictor's) importance by calculating the increase in the prediction error after permuting the feature (Molnar, 2019). See Breiman (2001) and Fisher, Rudin, and Dominici (2019) for further details.

[21] The temporal course of the 2.5% and 5% ES combination weights of the shrinkage models for the S&P 500 are given in Figure 3.B.8 and Figure 3.B.9, respectively.

(a) Ridge **(b)** eRidge

(c) LASSO **(d)** eLASSO

(e) Elastic net **(f)** eElasticNet

(g) peLASSO (eRidge) **(h)** peLASSO (eLASSO)

(i) peLASSO (Average) **(j)** Minimum loss

Figure 3.2: Shrinkage combination weights over time. This figure shows the estimated combination weights for the shrinkage methods' 1% ES forecasts over the period from June 3, 1999 to May 7, 2021. Given that the shrinkage methods are estimated pooled over all equity indices, the presented combination weights are the same for all 12 equity indices.

In contrast to the shrinkage models, the neural network combination model favors the standard forecasting methods, with highest importance scores for the GJR-GARCH-SSTD method. Still, the importance scores of the NN-1HL model have to be taken with a pinch of salt given the weaknesses of the permutation feature importance method[22] and the fact that we only calculate these scores ever four years (due to computational restrictions).

Overall, we document that the estimated importance scores of the machine learning combination models differ across time, suggesting that a data-driven selection of the individual models may offer advantages compared to a simple average forecast.

3.4.2. VaR and ES calibration backtests

We start evaluating the VaR and ES combination forecasts' accuracy by means of the unconditional and conditional calibration tests as outlined in Section 3.3.4.[23] Since presenting the detailed results of the 8 VaR and ES backtests for all 12 equity indices and three probability levels is not feasible, we condense the results by presenting the average VaR violation rates, the average passing rates of the calibration tests and the corresponding average p-values at the three probability levels, see Table 3.1.[24] The VaR violation rate divides the number of VaR violations by the sample size, whereas a calibration test counts as passed if the p-value is greater than 0.10, indicating no evidence against optimality at the 10% significance level.

Our main findings are as follows: First, we find the VaR violation rates of almost all methods close to the respective expected probability level. The best model differs by probability level: the LASSO and relative score methods are the closest to the expectation at the 1% level (0.98%), the ridge at the 2.5% level (2.50%) and the peLASSO (Average), trimmed average and trimmed best average at the 5% level (4.99% 5.01%, 5.01%). Second and more importantly, we document high test passing rates of over 70% across most methods and probability levels. This finding applies to both the machine learning and the competing combination methods. Nonetheless, LASSO, Elastic Net and Shrinkage-to-equal stay slightly behind for the VaR tests at the 1% probability level (64%, 67%, 64%). In a similar vein, the average combination model is slightly off for the VaR tests at the 5% level (62%). Third, the corresponding average p-values of around 0.5 or higher indicate that most of the VaR and

[22] See Molnar (2019) for details.

[23] We report the results of the calibration backtests for the individual VaR and ES methods in Table 3.A.4 in the appendix.

[24] Note that we do include the generalized residual test of Patton, Ziegel, and Chen (2019) at the 1% probability level. Our findings indicate the test has low power at this probability level due to the low number of tail events.

Table 3.1: VaR and ES calibration backtesting

	1% probability level					2.5% probability level					5% probability level				
	Viol	VaR	$\overline{p}_{\text{VaR}}$	ES	$\overline{p}_{\text{ES}}$	Viol	VaR	$\overline{p}_{\text{VaR}}$	ES	$\overline{p}_{\text{ES}}$	Viol	VaR	$\overline{p}_{\text{VaR}}$	ES	$\overline{p}_{\text{ES}}$
ML combination methods															
Minimum loss	0.97	97	0.55	100	0.69	2.51	94	0.54	99	0.60	4.97	88	0.50	97	0.55
Ridge	0.92	86	0.51	88	0.55	2.50	94	0.54	100	0.58	4.84	69	0.41	94	0.53
LASSO	0.98	64	0.29	80	0.41	2.53	81	0.44	93	0.46	4.84	69	0.39	90	0.46
ElasticNet	0.94	67	0.39	87	0.51	2.52	85	0.55	94	0.55	4.81	77	0.42	93	0.50
eRidge	0.89	89	0.49	93	0.63	2.48	90	0.54	99	0.61	4.83	83	0.44	85	0.45
eLASSO	0.92	81	0.50	93	0.57	2.57	94	0.54	99	0.65	4.76	73	0.39	79	0.38
eElasticNet	0.95	94	0.51	92	0.60	2.55	92	0.55	99	0.60	4.79	77	0.40	81	0.40
peLASSO (eRidge)	0.92	94	0.54	100	0.68	2.43	92	0.55	99	0.60	4.83	83	0.44	93	0.48
peLASSO (eLASSO)	0.93	92	0.55	98	0.67	2.44	94	0.52	99	0.55	4.82	81	0.47	90	0.50
peLASSO (Average)	0.95	89	0.53	98	0.74	2.44	94	0.54	97	0.55	4.99	88	0.52	97	0.57
NN-1HL	0.90	81	0.54	88	0.65	2.41	88	0.44	89	0.56	4.73	77	0.41	90	0.47
Competing combination methods															
Average	0.87	92	0.43	98	0.73	2.31	85	0.40	96	0.56	4.76	62	0.29	83	0.41
Trimmed average	1.03	97	0.53	97	0.59	2.51	81	0.49	94	0.52	5.01	88	0.51	93	0.52
Trimmed best-average	1.04	94	0.47	95	0.57	2.54	88	0.47	94	0.51	5.01	83	0.44	93	0.49
Inverse loss	0.88	92	0.45	98	0.73	2.32	90	0.45	99	0.61	4.77	71	0.31	88	0.46
Inverse rank	0.93	92	0.53	95	0.70	2.39	92	0.53	100	0.68	4.87	92	0.51	99	0.60
Difference spacing	0.97	94	0.60	95	0.63	2.53	94	0.50	96	0.57	4.96	88	0.51	96	0.57
Relative score	0.98	92	0.58	95	0.64	2.49	94	0.50	97	0.63	4.92	92	0.53	99	0.60
Shrinkage-to-equal	1.11	64	0.35	92	0.52	2.68	79	0.37	89	0.42	5.14	85	0.40	92	0.40
Single best	1.03	100	0.50	95	0.56	2.56	90	0.46	92	0.51	5.04	90	0.44	90	0.47

This table reports the results of the calibration backtests for evaluating the VaR and ES predictions calculated over the out-of-sample period from June 3, 1999 to May 7, 2021. *Viol* is the average VaR violation rate over all equity indices as percentages. *VaR* and *ES* are the average percentage of VaR and ES tests passed and $\overline{p}_{\text{VaR}}$ and $\overline{p}_{\text{ES}}$ are the corresponding average p-values over all VaR/ES tests and equity indices.

ES tests are not passed scarcely but with sufficiently large buffer, so that a change in the test significance level would not affect the results substantially.

3.4.3. Relative comparison of the combination approaches

In the previous section we have seen that the majority of (combination) methods pass most of the calibration tests—a finding often documented in the corresponding literature (see Nolde and Ziegel, 2017). Hence, this type of evaluation hardly helps to decide for a particular forecasting method. Instead, they primarily help to get confidence about the forecast accuracy in general. It is therefore of great importance to apply evaluation methods that allow for direct forecast comparison in order to be able to differentiate between various methods.

Table 3.2 presents the results of the relative comparison between the different forecast combination approaches.[25] Again, it is not feasible to present the detailed results of all equity indices. Hence, we aggregate the results and present averages or sums calculated over all equity indices. First, we assess the forecasts' accuracy using average realized losses,

[25] We report the results of the relative comparison between all forecasting methods in Table 3.A.5.

Table 3.2: Relative comparison of the forecast combination approaches

	1% probability level					2.5% probability level					5% probability level				
	rank	best	DM	SSM	$\overline{p}_{\text{MCS}}$	rank	best	DM	SSM	$\overline{p}_{\text{MCS}}$	rank	best	DM	SSM	$\overline{p}_{\text{MCS}}$
ML combination methods															
Minimum loss	10.4	1	0.6	12	0.93	9.2	0	1.3	12	0.99	7.7	0	1.3	12	1.00
Ridge	7.3	1	2.1	12	0.98	8.2	1	2.6	12	1.00	11.2	0	0.6	12	0.99
LASSO	12.8	1	0.5	11	0.74	12.9	0	0.5	12	0.87	10.8	0	0.2	12	0.90
Elastic net	11.4	1	0.7	10	0.80	9.4	0	1.4	12	0.94	7.0	1	1.3	12	0.98
eRidge	7.9	1	1.6	12	1.00	7.8	0	2.3	12	1.00	3.5	4	3.0	12	1.00
eLASSO	9.5	0	0.7	12	1.00	16.2	0	0.2	11	0.69	10.0	0	0.6	12	0.90
eElasticNet	14.2	0	0.3	11	0.79	14.8	0	0.3	12	0.89	8.5	0	1.0	12	0.99
peLASSO (eRidge)	7.4	1	1.4	12	0.99	5.3	1	2.8	12	1.00	6.2	1	0.8	12	1.00
peLASSO (eLASSO)	9.2	1	0.8	12	1.00	7.5	0	1.6	12	1.00	9.8	0	0.6	12	1.00
peLASSO (Average)	10.2	0	0.8	12	0.95	3.2	4	2.9	12	1.00	6.7	1	1.1	12	1.00
NN-1HL	8.2	3	1.5	12	0.96	9.0	0	0.7	12	0.94	11.4	1	0.2	11	0.86
Competing combination methods															
Average	14.4	0	0.3	11	0.74	15.5	0	0.0	10	0.65	18.0	0	0.0	10	0.64
Trimmed average	11.3	0	0.3	12	0.92	14.2	0	0.4	10	0.78	15.5	0	0.2	12	0.81
Trimmed best-average	11.6	0	0.4	12	0.86	14.1	0	0.3	10	0.75	16.7	0	0.1	11	0.77
Inverse loss	12.4	0	1.1	12	0.87	13.0	0	0.9	10	0.74	16.3	0	1.0	10	0.77
Inverse rank	5.5	0	2.3	12	1.00	5.8	3	2.4	12	0.99	8.3	1	1.9	12	1.00
Difference spacing	11.0	0	0.8	12	0.97	10.0	0	0.8	12	0.97	7.6	1	1.3	12	1.00
Relative score	6.8	1	1.4	12	1.00	4.4	3	2.0	12	1.00	4.3	2	1.8	12	1.00
Shrinkage-to-equal	16.4	0	0.0	9	0.62	16.2	0	0.0	11	0.71	13.5	0	0.2	12	0.95
Single best	12.0	1	0.6	12	0.90	13.1	0	0.2	10	0.76	17.0	0	0.1	10	0.64

This table reports the results of the relative comparison between the different forecast combination approaches over all 12 equity indices. *rank* is the average rank based on the average FZ loss and *best* is the number of times a method is the best method. *DM* is the average percentage of how often a specific method significantly outperforms another using pairwise modified Diebold-Mariano tests based on the FZ loss function. Averages are calculated over all equity indices. *SSM* is the number of indices a method is included in the superior set of models at the 75% level and $\overline{p}_{\text{MCS}}$ is the average over the 12 individual MCS p-values based on the Tmax statistics using 1000 iterations of the moving block bootstrap. The sample spans the period from June 3, 1999 to May 7, 2021.

similar to Taylor (2020) and Dimitriadis and Halbleib (2021). For each equity index, we rank the methods according to their average FZ loss and report the average rank across equity indices as well as the number of times a method is the best method. At the 1% probability level, the inverse rank method has the highest average rank and the NN-1HL model is the best model (i.e. with rank one) for three indices. Strong methods are also the ridge, eRidge and peLASSO (eRidge) from the shrinkage combination methods as well as the relative score approach from the competing methods. At the 2.5% probability level, the peLASSO (Average) is the best method according to these performance methods, with similar strong candidates as before: eRidge, peLASSO (eRidge), peLASSO (eLASSO), inverse rank and relative score. At the 5% probability level, the best method is again one of the shrinkage models. The eRidge method clearly dominates the other combination methods. Notably, for 8 of the 12 indices, one of the ML models is the best method.

Following Patton, Ziegel, and Chen (2019), we perform pairwise (modified) Diebold-Mariano (DM) tests based on the FZ loss function as a second performance measure.

Table 3.2 reports the average number of indices for which a specific method significantly outperforms another method (based on the DM tests). For instance, the eRidge significantly outperforms three methods (out of 19) on average at the 5% probability level, which is the highest number at this level. Other strong methods are those candidates that are already mentioned: ridge, eRidge, peLASSO (Average), peLASSO (eRidge), inverse rank.

Finally, we investigate the results of the model confidence set (MCS). Specifically, Table 3.2 reports the number of indices for which a method is included in the superior set of models (SSM) at the 75% confidence level and the corresponding average MCS p-values. Given the 12 equity indices we consider in our study, the best possible value for each probability level is 12. Similar to Bernardi and Catania (2016) and Bayer (2018), we can only eliminate a small number of models using the MCS testing framework. Nevertheless, we can document a slight superior performance of the ML combination methods. While the ML methods are included in the SSM for almost all of the 12 equity indices, the competing models show only similar results at the 1% probability level. This finding is corroborated by the average MCS p-values. These are generally higher for the ML combination methods, thus indicating a higher probability of these methods to be included in the SSM.

The relative comparisons between the different combination approaches reveal some further relevant insights, which we summarize in no particular order in the following:

(i) For each probability level, most shrinkage combination models outperform the unpenalized minimum loss method, which we consider as the benchmark for the shrinkage models.

(ii) The neural network combination model performs better than the majority of competing models. Also, it is among the best ML models at the 1% and 2.5% probability level, but is slightly trailing behind with respect to the shrinkage models at the 5% probability.

(iii) Non-parametric trimming methods show a poorer performance than data-driven selection via LASSO.

(iv) There is no clear winner combination method, but rather a small selection of models that are convincing. From the shrinkage models, the peLASSO models and the eRidge consistently stand out across equity indices and probability levels. Likewise, the inverse rank and relative score models from the competing combination methods are also performing consistently well.

Summing up the main results from the relative comparison, we find that the shrinkage combination models consistently exhibit a good relative accuracy in addition to convincing calibration backtest results. Also, neural network combination models show a good

performance with a few exemptions. Still, the excellent performance of the shrinkage models questions whether the additional complexity of the neural network models needs to be taken.

3.4.4. Forecasting performance in calm and recessionary periods

In this section, we investigate the forecasting performance during different market phases. Calm and recession periods are determined according to the National Bureau of Economic Research (NBER). Thus, recession periods comprise the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the COVID-19 crisis (March 2020 to sample end). Calm periods cover the remaining periods of the sample.

Table 3.3 shows the results of the calibration backtests as well as the relative comparison between the various forecasting methods for the calm and recession periods, summarized over all indices and probability levels. Similar to other studies (e.g. Halbleib and Pohlmeier, 2012), we find more VaR violations in recession than in calm periods across all forecasting methods. Together with the fact that the violations in recession periods predominately appear in clusters, this finding induces lower passing ratios of the VaR tests in recession periods. Strikingly, ES tests are not affected substantially: they exhibit higher ES passing ratios in recession periods. Evaluating the MCS reveals that all combination methods are (almost) always included in the 75% SSM for both calm and recession periods. Except for the average and the single best combination model, all methods are included the maximum of 36 times in the SSM. Hence, there are no significant differences between ML and the competing combination models. Among the individual methods, only Range-HAR models are always included in the 75% SSM. With some exceptions also Range-RGARCH and GJR-GARCH models show a good performance. Comparing across all methods, we document that the peLASSO (Average), Range-HAR-SSTD and eRidge are the best methods in calm periods, whereas the NN-1HL is the best model in recession periods.

3.4.5. Tail risk forecasting in the COVID-19 period

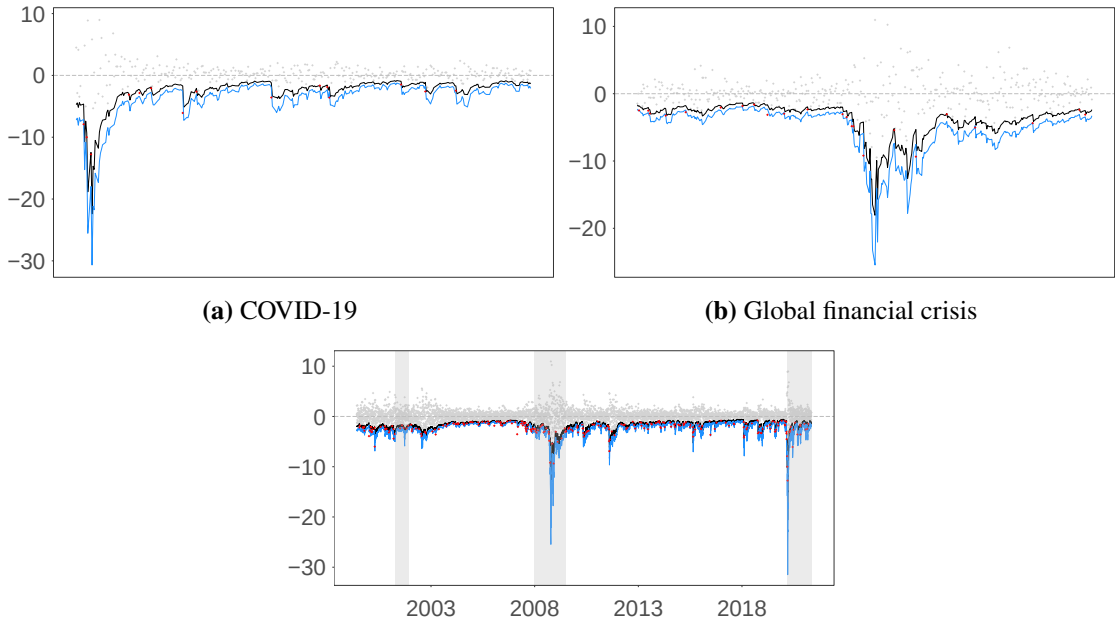
In this section, we take a closer look at how the VaR and ES forecasts behave in the still prevalent COVID-19 crisis period (March 1, 2020 to May 7, 2021). Figure 3.3a exhibits the corresponding temporal course of the 5% VaR and ES forecasts of the peLASSO (Ridge) combination method for the S&P 500 as an illustration. It is striking that the risk forecasting of the sharp slump at the beginning of the COVID-19 period worked reasonably well, although the drop came abruptly without a longer phase of announcement. This is evident from the fact that we do not observe a cluster of VaR violations around this tail event. Still, this abrupt drop results in the lowest VaR and ES figures over the full sample period (see Figure 3.3c), even lower than the minimum VaR and ES in the global financial crisis of

Table 3.3: Forecasting performance in calm and recessionary periods

	Calm periods						Recession periods					
	rel.Viol	VaR	ES	rank	best	SSM	rel.Viol	VaR	ES	rank	best	SSM
ML combination methods												
Minimum loss	1.0	91	88	11.1	1	36	1.2	70	90	11.4	2	36
Ridge	0.9	80	89	9.5	3	36	1.2	75	95	13.4	1	36
LASSO	0.9	70	81	14.8	2	36	1.4	71	92	15.8	2	36
Elastic net	0.9	74	85	11.3	2	36	1.3	68	92	13.9	1	36
eRidge	0.9	80	82	6.6	6	36	1.2	75	91	12.7	0	36
eLASSO	0.9	76	78	15.1	0	36	1.2	81	93	14.0	3	36
eElasticNet	0.9	84	81	15.1	0	36	1.2	83	93	14.4	1	36
peLASSO (eRidge)	0.9	84	84	6.9	1	36	1.2	75	92	13.5	0	36
peLASSO (eLASSO)	0.9	84	83	10.6	1	36	1.2	72	90	11.6	0	36
peLASSO (Average)	0.9	86	86	7.6	7	36	1.2	66	93	12.6	0	36
NN-1HL	0.9	78	73	14.5	1	36	1.1	75	91	7.2	6	36
Competing combination methods												
Average	0.9	67	79	19.7	0	35	1.2	79	94	17.9	0	36
Trimmed average	1.0	84	87	16.2	0	36	1.2	72	87	15.4	0	36
Trimmed best-average	1.0	84	85	16.9	0	36	1.2	66	85	15.0	1	36
Inverse loss	0.9	69	84	16.9	0	36	1.2	78	94	15.9	1	36
Inverse rank	0.9	87	90	7.5	0	36	1.1	72	92	9.6	1	36
Difference spacing	1.0	89	85	11.6	1	36	1.2	68	91	11.6	1	36
Relative score	1.0	88	90	6.1	3	36	1.2	66	91	10.4	1	36
Shrinkage-to-equal	1.0	83	75	18.5	0	36	1.3	66	90	16.4	1	36
Single best	1.0	86	83	18.1	0	36	1.2	68	88	13.6	1	36
Individual methods												
HS	0.8	9	18	34.9	0	0	3.1	2	8	35.0	0	7
WHS	1.0	37	25	33.0	0	0	1.1	45	74	33.8	0	5
RiskMetrics-N	1.6	2	1	33.6	0	0	1.6	39	21	31.9	0	26
GARCH-N	1.3	36	3	30.6	0	2	1.9	33	16	31.6	0	13
GJR-GARCH-SSTD	1.0	81	80	20.4	0	34	1.4	49	79	26.5	0	33
GJR-GARCH-FHS	1.0	78	77	22.9	0	33	1.4	55	81	25.2	0	34
GJR-GARCH-EVT	0.9	74	69	23.7	0	32	1.3	60	83	24.2	0	34
CAViaR-EVT	1.0	72	24	30.2	0	13	1.5	50	64	29.7	1	32
GAS	1.1	58	56	32.2	0	0	1.9	25	64	31.5	0	17
Range-RGARCH-SSTD	1.0	61	74	26.9	0	31	1.2	71	90	16.5	3	35
Range-RGARCH-FHS	1.0	62	72	25.8	0	34	1.1	74	89	16.1	3	36
Range-RGARCH-EVT	1.0	62	73	24.8	0	34	1.1	75	91	15.3	2	36
Range-HAR-SSTD	1.0	91	88	10.2	7	36	1.3	67	89	16.6	2	36
Range-HAR-FHS	1.0	89	95	13.8	0	36	1.2	73	90	15.8	0	36
Range-HAR-EVT	0.9	91	91	12.5	1	36	1.2	72	91	14.1	2	36

This table reports the results of the calibration backtests as well as the relative comparison between the various forecasting methods for calm and recession periods. Recession periods comprise the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the COVID-19 crisis (March 2020 to sample end). Calm periods cover the remaining periods of the sample. *rel.Viol* is the average relative VaR violation rate calculated as the number of realized divided by number of expected VaR violations. *VaR* and *ES* are the average percentage of VaR and ES tests passed. *rank* is the average rank based on the average FZ loss and *best* is the number of times a method is the best method. *SSM* is the number of indices a method is included in the superior set of models at the 75% level. Averages are calculated over the 12 equity indices and the three probability levels. For *best* and *SSM*, the maximum is 36.

2007-2008 (see Figure 3.3b). For this particular index and method, the minimum 5% ES is -31.5%, compared to -25.5% in the global financial crisis. In the phase after the initial shock we see that VaR and ES figures remain very volatile. Subsequently, we document further predicted tail events, which are, however, less severe that the initial one. These probably



(a) COVID-19

(b) Global financial crisis

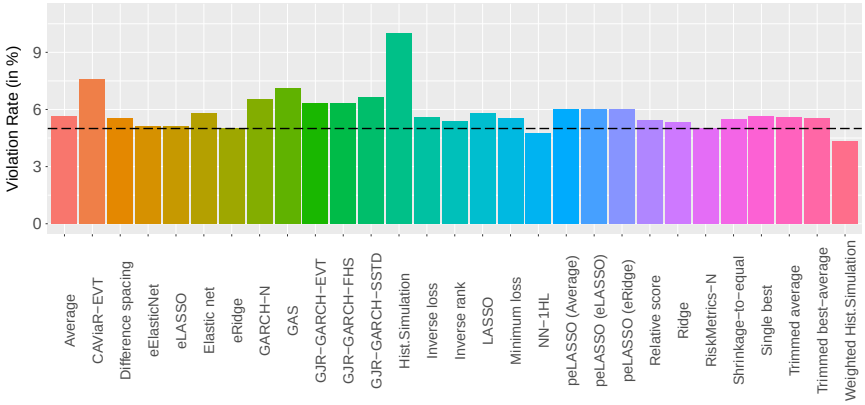
(c) Full sample

Figure 3.3: Tail risk forecasting in the COVID-19 period. This figure shows the daily 5% VaR forecasts (in black) and associated ES forecasts (in blue) of the peLASSO (eRidge) combination method as well as the realized returns of the S&P 500 (light-grey dots) over the COVID-19 period (March 1, 2020 to May 7, 2021). VaR violations are marked in red. At a probability level of 5%, a total of 15 violations are expected over the COVID-19 period. We document 17 realized violations when using the peLASSO (eRidge) method. For the purpose of comparison we also show the corresponding forecasts for the global financial crisis and the full sample period.

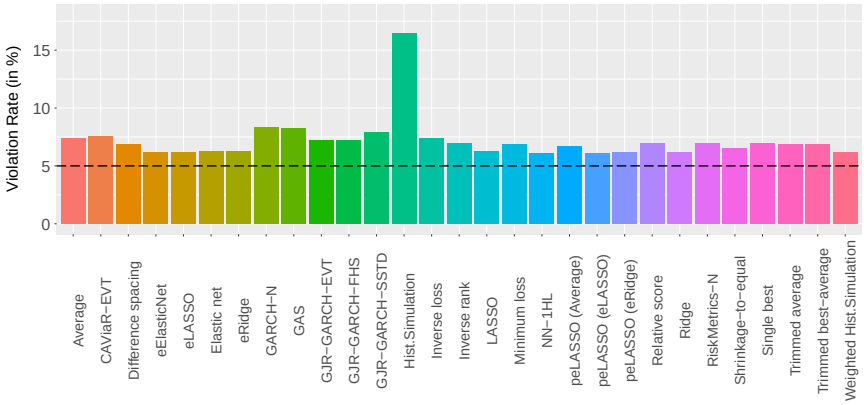
show the uncertainty in the stock market due to restrictions to the daily life (and thus also the economy) intended to stop the spreading of the virus.

