Risk and return of passive and active commodity futures strategies

Von der Fakultät für Wirtschaft, Recht und Gesellschaft
der Brandenburgischen Technischen Universität Cottbus–Senftenberg
zur Erlangung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften

genehmigte Dissertation

vorgelegt von

Dipl.-Math.
Julia Sophia Mehlitz
geb. Keitel am 02. Mai 1995 in Dresden

Vorsitzende/Vorsitzender:	Prof. Dr. David Müller
Gutachterin/Gutachter:	Prof. Dr. Benjamin R. Auer
Gutachterin/Gutachter:	Prof. Dr. Frank Schuhmacher
Tag der mündlichen Prüfung:	19. November 2021

Contents

List of Tables

List of Figures

Introduction

In recent decades, commodity futures investments have become more popular than ever – for instance, according to the commodity futures trading commission (CFTC), the open interest (that is, the number of open future contracts) of the Goldman Sachs Commodity Index increased tenfold between the ends of 1992 and 2016.[1] While several irrational bubbles and unforeseeable crises (such as the 2000 dot-com bubble, the 2008 global financial crisis or the 2020 stock market crash and other effects of the COVID-19 pandemic) have affected the common stock market, research on alternative investments was intensified and pointed out the general potential of commodity futures (see Jensen et al., 2000; Gorton and Rouwenhorst, 2006; Bhardwaj et al., 2015; Narayan et al., 2013; Daskalaki et al., 2014, 2017). Thus, commodity futures evolved from a pure hedging instrument for commodity risk managers into a popular liquid asset class (see Rouwenhorst and Tang, 2012; Tang and Xiong, 2012; Cheng and Xiong, 2014; Henderson et al., 2015). Because of the absence of restrictions for short sellers, their negligible transaction costs, manageable extent and high liquidity, commodity futures also offer attractive conditions for cross-sectional, time-variable investment strategies, as studied by Shen et al. (2007); Szakmary et al. (2010); Fuertes et al. (2010); Bianchi et al. (2015) or Fernandez-Perez et al. (2018a).

Motivated by the popularity of commodity futures investments in practice and academia, this thesis examines passive and active investment strategies in commodity futures markets – especially, the analysis of their risk and returns. Potential investors may wonder what level of risk to expect when investing in commodity futures or, (in view of the recent financial market troubles) about the extent to which these risks will be influenced by financial crises and which commodity investment strategies will still yield a good return. To evaluate the former questions, following standards of the Basel Committee on Banking Supervision, the expected shortfall (ES) measure found its way into general risk management applications. It was intended to replace the related value at risk (VaR) measure in the banking sector, as it fulfills the property of sub-additivity and allows not just for the frequency, but also for the magnitude of shortfalls (see Artzner et al., 1999; Acerbi and Tasche, 2002a,b; Basel Committee of Banking Supervision, 2012). Relying on this risk measure, the following questions arise for investors in commodity futures markets (or likewise, for financial trading institutions that have to compute for example margins of exchanges by means of ES): What are common levels of ES? On which method should ES estimation be based? How can estimation quality be evaluated ex post in order to readjust estimation processes if required?[2]

This thesis responds to these questions in the following way: First in Chapter 1, we evaluate the qualities of several popular ES estimators in a general context. We focus on non-parametric ES estimators, relying on some classic, weighted and outlier-robust versions of historic estimators (see Inui and Kijima, 2005; Peracchi and Tanase, 2008; Jadhav et al., 2009; Nadarajah et al., 2014) and represents of kernel density techniques (see Nadaraya, 1964; Scaillet, 2004; Chen, 2008). Additionally, we consider two parametric benchmarks: the normal distribution approach and the

[1] The detailed data are available on https://www.cftc.gov/MarketReports/CommitmentsofTraders/HistoricalCompressed/index.htm. Their computation of open interest follows https://www.cftc.gov/MarketReports/CommitmentsofTraders/index.htm.

[2] Indeed, several (parametric and non-parametric) ES estimators already exist, but no general convention how to estimate ES best became established. Moreover and in contrast to VaR (for which the backtests of Kupiec (1995); Christoffersen (1998) and Berkowitz (2001) turned out to be standards), the ES' lack of identifiability rules any ideal direct method of ex post quality evaluation out.

peak over threshold method (based on extreme value theory; see McNeil and Frey, 2000) and analyze several combined estimators. To exclude distorting influences on data and hence, prevent extra estimation errors, we work with several simulated settings that enable us to perform our evaluation without any additional time-varying mean and volatility models or backtests (which is not possible when working with commodity futures data directly). Rather, we analyze *pure* ES estimator characteristics in terms of certain performance plots, error and risk measures for our diverse simulated market settings.

After this, in Chapter 2, we turn our attention to ES estimation on commodity futures markets especially. We begin by analyzing time-dependent ES levels of passive commodity investments over nearly forty years. Then, applying two versions of a modern backtest of Du and Escanciano (2017) that adapt the established VaR backtest standards to ES, we compare the time-varying qualities of various ES estimators in the commodity futures market. In that process, we also analyze the extent to which the estimated ES behavior differs when historical market phases change or crises occur. In order to secure a correct validation of estimation results, we concentrate on ES estimators that are based on invertible distribution functions, such as parametric estimators (assuming strictly monotonic distribution functions) or semi-/ non-parametric approaches that base on smoothed and data-dependent distributions. In addition to peak over threshold and kernel density estimators, which we already introduced in Chapter 1, we expand our focus by allowing for the skewed t distribution of Hansen (1994) that incorporates both skewness and kurtosis, and further parametric methods that performed well in a commodity context, namely, the g-and-h distribution of Tukey (1977), the Johnson (1949) system and a bivariate Gaussian mixture distribution. With this proceeding, we can identify the sector-specific qualities of ES estimators in commodity markets.

After studying risk measurement with ES in general and for passive commodity futures investments in particular, we turn to the analysis of risk and return of active commodity futures market strategies. In this context, one of the most popular representatives are cross-sectional momentum strategies (see Rouwenhorst and Tang, 2012; Miffre, 2016), which emerged after the seminal work of Jegadeesh and Titman (1993, 2001) in stocks sector and were extended to commodity futures markets from Erb and Harvey (2006); Miffre and Rallis (2007) and others. However, for the last two decades, traditional momentum strategies were found to exhibit decreasing performance in stock markets (see Chordia et al., 2014; Hwang and Rubesam, 2015). In regard to that observation, the following questions immediately arise for investors in commodity futures markets: Was the profitability of momentum strategies adversely affected, too? And is there still some potential to enhance strategy performance?

In Chapter 3, we answer the first question with our performance analysis of several momentum strategies during the last quarter-century. Regarding the second, we investigate whether the negative impacts of crises or other kinds of market turmoil on strategy performance can be avoided by considering memory-enhanced momentum strategies. For that, we utilize short and long memory measures, namely variance ratios and Hurst exponents (based on Lo and MacKinlay (1988) and Hurst (1951), respectively), which can be estimated without further characteristics than past returns and deliver additional information about market stability to momentum strategies. For our new, memory-enhanced momentum strategies, we also study which commodity sectors remain most gainful for traders and which trading behavior prepares investors best to withstand portfolio losses or even increase wealth despite recent crises.

The remainder of this cumulative thesis is organized as follows: It is subdivided into three main parts (according to Chapters 1 to 3); Chapters 1 and 2 are based on Mehlitz and Auer (2020, 2021), respectively, whereas Chapter 3 is still under review. Each part starts with a short motivation that brings its particular approach into context of the thesis. In addition to the introductory and concluding sections of each chapter, which discuss our research questions, relate it to the literature and summarize our answers to them in detail, we conclude this thesis with a Summary of its most important findings.

Part I

Motivation

Financial trading institutions or private investors that plan to measure financial risks with ES have to select an appropriate method to estimate the expected shortfall of their data. For this purpose, we provide a structured comparison of several approved ES estimators in the first chapter of that thesis that will help risk managers to decide on suitable ES estimation methods, conditional on their respective market settings.

The variety of contemplable ES estimators can be subdivided into two classes: parametric (assuming underlying pre-specified probability distributions with a fixed set of parameters) and non-parametric (number and nature of factors in the assumed model structure are determined from data sample). The appropriateness of parametric estimation methods strongly depends on the underlying population properties, whereas an application of non-parametric methods works more generally. Therefore, we mainly concentrate on non-parametric estimators of ES in the following chapter and give an overview of their estimation qualities by means of classic error measures, relative standard deviation and certain performance profiles.

In order to obtain universally valid results that avoid misleading influences of past crises or other real-world phenomena, we base our evaluation on simulated settings with zero mean and unit variance. Practitioners can receive such data by extracting the underlying time-varying mean and volatility processes with appropriate filter models. As we work with simulated mean- and volatility-adjusted data directly, we prevent that our evaluation results can be distorted from chosen filter models.

The evaluation is not only limited to the commodity context, but also can be useful for other financial sectors. Indeed, we extend our focus to samples that originate from several skewed and tailed distributional settings, such that appliers from commodity, stocks and other finance sectors or (insurance) industries are enabled to identify the best-performing estimators to their setting (or this one that matches best to their data). To complete our overview, we also incorporate several common levels of ES and sample sizes (from one month to four years of daily data) that might result from different kinds of application.

1. A Monte Carlo evaluation of non-parametric estimators of expected shortfall

Abstract: Motivated by the growing importance of the expected shortfall in banking and finance, this study compares the performance of popular non-parametric estimators of expected shortfall (i. e., different variants of historical, outlier-adjusted and kernel methods) to each other, selected parametric benchmarks and estimates based on the idea of forecast combination. Within a multi-dimensional simulation setup (spanned by different distributional settings, sample sizes and confidence levels), we rank the estimators based on classic error measures as well as an innovative performance profile technique, which we adapt from the mathematical programming literature. Our rich set of results supports academics and practitioners in the search for an answer to the question of which estimators are preferable under which circumstances.

1.1. Introduction

Over decades, the value at risk (VaR) played a dominant role in quantifying asset and portfolio risks of commercial banks, insurance companies and non-financial firms (see Basel Committee of Banking Supervision, 1996, 2004; Jorion, 2007; Daníelsson, 2011). However, in recent years, the growing awareness of the measure's theoretical deficiencies has led researchers and practitioners to rethink its application. The VaR captures the loss that is not exceeded with a certain confidence level and thus naturally does not look at the loss we have to expect if a tail event occurs. In other words, it considers only the likelihood but not the size of large losses. Furthermore, it fails to account appropriately for portfolio risk diversification because it does not fulfill the important properties of subadditivity (see Artzner et al., 1997, 1999) and convexity (see Basak and Shapiro, 2001). In contrast, the expected shortfall (ES) does not suffer from such shortcomings because it is defined as the expected value of losses exceeding the VaR (see Acerbi and Tasche, 2002a,b; Tasche, 2002; Yamai and Yoshiba, 2005). Consequently, regulators have suggested phasing out the VaR and replacing it with the ES in the calculation of capital requirements of banks (see Basel Committee of Banking Supervision, 2012). Moreover, the usage of ES in the construction and evaluation of stock, bond, commodity, currency and even bitcoin portfolios is on the rise (see Harmantzis et al., 2006; Reboredo, 2013; Yao et al., 2013; Paraschiv et al., 2015; Degiannakis and Potamia, 2017; Stavroyiannis, 2018).

The literature has brought forth a wide variety of methods to estimate the ES of an asset or portfolio (see Nadarajah et al., 2014). Therefore, risk managers are faced with the question of which estimator they should choose. In general, there are two approaches that can provide an answer: backtesting and simulation. The problems with backtesting are that (i) its results will be bound to the features of the analyzed empirical dataset and (ii) sophisticated backtesting procedures for ES, which have become available just recently (see Du and Escanciano, 2017; Lösner et al., 2019), are limited to specific classes of ES estimators.[1] In contrast, simulation settings are more flexible because they allow the construction of various distributional settings to see in which environment an estimator behaves best and/or better than others. In the previous literature, we can find several studies comparing different ES estimators in such a way (see Chen, 2008; Jadhav

[1] For a review of the properties of the most influential backtests, see Novales and Garcia-Jorcano (2019).

et al., 2009; Yu et al., 2010). While their results deliver important insights into, for example, small-sample properties of the estimators, they are limited by the fact that they tend to compare a newly proposed estimator to a simple benchmark method or a closely related estimator. Thus, to the best of our knowledge, no simulation study provides a structured comparison of the most popular estimators.

We fill this gap with a comprehensive analysis of the estimation error of non-parametric ES estimators, specifically, the classic historic estimator, several weighted historic estimators (see Inui and Kijima, 2005; Peracchi and Tanase, 2008), outlier-robust historic estimators (see Jadhav et al., 2009) and some kernel estimators (see Scaillet, 2004; Chen, 2008). We focus on this class of estimators because it does not require the assumption of a specific loss distribution and is thus less prone to misspecification error than parametric estimators. Furthermore, commercial banks have shown a preference for non-parametric methods of estimating VaR (see Pérignon and Smith, 2010) and therefore might also tend to non-parametric ES estimators. As parametric benchmarks we include the classic normal distribution approach and the well-known peak over threshold method (originating from extreme value theory and propagated by McNeil and Frey, 2000) because the former is the simplest available estimator and the latter has been shown to outperform many other parametric techniques (such as mixture distributions and other non-normal classes) in the context of VaR estimation (see Kuester et al., 2006; Abad et al., 2014). Our main goal is to derive a guide for selecting the appropriate ES estimator. In other words, our results support a decision maker in the process of finding the estimator most suitable for a given situation which is characterized by the properties of the available data (e. g., the degrees of asymmetry and tail strength), the sample size and the desired confidence level.

Alongside this main contribution, we present two additional original analyzes. First, we illustrate how a mathematical programming technique called performance profile (see Dolan and Moré, 2002), which was developed for benchmarking and comparing optimization software, can be applied to the evaluation of ES estimators. Specifically, we show that its comprehensive graphical outputs can supply valuable information not contained in simple estimator rankings based on standard measures of estimation error. Second, we analyze whether a result intensively studied in the forecasting literature can be used to construct better ES estimates. It has been shown that combining forecasts originating from different models via averaging can generate new forecasts that outperform the original models (see Timmermann, 2006; Weron, 2014). We investigate whether averaging the estimates of different ES estimation techniques has a similar effect.

The remainder of our study is structured as follows. Section 1.2 presents our selection of ES estimators subdivided in benchmark estimators, non-parametric estimators and a simple combined estimate. Section 1.3 outlines the simulation setup (including distribution and parameter choices) and the measures of estimation error (covering standard metrics and performance profiles). Section 1.4 discusses our simulation results by distinguishing between settings particularly relevant for banks (Basel framework) and non-banks (corporate framework). Moreover, it looks at additional variants of combined estimators and reports the outcomes of some robustness checks. Section 1.5 concludes and outlines directions for future research.

1.2. Estimators of expected shortfall

1.2.1. Preliminary definitions

Before discussing the nature of our different estimators, we have to introduce some notation and formally define ES.

Let (X_t) be a time series of negative asset returns (i. e., returns multiplied by -1) or losses X_t, $t = 1, \ldots, n$, which is assumed to be a sequence of independently and identically distributed (iid)

random variables. For a given high confidence level γ (for example, 97.5 %), VaR_γ is then defined as the γ-quantile of the cumulative distribution function (cdf) of (X_t).

With the probability density function (pdf) f of the losses, ES_γ is given by

$$\text{ES}_\gamma = \frac{1}{1-\gamma}\int_{\text{VaR}_\gamma}^{\infty} x f(x)\mathrm{d}x = \frac{1}{1-\gamma}\int_{\gamma}^{1}\text{VaR}_v\mathrm{d}v. \tag{1.1}$$

In the continuous case, this expression equates to the tail conditional expectation (see Artzner et al. (1999)),

$$\text{ES}_\gamma = \mathbb{E}(X_t | X_t \geq \text{VaR}_\gamma), \tag{1.2}$$

where $\mathbb{E}(.)$ denotes the mean function. Hence, the dedicated ES_γ is the expected value of all X_t exceeding VaR_γ. Since the probability of loss larger than VaR_γ equals $1-\gamma$, ES_γ represents the expected loss in the unlikely worst-case scenario of a tail-event.

1.2.2. Benchmark estimators

1.2.2.1. Normal method

One of the simplest techniques to estimate ES is to assume that losses are normally distributed with mean μ and standard deviation σ (see McNeil et al., 2005, chpt. 2.2.4). Using (1.1) and inverse integration by substitution then delivers

$$\text{ES}_\gamma^{ND} = \mu + \frac{\sigma}{1-\gamma}\phi\left(\Phi^{-1}(\gamma)\right), \tag{1.3}$$

where ϕ is the standard normal pdf and Φ^{-1} is the inverse standard normal cdf. Thus, to obtain an empirical estimate, we simply have to estimate μ and σ via their sample counterparts and plug the resulting values into Equation (1.3).