In the evaluation of the tail risk forecasts during the COVID-19 period we cannot resort to the statistical backtesting framework as applied in Section 3.4.2 and 3.4.3. Given the small number of observations in this period (309 return observations) and correspondingly little tail events—3 at the 1% and 15 at the 5% probability level, respectively—VaR and ES backtests have too little power and are thus not reliable. Still, we can examine the VaR violation rates, which give an indication on the forecast accuracy of the different forecasting methods. Figure 3.4 shows the average VaR violation rates (over all equity indices) in the COVID-19 period as well as in the global financial crisis and the full sample period for comparative purposes. While the VaR violation rates in the full sample period are mostly in line with expectation (as shown in Section 3.4.2), they are elevated in the two crisis periods. Interestingly, the violation rates in the COVID-19 period are higher mainly for the individual methods, whereas this exceedance holds for all methods in the global financial crisis. A potential explanation for this finding is that the combination methods have learned from the

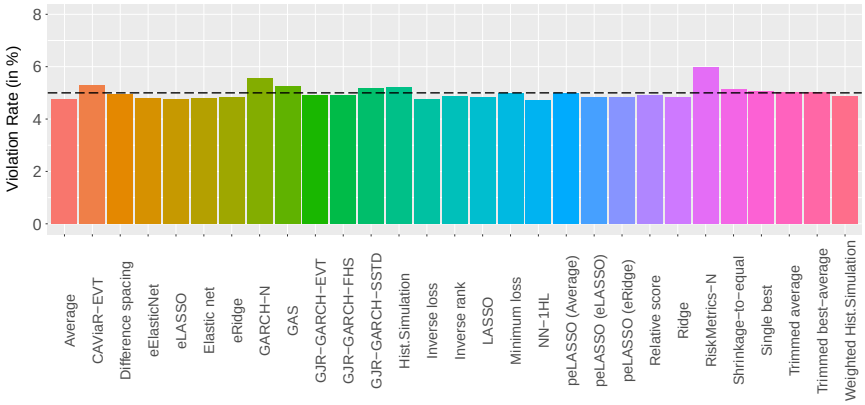
global financial crisis through the expanding window estimation and thus use this historical information when estimating the forecasts in the COVID-19 period.



(a) COVID-19



(b) Global financial crisis



(c) Full sample

Figure 3.4: VaR violations in the COVID-19 period. This figure shows the average VaR violation rates for the 5% VaR forecasts across all 12 equity indices for the COVID-19 period, the global financial crises and the full sample period. The black dashed line indicate the theoretically expected VaR violation rate.

3.4.6. Risk models in action: Quantifying the benefits from combination forecasts

From a practitioner's point of view it is a relevant question whether the superior statistical accuracy of the sophisticated combination forecasts also translates into a superior portfolio performance. To analyze this question we apply the different ES forecasts to a risk targeting strategy, outlined in Section 3.3.4. We start with analyzing the historical backtest of such a risk targeting strategy. That means, we assess how the risk targeting strategy would have performed when implemented over the whole out-of-sample period.

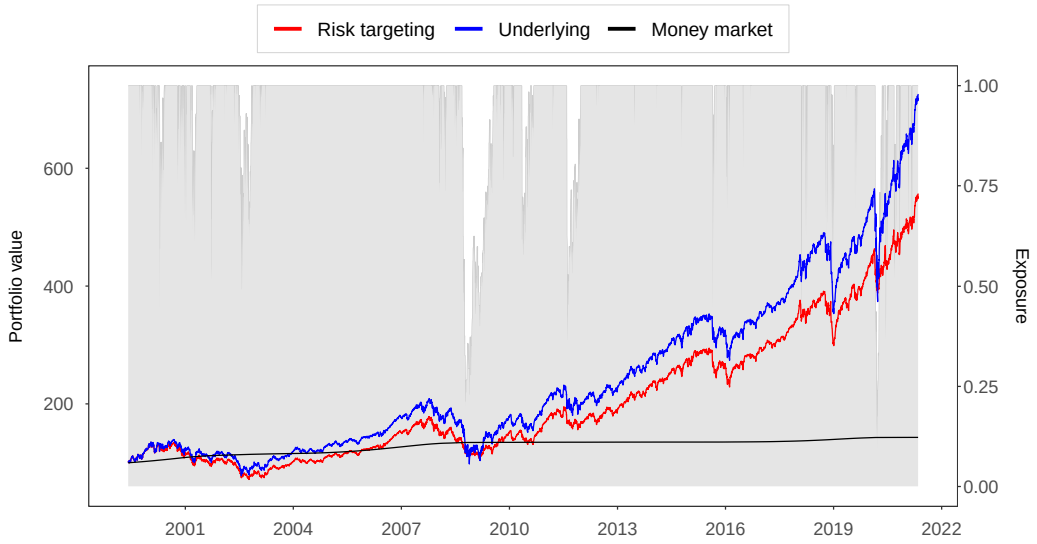


Figure 3.5: Evaluating ES combination forecasts in risk targeting strategies. This chart illustrates the performance of the ES targeting strategy with the S&P 500 as equity underlying. Specifically, we show the historical path of the protected portfolio (red line) over the sample period 1999–2021. Exposure is calculated based on the 5% ES of the peLASSO (Average) combination model. The target level is a 3% ES. For comparison, we include the performance of the underlying equity index (blue line) and a money market investment (black line; based on the US 3-month treasury bill).

Figure 3.5 illustrates the historical performance of the risk targeting strategy based on the peLASSO (eRidge) method over the out-of-sample period from 1999 to 2021.[26] The underlying in this example is the S&P 500 and we target an ES level of 3%, which is a reasonable assumption for tail risk-averse investors.[27] We observe a sharp decrease in exposure of the ES targeting strategy during the financial market crisis in 2008 and during the recent COVID-19 pandemic in early 2020, thus avoiding huge drawdowns of the underlying.

[26] Similar to Happersberger, Lohre, and Nolte (2020) and Dichtl and Drobetz (2011) we implement the risk targeting strategy without short sales or leverage and assume round-trip transaction costs of 10 basis points. To avoid portfolio shifts triggered by rather small market movements, we also apply a trading filter of 2%, which means that we act only on exposure changes in excess of 2%.

[27] Unreported results show that the results of the risk targeting strategies are robust to the ES target level and the probability level of the ES forecasts.

However, this comes with the cost of reducing upside participation at the end of the sample period.

Table 3.4: Combination forecasts in risk targeting strategies

Method	Return	SD	ES	MDD	Sharpe	Calmar	Sortino	Ret/ES	Part	TO
Risk targeting based on ML combination methods										
Minimum loss	5.01	16.65	-2.46	-56.72	0.21	0.10	0.43	2.08	92.59	1.72
Ridge	5.05	16.78	-2.48	-56.98	0.21	0.10	0.42	2.08	92.91	1.39
LASSO	5.01	16.85	-2.49	-57.70	0.21	0.10	0.42	2.05	93.14	1.54
Elastic net	5.03	16.80	-2.48	-57.36	0.21	0.10	0.42	2.07	93.01	1.50
eRidge	4.97	16.48	-2.43	-55.93	0.21	0.10	0.43	2.09	92.15	1.55
eLASSO	5.03	16.53	-2.44	-55.71	0.21	0.10	0.43	2.11	92.21	1.45
eElasticNet	5.01	16.52	-2.44	-55.78	0.21	0.10	0.43	2.10	92.20	1.46
peLASSO (eRidge)	5.04	16.55	-2.44	-56.06	0.21	0.10	0.43	2.11	92.34	1.50
peLASSO (eLASSO)	5.06	16.57	-2.44	-56.06	0.21	0.10	0.43	2.11	92.36	1.45
peLASSO (Average)	5.11	16.69	-2.46	-56.35	0.21	0.10	0.43	2.12	92.64	1.31
NN-1HL	4.96	16.66	2.46	-56.56	0.20	0.10	0.42	2.05	92.64	1.57
Risk targeting based on competing combination methods										
Average	5.11	16.75	-2.47	-56.15	0.21	0.10	0.43	2.11	92.69	1.21
Trimmed average	4.95	16.69	-2.46	-57.22	0.21	0.09	0.42	2.05	92.61	2.05
Trimmed best-average	4.94	16.69	-2.46	-57.27	0.21	0.09	0.42	2.05	92.61	2.04
Inverse loss	5.11	16.74	-2.47	-56.14	0.21	0.10	0.43	2.11	92.68	1.24
Inverse rank	5.02	16.69	-2.46	-56.64	0.21	0.10	0.43	2.08	92.66	1.63
Difference spacing	5.00	16.68	-2.46	-56.73	0.21	0.10	0.42	2.07	92.68	1.61
Relative score	4.99	16.66	-2.45	-56.77	0.21	0.10	0.42	2.07	92.62	1.77
Shrinkage-to-equal	4.97	16.61	-2.45	-56.98	0.21	0.10	0.42	2.07	92.49	1.96
Single best	4.93	16.70	-2.46	-57.10	0.20	0.10	0.42	2.05	92.63	2.08
Benchmarks investments										
Equity underlying	5.70	21.37	-3.26	-66.94	0.19	0.09	0.37	1.76	100	0
Money market	1.63	0.14	-0.00	-0.00	0.00	–	–	–	0	0

This table reports the backtesting results of the risk targeting strategy based on the various 5% ES combination forecasts over the out-of-sample period from June 3, 1999 to May 7, 2021. We target an ES of 3% over the whole out-of-sample period and report the average of the following performance measures over the 12 equity indices: the annualized mean return (Return), annualized standard deviation (SD), 5% ES, maximum drawdown (MDD), Sharpe ratio (SR), Calmar ratio, Sortino ratio, return-to-ES ratio, participation in the risky equity index (Part) and turnover (TO). Return, Sd, ES, MDD, Part and TO are given as percentages. For comparison, we include the average performance of the underlying equity indices and the performance of a money market investment (based on the US 3-month treasury bill).

Table 3.4 shows the corresponding estimation results. As before, we average all performance measures over all equity indices. As the objective of a risk targeting strategy is twofold—providing downside protection while still enjoying the upside potential of the risky underlying—the performance should be evaluated accordingly. Alongside standard measures like the Sharpe ratio and maximum drawdown, we therefore employ specific downside risk measures such as Calmar and Sortino ratios as well as the ratio of annualized return to absolute ES. Similar to Happersberger, Lohre, and Nolte (2020), we find that risk-targeting strategies based on all risk methods outperform the equity underlying. This is reflected in higher risk-adjusted returns (measured by the Sharpe ratio), higher Calmar, Sortino and return-to-ES ratios as well as lower maximum drawdowns and higher ES figures.

These results confirm the ability of the risk targeting strategy to reduce downside risk. This, however, comes with an insurance fee of between 59 and 92 basis points, depending on the used risk methods. Comparing across risk models, we observe the best performance for the egalitarian shrinkage models. In particular, the peLASSO (Average) method exhibits the highest average return-to-ES ratio (2.12). Other well performing methods are the average and the inverse loss methods (with average return-to-ES ratios of 2.11). Lowest risk figures are also found for the egalitarian shrinkage models (standard deviation of around 16.55% and 5% ES of around -2.44%). Given the superiority of the egalitarian models over the standard shrinkage models and the strong performance of the average combination method, we conclude that models related to averaging are particularly well-suited for their use in risk targeting strategies. A further observation is that the machine learning methods exhibit lower average turnovers figures than the competing combination approaches (except for average and inverse loss), which makes them appealing for their use in asset allocation strategies.

Given that analyzing the historical performance may suffer from path dependency, we additionally conduct a historical block-bootstrap analysis as robustness check.[28] The corresponding results can be found in Table 3.A.6. In a nutshell, the results of the block-bootstrap analysis corroborate our findings from the historical backtest. Tail risk models that help investors to achieve more accurate VaR and ES forecasts are associated with a (slightly) superior portfolio performance.

3.5. Conclusion

In this paper, we propose the combination of VaR and ES forecasts with machine learning techniques. In particular, we assess whether shrinkage and neural network combination models improve the predictive accuracy relative to a large set of competing combination approaches. The primary advantage of shrinkage models is that they are able to reduce overfitting caused by high multicollinearity of the individual predictors. Through their shrinkage and variable selection properties, these methods also stabilize the estimates of the combination weights and thereby improve VaR and ES forecasts. Neural networks are probably the most flexible machine learning models and thus well-suited to the problem of

[28] Following Annaert, Van Osselaer, and Verstraete (2009), Bertrand and Prigent (2011), Dichtl and Drobetz (2011), Dichtl, Drobetz, and Wambach (2017) and Happersberger, Lohre, and Nolte (2020), we draw blocks of 250 subsequent daily portfolio and risk-free returns on a rolling window basis and implement the risk targeting strategies in each draw. We thus obtain 5607 overlapping yearly returns as a basis for the comparison of our methods. Intuitively, this historical block-bootstrap approach enables us to assess a strategy's robustness with respect to alternative entry dates. Furthermore, the available data are used in the most efficient way while preserving all dependency effects in the series, such as autocorrelation and conditional heteroskedasticity (see Dichtl and Drobetz, 2011; Happersberger, Lohre, and Nolte, 2020).

forecast combination when the optimal combination of individual forecasts is potentially non-linear.

In the empirical application, we combine VaR and ES forecasts from 15 individual methods for a broad data set of 12 major equity indices over a period of 30 years. For training and hyperparameter tuning of the machine learning models, we use loss functions that overcome the lack of elicitability for ES by jointly modeling ES and VaR. Using a comprehensive VaR and ES backtesting framework, it turns out that the machine learning combination approaches dominate the set of competing combination approaches in modeling the tail of the return distribution. In particular, we demonstrate that egalitarian shrinkage models such as the egalitarian ridge or partially-egalitarian LASSO models exhibit an excellent forecasting performance in terms of statistical accuracy as well as economical relevance in risk targeting strategies. As for the neural network combination model, the results are less clear-cut. We provide evidence that this highly flexible machine learning method is able to outperform the majority of competing combination approaches, with particularly convincing results in periods of recessions. Still, the excellent performance of the shrinkage models questions whether the additional complexity of the neural network models needs to be taken. The overall success of machine learning techniques for combining VaR and ES predictions can be explained by the general advantages of forecast combinations, such as diversification gains or the robustness to structural breaks and misspecification risk.

When evaluating the combination forecasts during the recent COVID-19 period, we observe lower VaR violation rates than in the global financial crisis, indicating that the combination models have learned from previous recessions.

In future work, it would be interesting to consider other machine learning techniques such as random forests or gradient boosting for combining VaR and ES forecasts. An interesting extension would also be to investigate the use of machine learning techniques for multi-step-ahead VaR and ES prediction. In the spirit of Audrino, Sigrist, and Ballinari (2020b) one could also analyze whether the forecast accuracy may be further improved by expanding the predictor set by conditioning variables that could carry additional information for predicting short-term risk, such as news flow or index options data. Neural network models are particularly well-suited for this task as they avoid to specify a complex functional form to integrate a broad information set.

Appendix 3.A Tables

Table 3.A.1: The set of hyperparameters

This table reports the set of hyperparameters used to tune the shrinkage and neural network models.

	Shrinkage	NN-1HL
Number of nodes	—	{32}
Activation function	—	{ReLu}
Optimizer	—	{RMSprop, Adam}
Dropout rate	—	{0.1, 0.2, 0.3}
Regularization rate	$\left[10^{-5}, 10^{2}\right]$	{1e-01, 1e-02, 1e-03, 1e-04, 1e-05}
Number of epochs	—	{10, 20, 30, 40, 50, 75, 100}
Batch size	—	{32, 64, 128, 256}

Table 3.A.2: Summary statistics of the individual ES forecasts

This table reports the summary statistics of the 5% ES forecasts of the individual models for the S&P500. Specifically, we report the mean, standard deviation (Sd), minimum (Min), maximum (Max), skewness and kurtosis in percentage points. Mean and standard deviation are annualized for the daily return data. The sample spans the period from July 28, 1995 to May 7, 2021.

	Mean	Sd	Min	Max	Skewness	Kurtosis
HS	-2.60	0.89	-4.57	-1.26	-0.79	2.83
WHS	-2.69	1.26	-8.20	-1.00	-1.94	7.74
RiskMetrics-N	-2.12	1.27	-11.02	-0.60	-2.93	15.78
GARCH-N	-2.12	1.23	-18.74	-0.87	-3.84	28.27
GJR-GARCH-FHS	-2.46	1.65	-26.58	-0.89	-4.18	34.37
GJR-GARCH-EVT	-2.47	1.66	-26.88	-0.88	-4.19	34.69
GJR-GARCH-SSTD	-2.40	1.58	-24.52	-0.86	-3.99	30.96
CAViaR-EVT	-2.47	1.77	-33.59	-0.30	-5.56	58.92
GAS	-2.45	1.61	-31.63	-0.86	-4.66	42.37
Range-RGARCH-SSTD	-2.33	1.33	-14.05	-0.59	-2.87	16.99
Range-RGARCH-FHS	-2.34	1.32	-14.04	-0.63	-2.86	17.12
Range-RGARCH-EVT	-2.37	1.33	-14.12	-0.63	-2.84	16.94
Range-HAR-SSTD	-2.41	2.13	-54.30	-0.63	-9.98	167.69
Range-HAR-FHS	-2.46	2.21	-58.76	-0.66	-10.33	180.23
Range-HAR-EVT	-2.47	2.23	-59.09	-0.67	-10.29	179.23

Table 3.A.3: Correlation between the individual ES forecasts

This table presents a sub-set of the correlation matrix of the 5% ES forecasts for the S&P 500, derived from the 15 different individual models considered in this paper. The estimators in the columns correspond to standard choices in the extant literature (HS, WHS, GJR-GARCH-EVT, CAViar-EVT, GAS) as well as some models based on the realized range (Range-RGARCH-FHS, Range-HAR-SSTD). Given the estimation window of 1000 days, the ES forecasts range from July 28, 1995 to May 7, 2021.

	HS	WHS	GJR-GARCH-EVT	CAViaR-EVT	GAS	Range-RGARCH-FHS	Range-HAR-SSTD
HS	1	0.68	0.28	0.36	0.34	0.34	0.23
WHS	0.68	1	0.67	0.67	0.72	0.72	0.54
RiskMetrics-N	0.31	0.80	0.87	0.81	0.81	0.88	0.70
GARCH-N	0.33	0.72	0.96	0.92	0.90	0.94	0.82
GJR-GARCH-SSTD	0.30	0.68	0.99	0.95	0.92	0.93	0.86
GJR-GARCH-FHS	0.28	0.67	0.99	0.95	0.93	0.93	0.86
GJR-GARCH-EVT	0.28	0.67	1	0.95	0.93	0.93	0.86
CAViaR-EVT	0.36	0.67	0.95	1	0.88	0.87	0.82
GAS	0.34	0.72	0.93	0.88	1	0.85	0.80
Range-RGARCH-SSTD	0.37	0.74	0.92	0.87	0.85	0.99	0.81
Range-RGARCH-FHS	0.34	0.72	0.93	0.87	0.85	1	0.80
Range-RGARCH-EVT	0.35	0.73	0.93	0.87	0.85	0.99	0.80
Range-HAR-SSTD	0.23	0.54	0.86	0.82	0.80	0.80	1
Range-HAR-FHS	0.22	0.53	0.86	0.83	0.80	0.80	0.99
Range-HAR-EVT	0.22	0.53	0.86	0.83	0.81	0.80	0.99

Table 3.A.4: VaR and ES calibration backtesting of individual methods

This table reports the results of the calibration backtests for evaluating the VaR and ES predictions from the individual methods calculated over the out-of-sample period from June 3, 1999 to May 7, 2021. *Viol* is the average VaR violation rate over all equity indices in percentage points. *VaR* and *ES* are the average percentage of VaR and ES tests passed and $\overline{p}_{\text{VaR}}$ and $\overline{p}_{\text{ES}}$ are the corresponding average p-values over all VaR/ES tests and equity indices.

	1% probability level					2.5% probability level					5% probability level				
	Viol	VaR	$\overline{p}_{\text{VaR}}$	ES	$\overline{p}_{\text{ES}}$	Viol	VaR	$\overline{p}_{\text{VaR}}$	ES	$\overline{p}_{\text{ES}}$	Viol	VaR	$\overline{p}_{\text{VaR}}$	ES	$\overline{p}_{\text{ES}}$
HS	1.29	11	0.05	47	0.21	2.81	12	0.04	33	0.16	5.20	17	0.06	35	0.14
WHS	1.09	47	0.27	100	0.67	2.51	25	0.19	76	0.34	4.87	23	0.13	42	0.13
RiskMetrics-N	2.15	3	0.00	0	0.00	3.74	0	0.00	0	0.00	6.00	0	0.00	0	0.00
GARCH-N	1.76	22	0.12	0	0.00	3.31	21	0.09	0	0.00	5.54	42	0.14	0	0.00
GJR-GARCH-SSTD	1.02	97	0.47	98	0.55	2.56	79	0.49	86	0.46	5.19	79	0.44	82	0.45
GJR-GARCH-FHS	1.09	86	0.41	92	0.56	2.48	85	0.50	92	0.53	4.90	77	0.47	86	0.53
GJR-GARCH-EVT	0.95	92	0.49	92	0.55	2.37	73	0.40	83	0.44	4.92	75	0.48	83	0.46
CAViaR-EVT	1.15	72	0.25	50	0.18	2.77	54	0.22	22	0.10	5.28	65	0.33	36	0.16
GAS	1.43	28	0.08	47	0.23	2.79	48	0.26	61	0.28	5.23	65	0.27	83	0.36
Range-RGARCH-SSTD	1.10	69	0.34	98	0.51	2.63	60	0.26	82	0.36	5.18	50	0.22	71	0.32
Range-RGARCH-FHS	1.13	67	0.30	90	0.40	2.63	67	0.27	83	0.38	5.01	52	0.23	81	0.42
Range-RGARCH-EVT	1.03	72	0.42	95	0.49	2.55	62	0.25	83	0.41	5.00	56	0.24	82	0.41
Range-HAR-SSTD	1.04	92	0.49	95	0.58	2.62	90	0.53	97	0.53	5.15	92	0.49	96	0.52
Range-HAR-FHS	1.05	97	0.45	97	0.61	2.53	96	0.60	99	0.54	4.83	92	0.52	99	0.60
Range-HAR-EVT	0.97	97	0.57	97	0.62	2.47	96	0.53	100	0.57	4.97	96	0.61	99	0.60

Table 3.A.5: Relative comparison of all forecasting approaches

This table reports the results of the relative comparison between the all VaR and ES predictions over all 12 equity indices. *rank* is the average rank based on the average FZ loss and *best* is the number of times a method is the best method. *DM* is the average percentage of how often a specific method significantly outperforms another using pairwise modified Diebold-Mariano tests based on the FZ loss function. Averages are calculated over all equity indices. *SSM* is the number of indices a method is included in the superior set of models at the 75% level and $\overline{p}_{\text{MCS}}$ is the average over the 12 individual MCS p-values based on the Tmax statistics using 1000 iterations of the moving block bootstrap. The sample spans the period from June 3, 1999 to May 7, 2021.

	1% probability level					2.5% probability level					5% probability level				
	rank	best	DM	MCS	$\overline{p}_{\text{MCS}}$	rank	best	DM	MCS	$\overline{p}_{\text{MCS}}$	rank	best	DM	MCS	$\overline{p}_{\text{MCS}}$
Individual methods															
HS	35.0	0	0.0	0	0.00	35.0	0	0.0	0	0.00	35.0	0	0.0	0	0.00
WHS	32.8	0	1.0	0	0.00	33.8	0	1.0	0	0.00	34.0	0	1.0	0	0.00
RiskMetrics-N	33.9	0	0.2	0	0.00	33.0	0	0.8	0	0.00	32.8	0	1.3	0	0.00
GARCH-N	30.8	0	2.5	0	0.00	30.7	0	3.2	0	0.00	30.6	0	3.1	0	0.00
GJR-GARCH-SSTD	25.2	0	5.3	11	0.76	24.0	0	5.8	10	0.75	24.2	0	5.9	11	0.74
GJR-GARCH-FHS	26.1	0	5.3	11	0.74	25.5	0	5.8	10	0.70	24.2	0	5.8	11	0.72
GJR-GARCH-EVT	26.0	0	5.3	11	0.74	25.2	0	5.9	10	0.73	24.8	0	5.9	11	0.73
CAViaR-EVT	29.7	0	2.7	6	0.23	30.3	0	2.3	2	0.06	30.9	0	2.3	1	0.05
GAS	32.2	0	0.9	0	0.00	31.9	0	1.5	0	0.00	31.7	0	2.2	0	0.00
Range-RGARCH-SSTD	24.8	0	5.4	12	0.93	27.5	0	5.1	10	0.70	27.3	0	5.1	10	0.61
Range-RGARCH-FHS	24.9	0	5.3	12	0.92	25.8	0	5.4	11	0.79	25.3	0	5.4	9	0.62
Range-RGARCH-EVT	23.7	0	5.5	12	0.94	25.5	0	5.3	11	0.77	25.5	0	5.5	10	0.67
Range-HAR-SSTD	7.6	3	9.4	12	1.00	12.1	1	7.2	12	1.00	17.0	0	6.8	12	0.95
Range-HAR-FHS	11.7	0	6.7	12	0.98	13.2	0	7.1	12	0.95	17.9	0	6.9	12	0.96
Range-HAR-EVT	8.9	1	7.8	12	1.00	14.5	0	6.8	12	0.97	15.2	0	7.2	12	0.96
ML combination methods															
Minimum loss	12.7	1	7.0	12	1.00	10.3	0	8.6	12	1.00	8.1	0	9.8	12	1.00
Ridge	9.2	1	9.5	12	1.00	9.1	1	10.5	12	1.00	12.1	0	8.2	12	1.00
LASSO	15.8	1	6.9	12	0.87	14.7	0	7.3	12	0.96	11.8	0	7.9	12	0.97
Elastic net	14.2	0	7.2	11	0.86	10.9	0	8.7	12	0.95	7.6	1	9.7	12	1.00
eRidge	9.7	0	8.8	12	1.00	8.9	0	10.2	12	1.00	3.6	4	12.8	12	1.00
eLASSO	11.5	0	7.0	12	1.00	18.8	0	6.7	12	0.95	11.2	0	8.7	12	0.97
eElasticNet	16.8	0	6.3	12	0.96	16.7	0	6.9	12	1.00	9.2	0	9.3	12	1.00
peLASSO (eRidge)	9.2	1	8.2	12	1.00	5.8	1	11.4	12	1.00	6.5	1	9.7	12	1.00
peLASSO (eLASSO)	11.0	0	7.3	12	1.00	8.7	0	9.3	12	1.00	10.4	0	8.8	12	1.00
peLASSO (Average)	12.2	0	7.8	12	0.99	3.7	3	11.8	12	1.00	7.2	1	10.0	12	1.00
NN-1HL	10.2	3	8.7	12	0.99	10.6	0	8.0	12	0.95	12.7	1	8.2	11	0.91
Competing combination methods															
Average	17.5	0	6.7	12	0.90	18.1	0	6.5	11	0.81	20.5	0	6.2	10	0.81
Trimmed average	13.8	0	6.9	12	0.99	17.2	0	7.3	12	0.93	18.2	0	6.8	12	0.99
Trimmed best-average	14.5	0	6.9	12	0.99	16.9	0	7.1	11	0.87	19.7	0	6.8	12	0.96
Inverse loss	15.1	0	7.9	12	0.97	15.3	0	8.1	11	0.85	18.4	0	8.0	11	0.87
Inverse rank	6.8	0	9.7	12	1.00	6.5	3	10.6	12	1.00	8.8	1	10.6	12	1.00
Difference spacing	13.1	0	7.3	12	1.00	10.9	0	8.0	12	1.00	8.1	1	9.8	12	1.00
Relative score	8.2	0	8.8	12	1.00	4.6	3	10.4	12	1.00	4.4	2	10.9	12	1.00
Shrinkage-to-equal	20.8	0	5.8	12	0.88	18.5	0	6.3	12	0.95	14.5	0	7.1	12	1.00
Single best	14.7	1	6.8	12	0.98	15.9	0	6.8	11	0.90	20.6	0	6.5	11	0.83