1.2.2.2. Peak-over-threshold method

Because the ES focuses on extreme losses, extreme value theory is a particularly interesting tool for the derivation of new ES estimators. So far, most research in this field has focused on VaR estimation (see Brooks et al., 2005; Mögel and Auer, 2018) but can easily be extended to ES estimation (see McNeil and Frey, 2000; Martins-Filho and Yao, 2006; Martins-Filho et al., 2018). Motivated by its popularity and persuasive backtest performance for high confidence levels (see Gençay and Selçuk, 2004), we let the peak over threshold (POT) method represent the class of estimators based on extreme value theory.

The POT method builds on the limit theorem of Balkema and de Haan (1974) and Pickands (1975). In our context, this theorem states that, for (almost) any form of loss distribution, the distribution of excesses $Y_t := X_t - u$ over a large threshold u is well approximated by the generalized Pareto distribution (GPD). This result is important because it allows us to model the tail of the loss distribution without having to specify the specific form of the loss distribution. In other words, we can derive the ES based on the cdf of the excesses, which is given by

$$G(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0, \\ 1 - e^{-\frac{y}{\sigma}} & \text{if } \xi = 0, \end{cases} \tag{1.4}$$

where ξ and $\sigma > 0$ are shape and scale parameters, respectively (see McNeil, 1997). The support of this function is $y \geq 0$ when $\xi \geq 0$ and $0 \leq y \leq -\frac{\sigma}{\xi}$ when $\xi < 0$.

The cdf of the excesses implies the following cdf for the losses over u:

$$F(x) = \begin{cases} 1 - q\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0, \\ 1 - qe^{-\frac{x-u}{\sigma}} & \text{if } \xi = 0, \end{cases} \tag{1.5}$$

where $q > 1-\gamma$ is the percentage of losses X_t exceeding u. Consequently, the VaR can be obtained by inverting Equation (1.5), i. e.,

$$\mathrm{VaR}_\gamma^{POT} = \begin{cases} u + \frac{\sigma}{\xi}\left(\left(\frac{1-\gamma}{q}\right)^{-\xi} - 1\right) & \text{if } \xi \neq 0, \\ u - \sigma \ln\left(\frac{1-\gamma}{q}\right) & \text{if } \xi = 0, \end{cases} \tag{1.6}$$

and the ES for $\xi < 1$ by using (1.1) and integration by substitution (see McNeil and Frey, 2000):

$$\mathrm{ES}_\gamma^{POT} = \frac{\mathrm{VaR}_\gamma^{POT} - \xi u + \sigma}{1-\xi} \text{ for } \xi < 1. \tag{1.7}$$

With this knowledge at hand, we can estimate the ES by setting u, fitting a GPD to the corresponding excesses and plugging the estimated GPD parameters into Equation (1.7). However, note that the choice of u can be delicate because, if it is too high (low), the fit of the GPD will be poor (the above limit theorem will not be satisfied).

1.2.3. Non-parametric estimators

1.2.3.1. Historic methods and modifications

Historic estimators do not require fitting theoretical distributions to empirical data and are thus quite easy to implement. The only assumption which is implicitly made when using them is that a given data sample adequately represents the properties of the underlying sequence (X_t).

Using the historic VaR_γ^H in its established form (see Pritsker, 2006), the classic historic ES estimator can be defined as

$$\mathrm{ES}_\gamma^H = \mathbb{E}(X_t | X_t \geq \mathrm{VaR}_\gamma^H) = \frac{\sum_{t=1}^n X_t I(X_t \geq \mathrm{VaR}_\gamma^H)}{\sum_{t=1}^n I(X_t \geq \mathrm{VaR}_\gamma^H)}, \tag{1.8}$$

where $\mathrm{VaR}_\gamma^H = X_{(\lceil n\gamma \rceil)}$ and $I(.)$ is the mathematical indicator function which maps to 1 if its argument is true and to 0 otherwise. Here, $X_{(i)}$ denotes the ith order statistic of (X_t), which describes the ith-smallest value of (X_t), and $\lceil . \rceil$ is the ceiling function, which rounds its argument up to the next integer if it is not an integer. Because $n\gamma$ is rounded up even if $n\gamma$ has only a small decimal place, rounding up can introduce a non-negligible error compared to rounding down. To deal with this source of error, several modifications of Equation (1.8) have been proposed. If $n\gamma$ is natural, these variants reduce to the classic historic method.

The first modification dates back to Peracchi and Tanase (2008) and can be written as

$$\mathrm{ES}_\gamma^{H1} = \mathrm{ES}_\gamma^H + \left(1 - \frac{\lfloor n(1-\gamma) \rfloor}{n(1-\gamma)}\right) X_{(\lfloor n\gamma \rfloor)}, \tag{1.9}$$

where a correction term is added to the classic historic estimator and the floor function $\lfloor . \rfloor$ rounds down its argument if it is not an integer. The correction focuses on the loss $X_{(\lfloor n\gamma \rfloor)}$, which, if we sorted (X_t) in ascending order, would stand right before VaR_γ^H. Thus, it can be interpreted as the VaR we would use if we preferred underestimating risk. $X_{(\lfloor n\gamma \rfloor)}$ is weighted by the factor $\left(1 - \frac{\lfloor n(1-\gamma) \rfloor}{n(1-\gamma)}\right)$ which tends towards 1 (0) if $n\gamma$ has a small (large) non-zero decimal place.

A second modification is sketched in Nadarajah et al. (2014, sec. 4.5) and uses

$$\mathrm{ES}_\gamma^{H2} = \gamma \mathrm{ES}_\gamma^H + (1-\gamma)\mathbb{E}(X_t | X_t \geq X_{(\lfloor n\gamma \rfloor)}). \tag{1.10}$$

Here, the idea is to use a weighted sum of the classic ES_γ^H and the ES which results from underestimating risk. The weights are chosen to be γ and $1-\gamma$, respectively.

In a last variant by Inui and Kijima (2005), the weights γ and $1-\gamma$ in Equation (1.10) are replaced by $1-(\lceil n\gamma \rceil - n\gamma)$, which is the decimal place of $n\gamma$, and $\lceil n\gamma \rceil - n\gamma$, respectively. This gives

$$\mathrm{ES}_\gamma^{H3} = (1 - \lceil n\gamma \rceil + n\gamma)\mathrm{ES}_\gamma^H + (\lceil n\gamma \rceil - n\gamma)\mathbb{E}(X_t | X_t \geq X_{(\lfloor n\gamma \rfloor)}), \tag{1.11}$$

where we put more weight on the ES estimate corresponding to a smaller VaR if $n\gamma$ has a small decimal place.

Because all historic estimators discussed so far are (more or less) means of the largest losses, they are sensitive to outliers. To obtain more robust estimators, Jadhav et al. (2009) propose to eliminate outliers in the data by specifying a constant $a \in [0, 0.1]$, chosen by the user to express the risk of having outliers, and the function

$$k(t) = (n+1)\left(1 - \gamma - \frac{t(1-\gamma)}{\lfloor n(1-\gamma) \rfloor + 1}\right) \text{ for } t \in \mathbb{R}, \tag{1.12}$$

which yield the alternative estimator

$$\mathrm{ES}_\gamma^{J1} = \frac{1}{\lfloor n(1-\gamma)^{1+a} \rfloor + 2} \sum_{t=0}^{\lfloor n(1-\gamma)^{1+a} \rfloor + 1} X_{(n-\lfloor k(t) \rfloor)}. \tag{1.13}$$

If a is small enough, ES_γ^{J1} can reduce to the classic historic estimator. Otherwise, k excludes some of the largest losses in the estimation of the ES.

Using the weighting function $w(t) := k(t) - \lfloor k(t) \rfloor$ for $t \in \mathbb{R}$ we can obtain another version of this estimator which is

$$\mathrm{ES}_\gamma^{J2} = \frac{1}{\lfloor n(1-\gamma)^{1+a} \rfloor + 2} \sum_{t=0}^{\lfloor n(1-\gamma)^{1+a} \rfloor + 1} (1 - w(t))\, X_{(n-\lfloor k(t) \rfloor)} + w(t) X_{(n-1-\lfloor k(t) \rfloor)}. \tag{1.14}$$

Here, we compute a weighted sum of ES_γ^{J1} and a smaller ES which results from averaging smaller values of (X_t), i. e., ignoring additional large values of (X_t).

1.2.3.2. Kernel density methods

While historical methods capture ES by directly averaging discrete data, kernel density methods estimate the distribution function F via smoothing techniques and then use the results of this preliminary step to derive an ES estimate. In our comparative study, we consider two of the most popular approaches of this field, both of which require the specification of a kernel function K, which has to be a symmetric pdf, and a positive bandwidth h, which determines the degree of smoothing, but differ with respect to their specific form of estimating VaR.

Our first technique follows the seminal work of Nadaraya (1964) by estimating the relevant distribution function F via

$$\hat{F}(x) = \frac{1}{n}\sum_{t=1}^{n} G_h(x - X_t) \text{ with } G_h(x) = \int_{-\infty}^{\frac{x}{h}} K(u)\mathrm{d}u. \tag{1.15}$$

Here, G_h operates like a rescaled (in the sense of using the integration limit $\frac{x}{h}$ instead of x) cdf of the kernel density K. It sums larger values (from 0.5 up to one), if X_t is smaller than or equal to the argument x, and smaller values (between zero and 0.5), if X_t exceeds x. Thus, the resulting $\hat{F}$ describes an average of the rescaled cdf G_h over all data points X_t. Based on this estimated distribution function $\hat{F}$, VaR_γ^{K1} can be derived as the inverse solution of $\hat{F}(x) = \gamma$.

Our second method estimates the VaR as a kernel-weighted sum of order statistics, as suggested by Parzen (1979). In general, there are various possibilities to compute weights based on the kernel density function K. However, because Sheather and Marron (1990) show that many estimators resulting from different forms of weights $w(t)$ are asymptotically equivalent, we focus our attention on the most popular form of weighting:

$$w_\gamma(t) = \frac{1}{h} \int_{\frac{t-1}{n}}^{\frac{t}{n}} K\left(\frac{u-\gamma}{h}\right) \mathrm{d}u \text{ for } t = 1, \ldots, n. \tag{1.16}$$

As we can see, the weights in this specification originate from sub-areas of the integral of the compressed and shifted kernel density K between 0 and 1. The maximum weight $w_\gamma(t)$ is reached when $\frac{t-1/2}{n} = \gamma$. Because the single weights $w_\gamma(t)$ do not sum to unity, we have to divide them by their sum, which delivers the VaR estimator

$$\text{VaR}_\gamma^{K2} = \frac{1}{\sum_{t=1}^n w_\gamma(t)} \sum_{t=1}^n w_\gamma(t) X_{(t)}. \tag{1.17}$$

As shown by Scaillet (2004), in kernel density approaches, the ES can be estimated as a scaled, weighted sum of X_t. That is, plugging in either VaR_γ^{K1} or VaR_γ^{K2}, our two versions of kernel-based ES estimators are given by

$$\text{ES}_\gamma^{K1,K2} = \frac{1}{n(1-\gamma)} \sum_{t=1}^n X_t G_h(X_t - \text{VaR}_\gamma^{K1,K2}), \tag{1.18}$$

where the weights $G_h(X_t - \text{VaR}_\gamma^{K1,K2})$ are large if a data point X_t exceeds the estimated $\text{VaR}_\gamma^{K1,K2}$ and small otherwise.

The quality of a kernel estimate typically depends less on the form of K than on the magnitude of its bandwidth h (see Bowman and Azzalini, 1997, chpt. 1.2). Too-small values of h do not induce much smoothing (i. e., lead to spiky estimates which exaggerate some characteristics of a sample), whereas too-large values cause oversmoothing (i. e., obscuring most of the structure of the data). Consequently, an optimal bandwidth value should be chosen, which can be derived from data-dependent bandwidth selection techniques studied in the context of VaR_γ^{K2} (see Sheather and Marron, 1990; Cheng and Sun, 2006) as well as VaR_γ^{K1} and ES (see Sheather and Jones, 1991; Wand and Jones, 1995; Bowman et al., 1998; Cheng and Sun, 2006; Raykar and Duraiswami, 2006). In addition, it is instructive to know that under some data assumptions and if the bandwidth satisfies $h \to 0$, $nh^{3-\beta} \to \infty$ for any $\beta > 0$, and $nh^4 \log^2(n) \to 0$ as $n \to \infty$, the error behavior of ES_γ^{K1} and ES_γ^{H} is similar (see Chen, 2008).

1.2.4. Combined estimation

Studies from a variety of fields have shown that combining the predictions of different forecasting models often yields more accurate results than the best individual model (see Timmermann, 2006; Weron, 2014). Another interesting finding of this strand of the literature is, that when forming combinations via weighted averages, equal weights are superior to optimal weights estimated based on specific (potentially erroneous) criteria (see Genre et al., 2013; Claeskens et al., 2016).

Based on these results, we extend our set of ES estimators by another one. That is, we calculate the simple average

$$\mathrm{ES}_\gamma^{MV} = \frac{\mathrm{ES}_\gamma^{ND} + \mathrm{ES}_\gamma^{POT} + \mathrm{ES}_\gamma^{H} + \mathrm{ES}_\gamma^{H1} + \mathrm{ES}_\gamma^{H2} + \mathrm{ES}_\gamma^{H3} + \mathrm{ES}_\gamma^{J1} + \mathrm{ES}_\gamma^{J2} + \mathrm{ES}_\gamma^{K1} + \mathrm{ES}_\gamma^{K2}}{10}$$

across all of our estimators and compare its performance to the individual estimation methods. While the focus of our main analysis in Sections 1.4.1 and 1.4.2 will be on this overall combination ES_γ^{MV}, Section 1.4.3 looks at combinations formed by picking only a few of the ES estimators.

1.3. Simulation setup

1.3.1. General design

Simulation studies analyzing the quality of VaR or ES estimators typically assume that losses are generated by specific stochastic processes because this makes it possible to (i) calculate the true ES of the processes; and (ii) evaluate whether an estimator delivers values close to it in repeated random samples. A popular approach is to use a time series model with non-normal disturbances, where the equation structure incorporates empirical autocorrelation in returns and variances (see Campbell et al., 1993; Bollerslev et al., 1992) and the distribution of the disturbances reflects empirical skewness and fat-tails of returns (see Cont, 2001). Generalized autoregressive conditional heteroscedasticity models (see Manganelli and Engle, 2001; Chen, 2008) and related non-linear classes (see Martins-Filho and Yao, 2006; Martins-Filho et al., 2018) are typical examples of settings that have been used in previous research.

When evaluating ES estimators in such environments, we first have to estimate the parameters of the time series model based on a simulated sample, then apply the ES formulas discussed above to the (mean 0 and variance 1) model residuals and finally use the time-varying mean and variance predictions of the time series model to scale the obtained ES estimate to its proper level.[2] Consequently, the simulated estimation error of an ES estimator has two components: the non-negligible estimation error related to the time series model (see Mancini and Trojani, 2011; Kellner and Rösch, 2016; Pitera and Schmidt, 2018) and the actual (pure) error of the ES formula. Because we are interested in the pure error of our ES estimators, the design of our simulation study is close to Peracchi and Tanase (2008) and Yu et al. (2010) who simulate iid returns from normal, student t as well as normal and t mixture distributions. We differ from their approach by using the skewed t distribution of Hansen (1994) (see Section 1.3.2) because this distribution allows a flexible modeling of skewness and fat tails in a unified framework.[3]

We specify several settings with different distributional characteristics (see Section 1.3.3) and, within each setting, simulate m time series (X_t) of length n. We then use our estimators to produce ES estimates for each time series. Since we know the true ES, we can capture the estimation error over all time series with summary measures of distance (see Section 1.3.4). This allows us to identify the best and worst estimator(s) for a given distributional setting and to compare the performance of estimators across settings.

Deriving estimator rankings is the main goal of our study. Focusing on the pure error simplifies our simulation design but does not mean that it delivers unrealistic rankings in the light of typical empirical time series behavior (e. g. volatility clustering). Our simulated data is constructed such that it can be interpreted either as losses or the residuals of a standard time series model. This

[2] These steps are required because our ES formulas are designed for iid data and the residuals can be considered iid provided that the time series model is correctly specified (see Kuester et al., 2006). For an interesting new scaling approach, see Thavaneswaran et al. (2019).

[3] A similarly flexible alternative would be the family of g-and-h distributions, which just recently found its way into risk management applications (see Degen et al., 2007).

has two consequences. First, the rankings of our ES estimators are the same in an iid and a linked non-iid simulation setting. Second, decision makers working with non-iid data and established time series models can also use our results to look up the ideal ES estimator. They simply have to check which of our specified distributional settings fits their model residuals best.