Table 3.A.6: Combination forecasts in risk targeting strategies: Historical block-bootstrap

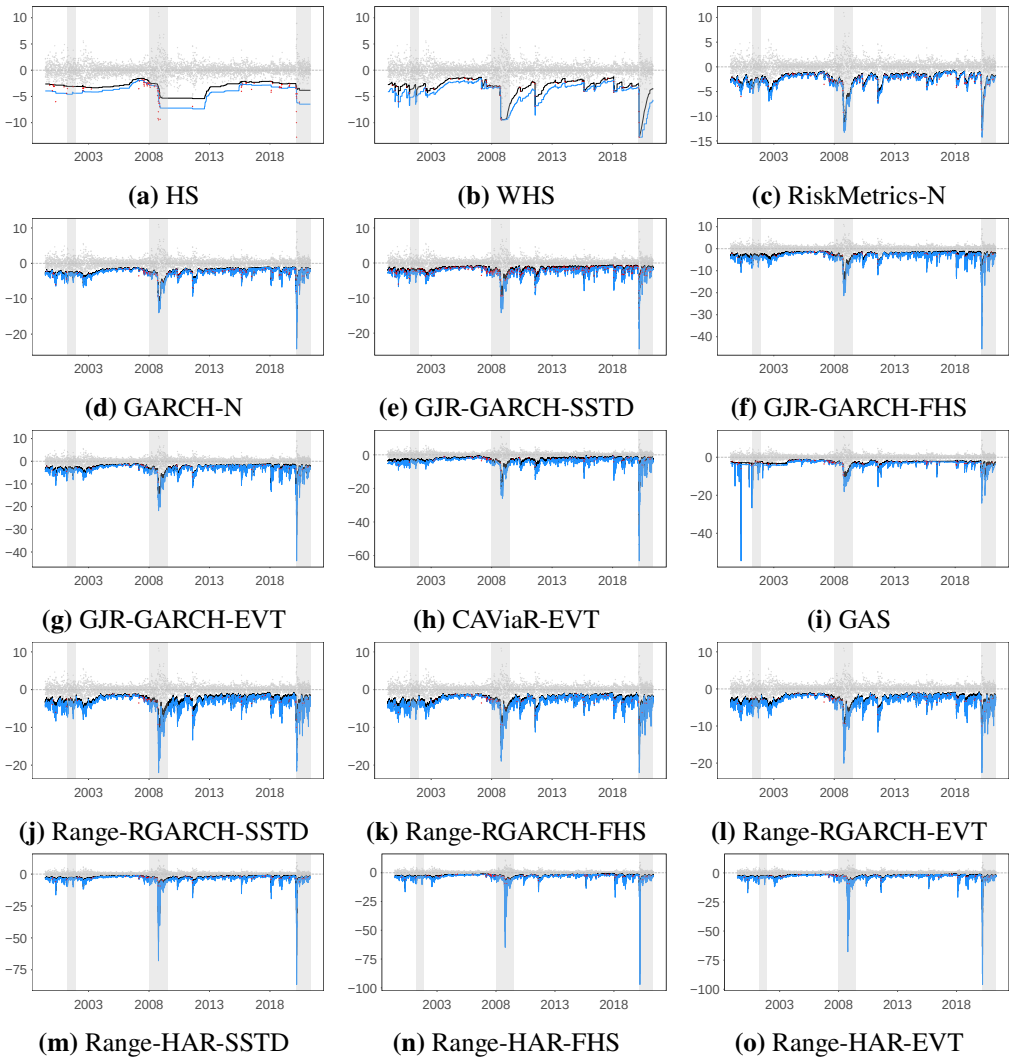
This table reports the results of the risk targeting strategy based on the various 5% ES combination forecasts for the historical block-bootstrap over the out-of-sample period from June 3, 1999 to May 7, 2021. We target an ES of 2.5% over the whole out-of-sample period and report the average of the following performance measures over the 12 equity indices: the annualized mean return (Return), annualized standard deviation (Sd), 5% ES, maximum drawdown (MDD), Sharpe ratio (SR), Calmar ratio, Sortino ratio, Omega ratio, return-to-ES ratio, participation in the risky equity index (Part) and turnover (TO). Return, Sd, ES, MDD, Part and TO are given in percentage points. The performance measures are based on the simulated yearly returns, except for MDD, Calmar ratio and participation. Those are based on the daily risky asset exposure of the corresponding draw and show the yearly mean of the specific measure. For comparison, we include the average performance of the underlying equity indices and the performance of a money market investment.

Method	Return	Sd	ES	MDD	Sharpe	Calmar	Sortino	Omega	Ret/ES	Part	TO
Risk targeting based on ML combination methods											
Minimum loss	3.82	18.19	37.63	-17.99	0.13	0.78	0.38	1.86	0.11	92.52	2.05
Ridge	3.84	18.25	37.73	-18.11	0.13	0.79	0.38	1.86	0.11	92.84	1.73
LASSO	3.81	18.39	38.27	-18.22	0.13	0.78	0.37	1.85	0.11	93.07	1.87
Elastic net	3.82	18.33	38.02	-18.15	0.13	0.78	0.38	1.85	0.11	92.93	1.84
eRidge	3.78	18.04	37.01	-17.85	0.13	0.78	0.38	1.86	0.11	92.07	1.89
eLASSO	3.84	18.04	36.94	-17.90	0.13	0.78	0.39	1.87	0.12	92.14	1.79
eElasticNet	3.82	18.03	36.91	-17.91	0.13	0.78	0.38	1.87	0.12	92.13	1.80
peLASSO (eRidge)	3.85	18.09	36.91	-17.87	0.13	0.78	0.39	1.87	0.12	92.25	1.83
peLASSO (eLASSO)	3.87	18.10	36.98	-17.88	0.13	0.78	0.39	1.87	0.12	92.27	1.79
peLASSO (Average)	3.91	18.20	37.32	-17.95	0.13	0.79	0.39	1.88	0.12	92.55	1.65
NN-1HL	3.76	18.27	37.52	-18.06	0.12	0.78	0.37	1.82	0.11	92.56	1.91
Risk targeting based on competing combination methods											
Average	3.91	18.15	37.44	-18.05	0.13	0.79	0.39	1.88	0.12	92.61	1.55
Trimmed average	3.76	18.32	38.13	-18.05	0.12	0.78	0.37	1.83	0.11	92.55	2.38
Trimmed best-average	3.75	18.31	38.09	-18.06	0.12	0.78	0.37	1.83	0.11	92.55	2.37
Inverse loss	3.90	18.13	37.38	-18.03	0.13	0.79	0.39	1.88	0.12	92.60	1.58
Inverse rank	3.83	18.22	37.70	-18.02	0.13	0.78	0.38	1.85	0.11	92.59	1.97
Difference spacing	3.80	18.20	37.59	-18.01	0.13	0.78	0.38	1.86	0.11	92.61	1.94
Relative score	3.79	18.22	37.81	-18.01	0.13	0.78	0.37	1.84	0.11	92.55	2.11
Shrinkage-to-equal	3.80	18.21	37.73	-17.97	0.13	0.78	0.38	1.86	0.11	92.44	2.29
Single best	3.74	18.35	38.21	-18.07	0.12	0.78	0.36	1.83	0.11	92.57	2.42
Benchmarks investments											
Equity underlying	4.40	21.24	50.83	-21.54	0.14	0.85	0.35	1.86	0.10	100	0
Money market	1.60	1.71	-0.03	-0.00	0.00	–	–	–	–	0	0

Appendix 3.B Figures

Figure 3.B.1: Individual methods' 1% VaR and ES forecasts over time

This figure shows the daily 1% VaR forecasts (in black) and associated ES forecasts (in blue) of the individual methods as well as the realized returns of the S&P 500 (light-grey dots) over the period from July 28, 1995 to May 7, 2021. The grey areas indicate recessions as determined by the National Bureau of Economic Research: the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the still prevalent COVID-19 crisis (March 2020 to sample end). VaR violations are marked in red. At a confidence level of 1%, a total of 67 violations are expected over the model period.



(a) HS **(b)** WHS **(c)** RiskMetrics-N

(d) GARCH-N **(e)** GJR-GARCH-SSTD **(f)** GJR-GARCH-FHS

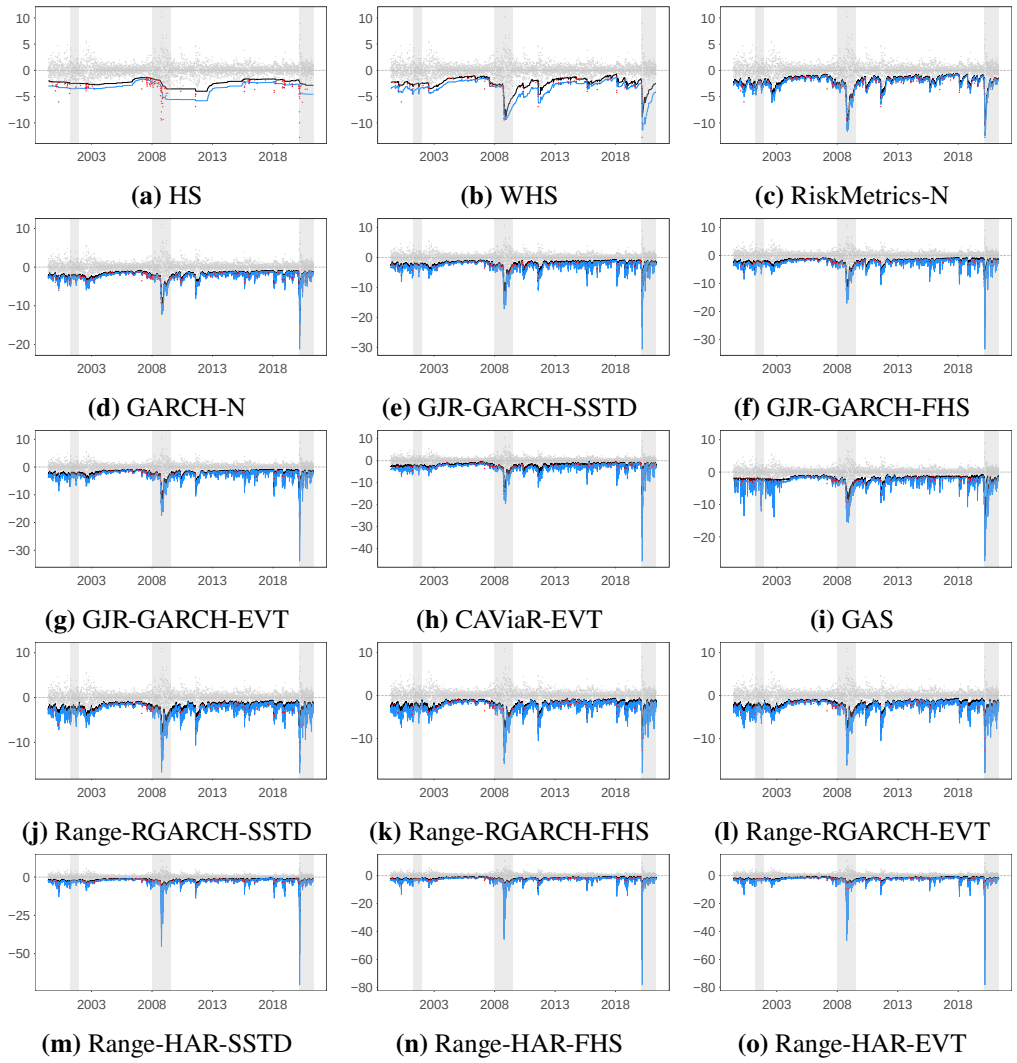
(g) GJR-GARCH-EVT **(h)** CAViaR-EVT **(i)** GAS

(j) Range-RGARCH-SSTD **(k)** Range-RGARCH-FHS **(l)** Range-RGARCH-EVT

(m) Range-HAR-SSTD **(n)** Range-HAR-FHS **(o)** Range-HAR-EVT

Figure 3.B.2: Individual methods' 2.5% VaR and ES forecasts over time

This figure shows the daily 2.5% VaR forecasts (in black) and associated ES forecasts (in blue) of the individual methods' as well as the realized returns of the S&P 500 (light-grey dots) over the period from July 28, 1995 to May 7, 2021. The grey areas indicate recessions as determined by the National Bureau of Economic Research: the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the still prevalent COVID-19 crisis (March 2020 to sample end). VaR violations are marked in red. At a confidence level of 2.5%, a total of 167 violations are expected over the model period.



(a) HS **(b)** WHS **(c)** RiskMetrics-N

(d) GARCH-N **(e)** GJR-GARCH-SSTD **(f)** GJR-GARCH-FHS

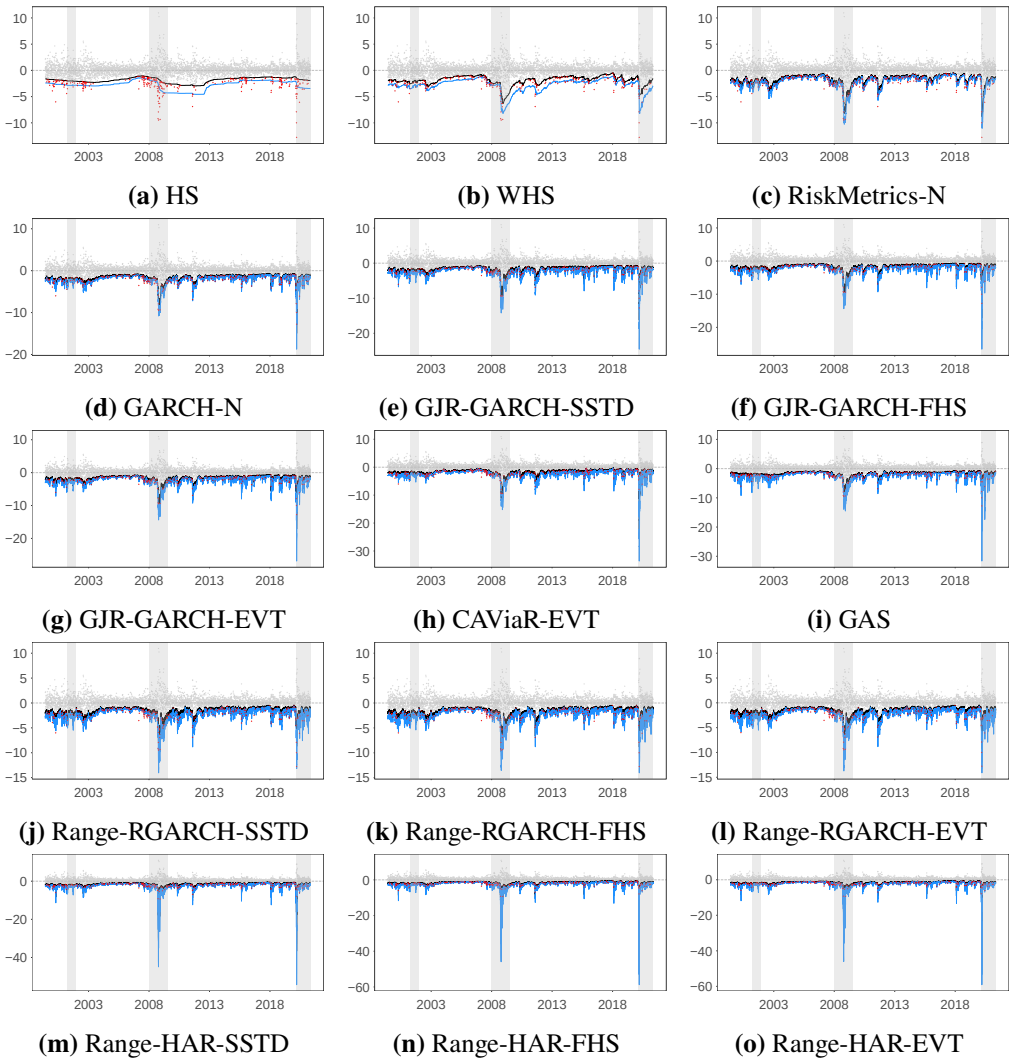
(g) GJR-GARCH-EVT **(h)** CAViaR-EVT **(i)** GAS

(j) Range-RGARCH-SSTD **(k)** Range-RGARCH-FHS **(l)** Range-RGARCH-EVT

(m) Range-HAR-SSTD **(n)** Range-HAR-FHS **(o)** Range-HAR-EVT

Figure 3.B.3: Individual methods' 5% VaR and ES forecasts over time

This figure shows the daily 5% VaR forecasts (in black) and associated ES forecasts (in blue) of the individual methods as well as the realized returns of the S&P 500 (light-grey dots) over the period from July 28, 1995 to May 7, 2021. The grey areas indicate recessions as determined by the National Bureau of Economic Research: the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the still prevalent COVID-19 crisis (March 2020 to sample end). VaR violations are marked in red. At a confidence level of 5%, a total of 335 violations are expected over the model period.



(a) HS **(b)** WHS **(c)** RiskMetrics-N

(d) GARCH-N **(e)** GJR-GARCH-SSTD **(f)** GJR-GARCH-FHS

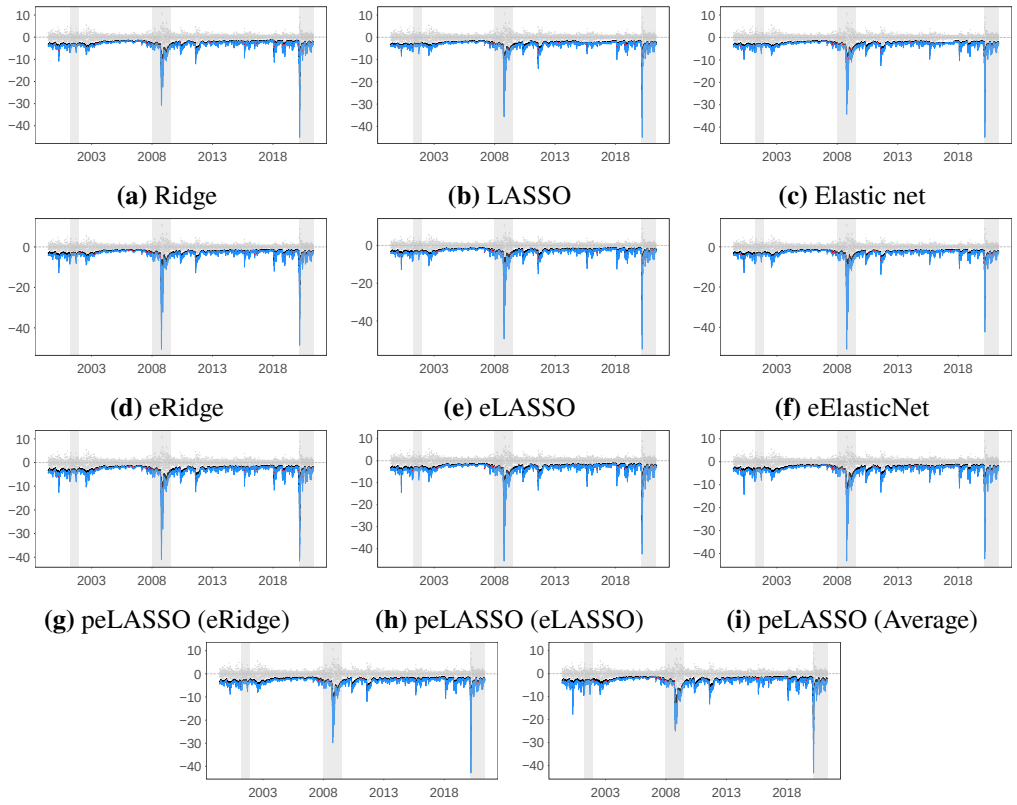
(g) GJR-GARCH-EVT **(h)** CAViaR-EVT **(i)** GAS

(j) Range-RGARCH-SSTD **(k)** Range-RGARCH-FHS **(l)** Range-RGARCH-EVT

(m) Range-HAR-SSTD **(n)** Range-HAR-FHS **(o)** Range-HAR-EVT

Figure 3.B.4: 1% VaR and ES ML combination forecasts over time

This figure shows the daily 1% VaR forecasts (in black) and associated ES forecasts (in blue) of the machine learning combination methods as well as the realized returns of the S&P 500 (light-grey dots) over the period from June 3, 1999 to May 7, 2021. The grey areas indicate recessions as determined by the National Bureau of Economic Research: the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the still prevalent COVID-19 crisis (March 2020 to sample end). VaR violations are marked in red. At a confidence level of 1%, a total of 57 violations are expected over the model period.



(a) Ridge **(b)** LASSO **(c)** Elastic net

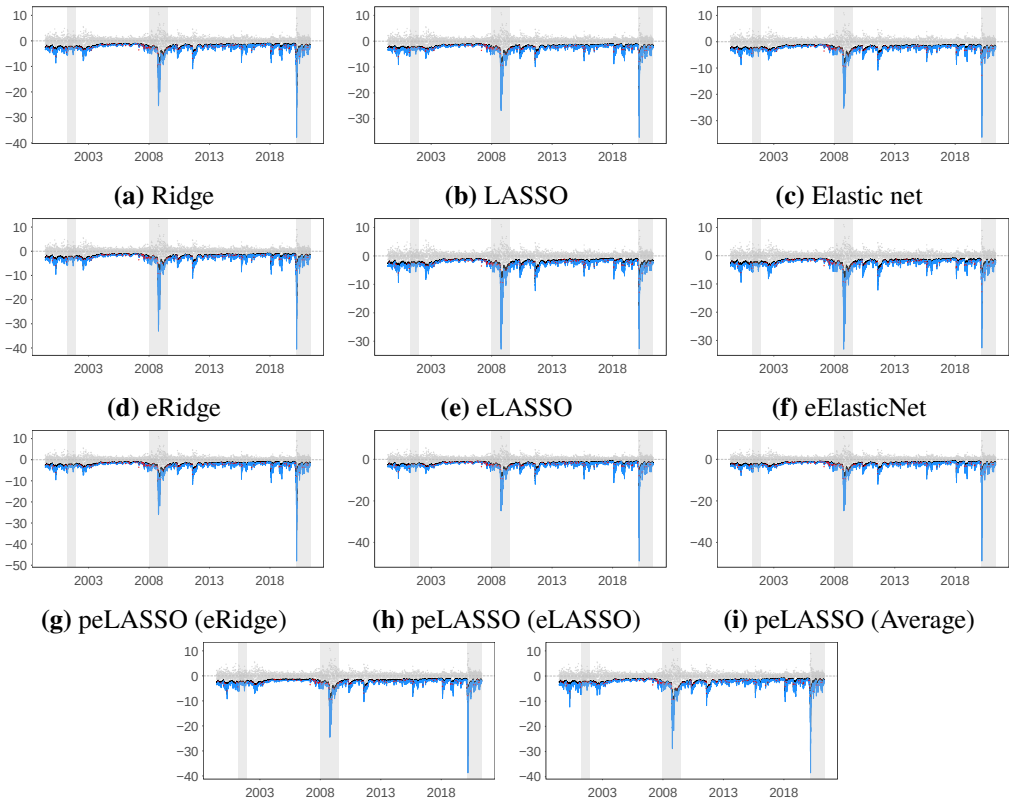
(d) eRidge **(e)** eLASSO **(f)** eElasticNet

(g) peLASSO (eRidge) **(h)** peLASSO (eLASSO) **(i)** peLASSO (Average)

(j) NN-1HL **(k)** Minimum loss

Figure 3.B.5: 2.5% VaR and ES ML combination forecasts over time

This figure shows the daily 2.5% VaR forecasts (in black) and associated ES forecasts (in blue) of the machine learning combination methods as well as the realized returns of the S&P 500 (light-grey dots) over the period from June 3, 1999 to May 7, 2021. The grey areas indicate recessions as determined by the National Bureau of Economic Research: the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the still prevalent COVID-19 crisis (March 2020 to sample end). VaR violations are marked in red. At a confidence level of 2.5%, a total of 142 violations are expected over the model period.



(a) Ridge **(b)** LASSO **(c)** Elastic net

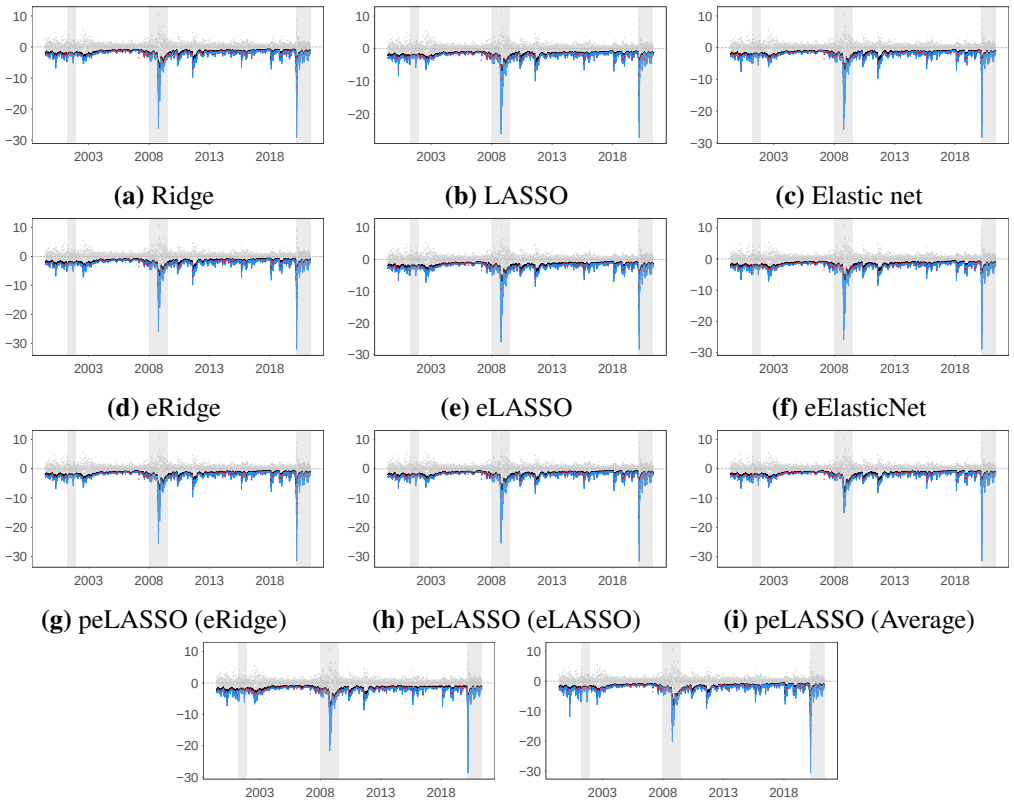
(d) eRidge **(e)** eLASSO **(f)** eElasticNet

(g) peLASSO (eRidge) **(h)** peLASSO (eLASSO) **(i)** peLASSO (Average)

(j) NN-1HL **(k)** Minimum loss

Figure 3.B.6: 5% VaR and ES ML combination forecasts over time

This figure shows the daily 5% VaR forecasts (in black) and associated ES forecasts (in blue) of the machine learning combination methods as well as the realized returns of the S&P 500 (light-grey dots) over the period from June 3, 1999 to May 7, 2021. The grey areas indicate recessions as determined by the National Bureau of Economic Research: the dot-com bubble (April 2001 to December 2001), the global financial crisis (January 2008 to July 2009) and the still prevalent COVID-19 crisis (March 2020 to sample end). VaR violations are marked in red. At a confidence level of 5%, a total of 285 violations are expected over the model period.