1.3.2. Data-generating process

To simulate loss data, we use the skewed t distribution of Hansen (1994), which is characterized by the pdf

$$f(x) = \begin{cases} bc\left(1 + \frac{1}{\nu-2}\left(\frac{bx+a}{1-\lambda}\right)^2\right)^{-\frac{\nu+1}{2}} & \text{if } x < -\frac{a}{b}, \\ bc\left(1 + \frac{1}{\nu-2}\left(\frac{bx+a}{1+\lambda}\right)^2\right)^{-\frac{\nu+1}{2}} & \text{if } x \geq -\frac{a}{b}, \end{cases} \tag{1.19}$$

where $2 < \nu < \infty$, $-1 < \lambda < 1$ and

$$a = 4\lambda c\frac{\nu-2}{\nu-1}, \qquad b^2 = 1 + 3\lambda^2 - a^2, \qquad c = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi(\nu-2)}\Gamma(\nu/2)}. \tag{1.20}$$

Per definition, a skewed t random variable has zero mean ($\mu_1 = 0$) and unit variance ($\mu_2 = 1$). If $\nu > 3$ and $\nu > 4$, respectively, its skewness (i. e., the third standardized moment) μ_3 and kurtosis (i. e., the fourth standardized moment) μ_4 are (see Jondeau and Rockinger, 2003):

$$\mu_3 = \left(m_3 - 3a\left(1 + 3\lambda^2\right) + 2a^3\right)/b^3, \qquad \mu_4 = \left(m_4 - 4am_3 + 6a^2\left(1 + 3\lambda^2\right) - 3a^4\right)/b^4, \tag{1.21}$$

where

$$m_3 = 16c\lambda\left(1 + \lambda^2\right)\frac{(\nu-2)^2}{(\nu-1)(\nu-3)}, \qquad m_4 = 3\frac{\nu-2}{\nu-4}\left(1 + 10\lambda^2 + 5\lambda^4\right). \tag{1.22}$$

If $\lambda = 0$, the skewed t distribution reduces to the student t distribution and if additionally $\nu \to \infty$, it converges to the standard normal distribution. Thus, the parameters λ and ν control the degree of skewness and kurtosis, respectively.

The true VaR_γ of a skewed t random variable can be obtained as the γ-quantile of the inverse cdf and the associated true ES_γ via (1.1).

1.3.3. Parameter specifications

To conduct our analysis, we need to specify the skewed t parameters used for simulating loss data and the parametrization of our ES estimators.

Simulation Based on typical empirical values of skewness and kurtosis (see, for example, Campbell et al., 1997; Peiró, 1999, tab. 1.1 and 1, respectively), we define five distributional settings: (a) positively skewed and light-tailed ($\mu_3 = 1$, $\mu_4 = 5$), (b) positively skewed and fat-tailed ($\mu_3 = 1$, $\mu_4 = 60$), (c) negatively skewed and light-tailed ($\mu_3 = -1$, $\mu_4 = 5$), (d) negatively skewed and fat-tailed ($\mu_3 = -1$, $\mu_4 = 60$) and (e) standard normal.[4] To obtain a general picture of estimation quality that is not linked to a specific parameter setting, we also look at a set of time series (f)

[4] In terms of λ and ν, i. e., with respect to (1.21), this means that we use the parameters (a) $\lambda = 0.4784$ and $\nu = 10.1389$, (b) $\lambda = 0.1575$ and $\nu = 4.1242$, (c) $\lambda = -0.4784$ and $\nu = 10.1389$ as well as (d) $\lambda = -0.1575$ and $\nu = 4.1242$.

resulting from a merger of all settings (a)–(e). We generally assume 252 trading days per year (as in Chan, 2013, chpt. 6) and simulate series of lengths $n \in \{21, 126, 252, 504, 1008\}$, which represent daily losses over one month, six months as well as one, two and four years, respectively. The number of time series simulated for each setting is $m = 10^5$.

Estimation We calculate the ES estimators of Section 1.2 for a confidence level of 97.5 %, as suggested by regulators (see Basel Committee of Banking Supervision, 2012), as well as alternative levels of 95 % and 99 %, which have been common in VaR calculation (see Basel Committee of Banking Supervision, 1996, 2004; Gilli and Këllezi, 2006). For J1 and J2, we represent the risk of having outliers via $a = 0.07$. In kernel-based estimation K1 and K2, we choose the standard normal pdf to be the kernel function K and follow bandwidth selection rules of Cheng and Sun (2006, Method 4) and Wand and Jones (1995) for choosing h. Finally, we set the threshold-exceeding percentage in the POT method to $q = 0.1$ because McNeil and Frey (2000) and Herrera (2013) show numerically that this choice can reduce potential errors.

1.3.4. Evaluation methods

1.3.4.1. Basic measures

To evaluate the reliability of an estimator $\hat{\theta}$ for a parameter θ researchers typically use its mean squared error $\text{MSE}(\hat{\theta}) = \text{Bias}(\hat{\theta}, \theta)^2 + \text{Variance}(\hat{\theta})$ because it summarizes its average deviation from the true parameter value and the risk of obtaining estimates crucially deviating from the mean estimate (see Greene, 2003, chpt. C.5). In the following, we have a closer look at measures capturing these two components.

Because it is debatable whether over- or underestimation is the smaller evil (see Mögel and Auer, 2018) and because we wish to derive a clear ranking of ES estimators, our main error measure is the mean absolute percentage error (MAPE), which, for an estimator j and m simulated time series, is defined as

$$\text{MAPE}^j = \frac{1}{m}\sum_{i=1}^{m}\text{APE}^j(i) = \frac{100\,\%}{m}\sum_{i=1}^{m}\left|\frac{\text{ES}_\gamma^j(i) - \text{ES}_\gamma}{\text{ES}_\gamma}\right|. \tag{1.23}$$

Because the MAPE is scale-independent, it can not only be compared across estimators but also across distributional settings (see Hyndman and Koehler, 2006). It illustrates the average absolute percentage deviation of an estimator j from the true (positive) ES_γ associated with a distributional setting. While our main analysis focuses on this measure, we also calculate the mean percentage error (MPE) via $\text{MPE}^j = \frac{100\,\%}{m}\sum_{i=1}^{m}\frac{\text{ES}_\gamma^j(i)-\text{ES}_\gamma}{\text{ES}_\gamma}$ and discuss its implications for overestimation ($\text{MPE}^j > 0$) and underestimation ($\text{MPE}^j < 0$) in Section 1.4.3.1.

To judge the variability of an ES estimator j, we additionally compute the relative standard deviation (RSD) of the estimates produced by it. That is, we use

$$\text{RSD}^j = \frac{100\,\%}{\mathbb{E}\left(\text{ES}_\gamma^j\right)}\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\text{ES}_\gamma^j(i) - \mathbb{E}\left(\text{ES}_\gamma^j\right)\right)^2}, \tag{1.24}$$

where $\mathbb{E}\left(\text{ES}_\gamma^j\right)$ is the mean of ES_γ^j over all m estimates.

1.3.4.2. Performance profile construction

If a simulated time series i strongly differs from the underlying theoretical distribution (for example, by having a large number of extreme values), the associated APE of an estimator j will be significantly higher than for other time series. Thus, if we rank estimators based on their APE,

the ranking in such extreme cases may differ from the ranking in regular situations. In an overall ranking based on MAPE, a small number of extreme time series may strongly influence (or dominate) general conclusions. In the worst case, just one crucial estimation failure could substantially downgrade an estimator in its overall rank.

The performance profile technique of Dolan and Moré (2002) can handle such complications because its different perspective makes it less sensitive to rare extreme errors than the simple averages used in the computation of error means or standard deviations.[5] In a first step, this approach − when transferred to our specific risk measurement application − uses APE values (as defined in Section 1.3.4.1) to derive a performance ratio

$$r_{i,j} = \frac{\mathrm{APE}^j(i)}{\min\limits_j\{\mathrm{APE}^j(i)\}} \tag{1.25}$$

for each time series i and each estimator j.[6] This ratio compares the error of an estimator j for a time series i with the error of the best estimator for this series. If j is the best estimator for i, we have $r_{i,j} = 1$, and $r_{i,j} > 1$ otherwise.

In the second step, we consider, for each estimator j, a function

$$\rho_j(x) = \frac{1}{m}|\{i|r_{i,j} \leq x\}| \text{ for } x \in \mathbb{R}, \tag{1.26}$$

which can be interpreted as the cumulative distribution function of estimator j's performance ratios because it captures the proportion of simulated time series with performance ratios less than or equal to the argument x. In other words, for each $x \in \mathbb{R}$, $\rho_j(x)$ is the probability that the APE of estimator j is within a factor x of the best APEs.