(a) Ridge **(b)** LASSO **(c)** Elastic net

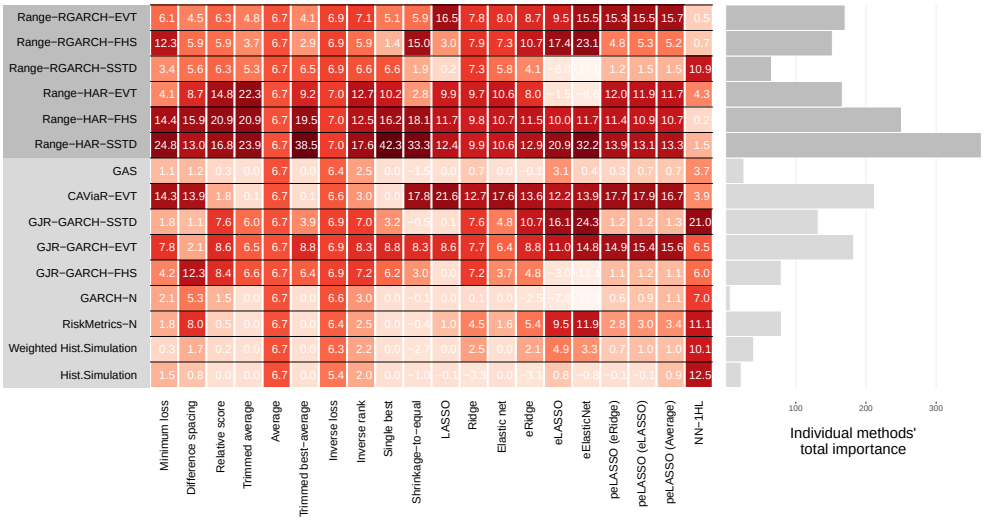
(d) eRidge **(e)** eLASSO **(f)** eElasticNet

(g) peLASSO (eRidge) **(h)** peLASSO (eLASSO) **(i)** peLASSO (Average)

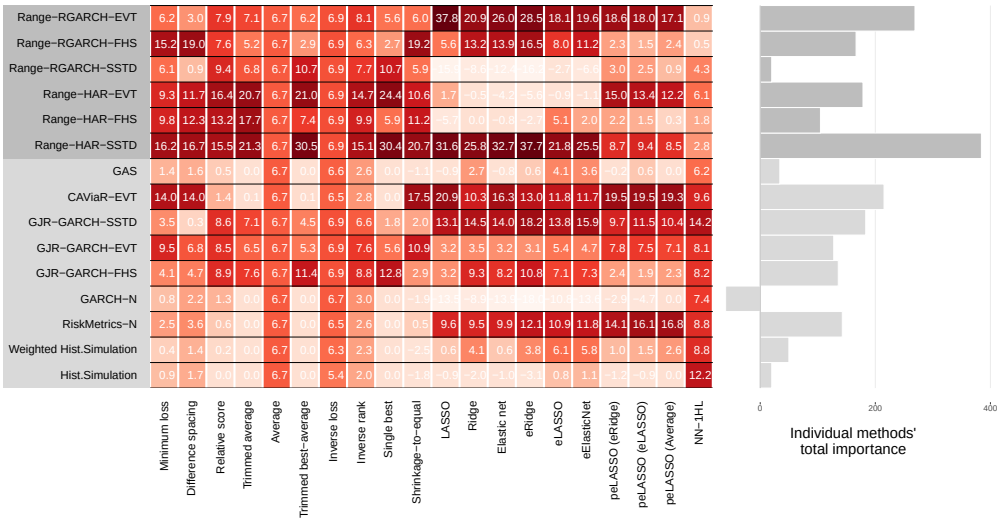
(j) NN-1HL **(k)** Minimum loss

Figure 3.B.7: Individual methods' importance at the 2.5% and 5% probability level

This figure shows the importance of the individual methods for all combination approaches at the 2.5% and 5% probability level. The figures are the average ES combination weights for the shrinkage and competing models across time and all 12 equity indices over the period from June 3, 1999 to May 7, 2021. For the neural network combination models, the figures are the mean of the permutation feature importance scores calculated every four years. The darker the red tone, the higher the importance. The right part of the figure shows the total importance of each individual method; the bars represent the sum of scores across all combination approaches.



(a) 2.5% probability level



(b) 5% probability level

Figure 3.B.8: Shrinkage combination weights of 2.5% forecasts over time

This figure shows the combination weights for the shrinkage methods' 2.5% ES forecasts over the period from June 3, 1999 to May 7, 2021. Given that the shrinkage methods are estimated pooled over all equity indices, the presented combination weights are the same for all 12 equity indices.

(a) Ridge

(b) eRidge

(c) LASSO

(d) eLASSO

(e) Elastic net

(f) eElasticNet

(g) peLASSO (eRidge)

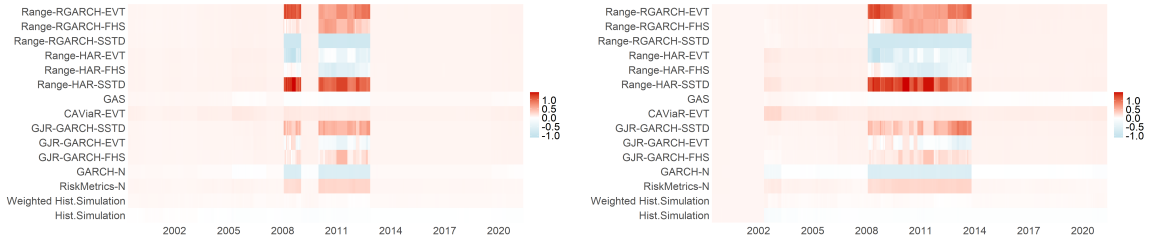
(h) peLASSO (eLASSO)

(i) peLASSO (Average)

(j) Minimum loss

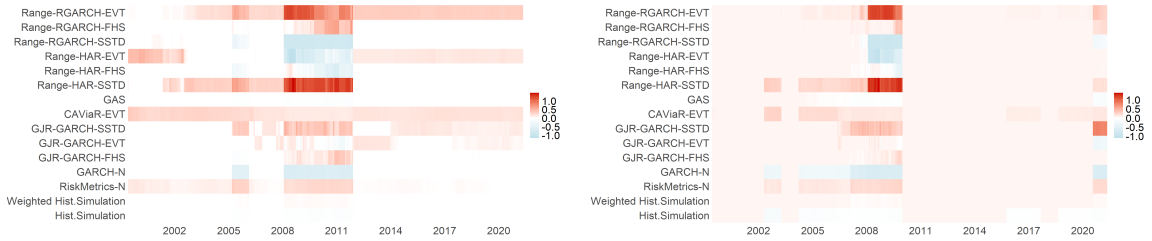
Figure 3.B.9: Shrinkage combination weights of 5% forecasts over time

This figure shows the combination weights for the shrinkage methods' 5% ES forecasts over the period from June 3, 1999 to May 7, 2021. Given that the shrinkage methods are estimated pooled over all equity indices, the presented combination weights are the same for all 12 equity indices.



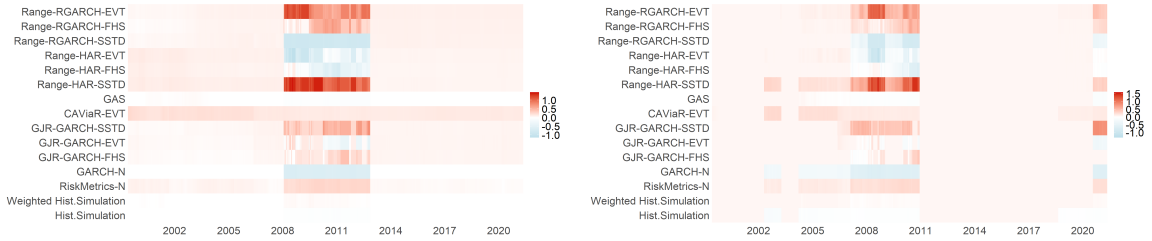
(a) Ridge

(b) eRidge



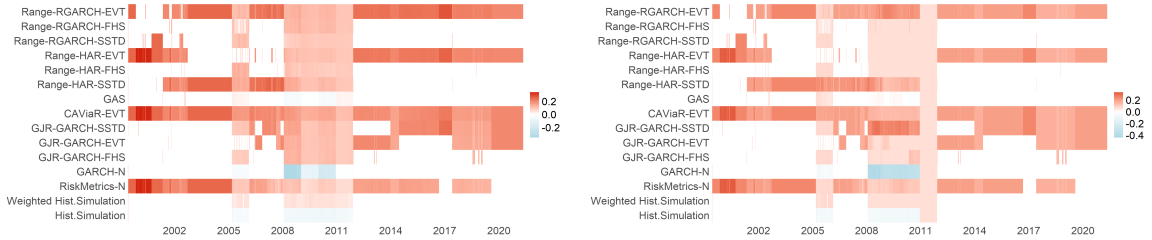
(c) LASSO

(d) eLASSO



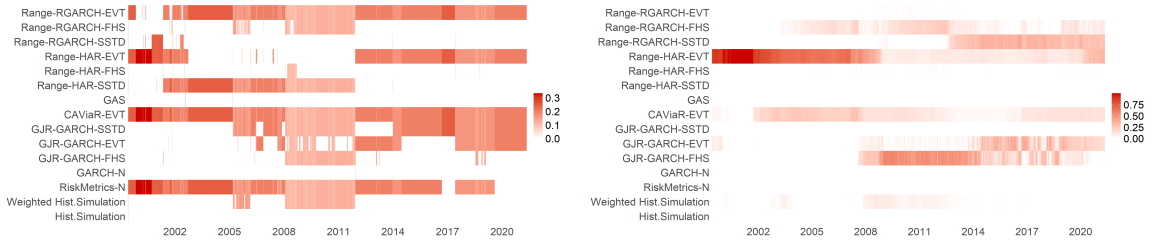
(e) Elastic net

(f) eElasticNet



(g) peLASSO (eRidge)

(h) peLASSO (eLASSO)



(i) peLASSO (Average)

(j) Minimum loss

References

Acerbi, Carlo and Balazs Szekely (2014). "Backtesting expected shortfall". *Risk* 27 (11), 76–81.

Acerbi, Carlo and Dirk Tasche (2002). "On the coherence of expected shortfall". *Journal of Banking & Finance* 26 (7), 1487–1503.

Alizadeh, Sassan, Michael W Brandt, and Francis X Diebold (2002). "Range-based estimation of stochastic volatility models". *Journal of Finance* 57 (3), 1047–1091.

Andersen, Torben G, Tim Bollerslev, Peter F Christoffersen, and Francis X Diebold (2006). "Volatility and correlation forecasting". In: *Handbook of Economic Forecasting*. Ed. by Elliott, G., Granger, Clive W.J., and Timmermann, A. Vol. 1. Elsevier. Chap. 15, 777–878.

— (2013). "Financial risk measurement for financial risk management". In: *Handbook of the Economics of Finance*. Ed. by Constantinides, George M, Harris, Milton, and Stulz, René M. Vol. 2. Elsevier. Chap. 17, 1127–1220.

Annaert, Jan, Sofieke Van Osselaer, and Bert Verstraete (2009). "Performance evaluation of portfolio insurance strategies using stochastic dominance criteria". *Journal of Banking & Finance* 33 (2), 272–280.

Arlot, Sylvain, Alain Celisse, et al. (2010). "A survey of cross-validation procedures for model selection". *Statistics surveys* 4, 40–79.

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, and David Heath (1999). "Coherent measures of risk". *Mathematical Finance* 9 (3), 203–228.

Audrino, Francesco, Fabio Sigrist, and Daniele Ballinari (2020b). "The impact of sentiment and attention measures on stock market volatility". *International Journal of Forecasting* 36 (2), 334–357.

Barone-Adesi, Giovanni, Kostas Giannopoulos, and Les Vosper (1999). "VaR without correlations for portfolios of derivative securities". *Journal of Futures Markets* 19 (5), 583–602.

Basel Committee on Banking Supervision (2016). *Minimum capital requirements for market risk*. https://www.bis.org/bcbs/publ/d352.pdf.

Bates, John M and Clive WJ Granger (1969). "The combination of forecasts". *Journal of the Operational Research Society* 20 (4), 451–468.

Bayer, Sebastian (2018). "Combining Value-at-Risk forecasts using penalized quantile regressions". *Econometrics and Statistics* 8, 56–77.

Bayer, Sebastian and Timo Dimitriadis (Sept. 2020). "Regression-based expected shortfall backtesting". *Journal of Financial Econometrics*. nbaa013.

Bergstra, James and Yoshua Bengio (2012). "Random search for hyper-parameter optimization". *Journal of Machine Learning Research* 13 (1), 281–305.

Bernardi, Mauro and Leopoldo Catania (2016). "Comparison of Value-at-Risk models using the MCS approach". *Computational Statistics* 31 (2), 579–608.

Bertrand, Philippe and Jean-Luc Prigent (2011). "Omega performance measure and portfolio insurance". *Journal of Banking & Finance* 35 (7), 1811–1823.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni (2021). "Bond risk premiums with machine learning". *The Review of Financial Studies* 34 (2), 1046–1089.

Bollerslev, Tim (1987). "A conditionally heteroskedastic time series model for speculative prices and rates of return". *Review of Economics and Statistics*, 542–547.

Bollerslev, Tim, Benjamin Hood, John Huss, and Lasse Heje Pedersen (2018b). "Risk everywhere: Modeling and managing volatility". *Review of Financial Studies* 31 (7), 2729–2773.

Boudoukh, Jacob, Matthew Richardson, and Robert Whitelaw (1998). "The best of both worlds". *Risk* 11 (5), 64–67.

Breiman, Leo (2001). "Random forests". *Machine Learning* 45 (1), 5–32.

Brownlees, Christian T and Giampiero M Gallo (2010). "Comparison of volatility measures: a risk management perspective". *Journal of Financial Econometrics* 8 (1), 29–56.

Chollet, Francois et al. (2015). *Keras*. https://github.com/fchollet/keras.

Christoffersen, Peter and Denis Pelletier (2004). "Backtesting value-at-risk: A duration-based approach". *Journal of Financial Econometrics* 2 (1), 84–108.

Christoffersen, Peter F (1998). "Evaluating interval forecasts". *International Economic Review* 39 (4), 841–862.

Clemen, Robert T (1989). "Combining forecasts: A review and annotated bibliography". *International Journal of Forecasting* 5 (4), 559–583.

Corsi, Fulvio (2009). "A simple approximate long-memory model of realized volatility". *Journal of Financial Econometrics* 7 (2), 174–196.

Corsi, Fulvio and Roberto Renò (2012). "Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling". *Journal of Business & Economic Statistics* 30 (3), 368–380.

Creal, Drew, Siem Jan Koopman, and André Lucas (2013). "Generalized autoregressive score models with applications". *Journal of Applied Econometrics* 28 (5), 777–795.

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals and Systems* 2 (4), 303–314.

Dichtl, Hubert and Wolfgang Drobetz (2011). "Portfolio insurance and prospect theory investors: Popularity and optimal design of capital protected financial products". *Journal of Banking & Finance* 35 (7), 1683–1697.

Dichtl, Hubert, Wolfgang Drobetz, and Martin Wambach (2017). "A bootstrap-based comparison of portfolio insurance strategies". *European Journal of Finance* 23 (1), 31–59.

Diebold, Francis X (1989). "Forecast combination and encompassing: Reconciling two divergent literatures". *International Journal of Forecasting* 5 (4), 589–592.

Diebold, Francis X and Robert S Mariano (1995). "Comparing predictive accuracy". *Journal of Business & Economic Statistics* 13 (3), 253–263.

Diebold, Francis X and Minchul Shin (2019). "Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives". *International Journal of Forecasting* 35 (4), 1679–1691.

Dietterich, Thomas G (2000). "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer, 1–15.

Dimitriadis, Timo and Roxana Halbleib (2021). "Realized Quantiles". *Journal of Business & Economic Statistics* 0 (0), 1–16.

Donaldson, R Glen and Mark Kamstra (1996). "Forecast combining with neural networks". *Journal of Forecasting* 15 (1), 49–61.

Embrechts, Paul and Marius Hofert (2014). "Statistics and quantitative risk management for banking and insurance". *Annual Review of Statistics and Its Application* 1, 493–514.

Emmer, Susanne, Marie Kratz, and Dirk Tasche (2015). "What is the best risk measure in practice? A comparison of standard measures". *Journal of Risk* 18 (2), 31–60.

Engle, Robert F and Simone Manganelli (2004). "CAViaR: Conditional autoregressive value at risk by regression quantiles". *Journal of Business & Economic Statistics* 22 (4), 367–381.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici (2019). "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *J. Mach. Learn. Res.* 20 (177), 1–81.

Fissler, Tobias and Johanna F Ziegel (2016). "Higher order elicitability and Osband's principle". *Annals of Statistics* 44 (4), 1680–1707.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Gerlach, Richard and Cathy WS Chen (2015). "Bayesian expected shortfall forecasting incorporating the intraday range". *Journal of Financial Econometrics* 14 (1), 128–158.

Giot, Pierre and Sébastien Laurent (2004). "Modelling daily value-at-risk using realized volatility and ARCH type models". *Journal of Empirical Finance* 11 (3), 379–398.

Glosten, Lawrence R, Ravi Jagannathan, and David E Runkle (1993). "On the relation between the expected value and the volatility of the nominal excess return on stocks". *Journal of Finance* 48 (5), 1779–1801.

Gneiting, Tilmann (2011). "Making and evaluating point forecasts". *Journal of the American Statistical Association* 106 (494), 746–762.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". *Journal of the American Statistical Association* 102 (477), 359–378.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016). *Deep learning*. Vol. 1. 2. MIT press Cambridge.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). "Empirical asset pricing via machine learning". *Review of Financial Studies* 33 (5), 2223–2273.

Halbleib, Roxana and Winfried Pohlmeier (2012). "Improving the value at risk forecasts: Theory and evidence from the financial crisis". *Journal of Economic Dynamics and Control* 36 (8), 1212–1228.

Hansen, Bruce E (2008). "Least-squares forecast averaging". *Journal of Econometrics* 146 (2), 342–350.

Hansen, Lars Kai and Peter Salamon (1990). "Neural network ensembles". *IEEE transactions on pattern analysis and machine intelligence* 12 (10), 993–1001.

Hansen, Peter R, Asger Lunde, and James M Nason (2011). "The model confidence set". *Econometrica* 79 (2), 453–497.

Hansen, Peter Reinhard, Zhuo Huang, and Howard Howan Shek (2012). "Realized GARCH: a joint model for returns and realized measures of volatility". *Journal of Applied Econometrics* 27 (6), 877–906.

Happersberger, David, Harald Lohre, and Ingmar Nolte (2020). "Estimating portfolio risk for tail risk protection strategies". *European Financial Management* 26 (4), 1107–1146.

Hart, Jeffrey D (1994). "Automated kernel smoothing of dependent data by using time series cross-validation". *Journal of the Royal Statistical Society: Series B (Methodological)* 56 (3), 529–542.

Hart, Jeffrey D and Cherng-Luen Lee (2005). "Robustness of one-sided cross-validation to autocorrelation". *Journal of Multivariate Analysis* 92 (1), 77–96.

Harvey, A C (2013). "Dynamic models for volatility and heavy tails: with applications to financial and economic time series". In: *Econometric Society Monographs*. Vol. 52. Cambridge University Press.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2011). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, second Edition.

Hinton, Geoffrey (2012). *Neural networks for machine learning*. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". *Science* 313 (5786), 504–507.

Hocquard, Alexandre, Sunny Ng, and Nicolas Papageorgiou (2013). "A constant-volatility framework for managing tail risk". *Journal of Portfolio Management* 39 (2), 28–40.

Hoerl, Arthur E and Robert W Kennard (1970a). "Ridge regression: applications to nonorthogonal problems". *Technometrics* 12 (1), 69–82.

— (1970b). "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics* 12 (1), 55–67.

Hornik, Kurt, Maxwell Stinchcombe, Halbert White, et al. (1989). "Multilayer feedforward networks are universal approximators." *Neural Networks* 2 (5), 359–366.

Kingma, Diederik P. and Jimmy Ba (2015). *Adam: A Method for Stochastic Optimization*. Conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Kuan, Chung-Ming and Halbert White (1994). "Artificial neural networks: An econometric perspective". *Econometric Reviews* 13 (1), 1–91.

Kuester, Keith, Stefan Mittnik, and Marc S Paolella (2006). "Value-at-risk prediction: A comparison of alternative strategies". *Journal of Financial Econometrics* 4 (1), 53–89.

Kupiec, Paul H (1995). "Techniques for verifying the accuracy of risk measurement models". *Journal of Derivatives* 3 (2), 73–84.

Liu, Lily Y, Andrew J Patton, and Kevin Sheppard (2015). "Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes". *Journal of Econometrics* 187 (1), 293–311.

Louzis, Dimitrios P, Spyros Xanthopoulos-Sisinis, and Apostolos P Refenes (2012). "Stock index realized volatility forecasting in the presence of heterogeneous leverage effects and long range dependence in the volatility of realized volatility". *Applied Economics* 44 (27), 3533–3550.

— (2014). "Realized volatility models and alternative Value-at-Risk prediction strategies". *Economic Modelling* 40, 101–116.

Manganelli, Simone and Robert F Engle (2004). "A comparison of value-at-risk models in finance". *Risk measures for the 21st century*, 123–44.

McNeil, Alexander J and Rüdiger Frey (2000). "Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach". *Journal of Empirical Finance* 7 (3), 271–300.

Molnar, Christoph (2019). *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/.

Nieto, Maria Rosa and Esther Ruiz (2016). "Frontiers in VaR forecasting and backtesting". *International Journal of Forecasting* 32 (2), 475–501.

Nolde, Natalia and Johanna F Ziegel (2017). "Elicitability and backtesting: Perspectives for banking regulation". *Annals of Applied Statistics* 11 (4), 1833–1874.

Parkinson, Michael (1980). "The extreme value method for estimating the variance of the rate of return". *Journal of Business*, 61–65.

Patton, Andrew J, Johanna F Ziegel, and Rui Chen (2019). "Dynamic semiparametric models for expected shortfall (and value-at-risk)". *Journal of Econometrics* 211 (2), 388–413.

Perchet, Romain, Raul Leote De Carvalho, Thomas Heckel, and Pierre Moulin (2015). "Predicting the success of volatility targeting strategies: Application to equities and other asset classes". *Journal of Alternative Investments* 18 (3), 21–38.

Righi, Marcelo Brutti and Paulo Sergio Ceretta (2015). "A comparison of expected shortfall estimation models". *Journal of Economics and Business* 78, 14–47.

RiskMetrics Group (1996). "Riskmetrics - technical document". *J. P. Morgan and Reuters*.

Shan, Kejia and Yuhong Yang (2009). "Combining regression quantile estimators". *Statistica Sinica* 19 (3), 1171–1191.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". *Journal of Machine Learning Research* 15 (1), 1929–1958.

Stock, James H and Mark W Watson (2004). "Combination forecasts of output growth in a seven-country data set". *Journal of Forecasting* 23 (6), 405–430.

Taylor, James W (2020). "Forecast combinations for value at risk and expected shortfall". *International Journal of Forecasting* 36 (2), 428–441.

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

Timmermann, Allan (2006). "Forecast combinations". In: *Handbook of Economic Forecasting*. Ed. by Elliott, G., Granger, C.W., and Timmermann, A. Vol. 1. Elsevier. Chap. 4, 135–196.

Ye, Yinyu (1987). "Interior algorithms for linear, quadratic, and linearly constrained nonlinear programming". PhD thesis. Department of ESS, Stanford University.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2), 301–320.

Appendix A

Supplementary Research Papers to Chapter 2

This appendix consists of three research papers that supplement Chapter 2 of this dissertation. All papers are joint work with colleagues from Invesco Quantitative Strategies and have been published in Risk & Reward, Invesco's flagship publication for genuine investment research (Kolrep, Lohre, and Happersberger, 2017; Lohre, Happersberger, and Radatz, 2018; Lohre, Happersberger, and Cherkezov, 2018).

In the first article, *Theory and Practice of Portfolio Insurance*, we analyze various portfolio insurance strategies, contrasting dynamic portfolio insurance strategies such as CPPI and DPPI (see Chapter 2.2.2) with the static stop-loss concept and option-based strategies. Our findings suggest that an active approach on the basis of dynamic risk forecasts is an effective contender.

The second article, *Evaluating Risk Mitigation Strategies*, discusses how to appropriately calibrate and assess portfolio insurance strategies based on the ensuing return distribution to best match a given client's risk preferences.

Based on the proposed methodology, the third article, *The Use of Equity Factor Investing for Portfolio Insurance*, shows that the choice of the equity underlying is important when designing portfolio insurance strategies. In particular, low-volatility underlyings are to be preferred, with other multi-factor propositions forming suitable alternatives when considering additional elements of dynamic risk management.

A.1. Theory and Practice of Portfolio Insurance

Theory and practice of portfolio insurance

By Dr. Martin Kolrep, Dr. Harald Lohre and David Happersberger

In brief
To limit the maximum loss of a portfolio, investment strategies can be enhanced by adding a portfolio insurance component. We have analyzed various portfolio insurance strategies – from the static stop-loss concept to option-based strategies and dynamic portfolio insurance strategies. The findings suggest that an active approach on the basis of dynamic risk forecasts is an effective alternative.

In order to achieve their performance goals, many investors are allocating towards more risky assets. In many cases, these investors can quickly find themselves in a tight spot if the risk budget is not expanded accordingly. This is where strict risk control via portfolio insurance can come into play. But, which portfolio insurance strategy proves to be most effective in historical simulations?