Plotting such functions for a given simulation setting enables us to compare the quality of our ES estimators in a comprehensive and elegant fashion. On the one hand, $\rho_j(1)$ tells us the probability that no other method has APE values better than estimator j. This implies that, if we are only interested in a ranking based on the number of wins (i. e. , if we are searching for the estimator which most frequently delivers the smallest APE), we have to compare the $\rho_j(1)$ values of the estimators. On the other hand, $\rho_j(x)$ with $x > 1$ helps us to identify estimators which may not be the very best in the majority of problems, but offer good estimation results (i. e., belong to the top estimators) in situations the other methods fail. Such estimators have a good overall assessment and therefore a high statistical efficiency. If the function ρ_j is steep, the associated estimator j can be considered very robust.

1.4. Results

1.4.1. Basel framework

We start our discussion of results with a parameter constellation particularly relevant for banks because it covers the suggestions of the Basel Committee of Banking Supervision (2012). That is, we focus on a 97.5 % confidence level and a one-year time horizon.

1.4.1.1. Basic assessment

For the Basel constellation and our different distributional settings, Table 1.1 presents the true ES values and the means of the estimates produced by our set of ES estimators. Furthermore, it reports the associated MAPE and RSD values of each estimator.

[5] For a general discussion of outlier sensitivity, see Wilcox (2012).
[6] Note that we add a small constant (10^{-5}) to APE to guarantee numerical feasibility.

Table 1.1.: ES estimates, MAPE and RSD for $\gamma = 97.5\,\%$ and $n = 252$

	ES	ND	POT	H	H1	H2	H3	J1	J2	K1	K2	MV
True ES and mean of ES estimates												
(a)	3.08	2.33	3.05	2.95	3.06	2.95	2.93	2.76	2.63	3.03	3.19	2.91
(b)	3.14	2.32	3.13	2.98	3.08	2.98	2.94	2.66	2.50	3.02	3.17	2.90
(c)	1.76	2.33	1.74	1.71	1.78	1.71	1.70	1.64	1.60	1.73	1.92	1.78
(d)	2.44	2.32	2.42	2.33	2.41	2.33	2.31	2.12	2.01	2.36	2.43	2.32
(e)	2.34	2.34	2.31	2.27	2.36	2.27	2.25	2.18	2.11	2.31	2.51	2.29
MAPE of ES estimates												
(a)	-	24.28	11.40	**10.71**	***10.22***	10.72	10.82	12.96	15.58	**10.67**	21.23	13.86
(b)	-	26.23	19.69	**15.55**	***14.93***	15.56	15.61	17.90	21.28	**15.20**	24.90	18.68
(c)	-	32.77	6.62	**6.60**	***6.30***	**6.60**	6.68	8.03	9.67	6.93	21.21	11.14
(d)	-	***8.71***	17.49	**13.54**	**12.86**	**13.54**	13.60	15.73	18.74	14.06	21.49	14.98
(e)	-	***4.17***	**6.82**	7.04	**6.75**	7.04	7.15	8.61	10.44	6.99	19.98	8.50
(f)	-	19.23	12.40	**10.69**	***10.21***	**10.69**	10.77	12.65	15.14	10.77	21.76	13.43
RSD of ES estimates												
(a)	-	***7.97***	45.32	13.11	12.93	13.09	12.96	12.48	**12.12**	13.48	46.11	**11.37**
(b)	-	***12.02***	97.21	20.02	19.66	20.00	19.73	**16.80**	**15.77**	19.37	50.88	17.96
(c)	-	***5.42***	8.85	7.96	7.83	7.95	7.85	**7.47**	**7.19**	8.75	54.32	7.74
(d)	-	***10.47***	310.68	17.23	16.89	17.20	16.95	**14.47**	**13.52**	17.95	43.71	34.84
(e)	-	***5.22***	8.61	8.45	8.36	8.44	8.38	8.34	**8.24**	8.73	46.07	**8.01**
(f)	-	***8.22***	94.13	13.35	13.14	13.34	13.17	**11.91**	**11.37**	13.65	48.22	15.98

For a confidence level of $\gamma = 97.5\,\%$ and a sample size of $n = 252$, this table presents the true expected shortfall (ES) as well as the average, the mean absolute percentage error (MAPE) and the relative standard deviation (RSD) of the ES estimates produced by our parametric and non-parametric techniques (defined in Section 1.2) in different simulation settings (specified in Section 1.3.3). The lowest (second and third lowest) MAPE and RSD values within each setting are marked in italics (bold). The estimators are abbreviated as follows: ND $\widehat{=}$ normal distribution; POT $\widehat{=}$ peak over threshold; H $\widehat{=}$ classic historic method; H1, H2, H3 $\widehat{=}$ modified historic methods related to Peracchi and Tanase (2008), Nadarajah et al. (2014) and Inui and Kijima (2005), respectively; J1, J2 $\widehat{=}$ outlier-adjusted historic methods of Jadhav et al. (2009); K1, K2 $\widehat{=}$ kernel-oriented estimators of Scaillet (2004) based on Nadaraya (1964) and Parzen (1979), respectively; MV $\widehat{=}$ combined estimate averaging all estimators. The distributional settings are labeled as follows: (a) positive skew, light tail; (b) positive skew, fat tail; (c) negative skew, light tail; (d) negative skew, fat tail; (e) normally distributed; (f) merger of settings (a)–(e). With the exception of (f), all settings simulate $m = 10^5$ time series.

A first look at the true ES values shows that, not surprisingly, they are larger for positively skewed (a, b) than for negatively skewed losses (c, d). Similarly, the ES is higher for fat-tailed (b, d) than for light-tailed distributions (a, c). The ES of normally distributed data lies between the ones for light-tailed data with positive (a) and negative (c) skewness. Turning to the estimators, we see that the normal method does not estimate case-sensitive. That is, because of its focus on mean and variance (which is the same in all distributional settings) its average estimate is similar across (a)–(e). Another noteworthy observation is that, while the estimator K2 tends to overestimate the ES in the majority of situations, the other estimators tend to underestimate.

As far as the MAPE and RSD values of the estimators are concerned, we can identify H1 as a favorable estimator in settings (a)–(c), whereas ND outperforms in settings (d) and (e). While the reason for the good performance of ND in setting (e) is obvious, the result in the non-normal environment (d) may appear puzzling at first glance. It can be explained by the fact that our parametrization has unintentionally generated a situation where the case-insensitive average estimates of ND are close to the true ES of the non-normal data. Thus, we should not conclude that ND is generally preferable in negatively skewed and fat-tailed data. However, we can nicely see that the very simple ND approach can have benefits in non-normal data especially because, in our setting, skewness and kurtosis levels are set to values typically observed in practice.

Most ES estimators perform best in the negatively skewed and light-tailed environment (c) because here, in contrast to the other settings, the tail data occurs with higher probability within a small interval such that even small samples contain sufficient tail information for high quality

estimates.[7] Because of the reverse reasoning, the POT method performs particularly poor in situations when insufficient tail data is available to adequately fit the GPD distribution. Especially in settings (b) and (d), we can observe large MAPE and strikingly high RSD values. A closer inspection of the detailed simulation outcomes tells us that the POT suffers from some rare but drastic misestimations. Consequently, proponents of the extreme value theory strand have two options. On the one hand, they could closely monitor the results of the method to identify outcomes of obviously unrealistic magnitude. In other words, they could rely on the technique in the situations it works well and switch to another estimator when there is indisputable evidence of failure. On the other hand, they could choose to generally work with a combination of the POT approach and other estimation methods.[8] MV is a simple representative of this option. Especially for settings (b) and (d), its MAPE and RSD values are significantly lower than the ones of POT.

In the overall picture, i. e., the merged setting (f), the H1 estimates stand out because, on average, they deviate only by 10.21 % from the true ES values and fluctuate by only 13.14 % around their mean. The other historic methods and the estimator K1 have MAPE and RSD values of similar magnitude, whereas most remaining non-parametric and parametric methods perform worse. With somewhat lower RSD, the MAPE values of J1 and J2 significantly exceed that of H1. K2 disappoints with the highest MAPE of all estimators. Unlike the POT method, where rare events caused some large misestimations, the error of K2 might be systematic because MAPE and RSD have a large magnitude in all settings. To reinforce our POT argumentation and to check whether our presumption for K2 is true, we apply the performance profile technique.