Investors' objectives are generally expressed as a combination of risk and return targets. Defining the return target is usually relatively simple – but the definition of risk targets is less straightforward. One conventional approach is to consider "volatility", that is, the average variation of portfolio return over time. For many investors, however, "maximum drawdown" is a more relevant statistic, as it points to the maximum loss of value. To limit the maximum drawdown, investors typically follow broadly diversified investment strategies that include a tactical asset allocation component designed to avoid losses as often as possible.

However, to effectively limit maximum drawdown, a given investment strategy could implement some form of portfolio insurance. Portfolio insurance strategies aim primarily to improve the downside risk profile of an investment without jeopardizing long-term return potential. In this article, we will present various portfolio insurance strategies and analyze their strengths and weaknesses.

1. Static portfolio insurance using "stop-loss"
The stop-loss strategy is an example of a basic portfolio insurance strategy: when the portfolio value falls below a certain threshold (or floor), all risk positions are sold and replaced by risk-free assets (cf. Rubinstein, 1985).

This can be illustrated by looking at a conservative multi-asset portfolio comprising 33.3% equities, 16.7% commodities and 50% fixed income assets.[1] Despite this conservative allocation, with 3.9% annualized return and 6.4% annualized volatility in the sample period (July 2003 to November 2016), the maximum drawdown during the 2008 financial crisis was as much as -27.2% (see table 1 at the end of the article).[2] To mitigate such losses, we added a stop-loss rule, setting the trigger at a floor of 95% per calendar year (figure 1).

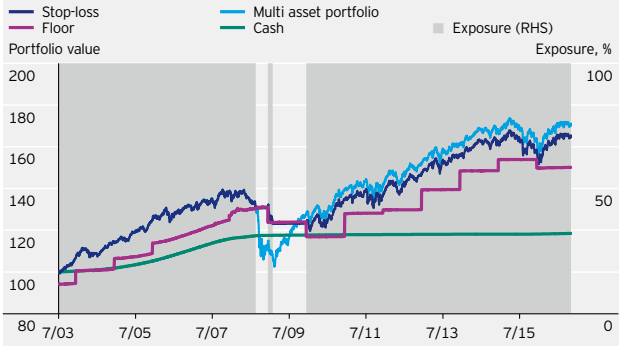
If interest rates are positive, a buffer of more than 5% can be implemented at the beginning of the relevant year; conversely, negative interest rates result in a smaller buffer. The targeted floor is marked by the purple line. It is easy to see that this floor would have been breached from 9 September 2008 onwards – triggering a full reallocation of the portfolio to cash.

This observation reveals a fundamental problem: would a timely exit really have been possible on reaching the 95% threshold in such a volatile period? Moreover, the simple nature of the stop-loss strategy does not envisage a re-entry to the market. In our model, we assume reinvestment at the beginning of the following year. And, although the trigger value is lowered, the marked declines in early 2009 would mean that the portfolio was once again "stopped-out" from 17 February 2009 until the end of the year – precluding participation in the significant recovery that followed.

2. Option-based portfolio insurance
Another static portfolio insurance strategy is the purchase of a European put option.[3] Unlike the stop-loss strategy, the put option ensures that the portfolio value will not breach the targeted floor at expiry.

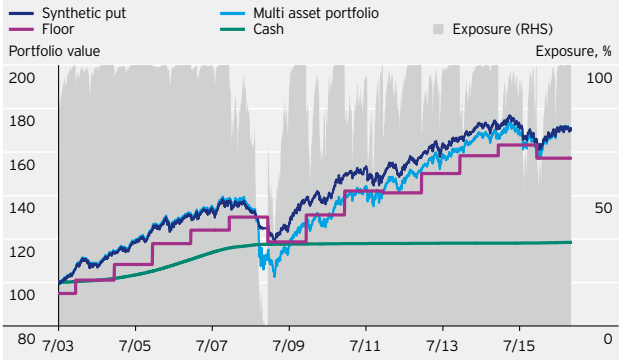
But such a strategy can be expensive, since the option premium is payable on a yearly basis, although the portfolio insurance proves unnecessary in the majority

Figure 1
Performance and allocation of the stop-loss strategy



The chart shows the performance of a conservative multi-asset portfolio using a stop-loss strategy (dark blue line) in relation to the floor (purple line) over time. If the portfolio value falls below the floor (here: 95% of the initial annual portfolio value), all risky assets are liquidated. The portfolio is reinvested, if necessary, at the start of the next year. At all events, the floor value is adjusted at the start of each year to accommodate investment. For comparison, we have included the performance of the underlying conservative multi-asset strategy (light blue line) and a money market investment (green line).
Sources: Bloomberg, Invesco. Period: 23 July 2003 to 22 November 2016; 23 July 2003 = 100.

Figure 2
Performance and allocation of the synthetic put strategy

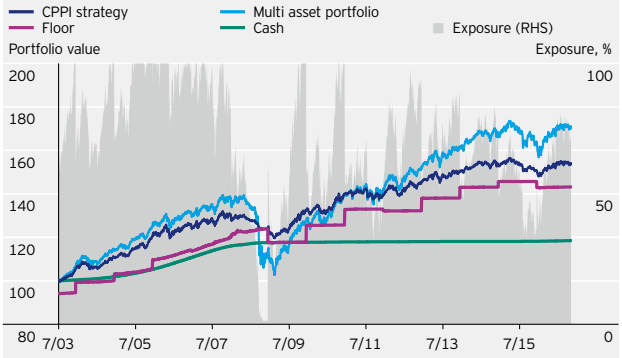


The chart shows the performance of a conservative multi-asset portfolio using a synthetic put strategy (dark blue line) in relation to the floor (purple line) over time. Participation in the risky asset's performance is calculated using a classic Black-Scholes formula measuring the sensitivity of a synthetic put with a strike price matching the floor value (here: 95% of the initial annual portfolio value). For comparison, we have included the performance of the underlying conservative multi-asset strategy (light blue line) and a money market investment (green line).
Sources: Bloomberg, Invesco. Period: 23 July 2003 to 22 November 2016; 23 July 2003 = 100.

of cases. Moreover, it is often not easy to find option contracts that fit the needs of the portfolio – particularly when it comes to complex investment vehicles like the proposed multi-asset portfolio. Yet, both of these problems can be addressed by synthetically replicating the necessary European put option, which ultimately consists in dynamically adjusting the investment exposure of the multi-asset portfolio.[4]

Figure 2 charts the evolution of the synthetic put strategy over time. We note that the rate of investment (exposure) varies significantly, depending on the difference between the portfolio value and the strike price, as well as expected volatility.[5] Unlike the stop-loss strategy, exposure would have been reduced early enough in 2008 to avoid a massive drawdown. Yet, it was still at 44% when the floor was first breached in 2008; by the end of the year, the portfolio value would have been 4% below the floor value. This demonstrates one weakness of a synthetic put strategy, which also has the disadvantage of frequent portfolio reallocation. Nonetheless, the synthetic put strategy would have made far better use of the subsequent market recovery than the stop-loss strategy. Ultimately, performance would have matched that of the underlying multi-asset portfolio – with substantially less volatility and a lower maximum drawdown.

Figure 3
Performance and allocation of the CPPI strategy



The chart shows the performance of a conservative multi-asset portfolio using a CPPI strategy (dark blue line) in relation to the floor (purple line) over time. Exposure is calculated using the cushion (difference between the portfolio value and the floor; here: 95% of the initial annual portfolio value) and the multiplier that is based on daily risk forecasts of the historically simulated ES of the multi-asset portfolio (3%). For comparison, we have included the performance of the underlying conservative multi-asset strategy (light blue line) and a money market investment (green line).
Sources: Bloomberg, Invesco. Period: 23 July 2003 to 22 November 2016; 23 July 2003 = 100.

3. CPPI and related dynamic portfolio insurance strategies

Given the shortcomings of option-based portfolio insurance, an alternative can be found in a dynamic variant of the classic CPPI (constant proportion portfolio insurance[6]) strategy. First, we will examine the CPPI concept itself, before looking deeper into dynamic portfolio insurance.

3.1 CPPI

At the heart of the classic CPPI strategy is the so-called cushion C_t, i.e. the difference between the invested capital (or wealth), W_t and the net present value of the floor $NPV(F_T)$:

$$(1) \quad C_t = W_t - NPV\left(F_T\right)$$

In order to avoid a breach of the floor, the risky investment $E_t = e_t \times W_t$ (with investment exposure e_t) should be set such that:

$$(2) \quad C_t \geq W_t \times \text{MaxLoss}\left(W_t\right)$$
$$C_t \geq e_t \times W_t \times \text{MaxLoss}\left(risky\ asset\right)$$
$$E_t \leq \frac{C_t}{\text{MaxLoss}\left(risky\ asset\right)} = m \times C_t$$

In this context, the multiplier

$$m := \frac{1}{\text{MaxLoss}\left(risky\ asset\right)}$$

allows for a neat interpretation: it indicates how often a given cushion can be invested in the risky asset without breaching the floor assuming that the maximum loss assumption of the risky asset is not violated.

The classic CPPI strategy is based on a static multiplier – often reflecting a constant worst-case scenario. Figure 3 illustrates the performance and exposure of a CPPI strategy, which assumes a constant maximum overnight loss of 3%, which is equivalent to the historically simulated expected shortfall (ES) of the multi-asset portfolio. Although this very conservative position would have prevented catastrophic drawdowns during the financial market crisis, it would also have left significant return potential unused over the long term. This is reflected in the average investment exposure of just 70.2% – pushing annualized returns down a full 75 bp to a mere 3.14% p.a. (see table 1 at the end of the article).

3.2 DPPI

This is where dynamic proportion portfolio insurance (DPPI) proves its effectiveness. Instead of using a static multiplier, the risk budget adapts dynamically to changes in expected shortfall (ES). Exposure is set such that:

$$(3) \quad E_t \leq \frac{C_t}{\text{MaxLoss}_t\left(risky\ asset\right)} = m_t \times C_t$$

with the multiplier

$$m_t := \frac{1}{ES_t^{99\%}\left(risky\ asset\right)}$$

In this way, the exposure of the portfolio reacts to changes in the risk forecast – ensuring that it does not remain artificially low as a result of a constant conservative risk assumption. For this to work in practice, the risk model must be capable of quickly homing in on volatility spikes, and just as quickly readjusting to a normalization of market volatility. To this end, a Copula-GARCH model is extremely useful for forecasting ES (see box: Risk forecasting for dynamic portfolio insurance strategies).

We start by setting the exposure in accordance with equation (3). Figure 4 shows that, although the DPPI

strategy actively adjusts exposure, it fluctuates to a lesser degree than with the synthetic put. With the onset of the financial market crisis, exposure dropped to zero, so that the portfolio value at the end of 2008 was equal to the floor. Then, even with the V formation (steep decline followed by a rapid recovery) in early 2009, which is a major pitfall for portfolio insurance, the DPPI portfolio did not end up like the stop-loss in a "cash lock" within the money market. It participated in at least part of the subsequent recovery.

On the whole, the DPPI strategy actually delivered a marginal excess return compared with the pure

Box
Risk forecasting for dynamic portfolio insurance strategies

Modern risk modelling is guided by empirical patterns, which cannot be adequately captured using a conventional approach with an assumption of normal distributions. In particular, extreme events occur substantially more often than postulated by a normal distribution. Volatility and correlations are not constant, and volatility-clustering is not uncommon.

An effective method of understanding empirical risk is the Copula-GARCH model, as proposed by Patton (2006) or Jondeau and Rockinger (2006): the GARCH component measures the risk dynamics, while the copula estimation permits adequate modelling of the dependence structure.

Another matter to consider, in addition to the structure of the model itself, is the question of an appropriate risk measure. Whereas many risk management approaches rely on value-at-risk (VaR), portfolio insurance strategies naturally lend themselves to using expected shortfall (ES) to measure risk. In the case of VaR, it indicates the maximum possible loss at a given confidence level (usually 95% or 99%). However, VaR is silent with respect to the losses beyond the VaR threshold. Conversely, the ES measures the expected loss in the event of a VaR violation.

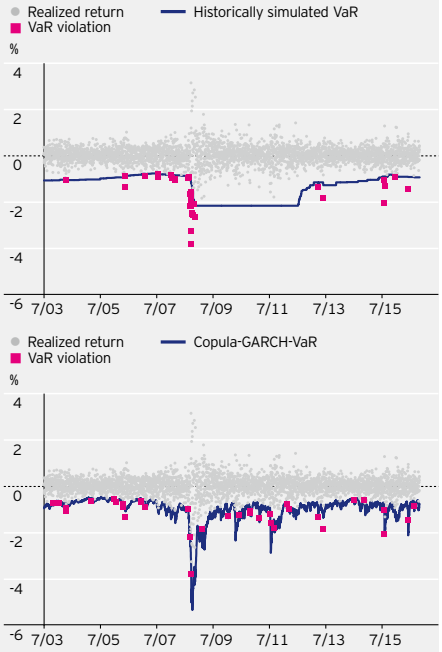
Validity of VaR and ES forecasts
The validity of Copula-GARCH risk forecasts can be demonstrated using various statistical tests. In order to have a sound basis for the estimated ES, the corresponding VaR quantile must be correctly specified. In a set of 260 forecasts of 1-day VaR (99% confidence) per year, there should theoretically be 2.6 violations. The upper panel of the chart shows a very simple VaR forecast as given by the empirical VaR over a sliding 1,000-day window. As expected, the majority of realized returns were higher than the forecasted VaR. In the sample period from July 2003 to November 2016, there were only 32 violations (pink dots) – which is nearly the same as the 35 expected (= 1% of 3.479).

An analysis of the VaR violations throughout time is sufficient to call into doubt the utility of the historically simulated VaR – given that nearly all of them occurred during the 2008 financial market crisis due to a latent underestimation of risk. Subsequently, the historically simulated VaR forecast was overly conservative, and there were no more violations for five years. Thus, a portfolio insurance strategy on this basis would have held investment exposure much too low over time.

This conclusion is confirmed by rigorous statistical testing. Using the unconditional coverage test (Kupiec, 1995), the historically simulated VaR does indeed deliver a conclusive number of violations over the entire period. But, based on the test for correct coverage and independence (Christoffersen, 1998) and the duration test (Christoffersen and Pelletier, 2004), it is clear that the violations are not independently occurring, but rather appear in clusters.

The lower panel of the chart shows the VaR forecast on the basis of the Copula-GARCH model, which is much more sensitive and quick to react to the prevailing risk environment. The 35 violations over the entire period are precisely in line with the theoretical expectation; moreover, their occurrence is markedly less clustered – as confirmed by the statistical tests. And: the ES estimator corresponding to the Copula-GARCH VaR quantile also passes the so-called "zero mean" test proposed by McNeil and Frey (2000), i.e. the excess losses are independently distributed around a mean of zero.

VaR-forecasts and realized returns of the multi-asset portfolio



The chart shows the daily VaR forecasts (blue line) and realized returns of the multi-asset portfolio (grey dots) over time. VaR violations are marked in pink. At a confidence level of 99%, a total of 35 violations are expected over the model period. Both historically simulated VaR (above) and Copula-GARCH VaR (below) exhibit the expected number of violations on average – but only under the Copula-GARCH VaR forecast are these violations independent and non-clustered. Sources: Bloomberg, Invesco. Period: 23 July 2003 to 22 November 2016.

multi-asset strategy (3.98% return; 4.69% volatility – see table 1 at the end of the article). Compared to the stop-loss and synthetic put, the maximum drawdown is significantly lower (by approx. 4 percentage points). Thus, the portfolio insurance can be achieved without the purchase or replication of an option, and can also be easily and flexibly adapted to accommodate changing investment demands.

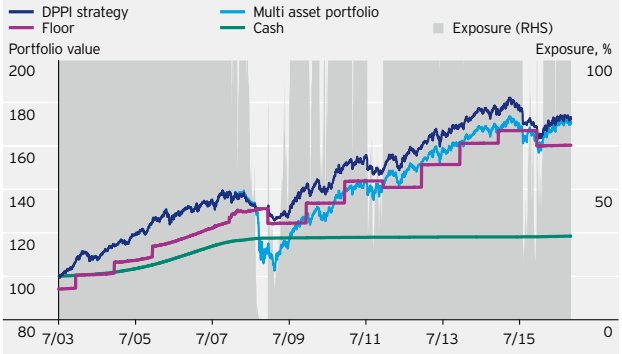
4. Dynamic portfolio insurance with a "ratchet floor": the TIPP

A more conservative alternative to the CPPI strategy is the so-called TIPP (time invariant portfolio protection) strategy. In essence, it complements the CPPI strategy by locking in a portion of gains achieved with the portfolio. The floor is "ratcheted-up" as soon as a new high is reached in portfolio value. Figure 5 shows the development of a dynamic TIPP strategy (dTIPP), based on the identical ES risk forecast as the DPPI strategy. Exposure over the entire period is roughly 10 percentage points lower than that of the DPPI strategy – a consequence of the floor always being closer to the portfolio value so that no additional cushion can be built up. This implies a clear reduction of returns vs. DPPI – but one that is less dramatic in risk-adjusted terms.

Conclusion

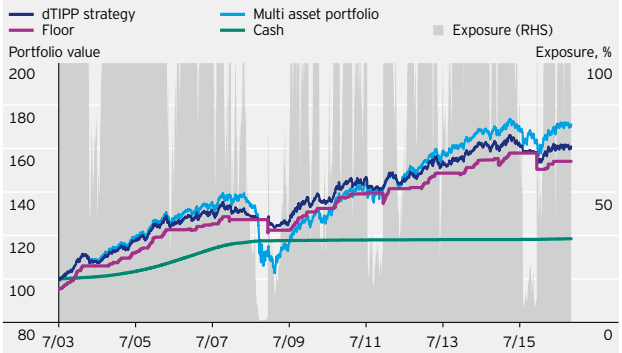
Our examination has shown that dynamic portfolio insurance could be useful in improving the risk-return profile of an investment (table 1). The most attractive alternative we have found was the DPPI strategy – an improvement on the classic CPPI strategy. Because DPPI works with a dynamic measure of risk, it adapts much more readily to the market environment than the CPPI approach with its constant multiplier. Moreover, in terms of the Sharpe ratio, maximum drawdown and investment exposure, the DPPI strategy outperformed the stop-loss, the synthetic put and the dTIPP strategy.

Figure 4
Performance and allocation of the DPPI strategy



The chart shows the performance of a conservative multi-asset portfolio using a DPPI strategy in relation to the floor over time. Exposure is calculated using the cushion (difference between the portfolio value and the floor; here: 95% of the initial annual portfolio value) and the multiplier (based on daily risk forecasting; here: Copula-GARCH 99%-ES). For comparison, we have included the performance of the underlying conservative multi-asset strategy and a money market investment.
Sources: Bloomberg, Invesco. Period: 23 July 2003 to 22 November 2016; 23 July 2003 = 100.

Figure 5
Performance and allocation of the dTIPP strategy



The chart shows the performance of a conservative multi-asset portfolio using a dTIPP strategy in relation to the floor over time. Exposure is calculated using the cushion (difference between the portfolio value and the floor; here: 95% of the initial annual portfolio value each year) and the multiplier (based on daily risk forecasting; here: Copula-GARCH 99%-ES). The key characteristic of the dTIPP strategy lies in the "ratcheting-up" of the floor (95%) once a new high is achieved. For comparison, we have included the performance of the underlying conservative multi-asset strategy and a money market investment.
Sources: Bloomberg, Invesco. Period: 23 July 2003 to 22 November 2016; 23 July 2003 = 100.

Table 1
Figures for the conservative multi-asset portfolio with and without portfolio insurance

	Multi asset portfolio	Money market investment	Stop loss	Synthetic put	DPPI	dTIPP
Return p.a. (%)	3.89	1.23	3.65	3.89	3.98	3.45
Volatility p.a. (%)	6.40	0.11	5.04	4.71	4.69	4.05
Sharpe ratio	0.42	0.00	0.48	0.56	0.59	0.55
Maximum drawdown (%)	-27.16	0.00	-14.49	-14.28	-10.43	-8.82
Exposure (%)	100.00	0.00	91.09	89.58	90.37	80.38

The table shows the performance figures for the various portfolio insurance strategies in combination with a multi asset portfolio: stop-loss, synthetic put, constant proportion portfolio insurance (CPPI), dynamic proportion portfolio insurance (DPPI) and dynamic time invariant portfolio protection (dTIPP). In each calendar year, a floor of 95% of the initial portfolio value is targeted. For comparison, we have included the performance figures for the underlying conservative multi-asset strategy and a money market investment. Sources: Bloomberg, Invesco. Period: 23 July 2003 to 22 November 2016.

Bibliography

- Black, F. and Jones, R. (1987). Simplifying portfolio insurance. Journal of Portfolio Management 14, 48-51.
- Black, F. and Jones, R. (1988). Simplifying portfolio insurance for corporate pension plans. Journal of Portfolio Management 14, 33-37.
- Christoffersen, P. (1998). Evaluating interval forecasts. International Economic Review 39, 841-862.
- Christoffersen, P. and Pelletier, D. (2004). Backtesting value-at-risk: A duration-based approach. Journal of Financial Econometrics 2(1), 84-108.
- Dichtl, H., and Drobetz, W. (2011). Portfolio insurance and prospect theory investors: Popularity and optimal design of capital protected financial products. Journal of Banking and Finance 35, 1683-1697.
- Estep, T. and Kritzman, M. (1988). TIPP: Insurance without complexity. Journal of Portfolio Management 14, 38-42.
- Jondeau, E. and Rockinger, M. (2006). The Copula-GARCH model of conditional dependencies: An international stock market application. Journal of International Money and Finance 25, 827-853.
- Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. The Journal of Derivatives 3(2), 73-84.
- McNeil, A. J., and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. Journal of Empirical Finance 7(3), 271-300.
- Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. International Economic Review 47(2), 527-556.
- Perold, A. F. (1986). Constant proportion portfolio insurance. Working paper, Harvard Business School.
- Perold, A. F. and Sharpe, W. F. (1988). Dynamic strategies for asset allocation. Financial Analysts Journal 44, 16-27.
- Rubinstein, M. (1985). Alternative paths to portfolio insurance. Financial Analysts Journal 41, 42-52.
- Rubinstein, M. and Leland, H. E. (1981). Replicating options with positions in stock and cash. Financial Analysts Journal 37, 63-72.

About the authors

Dr. Martin Kolrep
Senior Portfolio Manager,
Invesco Quantitative Strategies
Dr. Martin Kolrep is involved in the development of client solutions and the management of multi asset strategies.

Dr. Harald Lohre
Senior Research Analyst,
Invesco Quantitative Strategies
Dr. Harald Lohre develops quantitative models to forecast risk and return used in the management of multi-asset strategies.

David Happersberger
PhD Candidate Lancaster University,
Invesco Quantitative Strategies
As part of a joint research initiative between Lancaster University and Invesco Quantitative Strategies, David Happersberger is pursuing post-graduate work on practice-oriented issues of financial market econometrics. At the same time, he is actively supporting the transfer of research results into the multi-asset investment process.

Notes

1 Throughout the article and in all figures and tables, the multi-asset data set consists of the following series (portfolio weights are given in parentheses): EuroStoxx 50 Future (5.8%), FTSE 100 Index Future (5.8%), S&P500 Future (15%), Nikkei 225 Future (6.7%), Euro-Bund Future (16.7%), US 10YR Note Future (16.7%), JPN 10Y Bond Future (16.7%), S&P GSCI Crude Oil (3.5%), S&P GSCI Gold (5.8%), Bloomberg Agriculture Subindex (3.8%), Bloomberg Copper Subindex (3.5%). For money market investments we use the 3-month US Treasury bill. All asset returns are in local currency. Portfolio returns and values are computed from the perspective of an U.S. investor who is hedging any currency exposure. Furthermore, all simulations in this article are provided for illustrative purposes only and are subject to limitations. Unlike actual portfolio outcomes, the model outcomes do not reflect actual trading, liquidity constraints, fees, expenses, taxes and other factors that could impact future returns.

2 Table 1 at the end of the article shows the performance figures for all of the strategies presented.

3 A European option can only be exercised at expiry (unlike an American option, which can be exercised at any time during its term).

4 Delta, i.e. the sensitivity of the synthetic put option to changes in the underlying, is determined using the classic Black-Scholes model. The strike price is set to reflect the desired floor value (Rubinstein and Leland, 1981; Dichtl and Drobetz, 2011).

5 A volatility forecast is necessary to determine delta and we build on a Copula-GARCH model (see box: Risk forecasting for dynamic portfolio insurance).

6 For more on CPPI strategies, cf. Perold (1986), Black and Jones (1987, 1988), Perold and Sharpe (1988).

A.2. Evaluating Risk Mitigation Strategies

Evaluating risk mitigation strategies

By Dr. Harald Lohre, David Happersberger and Erhard Radatz

In brief
Risk mitigation strategies seek to create an asymmetric risk-return profile. But benchmarking against the underlying investment is not a valid approach given the potentially stark difference in risk profiles. We discuss how to appropriately calibrate and assess portfolio insurance strategies based on the ensuing return distribution to better fit a given client's risk preferences.

In light of the sustained low yield environment, investors have increasingly taken on more risk to meet their return targets. Yet, their ability to cope with higher risk is limited, which is what makes strict risk management and suitable portfolio insurance techniques so important.

In a previous article[1], we discussed a variety of risk mitigation approaches for a given underlying investment strategy. In particular, we investigated portfolio insurance strategies ranging from static stop-loss techniques to option-based strategies and dynamic portfolio insurance techniques. We concluded that an active portfolio insurance strategy based on a dynamic risk forecast is a cost-effective way to limit a portfolio's maximum loss at a high probability.

In this article we go further and explain how to calibrate such a strategy to individual risk preferences. Since portfolio insurance is meant to accommodate conservative clients' need for an asymmetric return profile, adding a risk overlay ultimately boils down to reshaping the portfolio return distribution. Essentially, the aim is to significantly reduce the probability of suffering from severe tail events while sacrificing some of the underlying strategy's upside potential.

The mechanics of dynamic portfolio insurance
Our preferred dynamic portfolio insurance strategy is rooted in the classic CPPI (constant proportion portfolio insurance[2]) strategy. It typically sets the exposure in a given risky underlying in such a way that a chosen floor level is not breached within a specified investment period. Thus, it is essential to

closely monitor the cushion C_t that represents the difference between the invested wealth W_t and the net present value of the floor $NPV(F_T)$:

(1) $C_t = W_t - NPV(F_T)$

To effectively protect the floor,

$C_t \geq W_t * \text{MaxLoss}(W_t)$

must hold true. With the investment exposure e_t and the corresponding risky investment $E_t = e_t * W_t$ the above formula can be restated as

(2) $$C_t \geq e_t * W_t * \text{MaxLoss}\left(\text{risky asset}\right)$$
$$\Leftrightarrow E_t \leq \frac{C_t}{\text{MaxLoss}\left(\text{risky asset}\right)} = m * C_t$$

This reformulation brings in the notion of the CPPI multiplier m. The multiplier indicates how often the cushion can be invested in the risky underlying without breaching the floor provided the maximum loss assumption holds.

To be on the safe side, one could impose a static multiplier derived from a worst-case risk estimate. But, as we demonstrated in the previous article, such a conservative estimate would severely undermine participation in the underlying. To remedy this issue, we put forward the use of a dynamic forecast of maximum loss. That is, we make use of a dynamic multiplier

$$m_t := \frac{1}{ES_t^{99\%}\left(\text{risky asset}\right)}$$

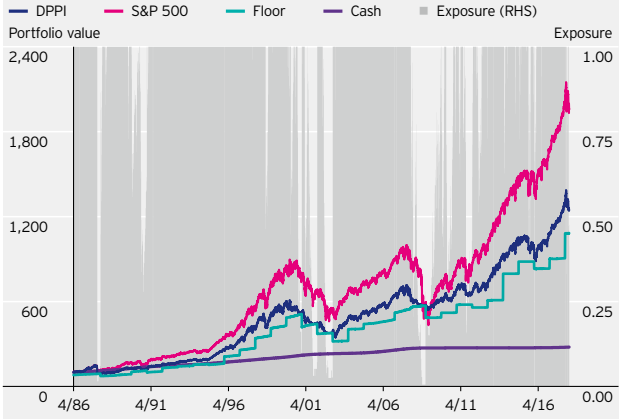
labelling this type of risk mitigation DPPI (dynamic proportion portfolio insurance). In this setting, the risk budget and investment exposure dynamically adjust to changes in the estimated expected shortfall (ES) forecast. In particular, participation in the underlying is higher in calmer risk environments, while a pick-up in risk leads to a reduction of investment exposure. Obviously, it is essential to rely on risk estimates that allow for timely modelling of tail risk within the portfolio return distribution.

Panel (a) of figure 1 charts the mechanics and evolution of a DPPI strategy applied to an S&P 500 underlying at an 85% floor level.[3] The dynamic adjustment of the time-varying multiplier m_t follows the expected shortfall forecast derived from a GARCH(1,1)-model. Clearly one can appreciate the role and interaction of floor and multiplier: if the underlying investment is far above the floor, the DPPI tends to have a high investment exposure more or less independent of the risk estimate. With less cushion, the DPPI strategy is more sensitive to risk changes, potentially leading to a complete de-investment.

Over the course of the 32-year backtest, we only observe a few periods of de-investment, of which only four ended in a cash-lock position. While one seeks to avoid cash-lock through the adaptive positioning based on the risk forecast, the success of this approach depends on the specific nature of the corresponding market setbacks. For instance, the minimum daily return of the S&P 500 (-28.6% on 19 October 1987) fully consumed a seemingly comfortable cushion of more than 25%, and induced

Figure 1
Performance and allocation of the DPPI strategy

Panel (a)



Panel (b)	S&P 500	Money market	DPPI
Return p.a. (%)	9.23	3.20	7.82
Volatility p.a. (%)	19.37	0.22	14.41
Sharpe ratio	0.31	0.00	0.32
Maximum drawdown (%)	-61.17	0.00	-45.80
Expected shortfall 99% (%)	-5.09	0.00	-3.66
Mean exposure (%)	100.00	0.00	86.18

The chart in Panel (a) shows the performance of an equity portfolio (S&P 500) using a DPPI strategy (blue line) in relation to the floor (green line) over time. Exposure is calculated using the cushion (difference between the portfolio value and the floor; here: 85% of the initial annual portfolio value) and the multiplier (based on daily risk forecasting; here: GARCH 99%-ES). For comparison. we have included the performance of the underlying S&P 500 (pink line) and a money market investment (purple line). Panel (b) shows the corresponding performance measures.
Period: 9 April 1986 to 9 April 2018; 9 April 1986 = 100.
Sources: Bloomberg. Invesco. This is simulated past performance and past performance is not a guide to future returns.

switching from a 100% investment exposure to cash-lock in just one day. However, in other periods of weak S&P 500 performance, market drawdowns evolved more gradually, allowing the DPPI portfolio time to de-invest and re-invest. The last complete de-investment occurred during the global financial crisis. In the aftermath, interest rates have come down, implicitly elevating the floor level. During high volatility episodes in the equity market, we could observe similar de-risking events within the last decade. Yet these only served to reduce portfolio volatility given quick recoveries in the S&P 500.

Examining the whole sample path, we learn that the DPPI strategy was indeed able to mitigate downside risk. Compared to the underlying investment, the maximum drawdown decreases by approximately 15 percentage points, volatility by 5 percentage points and expected shortfall by 1.5 percentage points under the DPPI strategy (cf. panel (b)). Although these reductions come at the cost of some return potential – the DPPI portfolio earns 141bps less than the underlying – , risk-adjusted measures are in favour of the DPPI strategy.

Designing DPPI strategies
The preceding example illustrates an important caveat in evaluating a given DPPI strategy, namely, its inherent path dependency. To avoid assessing the strategy based on just one historical path, we rather simulate a large number of alternative price paths and apply the given DPPI-setup. Hence, instead of just one risk and return combination, we obtain a full return distribution.[4] Figure 2 shows portfolio return distributions of yearly returns based on 5,000 simulations, for the portfolio fully invested in the (simulated) underlying S&P 500 as well as for the corresponding DPPI strategy with an 85% floor. The risk estimates required for computation of the dynamic multiplier for the DPPI strategy are based on a simple GARCH(1,1)-model. This model captures the main empirical characteristics of asset returns, such as time-varying volatility, fat tails and volatility clustering.[5]

We observe a left-skewed distribution for the simulated equity underlying. There is tail risk with a non-negligible probability of yearly returns being less than -15%. Applying DPPI results in significantly less tail risk. Yet, one has to note that there is still a small probability of breaching the floor level given that the strategy is adjusted at discrete (daily) intervals.

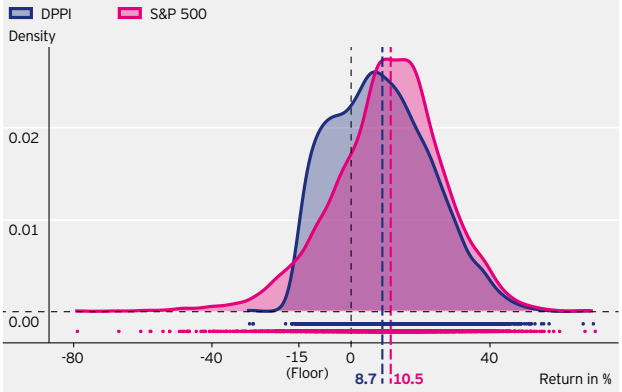
More importantly, however, figure 2 clearly demonstrates that tail risk reduction, on average, comes at the cost of reduced upside potential. While the historical backtest might suggest an outperformance of the DPPI strategy relative to its underlying, the simulated return distributions more readily articulate that portfolio insurance actually comes at an implicit insurance premium.

Judging by the mean yearly return difference of the two distributions, this premium would amount to some 1.8% (10.5% – 8.7% = 1.8%). At this premium, we can expect to avoid severe tail risk events, 29 of which could be worse than -40% (as simulated in our block-bootstrap analysis).

In the same vein, this framework clarifies the consequences of certain design choices (such as underlying and floor level) for the client's expected portfolio return distribution. For instance, a common theme is that floor levels are set too tight relative to the riskiness of the underlying. Put differently, investors often favour riskier underlyings to achieve certain return targets. Yet, absent a higher risk budget, a riskier strategy will frequently be prevented from breathing freely given that the available cushion is easily consumed. This leads to frequent de-investments or even cash-lock situations triggered by the DPPI mechanism.

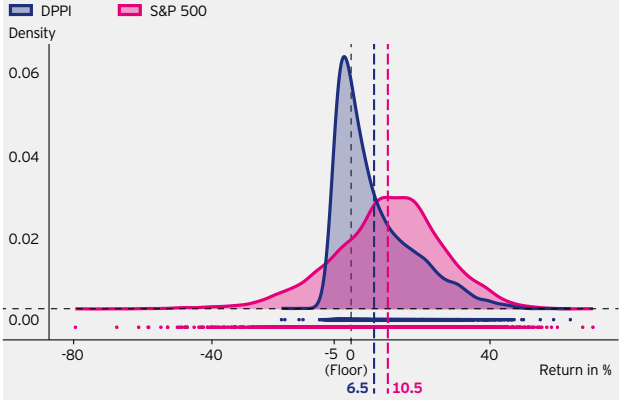
To illustrate this issue, figure 3 shifts the floor level from 85% to 95%. As a result, the DPPI return distribution is massively distorted with a lot of return realizations around -5%, i.e. rather close to the floor level. Obviously, this is reminiscent of the fact that, under a too tight floor level, the DPPI strategy frequently de-invests or ends up in cash-lock, disabling it from participating to a meaningful extent in equity markets. The corresponding statistics in table 1 show that the mean exposure reduces to 61%, leading to a significantly lower mean return (6.5% vs. 8.7%) and lower Sharpe ratio (0.24 vs. 0.35) when we shift the floor level from 85% to 95%.[6]

Figure 2
Comparing return distributions



The chart shows the distribution of block-bootstrapped yearly returns (M = 5,000 simulations) of the DPPI portfolio (blue shade) and the one of a pure buy-and-hold portfolio invested in the corresponding simulated S&P 500 (pink shade). The floor level of the DPPI strategy is 85%. Below the two density plots we have added the corresponding support and the mean levels of the return distributions.
Sources: Bloomberg, Invesco.

Figure 3
Comparing return distributions: tight floor levels



The chart shows the distribution of block-bootstrapped yearly returns of the DPPI portfolio (blue shade) and the one of a pure buy-and-hold portfolio invested in the corresponding simulated S&P 500 (pink shade). The floor level of the DPPI strategy is 95%. Below the two density plots we have added the corresponding support and the mean levels of the return distributions.
Sources: Bloomberg, Invesco.

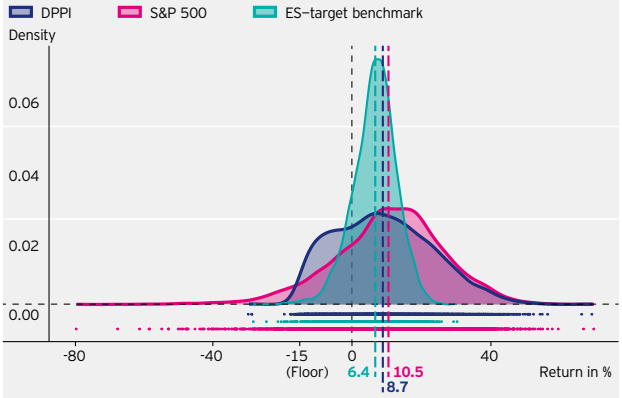
An alternative benchmark for DPPI strategies
Given the potential for considerable reshaping of the portfolio return distribution through portfolio insurance, it is evident that DPPI should not be benchmarked relative to its underlying. As an alternative, we construct a benchmark with similar risk characteristics. Because we are comparing an asymmetric distribution, a symmetric risk measure like volatility is not viable. Given that risk-averse investors are more concerned about the tails of a distribution, we will base our analysis on the expected shortfall (ES), using a 99% confidence level.

Given the potential for considerable reshaping of the portfolio return distribution through portfolio insurance, it is evident that DPPI should not be benchmarked relative to its underlying.

While there are numerous ways to create a benchmark with a given ES, we opt for an easy and replicable solution. We add cash to the underlying S&P 500 investment to scale down its risk to the pre-defined ES limit of 15%, corresponding to the floor level of the DPPI strategy. We will call this portfolio "ES-target benchmark".[7] As a result, we are comparing two different strategies with similar risk profiles (as defined by their 99%-ES): a portfolio dynamically allocating between cash and the risky underlying (DPPI portfolio) and a static mix of cash and underlying that has an ES similar to the DPPI portfolio (ES-target portfolio).

To achieve an ES of 15% over the sample period, a 39/61 mix of S&P 500 and cash is needed to compute the ES-target benchmark. In figure 4, the ensuing portfolio return distribution is contrasted to that of the underlying S&P 500 and the DPPI strategy with a floor level of 85%. Obviously, the ES-target benchmark return distribution is a compressed version of the underlying S&P 500 return distribution. Most importantly, although its mean return is smaller than the DPPI (6.4% vs. 8.7%), there is still a small probability of significant tail events attached to this strategy (cf. figure 4 and table 1).

Figure 4
Comparing return distributions



The chart shows the distribution of block-bootstrapped yearly returns of the DPPI portfolio (blue shade) and the one of a pure buy-and-hold portfolio invested in the corresponding simulated S&P 500 (pink shade). The floor level of the DPPI strategy is 85%. The third return distribution applies to a partial investment in the underlying that adds cash such that the average risk level (in terms of the 99%-ES) conforms to the floor level of the DPPI strategy (green shade). Below the density plots we have added the corresponding support and the mean levels of the return distributions.
Sources: Bloomberg, Invesco.

Conclusion
Many investors tend to benchmark the performance of their portfolio insurance strategy vis-à-vis the return of the underlying portfolio. Instead, we suggest the ES-target benchmark strategy. This tail risk-adjusted alternative transforms the underlying's return distribution to better fit the client's risk preferences. Of course, investigating the ensuing portfolio return distributions based on block-bootstrap resampling sheds even more light on the effects of a given portfolio insurance application. We seek to apply this methodology in a future article to investigate the merits of different underlyings in a portfolio insurance framework.

Table 1
Performance of DPPI strategies vis-à-vis the ES-target benchmark

	S&P 500	Money market	DPPI (95% Floor)	DPPI (85% Floor)	ES-Target
Return p.a. (%)	10.49	3.81	6.45	8.71	6.43
Volatility p.a. (%)	15.95	0.96	10.93	14.09	6.30
Sharpe ratio	0.42	0.00	0.24	0.35	0.42
Maximum drawdown (mean, %)	-14.98	0.00	-8.09	-11.77	-3.52
Expected shortfall 99% (%)	-43.83	1.42	-7.85	-16.83	-15.00
Mean exposure (%)	100.00	0.00	61.14	87.28	39.18

The table shows performance measures of a block-bootstrapped DPPI strategy based on an equity portfolio (S&P 500) using different floor levels (85% and 95%). For comparison, we have included the performance measures of an ES-target strategy, targeting the same level of expected shortfall as the DPPI, alongside the underlying S&P 500 and a money market investment. Reported are the mean return, volatility, Sharpe ratio and expected shortfall of the simulated yearly returns, as well as the mean of the maximum drawdowns (which are computed for each simulated path) and mean exposure.
Period: 9 April 1986 to 9 April 2018; 9 April 1986 = 100.
Sources: Bloomberg, Invesco. This is simulated past performance and past performance is not a guide to future returns.

Bibliography

- Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold (2013). Financial risk measurement for financial risk management, in Handbook of the Economics of Finance, ed. by G. M. Constantinides, M. Harris, and R. M. Stulz 2(17), 1127–1220.
- Ardia, D., Boudt, K., and Wauters, M. (2016). Smart beta and CPPI performance. Finance, 37(3), 31-65.
- Black, F. and Jones, R. (1987). Simplifying portfolio insurance. Journal of Portfolio Management 14(1), 48–51.
- Black, F. and Jones, R. (1988). Simplifying portfolio insurance for corporate pension plans. Journal of Portfolio Management 14(4), 33–37.
- Happersberger, D., H. Lohre, and I. Nolte (2018). Estimating portfolio risk for tail risk protection strategies. Working paper, Lancaster University Management School.
- Perold, A. F. (1986). Constant proportion portfolio insurance. Working paper, Harvard Business School.
- Perold, A. F. and Sharpe, W. F. (1988). Dynamic strategies for asset allocation. Financial Analysts Journal 44, 16–27.
- Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. Journal of the American Statistical Association 89(428), 1303–1313.

About the authors

Dr. Harald Lohre
Senior Research Analyst,
Invesco Quantitative Strategies and Visiting Research Fellow EMP/Lancaster University
Dr. Harald Lohre develops quantitative models to forecast risk and return used in the management of multi-asset strategies.

David Happersberger, PhD Candidate
Lancaster University and Invesco Quantitative Strategies
As part of a joint research initiative between Lancaster University and Invesco Quantitative Strategies, David Happersberger is pursuing postgraduate work on practice-oriented issues of financial market econometrics. At the same time, he is actively supporting the transfer of research results into the multi-asset investment process at Invesco Quantitative Strategies.

Erhard Radatz
Portfolio Manager,
Invesco Quantitative Strategies
In his role, Erhard Radatz manages multi-asset portfolios that include the elements factor-based investing, active asset allocation and downside risk management.

Notes

1 See Theory and practice of portfolio insurance, Risk & Reward #2/2017.
2 For more on CPPI strategies, cf. Perold (1986), Black and Jones (1987, 1988), Perold and Sharpe (1988).
3 Throughout the article, and in all figures and tables, we employ the S&P 500 Future as equity investment. For money market investments we use the 3-month US Treasury bill. All asset returns are in local currency. All simulations in this article are provided for illustrative purposes only and are subject to limitations. Unlike actual portfolio outcomes, the model outcomes do not reflect actual trading, liquidity constraints, fees, expenses, taxes or other factors that could impact future returns.
4 In simulating alternative price paths, we use the stationary block-bootstrap of Politis and Romano (1994). We follow Ardia, Boudt and Wauters (2016) in that block lengths are drawn from a geometric distribution with a minimum block length of one day and an average of 15 days.
5 For more on GARCH models, cf. Andersen et al. (2013).
6 As is common in academic literature, the annualized returns, volatilities, and Sharpe ratios shown in Table 1 are based on the 5,000 annual returns from the simulations. So, given the different frequencies, it is not surprising that the historical volatilities shown in Panel (b) of Figure 1 and that are based on historical daily returns, are slightly higher. This effect is exacerbated because, of course, the simulation paths are relatively rare in containing the extreme historical returns realizations, and thus there is a corresponding relativization.
7 See Happersberger, Lohre and Nolte (2018) for an empirical study of ES-target strategies in the context of tail risk protection.

The outputs of the assumptions are provided for illustration purposes only. Unlike actual portfolio outcomes, the model outcomes do not reflect actual trading, liquidity constraints, fees, expenses, taxes and other factors that could impact future return.

A.3. The Use of Equity Factor Investing for Portfolio Insurance

The use of equity factor investing for portfolio insurance

By Dr Harald Lohre, David Happersberger and Alexandar Cherkezov

In brief
Equity investments promise high expected returns, but not many investors can tolerate the associated risks. A possible solution may be to complement the equity strategy with a portfolio insurance element which ideally reduces the equity exposure whenever necessary to prevent the overall strategy from breaching a pre-defined floor. Based on a block-bootstrap methodology, we show that the choice of equity underlying is key in this context; in particular, low-volatility underlyings are to be preferred, with other multi-factor propositions forming suitable alternatives when considering additional elements of dynamic risk management.

Portfolio insurance techniques such as CPPI (constant proportion portfolio insurance) are commonly used to protect investments from downside risk.[1] This risk is obviously pronounced in the case of pure equity investments. We examine the interaction of CPPI with different equity underlyings, including standard market cap index, multi-factor and low-volatility investments.

To evaluate CPPI strategies for different equity underlyings, the usual way is to consider their historical performance. Yet, given the inherent path dependency of CPPI, any conclusion from this would be mostly anecdotal. Instead, we have earlier[2] suggested a block-bootstrap methodology which utilizes historical returns to simulate a large number of consistent alternative price paths and CPPI outcomes. Rather than evaluating just one price path, we base our analysis on the overall portfolio return distribution associated with a given portfolio insurance underlying. While the initial CPPI analysis is based on a static assumption for overnight risk (i.e. a constant multiplier) we go on to look for the incremental value of

including volatility targeting and dynamic risk forecast elements rendering the dynamic portfolio insurance dynamic (DPPI).

Rather than evaluating just one price path, we base our analysis on the overall portfolio return distribution associated with a given portfolio insurance underlying.

CPPI in a nutshell

For a given investment period, a CPPI[3] strategy seeks to respect a pre-specified floor by actively managing the exposure to the risky underlying. A key ingredient is the cushion C_t – i.e. the difference between the invested wealth W_t and the net present value of the floor $NPV(F_T)$:

$$(1) \quad C_t = W_t - NPV\left(F_T\right)$$

To maintain the floor,

$$(2) \quad C_t \geq W_t * \text{MaxLoss}\left(W_t\right)$$

must hold. Introducing the investment exposure e_t the associated risky investment can be written as $E_t = e_t * W_t$ so that the above condition (2) translates to

$$(3) \quad C_t \geq e_t * W_t * \text{MaxLoss}\left(risky\ asset\right)$$

$$\Leftrightarrow E_t \leq \frac{C_t}{\text{MaxLoss}\left(risky\ asset\right)} = m * C_t$$

This reformulation introduces a further key element: the CPPI multiplier m. It can be interpreted as the number of times the cushion can be invested in the risky underlying without risking a breach of the floor (provided the maximum loss assumption holds). To play it safe, a static multiplier derived from a worst-case risk estimate could be imposed. In our initial analysis of CPPI for equity style underlyings, we have chosen a constant multiplier of 6, which corresponds to an overnight risk assumption of 16.7%.[4]

Equity investing with style

Equity investments often closely follow broad market cap-weighted indices. Yet there are investment styles that differ from simple index investing, such as the popular styles value and momentum. For instance, a value investor would prefer stocks that are relatively cheap according to some measure of intrinsic value and would avoid relatively expensive stocks. While a value investor ultimately relies on stocks reverting to their fundamental value, a momentum investor would bet on the stocks' recent price momentum continuing. He would therefore be actively chasing recent winner stocks while cutting recent loser stocks.

These two investment philosophies are particularly common amongst quantitative factor-oriented managers. Alongside value and momentum, there are many more stock characteristics deemed relevant in explaining the cross-section of equity returns. For the subsequent analysis, we are particularly interested in capturing the most salient equity styles and additionally consider a "quality" style as well as defensive "low-volatility" style. While quality would favour companies with healthy balance sheet ratios and/or sustainable investment and financing activities, the low-volatility style seeks to improve a portfolio's risk-adjusted returns by avoiding highly volatile stocks.

Table 1 illustrates the performance of these various equity style investments for a European investment universe over the period 31 October 2006 to 31 May 2018.[5] As the sample period begins with the onset of the global financial crisis (GFC), the overall equity index performance is moderate. The MSCI Europe returned 3.25% p.a. at 19.4% annualized volatility while suffering a maximum drawdown of -58.5% over the course of the GFC. Yet the style returns differ considerably, ranging from 1.29% (value) to 6.53% (momentum). Notably, value investing was the most risky style over the sample period in terms of volatility (21.7%) and maximum drawdown (-65.1%). Quality and minimum volatility investments have been more resilient, as characterized by maximum drawdowns of -46.8% (quality) and -50.5% (minimum volatility).

Table 1
Performance of various equity style investments

	Index	Cash	Value	Momentum	Quality	QMV	Min-Vol	Active Low-Vol
Return p.a. (%)	3.25	0.84	1.29	6.53	6.22	4.88	4.33	5.87
Volatility p.a. (%)	19.4	0.1	21.7	18.6	18.4	18.3	14.9	16.0
Sharpe ratio	0.12		0.02	0.31	0.29	0.22	0.24	0.32
Maximum drawdown (%)	-58.5		-65.1	-54.9	-46.8	-55.6	-50.5	-46.3

The table shows performance measures of equity style investments; for Index we use the MSCI Europe. Value, Momentum and Quality are the respective MSCI Europe Value, MSCI Europe Momentum and MSCI Europe Quality indices. QMV represents an equally-weighted combination of Quality, Value and Momentum based on the corresponding MSCI indices. Minimum volatility is the MSCI Europe Min Vol index. All MSCI indices give net total returns in EUR. Active Low-Vol is based on backtested returns of an integrated multi-factor equity portfolio optimized according to quality, momentum and value signals but targeting a considerably lower risk than the market. Cash returns are based on EONIA. Reported are the annualized return and volatility figures, the corresponding Sharpe ratios and maximum drawdowns. Sources: MSCI, Bloomberg, Deutsche Bundesbank. Period: 31 October 2006 to 31 May 2018. **This is simulated past performance and past performance is not a guide to future returns.**

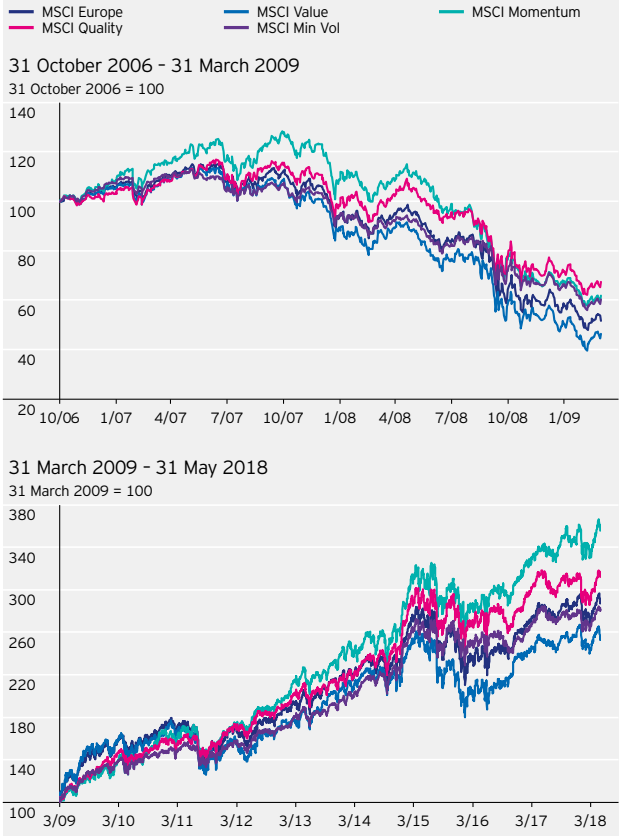
Nevertheless, quality investing comes with an annualized volatility of 18.4%, whereas minimum-volatility investing actually has the smallest realized volatility of 14.9%, reducing volatility by at least 20%.

While these figures refer to the full sample performance, it is worthwhile to investigate equity style performance in two sub-periods. We thus divide the sample into two before and after investigate March 2009, when global equity markets reached their lows during the global financial crisis. In the volatile first sub-period, we find that quality and low-volatility faired particularly well compared to value which performed even worse than the market index (figure 1). Interestingly, value was also lagging in the subsequent bull market, again finishing last in the league table of equity style factors. Momentum was the best performing style from March 2009 onwards, with quality coming in second place. Most interestingly, we find that minimum-volatility exhibited index-like returns, yet at a lower volatility.