1.4.1.2. Performance profiles

To avoid profiles, which are overfilled with detail, we use the results of Section 1.4.1.1 to group estimators by type and widely similar performance. That is, while ND, POT, K1, K2 and MV are left as they are, the historic (H, H1, H2, H3) and outlier-adjusted (J1, J2) methods are put in two summary categories. Figure 1.1 presents the results, i.e., the functions $\rho_j(x)$ for $x \in [1, 3]$.

The performance plots confirm two of our results derived from Table 1.1. First, ND is again the superior estimator in settings (d) and (e). Its probability $\rho_j(1)$ of being the best method is the highest across all estimators and takes values of about 38 % and 45 %, respectively. Furthermore, ND estimates ES with the greatest efficiency. This is because its profile function ρ_j is significantly above the ones for the other estimators over the entire interval $[1, 3]$ and reaches maxima of about 57 % and 62 %, respectively. Second, despite of having large MAPE and RSD values in comparison to other estimators in most skewed settings, POT reveals its strengths in the performance profiles. In setting (c), it has the highest probability of being the best estimator. Even in the cases where it is not the leading estimator, it belongs to the best methods with a high probability of up to about 55 %. In settings (a), (b) and (d), it performs similar to K1 and mostly better than the historic methods. Thus, its extraordinary misestimations do not cause a downgrade in the profile method. Especially in the overall setting (f), we can see that POT is highly competitive.

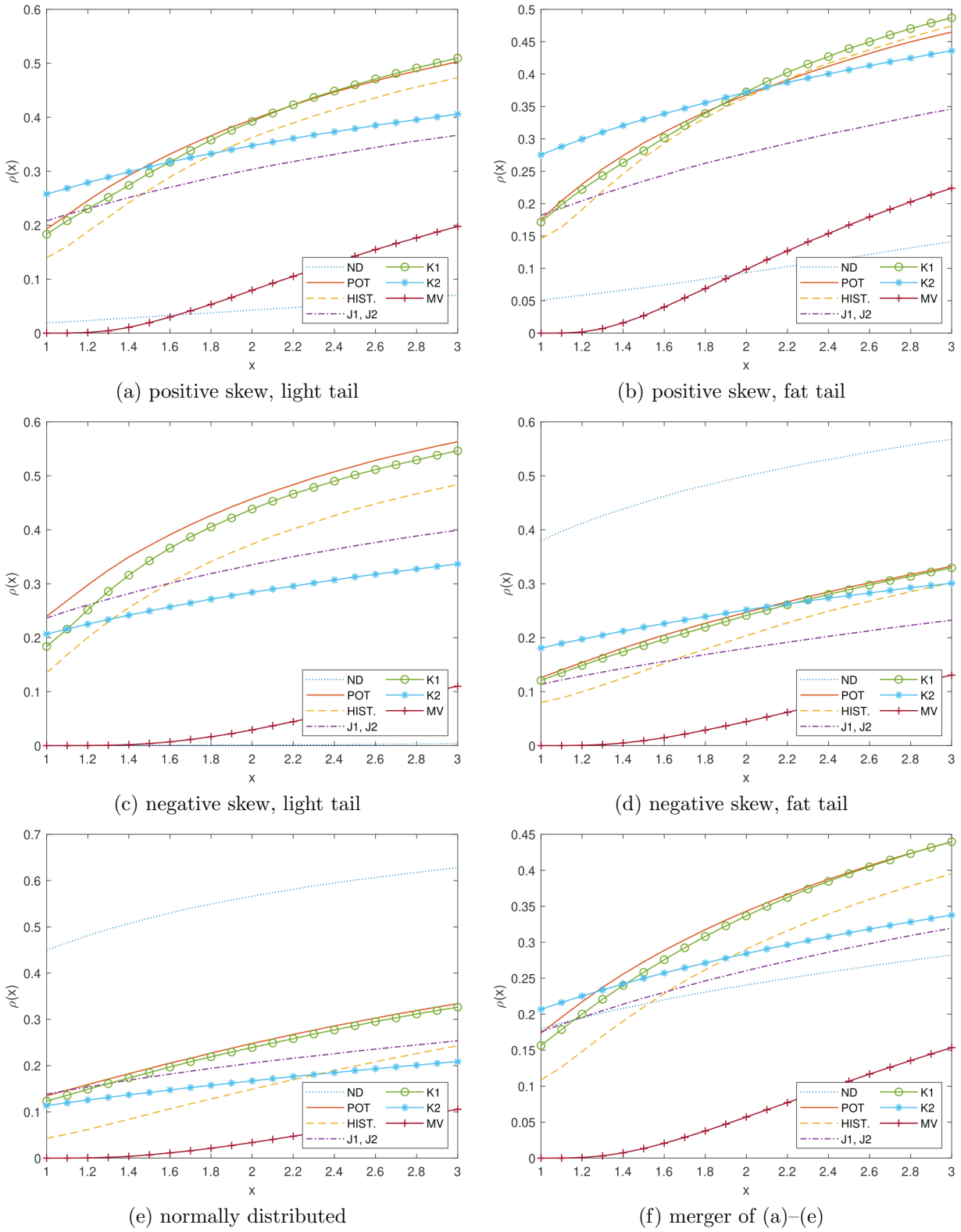
Quite interesting observations can be made for K2. While, in Table 1.1, its performance is not very persuasive, Figure 1.1 reveals that it has the highest $\rho_j(1)$ of about 25 % in settings (a) and (b). Furthermore, with respect to this criterion, it ranks first in the overall setting (f). However, the performance curves of K2 are not as steep as those of other estimators. This tells us that K2 often does not belong to the top estimators when it does not deliver the lowest error.

Finally, two other points are noteworthy. First, our combined estimator MV always either performs worst or only manages to outperform ND. This result calls for an analysis of combinations that go beyond our simple ad-hoc average. Second, when it comes to overall performance in

[7] Figure A.1 of the appendix illustrates and compares the pdfs of our different distributional settings.

[8] This is in fact a strategy which can often be found in the literature (see, for example, Taylor, 2008a; Schaumburg, 2012; Novales and Garcia-Jorcano, 2019).

Figure 1.1.: Performance profiles for $\gamma = 97.5\,\%$ and $n = 252$



(a) positive skew, light tail

(b) positive skew, fat tail

(c) negative skew, light tail

(d) negative skew, fat tail

(e) normally distributed

(f) merger of (a)–(e)

For a confidence level of $\gamma = 97.5\,\%$ and a sample size of $n = 252$, this figure plots the performance profiles (defined in Section 1.3.4.2) of our expected shortfall estimators. Each subfigure concentrates on a specific distributional setting (specified in Section 1.3.3). The estimators are abbreviated as in Table 1.1 and, for better visualization, the historic methods (H, H1, H2, H3) and the outlier-adjusted methods (J1, J2) have been summarized in groups.

setting (f), POT and K1 shine relative to the historic methods. This is in contrast to Table 1.1, which emphasizes H1. Thus, depending on whether we prefer a MAPE and RSD evaluation or a performance profile perspective, we end up with a different ideal estimator.

1.4.2. Corporate framework

1.4.2.1. Varied time horizon

While banks have to estimate ES based on an one-year horizon, risk managers of other industries may use different time periods. Therefore, it is instructive to analyze the effects of higher (two and four years) and lower (one and six months) sample size n on the ranking of our estimators.

Tables 1.2 and 1.3 report the MAPE and RSD values resulting from modifying the sample size in the Basel framework. For better comparison, the Basel framework itself is also included. We can observe that the historic methods tend to produce the most reliable estimation results. In the overall setting (f) and with the exception of $n = 21$, H and H1 rank highest in terms of MAPE. In the same distributional setting, the RSD values of H1 are lower than those of most other non-parametric methods and the POT approach. Because of their construction, ND, J1 and J2 tend to have smaller RSD values than H1. However, especially in large samples, this comes at the cost of distinctively higher MAPE.