In quantitative investing, it is common to combine different investment styles to create a more diversified multi-factor portfolio.[6] In that regard, a typical combination would include quality, momentum and value to obtain a core equity proposition with similar risk characteristics as the market index but potentially better returns. Indeed, a simple equal weighting of these three styles (labelled QMV in table 1) would have outperformed the MSCI Europe by 1.63 percentage points p.a.

To further exploit the notion of defensive investing, we also consider an integrated multi-factor approach that optimizes an equity portfolio according to quality, momentum and value signals, but targeting a considerably lower risk target than the market (labelled Active Low-Vol in table 1). Such an active low-volatility proposition would indeed have been highly attractive, with a 5.87% return at 16.0% volatility (i.e. a Sharpe ratio of 0.32). Moreover, its maximum drawdown is even less than that of the quality style investment (-46.3%).

Figure 1
Equity style investments over time



The charts show performance of equity style investments over time; as index we use the MSCI Europe. Value, Momentum and Quality are the respective MSCI Europe Value, MSCI Europe Momentum and MSCI Europe Quality indices. Minimum-volatility is the MSCI Europe Min Vol. All MSCI indices give net total returns in EUR. Cash is EONIA.
Sources: MSCI, Bloomberg. Data as at 31 May 2018. **Past performance is not a guide to future returns.**

Factor investing and CPPI

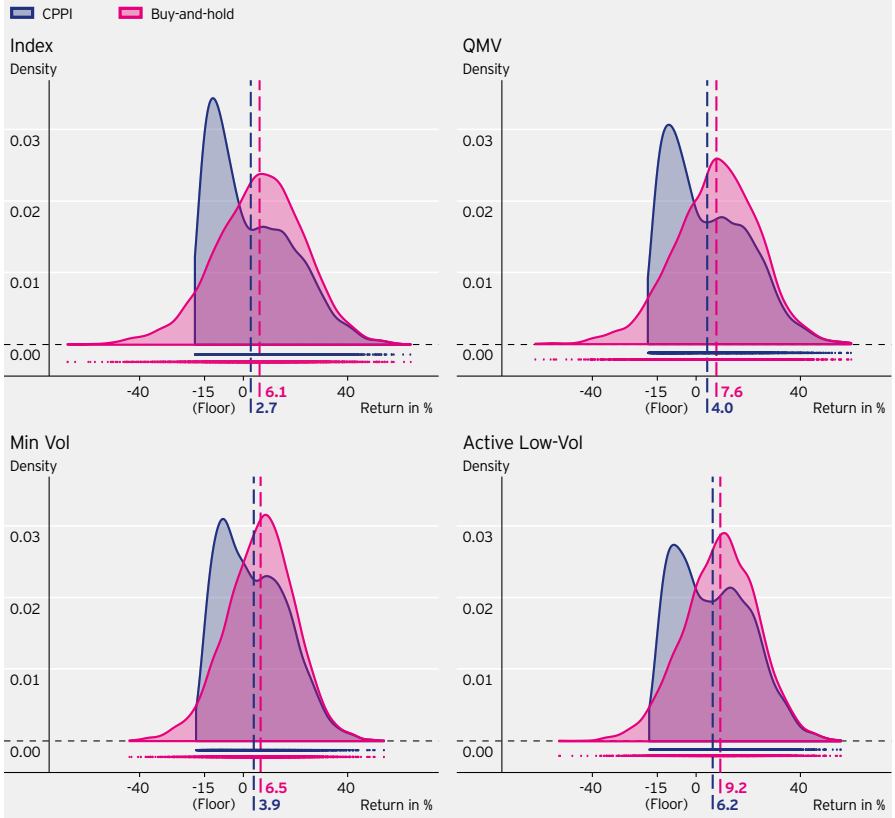
In light of the stark differences in equity style performance, the corresponding CPPI strategies may also differ. Yet the bulk of the CPPI literature focuses on index investments as the equity underlying of choice. A notable exception is Ardia, Boudt and Wauters (2016), who provide a thorough treatment of the topic in question. In particular, they carefully examine CPPI strategies based on different equity underlyings, including standard market cap, fundamental and low-volatility weightings.

We follow Ardia, Boudt and Wouters (2016) in analyzing the equity style investments from the preceding section in the context of CPPI strategies. As in a previous article[7], we do not base our analysis on the historical CPPI performance but on 5,000 block-bootstrap samples.[8] Given the inherent path dependency of CPPI, this setup is a meaningful improvement over standard analyses, as we can assess the probable portfolio return distribution of a given equity factor underlying.

Figure 2 shows the results for a floor of 85% and a static multiplier of 6. As for the chosen equity underlyings, we focus on the most relevant, i.e. index, multi-factor core (QMV), minimum-volatility and an active low-volatility investment.[9] Needless to say, all equity underlyings exhibit significant tail risk, but this is less pronounced for the two low-volatility propositions. Interestingly, the post-CPPI return distributions are quite different across the board.

To allow for a direct comparison of the return distributions, we have merged these into one chart (figure 3): for the index underlying, CPPI produces a relatively large number of outcomes rather close to the floor, for QMV this effect is less pronounced. Minimum-volatility or active low-volatility underlyings better transform the tail shape of the ensuing CPPI return distribution. These first-glance conclusions from figure 3 are by and large backed by the statistics in table 2: panel B for the static CPPI strategy clearly supports the above ranking from active low-volatility down to index investments, in terms of return, Sharpe ratio and Calmar ratio.

Figure 2
Factor investing and CPPI



The chart shows the distribution of block-bootstrapped yearly returns of the CPPI portfolio (blue shading) and that of a pure buy-and-hold portfolio invested in the simulated equity underlying (pink shading). The floor level of the CPPI strategy is 85%. Below the two density plots we have added the corresponding support and the mean levels of the return distributions. The upper left is for the index investment, the upper right is for the equally weighted multi-factor investment in quality, momentum and value (QMV), the lower left is for minimum-volatility and the lower right is for the active low-volatility strategy.
Sources: MSCI, Bloomberg, Invesco.

Introducing a volatility target in equity factor underlyings

Ex ante, low-volatility underlyings are expected to outperform given the negative vega of the CPPI strategy. As the underlying's volatility increases, the CPPI payoff declines, as shown in Black and Jones (1987). Thus, one may question whether it is merely the lower volatility of the low-volatility investments that makes CPPI so promising. Therefore, we will now investigate whether an explicit volatility targeting element can help index and multi-factor core investments to close the gap versus the low-volatility underlyings. With volatility targeting, the exposure to index or multi-factor core investments is reduced while one dynamically replicates the volatility of the minimum-volatility strategy.

Indeed, panel C of table 2 reveals that volatility targeting is beneficial for index and other core underlyings: we observe an increase in returns and a decrease in volatility, helping to reduce the gap in risk-adjusted performance. The middle chart in figure 3 visualizes the close alignment of the return distributions. Nevertheless, tail risk statistics are barely altered by the volatility adjustment. It is evident that the CPPI performance wedge is not only driven by the reduced volatility of the underlyings but also by the distinctive relative return pattern of low-volatility strategies in bearish markets. Given these results, it is straightforward to additionally consider a dynamic risk forecasting element which allows the investment exposure to be actively managed in order to further smooth tail risks.

What about adding a dynamic portfolio insurance element?

A conservative multiplier assumption might severely undermine participation in any given underlying. Alternatively, we consider a dynamic multiplier

$$m = m_t := \frac{1}{ES_t^{99\%}(risky\ asset)}$$

governed by an estimate of the underlying's expected shortfall. In such a DPPI (dynamic proportion portfolio insurance) setting, investment exposure will be higher in calmer periods, while more volatile episodes witness a reduction of investment exposure. Obviously, it is essential to rely on risk models that allow for timely modelling of tail risk within the portfolio return distribution, and we will build on expected shortfall forecasts derived from GARCH(1,1)-models.[10]

From figure 3 (bottom chart) and panel D of table 2, we conclude that introducing a dynamic risk forecasting

Table 2
Performance of simulated strategies

		Underlying index	Underlying QMV	Underlying Min-Vol	Underlying Active Low-Vol
Panel A: Pure equity	Return p.a. (%)	6.10	7.57	6.51	9.17
	Volatility p.a. (%)	17.22	16.19	13.23	14.24
	Sharpe ratio	0.31	0.42	0.43	0.58
	Mean annual maximum drawdown (%)	-18.79	-17.48	-14.24	-14.59
	Mean annual Calmar ratio	0.75	0.87	0.90	1.08
	Mean exposure (%)	100.00	100.00	100.00	100.00
	Value-at-risk 99% (%)	39.45	34.38	26.81	25.95
	Expected shortfall 99% (%)	46.62	41.49	32.17	32.32
Panel B: CPPI	Return p.a. (%)	2.73	4.03	3.88	6.25
	Volatility p.a. (%)	15.72	15.49	12.95	14.42
	Sharpe ratio	0.12	0.21	0.23	0.37
	Mean annual maximum drawdown (%)	-14.34	-13.80	-11.69	-12.24
	Mean annual Calmar ratio	0.44	0.57	0.65	0.83
	Mean exposure (%)	72.41	74.95	78.97	80.83
	Value-at-risk 99% (%)	17.83	17.60	16.25	15.98
	Expected shortfall 99% (%)	18.14	17.97	16.97	16.77
Panel C: CPPI with volatility targeting	Return p.a. (%)	3.15	4.25	3.88	6.17
	Volatility p.a. (%)	13.33	13.34	12.95	13.44
	Sharpe ratio	0.17	0.26	0.23	0.40
	Mean annual maximum drawdown (%)	-12.50	-12.24	-11.69	-11.61
	Mean annual Calmar ratio	0.60	0.72	0.65	0.89
	Mean exposure (%)	63.34	66.66	78.97	76.70
	Value-at-risk 99% (%)	17.64	17.37	16.25	15.88
	Expected shortfall 99% (%)	17.98	17.84	16.97	16.68
Panel D: DPPI with volatility targeting	Return p.a. (%)	3.22	4.43	4.17	5.56
	Volatility p.a. (%)	11.95	12.16	12.75	13.25
	Sharpe ratio	0.20	0.29	0.26	0.36
	Mean annual maximum drawdown (%)	-11.33	-11.20	-11.45	-11.58
	Mean annual Calmar ratio	0.58	0.72	0.69	0.84
	Mean exposure (%)	64.43	68.16	84.95	78.98
	Value-at-risk 99% (%)	15.36	15.23	15.63	15.65
	Expected shortfall 99% (%)	15.82	15.75	16.11	16.01

The table shows average performance measures based on block-bootstrapped equity style investments (panel A), and variants thereof based on CPPI (panels B and C) and DPPI (panel D). The floor for both, CPPI and DPPI, is 85%. Reported are the mean return, volatility, Sharpe ratio and expected shortfall of the simulated yearly returns, as well as the mean of the maximum drawdowns (which are computed for each simulated path) and mean exposure.
Sources: MSCI, Bloomberg, Invesco. Block-bootstrapping period: 31 October 2006 to 31 May 2018. **This is simulated past performance and past performance is not a guide to future returns.**

element helps index and multi-factor core (QMV) strategies to reduce the remaining performance wedge relative to low-volatility alternatives. All three – index, QMV and minimum-volatility – experience a slight increase in average returns. Yet the relative reduction in volatility and downside risk is more pronounced for the index and QMV alternatives, leading to a closer alignment of risk-adjusted performance. Nevertheless, an active low-volatility approach with dynamic proportion portfolio insurance (DPPI) still produced the best outcome.

Introducing a dynamic risk forecasting element helps index and multi-factor core strategies to reduce the remaining performance wedge relative to low-volatility alternatives.

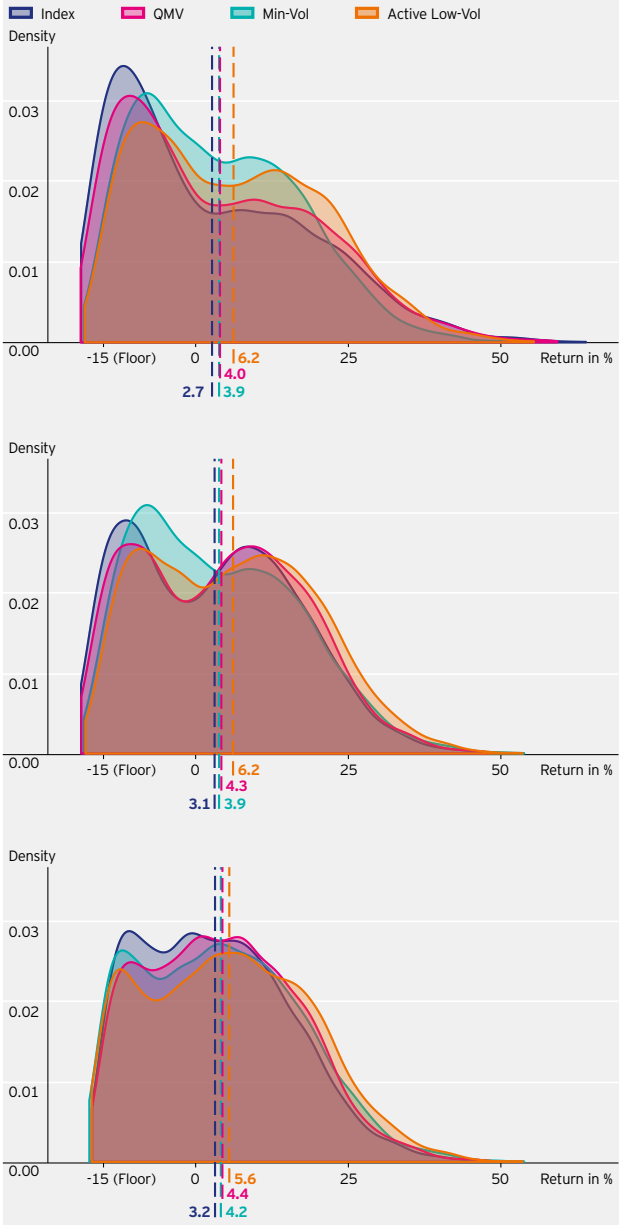
Conclusion

The choice of equity underlying is important when designing portfolio insurance strategies, especially when simple protection mechanisms are applied. We have shown that low-volatility underlyings are particularly useful for downside protection mechanisms, given their lower volatility and more favourable relative return patterns in downside markets. Using a block-bootstrap methodology to simulate the portfolio return distribution, we show that volatility targeting and dynamic risk forecasting elements can improve the portfolio insurance results for index-like alternatives. Still, investing in an active low-volatility underlying while closely managing its investment exposure with suitable dynamic risk forecasts can be the method of choice.

Investing in an active low-volatility underlying while closely managing its investment exposure with suitable dynamic risk forecasts can be the method of choice.

Figure 3

CPPI and factor investing: the role of volatility targeting and DPPI



The charts compare the distribution of block-bootstrapped yearly returns of portfolio insurance strategies based on index, multi-factor core, minimum-volatility, and active low-volatility underlyings. The floor level of the portfolio insurance strategies is 85%. The upper chart is for the static CPPI strategies with a multiplier of 6. The middle chart is for static CPPI with a multiplier of 6 where the underlying style factor investments are subject to targeting the volatility of the minimum-volatility index. The bottom chart considers these refined underlyings in a portfolio insurance strategy based on a dynamic multiplier (DPPI).
Sources: MSCI, Bloomberg, Invesco.

Bibliography
Andersen, T. G., T. Bollerslev, P. F. Christoffersen and F. X. Diebold (2013): Financial risk measurement for financial risk management, in Handbook of the Economics of Finance, ed. by G. M. Constantinides, M. Harris and R. M. Stulz 2 (17), 1127–1220.

Ardia, D., Boudt, K. and Wauters, M. (2016): Smart beta and CPPI performance. Finance, 37(3), 31-65.

Black, F. and Jones, R. (1987): Simplifying portfolio insurance. Journal of Portfolio Management 14, 48–51.

Black, F. and Jones, R. (1988): Simplifying portfolio insurance for corporate pension plans. Journal of Portfolio Management 14, 33–37.

Happersberger, D., H. Lohre and I. Nolte (2018): Estimating portfolio risk for tail risk protection strategies. Working paper, Lancaster University Management School.

Perold, A. F. (1986): Constant proportion portfolio insurance. Working paper, Harvard Business School.

Perold, A. F. and Sharpe, W. F. (1988): Dynamic strategies for asset allocation. Financial Analysts Journal 44, 16–27.

Politis, D. N. and Romano, J. P. (1994): The stationary bootstrap. Journal of the American Statistical Association 89 (428), 1303–1313.

About the authors

Dr Harald Lohre
Senior Research Analyst,
Invesco Quantitative Strategies
Visiting Research Fellow, EMP at Lancaster University Management School
Dr. Harald Lohre develops quantitative models to forecast risk and return used in the management of multi-asset strategies.

David Happersberger
PhD Candidate, Lancaster University, and Invesco Quantitative Strategies
As part of a joint research initiative between Lancaster University and Invesco Quantitative Strategies, David Happersberger is pursuing postgraduate work on practice-oriented issues of financial market econometrics. At the same time, he actively supports the transfer of research results into the multi-asset investment process at Invesco Quantitative Strategies.

Alexandar Cherkezov, CFA
Portfolio Manager,
Invesco Quantitative Strategies
Alexandar Cherkezov manages multi-asset portfolios that include the elements factor-based investing, active asset allocation and downside risk management.

Notes
1 In "The theory and practice of portfolio insurance", Risk & Reward #2/2017 we investigated portfolio insurance strategies ranging from static stop-loss techniques to option-based strategies and dynamic portfolio insurance techniques.
2 See "Evaluating risk mitigation strategies", Risk & Reward #2/2018.
3 For more on CPPI strategies, cf. Perold (1986), Black and Jones (1987, 1988), Perold and Sharpe (1988).
4 See Ardia, Boudt and Wauters (2016) for an overview of different CPPI studies and the choice of multipliers.
5 Throughout the article, all asset returns are in EUR. For money market investments we use EONIA. All simulations in this article are provided for illustrative purposes only and are subject to limitations. Unlike actual portfolio outcomes, the model outcomes do not reflect actual trading, liquidity constraints, fees, expenses, taxes or other factors that could impact future returns.
6 See "Factor investing: building a balanced factor portfolio", Risk & Reward #1/2017.
7 See "Evaluating risk mitigation strategies", Risk & Reward #2/2018.
8 In simulating alternative price paths, we use the stationary block-bootstrap of Politis and Romano (1994). We follow Ardia, Boudt and Wauters (2016) to the extent that block lengths are drawn from a geometric distribution with a minimum block length of one day and an average of 15 days.
9 Investors are more likely to use single-factor portfolios as complementing building blocks or to express an investment view. These are rarely the underlyings of portfolio insurance strategies.
10 The GARCH(1,1)-model captures the main empirical characteristics of asset returns, such as time-varying volatility, fat tails and volatility clustering. For more on GARCH models, cf. Andersen et al. (2013).

The outputs of the assumptions are provided for illustration purposes only. Unlike actual portfolio outcomes, the model outcomes do not reflect actual trading, liquidity constraints, fees, expenses, taxes and other factors that could impact future return.

Complete References

Acerbi, Carlo and Balazs Szekely (2014). "Backtesting expected shortfall". *Risk* 27 (11), 76–81.

Acerbi, Carlo and Dirk Tasche (2002). "On the coherence of expected shortfall". *Journal of Banking & Finance* 26 (7), 1487–1503.

Alizadeh, Sassan, Michael W Brandt, and Francis X Diebold (2002). "Range-based estimation of stochastic volatility models". *Journal of Finance* 57 (3), 1047–1091.

Andersen, Torben G, Tim Bollerslev, Peter F Christoffersen, and Francis X Diebold (2006). "Volatility and correlation forecasting". In: *Handbook of Economic Forecasting*. Ed. by Elliott, G., Granger, Clive W.J., and Timmermann, A. Vol. 1. Elsevier. Chap. 15, 777–878.

— (2013). "Financial risk measurement for financial risk management". In: *Handbook of the Economics of Finance*. Ed. by Constantinides, George M, Harris, Milton, and Stulz, René M. Vol. 2. Elsevier. Chap. 17, 1127–1220.

Annaert, Jan, Sofieke Van Osselaer, and Bert Verstraete (2009). "Performance evaluation of portfolio insurance strategies using stochastic dominance criteria". *Journal of Banking & Finance* 33 (2), 272–280.

Ardia, David, Kris Boudt, and Marjan Wauters (2016). "Smart beta and CPPI performance". *Finance* 37 (3), 31–65.

Arlot, Sylvain, Alain Celisse, et al. (2010). "A survey of cross-validation procedures for model selection". *Statistics surveys* 4, 40–79.

Arnott, Rob, Noah Beck, Vitali Kalesnik, and John West (2016). "How can 'smart beta' go horribly wrong?" *Research Affiliates, February*.

Arnott, Rob, Campbell R Harvey, and Harry Markowitz (2019). "A backtesting protocol in the era of machine learning". *Journal of Financial Data Science* 1 (1), 64–74.

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, and David Heath (1999). "Coherent measures of risk". *Mathematical Finance* 9 (3), 203–228.

Asness, Clifford S (2016). "The siren song of factor timing aka "smart beta timing" aka "style timing"". *Journal of Portfolio Management* 42 (5), 1–6.

Audrino, Francesco and Peter Bühlmann (2004). "Synchronizing multivariate financial time series". *Journal of Risk* 6 (2), 81–106.

Audrino, Francesco, Fabio Sigrist, and Daniele Ballinari (2020a). "The impact of sentiment and attention measures on stock market volatility". *International Journal of Forecasting* 36 (2), 334–357.

— (2020b). "The impact of sentiment and attention measures on stock market volatility". *International Journal of Forecasting* 36 (2), 334–357.

Avramov, Doron, Si Cheng, Amnon Schreiber, and Koby Shemer (2017). "Scaling up market anomalies". *Journal of Investing* 26 (3), 89–105.

Baker, Malcolm and Jeffrey Wurgler (2006). "Investor sentiment and the cross-section of stock returns". *Journal of Finance* 61 (4), 1645–1680.

Balder, Sven, Michael Brandl, and Antje Mahayni (2009). "Effectiveness of CPPI strategies under discrete-time trading". *Journal of Economic Dynamics and Control* 33 (1), 204–220.

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

Barber, Brad M and Terrance Odean (2008). "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors". *Review of Financial Studies* 21 (2), 785–818.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny (1998). "A model of investor sentiment". *Journal of Financial Economics* 49 (3), 307–343.

Barone-Adesi, Giovanni, Kostas Giannopoulos, and Les Vosper (1999). "VaR without correlations for portfolios of derivative securities". *Journal of Futures Markets* 19 (5), 583–602.

Basak, Suleyman (2002). "A comparative study of portfolio insurance". *Journal of Economic Dynamics and Control* 26 (7), 1217–1241.

Basel Committee on Banking Supervision (2016). *Minimum capital requirements for market risk*. https://www.bis.org/bcbs/publ/d352.pdf.

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

Bates, John M and Clive WJ Granger (1969). "The combination of forecasts". *Journal of the Operational Research Society* 20 (4), 451–468.

Bayer, Sebastian (2018). "Combining Value-at-Risk forecasts using penalized quantile regressions". *Econometrics and Statistics* 8, 56–77.

Bayer, Sebastian and Timo Dimitriadis (Sept. 2020). "Regression-based expected shortfall backtesting". *Journal of Financial Econometrics*. nbaa013.

Ben Ameur, Hachmi and Jean-Luc Prigent (2007). "Portfolio insurance: Determination of a dynamic CPPI multiple as function of state variables". *Working paper* THEMA (University of Cergy) and ISC (Paris).

— (2014). "Portfolio insurance: Gap risk under conditional multiples". *European Journal of Operational Research* 236 (1), 238–253.

Benartzi, Shlomo and Richard H Thaler (1995). "Myopic loss aversion and the equity premium puzzle". *Quarterly Journal of Economics* 110 (1), 73–92.

Bender, Jennifer, Xiaole Sun, Ric Thomas, and Volodymyr Zdorovtsov (2018). "The promises and pitfalls of factor timing". *Journal of Portfolio Management* 44 (4), 79–92.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1), 289–300.

Benjamini, Yoav and Daniel Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency". *Annals of Statistics*, 1165–1188.

Benninga, Simon (1990). "Comparing portfolio insurance strategies". *Financial Markets and Portfolio Management* 4 (1), 20–30.

Bergstra, James and Yoshua Bengio (2012). "Random search for hyper-parameter optimization". *Journal of Machine Learning Research* 13 (1), 281–305.

Berkowitz, Jeremy, Peter Christoffersen, and Denis Pelletier (2011). "Evaluating value-at-risk models with desk-level data". *Management Science* 57 (12), 2213–2227.

Bernardi, Mauro and Leopoldo Catania (2016). "Comparison of Value-at-Risk models using the MCS approach". *Computational Statistics* 31 (2), 579–608.

Bertrand, Philippe and Jean-Luc Prigent (2002). "Portfolio insurance: The extreme value approach to the CPPI method". *Finance* 23 (2), 69–86.

— (2011). "Omega performance measure and portfolio insurance". *Journal of Banking & Finance* 35 (7), 1811–1823.

Beschwitz, Bastian von, Donald B Keim, and Massimo Massa (2020). "First to "read" the news: News analytics and algorithmic trading". *Review of Asset Pricing Studies* 10 (1), 122–178.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni (2021). "Bond risk premiums with machine learning". *The Review of Financial Studies* 34 (2), 1046–1089.

Black, Fischer and Robert W Jones (1987). "Simplifying portfolio insurance". *Journal of Portfolio Management* 14 (1), 48–51.

— (1988). "Simplifying portfolio insurance for corporate pension plans". *Journal of Portfolio Management* 14 (4), 33–37.

Black, Fischer and AndréF Perold (1992). "Theory of constant proportion portfolio insurance". *Journal of Economic Dynamics and Control* 16 (3-4), 403–426.

Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". *Journal of Econometrics* 31 (3), 307–327.

— (1987). "A conditionally heteroskedastic time series model for speculative prices and rates of return". *Review of Economics and Statistics*, 542–547.

Bollerslev, Tim, Benjamin Hood, John Huss, and Lasse Heje Pedersen (2018a). "Risk everywhere: Modeling and managing volatility". *Review of Financial Studies* 31 (7), 2729–2773.

— (2018b). "Risk everywhere: Modeling and managing volatility". *Review of Financial Studies* 31 (7), 2729–2773.

Bonferroni, Carlo E (1936). "Teoria statistica delle classi e calcolo delle probabilita'." *Liberia Internazionale Seeber*.

Boudoukh, Jacob, Matthew Richardson, and Robert Whitelaw (1998). "The best of both worlds". *Risk* 11 (5), 64–67.

Boudt, Kris, B Peterson, and Christophe Croux (2008). "Estimation and decomposition of downside risk for portfolios with non-normal returns". *Journal of Risk* 11 (2), 79–103.

Brandt, Michael W, Pedro Santa-Clara, and Rossen Valkanov (2009). "Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns". *Review of Financial Studies* 22 (9), 3411–3447.

Breiman, Leo (2001). "Random forests". *Machine Learning* 45 (1), 5–32.

Brownlees, Christian T and Giampiero M Gallo (2010). "Comparison of volatility measures: a risk management perspective". *Journal of Financial Econometrics* 8 (1), 29–56.

Burns, Patrick, Robert F Engle, and Joseph J Mezrich (1998). "Correlations and volatilities of asynchronous data". *Journal of Derivatives* 5 (4), 7–18.

Cahan, Rochester and Yin Luo (2013). "Standing Out From the Crowd: Measuring Crowding in Quantitative Strategies". *Journal of Portfolio Management* 39, 14–23.

Cappiello, Lorenzo, Robert F Engle, and Kevin Sheppard (2006). "Asymmetric dynamics in the correlations of global equity and bond returns". *Journal of Financial Econometrics* 4 (4), 537–572.

Chan, Wesley S (2003). "Stock price reaction to news and no-news: drift and reversal after headlines". *Journal of Financial Economics* 70 (2), 223–260.

Chen, Jiah-Shing, Chia-Lan Chang, Jia-Li Hou, and Yao-Tang Lin (2008). "Dynamic proportion portfolio insurance using genetic programming with principal component analysis". *Expert Systems with Applications* 35 (1), 273–278.

Chollet, Francois et al. (2015). *Keras*. https://github.com/fchollet/keras.

Christoffersen, Peter and Denis Pelletier (2004). "Backtesting value-at-risk: A duration-based approach". *Journal of Financial Econometrics* 2 (1), 84–108.

Christoffersen, Peter F (1998). "Evaluating interval forecasts". *International Economic Review* 39 (4), 841–862.

Clemen, Robert T (1989). "Combining forecasts: A review and annotated bibliography". *International Journal of Forecasting* 5 (4), 559–583.

Cochrane, John H (2009). *Asset pricing*. Princeton University Press.

Cohen, Randolph B, Paul A Gompers, and Tuomo Vuolteenaho (2002). "Who underreacts to cash-flow news? Evidence from trading between individuals and institutions". *Journal of Financial Economics* 66 (2), 409–462.

Cont, Rama and Peter Tankov (2009). "Constant proportion portfolio insurance in the presence of jumps in asset prices". *Mathematical Finance* 19 (3), 379–401.

Cooper, Tony (2010). "Alpha generation and risk smoothing using managed volatility". *Working Paper*.

Coqueret, Guillaume (2020). "Stock-specific sentiment and return predictability". *Quantitative Finance* 20 (9), 1531–1551.

Cornish, Edmund A and Ronald A Fisher (1938). "Moments and cumulants in the specification of distributions". *Revue de l'Institut International de Statistique* 5 (4), 307–320.

Corsi, Fulvio (2009). "A simple approximate long-memory model of realized volatility". *Journal of Financial Econometrics* 7 (2), 174–196.

Corsi, Fulvio and Roberto Renò (2012). "Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling". *Journal of Business & Economic Statistics* 30 (3), 368–380.

Creal, Drew, Siem Jan Koopman, and André Lucas (2013). "Generalized autoregressive score models with applications". *Journal of Applied Econometrics* 28 (5), 777–795.

Cutler, David M, James M Poterba, and Lawrence H Summers (1989). "What moves stock prices?" *Journal of Portfolio Management* 15 (3), 4–12.

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals and Systems* 2 (4), 303–314.

Da, Zhi, Joseph Engelberg, and Pengjie Gao (2011). "In search of attention". *Journal of Finance* 66 (5), 1461–1499.

Da, Zhi and Mitchell Craig Warachka (2009). "Cashflow risk, systematic earnings revisions, and the cross-section of stock returns". *Journal of Financial Economics* 94 (3), 448–468.

Dang, Tung Lam, Fariborz Moshirian, and Bohui Zhang (2015). "Commonality in news around the world". *Journal of Financial Economics* 116 (1), 82–110.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam (1998). "Investor psychology and security market under- and overreactions". *Journal of Finance* 53 (6), 1839–1885.

Daniel, Kent, Sheridan Titman, and KC John Wei (2001). "Explaining the cross-section of stock returns in Japan: factors or characteristics?" *Journal of Finance* 56 (2), 743–766.

Daniel, Kent D, David Hirshleifer, and Avanidhar Subrahmanyam (2001). "Overconfidence, arbitrage, and equilibrium asset pricing". *Journal of Finance* 56 (3), 921–965.

De Bondt, Werner FM and Richard Thaler (1985). "Does the stock market overreact?" *Journal of Finance* 40 (3), 793–805.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal (2009). "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?" *Review of Financial Studies* 22 (5), 1915–1953.

DeMiguel, Victor, Alberto Martin-Utrera, Francisco J Nogales, and Raman Uppal (2020). "A transaction-cost perspective on the multitude of firm characteristics". *Review of Financial Studies* 33 (5), 2180–2222.

Dichtl, Hubert and Wolfgang Drobetz (2011). "Portfolio insurance and prospect theory investors: Popularity and optimal design of capital protected financial products". *Journal of Banking & Finance* 35 (7), 1683–1697.

Dichtl, Hubert, Wolfgang Drobetz, Harald Lohre, Carsten Rother, and Patrick Vosskamp (2019). "Optimal timing and tilting of equity factors". *Financial Analysts Journal* 75 (4), 84–102.

Dichtl, Hubert, Wolfgang Drobetz, and Martin Wambach (2017). "A bootstrap-based comparison of portfolio insurance strategies". *European Journal of Finance* 23 (1), 31–59.

Diebold, Francis X (1989). "Forecast combination and encompassing: Reconciling two divergent literatures". *International Journal of Forecasting* 5 (4), 589–592.

Diebold, Francis X and Robert S Mariano (1995). "Comparing predictive accuracy". *Journal of Business & Economic Statistics* 13 (3), 253–263.

Diebold, Francis X and Minchul Shin (2019). "Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives". *International Journal of Forecasting* 35 (4), 1679–1691.

Dietterich, Thomas G (2000). "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer, 1–15.

Dimitriadis, Timo and Roxana Halbleib (2021). "Realized Quantiles". *Journal of Business & Economic Statistics* 0 (0), 1–16.

Donaldson, R Glen and Mark Kamstra (1996). "Forecast combining with neural networks". *Journal of Forecasting* 15 (1), 49–61.

Embrechts, Paul and Marius Hofert (2014). "Statistics and quantitative risk management for banking and insurance". *Annual Review of Statistics and Its Application* 1, 493–514.

Emmer, Susanne, Marie Kratz, and Dirk Tasche (2015). "What is the best risk measure in practice? A comparison of standard measures". *Journal of Risk* 18 (2), 31–60.

Engelberg, Joseph, R David McLean, and Jeffrey Pontiff (2018). "Anomalies and news". *Journal of Finance* 73 (5), 1971–2001.

Engle, Robert F (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". *Econometrica* 50 (4), 987–1007.

Engle, Robert F and Simone Manganelli (2004). "CAViaR: Conditional autoregressive value at risk by regression quantiles". *Journal of Business & Economic Statistics* 22 (4), 367–381.

Fama, Eugene F and Kenneth R French (1992). "The cross-section of expected stock returns". *Journal of Finance* 47 (2), 427–465.

Fama, Eugene F and Kenneth R French (2006). "Profitability, investment and average returns". *Journal of Financial Economics* 82 (3), 491–518.

— (2016). "Dissecting anomalies with a five-factor model". *Review of Financial Studies* 29 (1), 69–103.

Fang, Lily and Joel Peress (2009). "Media coverage and the cross-section of stock returns". *Journal of Finance* 64 (5), 2023–2052.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici (2019). "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *J. Mach. Learn. Res.* 20 (177), 1–81.

Fissler, Tobias and Johanna F Ziegel (2016). "Higher order elicitability and Osband's principle". *Annals of Statistics* 44 (4), 1680–1707.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Gerlach, Richard and Cathy WS Chen (2015). "Bayesian expected shortfall forecasting incorporating the intraday range". *Journal of Financial Econometrics* 14 (1), 128–158.

Gibbons, Michael R, Stephen A Ross, and Jay Shanken (1989). "A test of the efficiency of a given portfolio". *Econometrica*, 1121–1152.

Giese, Guido (2012). "Optimal design of volatility-driven algo-alpha trading strategies". *Risk* 25 (6), 68–73.

Giot, Pierre and Sébastien Laurent (2004). "Modelling daily value-at-risk using realized volatility and ARCH type models". *Journal of Empirical Finance* 11 (3), 379–398.

Glosten, Lawrence R, Ravi Jagannathan, and David E Runkle (1993). "On the relation between the expected value and the volatility of the nominal excess return on stocks". *Journal of Finance* 48 (5), 1779–1801.

Gneiting, Tilmann (2011). "Making and evaluating point forecasts". *Journal of the American Statistical Association* 106 (494), 746–762.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". *Journal of the American Statistical Association* 102 (477), 359–378.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016). *Deep learning*. Vol. 1. 2. MIT press Cambridge.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). "Empirical asset pricing via machine learning". *Review of Financial Studies* 33 (5), 2223–2273.

Hafez, P. A., J. A. Guerrero-Colon, and S. Duprey (2015). "Thematic alpha streams improve equity portfolio performance". *RavenPack Research Paper*.

Hafez, P. A. and Junqiang Xie (2011). *Introducing the RavenPack sentiment index*. Tech. rep. RavenPack Analytics.

Hafez, Peter Agder (2010). "News beta - A new measure for risk & stock analysis". *RavenPack Research Paper*.

Halbleib, Roxana and Winfried Pohlmeier (2012). "Improving the value at risk forecasts: Theory and evidence from the financial crisis". *Journal of Economic Dynamics and Control* 36 (8), 1212–1228.

Hallerbach, Winfried G (2012). "A proof of the optimality of volatility weighting over time". *Journal of Investment Strategies* 1 (4), 87–99.

— (2015). "Advances in portfolio risk control". In: *Risk-Based and Factor Investing*. Ed. by Jurczenko, Emmanuel. Vol. 1. Elsevier. Chap. 1, 1–30.

Hamidi, B, E Jurczenko, and B Maillet (2009). "A CAViaR modelling for a simple time-varying proportion portfolio insurance strategy". *Bankers, Markets & Investors* 102, 4–21.

Hamidi, Benjamin, Christophe Hurlin, Patrick Kouontchou, and Bertrand Maillet (2015). "A DARE for VaR". *Finance* 36 (1), 7–38.

Hamidi, Benjamin, Bertrand Maillet, and Jean-Luc Prigent (2014). "A dynamic autoregressive expectile for time-invariant portfolio protection strategies". *Journal of Economic Dynamics and Control* 46, 1–29.

Hamidi, Benjamin, Bertrand B Maillet, and Jean-Luc Prigent (2009). "A risk management approach for portfolio insurance strategies". In: *Proceedings of the 1st EIF International Financial Research Forum, Economica*.

Hanauer, Matthias (2014). "Is Japan different? Evidence on momentum and market dynamics". *International Review of Finance* 14 (1), 141–160.

Hansen, Bruce E (1994). "Autoregressive conditional density estimation". *International Economic Review* 35, 705–730.

— (2008). "Least-squares forecast averaging". *Journal of Econometrics* 146 (2), 342–350.

Hansen, Lars Kai and Peter Salamon (1990). "Neural network ensembles". *IEEE transactions on pattern analysis and machine intelligence* 12 (10), 993–1001.

Hansen, Peter R, Asger Lunde, and James M Nason (2011). "The model confidence set". *Econometrica* 79 (2), 453–497.

Hansen, Peter Reinhard, Zhuo Huang, and Howard Howan Shek (2012). "Realized GARCH: a joint model for returns and realized measures of volatility". *Journal of Applied Econometrics* 27 (6), 877–906.

Happersberger, David, Harald Lohre, and Ingmar Nolte (2020). "Estimating portfolio risk for tail risk protection strategies". *European Financial Management* 26 (4), 1107–1146.

Hart, Jeffrey D (1994). "Automated kernel smoothing of dependent data by using time series cross-validation". *Journal of the Royal Statistical Society: Series B (Methodological)* 56 (3), 529–542.

Hart, Jeffrey D and Cherng-Luen Lee (2005). "Robustness of one-sided cross-validation to autocorrelation". *Journal of Multivariate Analysis* 92 (1), 77–96.

Harvey, A C (2013). "Dynamic models for volatility and heavy tails: with applications to financial and economic time series". In: *Econometric Society Monographs*. Vol. 52. Cambridge University Press.

Harvey, Campbell R, Yan Liu, and Alessio Saretto (2020). "An evaluation of alternative multiple testing methods for finance applications". *The Review of Asset Pricing Studies* 10 (2), 199–248.

Harvey, Campbell R, Yan Liu, and Heqing Zhu (2016). "... and the cross-section of expected returns". *Review of Financial Studies* 29 (1), 5–68.

Harvey, David, Stephen Leybourne, and Paul Newbold (1997). "Testing the equality of prediction mean squared errors". *International Journal of Forecasting* 13 (2), 281–291.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2011). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, second Edition.

Haugen, Robert A and Nardin L Baker (1991). "The efficient market inefficiency of capitalization-weighted stock portfolios". *Journal of Portfolio Management* 17 (3), 35–40.

— (1996). "Commonality in the determinants of expected stock returns". *Journal of Financial Economics* 41 (3), 401–439.

Heston, Steven L and Nitish Ranjan Sinha (2017). "News versus sentiment: Predicting stock returns from news stories". *Financial Analysts Journal* 73 (3), 67–83.

Hillert, Alexander, Heiko Jacobs, and Sebastian Müller (2014). "Media makes momentum". *Review of Financial Studies* 27 (12), 3467–3501.

Hinton, Geoffrey (2012). *Neural networks for machine learning*. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". *Science* 313 (5786), 504–507.

Hocquard, Alexandre, Sunny Ng, and Nicolas Papageorgiou (2013). "A constant-volatility framework for managing tail risk". *Journal of Portfolio Management* 39 (2), 28–40.

Hoerl, Arthur E and Robert W Kennard (1970a). "Ridge regression: applications to nonorthogonal problems". *Technometrics* 12 (1), 69–82.

— (1970b). "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics* 12 (1), 55–67.

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure". *Scandinavian Journal of Statistics*, 65–70.

Hornik, Kurt, Maxwell Stinchcombe, Halbert White, et al. (1989). "Multilayer feedforward networks are universal approximators." *Neural Networks* 2 (5), 359–366.

Hou, Kewei, G Andrew Karolyi, and Bong-Chan Kho (2011). "What factors drive global stock returns?" *Review of Financial Studies* 24 (8), 2527–2574.

Hsu, Jason, Vitali Kalesnik, and Vivek Viswanathan (2015). "A framework for assessing factors and implementing smart beta strategies". *Journal of Index Investing* 6 (1), 89–97.

Huang, Ethan, Victor Liu, Li Ma, and James Osiol (2010). "Methods in dynamic weighting". *Capital IQ Working Paper*.

Ilmanen, Antti and Jared Kizer (2012). "The death of diversification has been greatly exaggerated". *Journal of Portfolio Management* 38 (3), 15–27.

Jacobs, Heiko and Sebastian Müller (2020). "Anomalies across the globe: Once public, no longer existent?" *Journal of Financial Economics* 135 (1), 213–230.

Jegadeesh, Narasimhan (1990). "Evidence of predictable behavior of security returns". *Journal of Finance* 45 (3), 881–898.

Jegadeesh, Narasimhan and Sheridan Titman (1993). "Returns to buying winners and selling losers: Implications for stock market efficiency". *Journal of Finance* 48 (1), 65–91.

Jensen, Michael C, Fischer Black, and Myron S Scholes (1972). "The capital asset pricing model: Some empirical tests". *Studies in the Theory of Capital Markets. Praeger Publishers*.

Jiang, Chonghui, Yongkai Ma, and Yunbi An (2009). "The effectiveness of the VaR-based portfolio insurance strategy: An empirical analysis". *International Review of Financial Analysis* 18 (4), 185–197.

Jondeau, Eric and Michael Rockinger (2006). "The Copula-GARCH model of conditional dependencies: An international stock market application". *Journal of International Money and Finance* 25 (5), 827–853.

Kan, Raymond and GuoFu Zhou (2012). "Tests of mean-variance spanning". *Annals of Economics and Finance* 13 (1), 139–187.

Ke, Zheng Tracy, Bryan T Kelly, and Dacheng Xiu (2019). "Predicting Returns with Text Data". *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2019-69).

Kingma, Diederik P. and Jimmy Ba (2015). *Adam: A Method for Stochastic Optimization*. Conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Kirby, Chris and Barbara Ostdiek (2012). "It's all in the timing: Simple active portfolio strategies that outperform naive diversification". *Journal of Financial and Quantitative Analysis* 47 (2), 437–467.

Koenker, Roger and Gilbert Bassett (1978). "Regression quantiles". *Econometrica* 46 (1), 33–50.

Kolasinski, Adam C, Adam V Reed, and Matthew C Ringgenberg (2013). "A multiple lender approach to understanding supply and search in the equity lending market". *Journal of Finance* 68 (2), 559–595.

Kolrep, Martin, Harald Lohre, and David Happersberger (2017). "Theory and Practice of Portfolio Insurance". *Risk & Reward*, 4–9.

Kuan, Chung-Ming and Halbert White (1994). "Artificial neural networks: An econometric perspective". *Econometric Reviews* 13 (1), 1–91.

Kuester, Keith, Stefan Mittnik, and Marc S Paolella (2006). "Value-at-risk prediction: A comparison of alternative strategies". *Journal of Financial Econometrics* 4 (1), 53–89.

Kupiec, Paul H (1995). "Techniques for verifying the accuracy of risk measurement models". *Journal of Derivatives* 3 (2), 73–84.

La Porta, Rafael, Josef Lakonishok, Andrei Shleifer, and Robert Vishny (1997). "Good news for value stocks: Further evidence on market efficiency". *Journal of Finance* 52 (2), 859–874.

Lee, Wai (2017). "Factors timing factors". *Journal of Portfolio Management* 43 (5), 66–71.

Lehmann, Bruce N (1990). "Fads, Martingales, and Market Efficiency". *Quarterly Journal of Economics* 105 (1), 1–28.

Leinweber, David and Jacob Sisk (2011). "Event driven trading and the 'new news'". *Journal of Portfolio Management* 38 (1).

Lintner, John (1965). "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets". *Review of Economics and Statistics* 47 (1), 13–37.

Liu, Lily Y, Andrew J Patton, and Kevin Sheppard (2015). "Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes". *Journal of Econometrics* 187 (1), 293–311.

Lo, Andrew W and A Craig MacKinlay (1990a). "An econometric analysis of nonsynchronous trading". *Journal of Econometrics* 45 (1-2), 181–211.

— (1990b). "Data-snooping biases in tests of financial asset pricing models". *Review of Financial Studies* 3 (3), 431–467.

Lohre, Harald, David Happersberger, and Alexandar Cherkezov (2018). "The Use of Equity Factor Investing for Portfolio Insurance". *Risk & Reward*, 32–38.

Lohre, Harald, David Happersberger, and Erhard Radatz (2018). "Evaluating risk mitigation strategies". *Risk & Reward*, 27–31.

Longin, Francois and Bruno Solnik (1995). "Is the correlation in international equity returns constant: 1960–1990?" *Journal of International Money and Finance* 14 (1), 3–26.

Louzis, Dimitrios P, Spyros Xanthopoulos-Sisinis, and Apostolos P Refenes (2012). "Stock index realized volatility forecasting in the presence of heterogeneous leverage effects and long range dependence in the volatility of realized volatility". *Applied Economics* 44 (27), 3533–3550.

Louzis, Dimitrios P, Spyros Xanthopoulos-Sisinis, and Apostolos P Refenes (2014). "Realized volatility models and alternative Value-at-Risk prediction strategies". *Economic Modelling* 40, 101–116.

Malkiel, Burton G and Eugene F Fama (1970). "Efficient capital markets: A review of theory and empirical work". *Journal of Finance* 25 (2), 383–417.

Manganelli, Simone and Robert F Engle (2004). "A comparison of value-at-risk models in finance". *Risk measures for the 21st century*, 123–44.

Martin, R Douglas and Rohit Arora (2017). "Inefficiency and bias of modified value-at-risk and expected shortfall". *Journal of Risk* 19 (6), 59–84.

McLean, R David and Jeffrey Pontiff (2016). "Does academic research destroy stock return predictability?" *Journal of Finance* 71 (1), 5–32.

McNeil, Alexander J and Rüdiger Frey (2000). "Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach". *Journal of Empirical Finance* 7 (3), 271–300.

Molnar, Christoph (2019). *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/.

Mossin, Jan (1966). "Equilibrium in a capital asset market". *Econometrica*, 768–783.

Nadarajah, Saralees, Bo Zhang, and Stephen Chan (2014). "Estimation methods for expected shortfall". *Quantitative Finance* 14 (2), 271–291.

Newey, Whitney K and James L Powell (1987). "Asymmetric least squares estimation and testing". *Econometrica* 55, 819–847.

Nieto, Maria Rosa and Esther Ruiz (2016). "Frontiers in VaR forecasting and backtesting". *International Journal of Forecasting* 32 (2), 475–501.

Nolde, Natalia and Johanna F Ziegel (2017). "Elicitability and backtesting: Perspectives for banking regulation". *Annals of Applied Statistics* 11 (4), 1833–1874.

Novy-Marx, Robert (2013). "The other side of value: The gross profitability premium". *Journal of Financial Economics* 108 (1), 1–28.

Parkinson, Michael (1980). "The extreme value method for estimating the variance of the rate of return". *Journal of Business*, 61–65.

Patton, Andrew J (2006). "Modelling asymmetric exchange rate dependence". *International Economic Review* 47 (2), 527–556.

Patton, Andrew J, Johanna F Ziegel, and Rui Chen (2019). "Dynamic semiparametric models for expected shortfall (and value-at-risk)". *Journal of Econometrics* 211 (2), 388–413.

Perchet, Romain, Raul Leote De Carvalho, Thomas Heckel, and Pierre Moulin (2015). "Predicting the success of volatility targeting strategies: Application to equities and other asset classes". *Journal of Alternative Investments* 18 (3), 21–38.

Perold, Andre (1986). "Constant proportion portfolio insurance". *Harvard Business School*.

Perold, Andre F and William F Sharpe (1988). "Dynamic strategies for asset allocation". *Financial Analysts Journal* 44 (1), 16–27.

Pritsker, Matthew (2006). "The hidden dangers of historical simulation". *Journal of Banking & Finance* 30 (2), 561–582.

RavenPack Analytics (2017). *User Guide and Service Overview*. Tech. rep. RavenPack Analytics.

Righi, Marcelo Brutti and Paulo Sergio Ceretta (2015). "A comparison of expected shortfall estimation models". *Journal of Economics and Business* 78, 14–47.

RiskMetrics Group (1996). "Riskmetrics - technical document". *J. P. Morgan and Reuters*.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf (2008). "Formalized data snooping based on generalized error rates". *Econometric Theory* 24 (2), 404–447.

Romano, Joseph P and Michael Wolf (2005). "Stepwise multiple testing as formalized data snooping". *Econometrica* 73 (4), 1237–1282.

Romano, Joseph P, Michael Wolf, et al. (2007). "Control of generalized error rates in multiple testing". *Annals of Statistics* 35 (4), 1378–1408.

Ross, Stephen A (1976). "The arbitrage theory of capital asset pricing". *Journal of Economic Theory* 13 (3), 341–360.

Santos, André AP, Francisco J Nogales, and Esther Ruiz (2012). "Comparing univariate and multivariate models to forecast portfolio value-at-risk". *Journal of Financial Econometrics* 11 (2), 400–441.

Scherer, B (2013). "Synchronize your data or get out of step with your risks". *Journal of Derivatives* 20 (3), 75–84.

Scholes, Myron and Joseph Williams (1977). "Estimating betas from nonsynchronous data". *Journal of Financial Economics* 5 (3), 309–327.

Shan, Kejia and Yuhong Yang (2009). "Combining regression quantile estimators". *Statistica Sinica* 19 (3), 1171–1191.

Sharpe, William F (1964). "Capital asset prices: A theory of market equilibrium under conditions of risk". *Journal of Finance* 19 (3), 425–442.

Sloan, R (1996). "Do stock prices fully reflect information in accruals and cash flows about future earnings?" *Accounting Review* 71 (3), 289–315.

Soupé, François, Thomas Heckel, and Raul Leote De Carvalho (2014). "Portfolio insurance with adaptive protection (PIWAP)". *Journal of Investment Strategies* 5 (3), 1–15.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". *Journal of Machine Learning Research* 15 (1), 1929–1958.

Stock, James H and Mark W Watson (2004). "Combination forecasts of output growth in a seven-country data set". *Journal of Forecasting* 23 (6), 405–430.

Taylor, James W (2008). "Estimating value at risk and expected shortfall using expectiles". *Journal of Financial Econometrics* 6 (2), 231–252.

— (2019). "Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution". *Journal of Business & Economic Statistics* 37 (1), 121–133.

— (2020). "Forecast combinations for value at risk and expected shortfall". *International Journal of Forecasting* 36 (2), 428–441.

Taylor, Stephen J (1986). *Modelling Financial Time Series*. Chichester: Wiley.

Tetlock, Paul C (2007). "Giving content to investor sentiment: The role of media in the stock market". *Journal of Finance* 62 (3), 1139–1168.

Tetlock, Paul C, Maytal Saar-Tsechansky, and Sofus Macskassy (2008). "More than words: Quantifying language to measure firms' fundamentals". *Journal of Finance* 63 (3), 1437–1467.

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

Timmermann, Allan (2006). "Forecast combinations". In: *Handbook of Economic Forecasting*. Ed. by Elliott, G., Granger, C.W., and Timmermann, A. Vol. 1. Elsevier. Chap. 4, 135–196.

Tversky, Amos and Daniel Kahneman (1992). "Advances in prospect theory: Cumulative representation of uncertainty". *Journal of Risk and Uncertainty* 5 (4), 297–323.

Uhl, Matthias W, Mads Pedersen, and Oliver Malitius (2015). "What's in the news? Using news sentiment momentum for tactical asset allocation". *Journal of Portfolio Management* 41 (2), 100.

Wang, Ying, Bohui Zhang, and Xiaoneng Zhu (2018). "The momentum of news". *SSRN Working Paper*.

White, Halbert (2000). "A reality check for data snooping". *Econometrica* 68 (5), 1097–1126.

Yamai, Yasuhiro and Toshinao Yoshiba (2002). "On the validity of value-at-risk: Comparative analyses with expected shortfall". *Monetary and Economic Studies* 20 (1), 57–85.

Ye, Yinyu (1987). "Interior algorithms for linear, quadratic, and linearly constrained non-linear programming". PhD thesis. Department of ESS, Stanford University.

Zangari, Peter (1996). "A VaR methodology for portfolios that include options". *RiskMetrics Monitor* 1, 4–12.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2), 301–320.