Table 1.2.: MAPE of ES estimates for $\gamma = 97.5\,\%$ and varied n

	ND	POT	H	H1	H2	H3	J1	J2	K1	K2	MV
$n = 21$											
(a)	***28.49***	38.18	33.78	38.75	33.83	35.72	**33.77**	35.86	**33.71**	288.55	60.06
(b)	***32.82***	46.37	41.72	**37.58**	41.75	43.04	41.88	43.20	**41.19**	242.66	61.22
(c)	31.59	24.47	***21.76***	51.54	21.81	23.30	**21.78**	23.42	**21.80**	307.29	54.88
(d)	***20.98***	41.25	37.01	**36.29**	37.04	38.28	37.10	38.44	**36.89**	281.84	60.51
(e)	***14.51***	28.06	**24.59**	45.84	24.65	26.59	24.75	26.97	**24.51**	336.83	57.73
(f)	***25.68***	35.67	**31.77**	42.00	31.82	33.39	31.86	33.58	**31.62**	291.86	58.88
$n = 126$											
(a)	24.43	27.87	**15.11**	***14.15***	15.14	15.25	15.14	18.26	**14.99**	49.74	21.01
(b)	26.97	70.50	21.12	***20.14***	21.13	21.20	**21.06**	24.00	**21.00**	54.98	30.21
(c)	32.56	17.09	**9.40**	***8.38***	**9.42**	9.50	**9.42**	11.43	9.54	33.03	14.98
(d)	***10.96***	65.55	18.59	**17.52**	18.61	18.68	**18.55**	21.13	19.08	41.46	25.01
(e)	***5.90***	15.63	10.24	**9.27**	10.26	10.38	10.19	12.67	**9.98**	35.65	13.02
(f)	20.16	39.33	**14.89**	***13.89***	14.91	15.00	**14.87**	17.50	14.92	42.97	20.85
$n = 252$											
(a)	24.28	11.40	**10.71**	***10.22***	10.72	10.82	12.96	15.58	**10.67**	21.23	13.86
(b)	26.23	19.69	**15.55**	***14.93***	15.56	15.61	17.90	21.28	**15.20**	24.90	18.68
(c)	32.77	6.62	**6.60**	***6.30***	**6.60**	6.68	8.03	9.67	6.93	21.21	11.14
(d)	***8.71***	17.49	**13.54**	**12.86**	**13.54**	13.60	15.73	18.74	14.06	21.49	14.98
(e)	***4.17***	**6.82**	7.04	**6.75**	7.04	7.15	8.61	10.44	6.99	19.98	8.50
(f)	19.23	12.40	**10.69**	***10.21***	**10.69**	10.77	12.65	15.14	10.77	21.76	13.43
$n = 504$											
(a)	24.18	7.61	***7.58***	7.65	***7.58***	7.65	10.29	12.51	**7.59**	12.05	10.47
(b)	25.82	11.87	11.23	***11.04***	**11.22**	**11.22**	14.81	17.81	**11.05**	14.43	14.05
(c)	32.91	***4.61***	**4.64**	5.15	**4.64**	4.69	6.33	7.70	5.50	14.83	9.10
(d)	***7.09***	10.29	9.87	**9.70**	**9.86**	9.87	12.99	15.63	10.90	14.23	11.04
(e)	***2.93***	**4.75**	**4.88**	5.36	4.89	4.96	6.66	8.12	4.91	13.43	6.09
(f)	18.59	7.83	***7.64***	7.78	***7.64***	**7.68**	10.22	12.35	7.99	13.79	10.15
$n = 1008$											
(a)	24.17	**5.33**	5.37	***5.30***	5.37	5.38	10.44	11.42	**5.35**	8.36	8.65
(b)	25.65	8.20	**8.08**	***8.01***	**8.08**	8.09	15.59	16.88	9.06	10.39	11.80
(c)	32.94	**3.26**	**3.30**	***3.24***	**3.30**	3.31	6.37	6.98	5.79	10.01	7.85
(d)	***5.93***	7.13	**7.08**	**7.00**	**7.08**	**7.08**	13.63	14.77	10.10	10.47	9.03
(e)	***2.07***	**3.35**	3.47	**3.41**	3.47	3.48	6.57	7.22	3.58	9.30	4.59
(f)	18.15	**5.45**	**5.46**	***5.39***	**5.46**	5.47	10.52	11.45	6.78	9.71	8.38

For a confidence level of $\gamma = 97.5\,\%$ and varied sample sizes n, this table presents the mean absolute percentage error (MAPE) of the expected shortfall (ES) estimates produced by our parametric and non-parametric techniques. Simulation settings and methods are specified and abbreviated as in Table 1.1. The lowest (second and third lowest) MAPE values within each setting are again marked in italics (bold).

Table 1.3.: RSD of ES estimates for $\gamma = 97.5\,\%$ and varied n

	ND	POT	H	H1	H2	H3	J1	J2	K1	K2	MV
$n = 21$											
(a)	***27.22***	47.24	40.80	**35.56**	40.65	37.73	40.83	37.60	40.68	60.47	**28.72**
(b)	**34.81**	62.72	60.04	**47.45**	59.74	53.51	58.67	52.19	52.54	66.18	***34.46***
(c)	***18.21***	26.39	23.74	**20.15**	23.63	21.59	23.68	**21.43**	23.91	79.40	27.97
(d)	***28.91***	52.90	49.12	**38.35**	48.85	43.45	49.97	43.84	46.66	68.59	**31.72**
(e)	***18.29***	31.79	27.34	**24.80**	27.25	**25.63**	27.49	25.72	27.35	63.79	32.63
(f)	***25.49***	44.21	40.21	**33.26**	40.02	36.38	40.13	36.16	38.23	67.69	**31.10**
$n = 126$											
(a)	***11.21***	423.56	17.89	**17.66**	17.87	17.74	17.87	**16.67**	19.15	80.34	46.28
(b)	***16.47***	1338.73	27.10	**26.61**	27.05	26.78	27.03	**22.27**	27.56	84.10	134.46
(c)	***7.57***	783.99	10.73	**10.56**	10.71	10.62	10.77	**9.89**	12.10	63.98	77.54
(d)	***14.01***	1799.47	23.30	**22.85**	23.25	23.00	23.24	**19.00**	25.37	79.50	199.64
(e)	***7.40***	348.77	11.67	**11.54**	11.65	11.59	11.62	**11.38**	12.53	63.78	37.03
(f)	***11.33***	938.90	18.14	**17.84**	18.10	17.94	18.11	**15.84**	19.34	74.34	98.99
$n = 252$											
(a)	***7.97***	45.32	13.11	12.93	13.09	12.96	12.48	**12.12**	13.48	46.11	**11.37**
(b)	***12.02***	97.21	20.02	19.66	20.00	19.73	**16.80**	**15.77**	19.37	50.88	17.96
(c)	***5.42***	8.85	7.96	7.83	7.95	7.85	**7.47**	**7.19**	8.75	54.32	7.74
(d)	***10.47***	310.68	17.23	16.89	17.20	16.95	**14.47**	**13.52**	17.95	43.71	34.84
(e)	***5.22***	8.61	8.45	8.36	8.44	8.38	8.34	**8.24**	8.73	46.07	**8.01**
(f)	***8.22***	94.13	13.35	13.14	13.34	13.17	**11.91**	**11.37**	13.65	48.22	15.98
$n = 504$											
(a)	***5.63***	9.62	9.48	9.35	9.47	9.36	8.84	**8.66**	9.54	25.06	**7.01**
(b)	***9.23***	20.51	14.56	14.29	14.55	14.31	11.85	**11.36**	13.52	24.80	**10.19**
(c)	***3.84***	5.81	5.77	5.68	5.76	5.68	**5.36**	**5.21**	6.68	48.95	6.21
(d)	***8.21***	14.29	12.78	12.53	12.77	12.54	10.24	**9.78**	13.13	28.13	**8.89**
(e)	***3.68***	**5.96**	6.06	6.00	6.06	6.00	6.00	**5.94**	6.14	37.71	5.99
(f)	***6.12***	11.24	9.73	9.57	9.72	9.58	8.46	**8.19**	9.80	32.93	**7.66**
$n = 1008$											
(a)	***3.99***	6.70	6.67	6.66	6.67	6.66	6.24	**6.20**	6.70	19.02	**4.96**
(b)	***6.80***	10.50	10.29	10.26	10.29	10.26	8.22	**8.11**	9.94	17.16	**7.10**
(c)	***2.71***	4.09	4.08	4.07	4.08	4.07	**3.71**	**3.68**	5.65	37.66	4.60
(d)	***5.84***	9.10	8.99	8.96	8.98	8.96	7.10	**6.99**	10.23	23.01	**6.36**
(e)	***2.60***	**4.20**	4.29	4.28	4.29	4.28	4.23	**4.22**	4.44	30.59	4.50
(f)	***4.39***	6.92	6.86	6.84	6.86	6.85	5.90	**5.84**	7.40	25.49	**5.50**

For a confidence level of $\gamma = 97.5\,\%$ and varied sample sizes n, this table presents the relative standard deviation (RSD) of the expected shortfall (ES) estimates produced by our parametric and non-parametric techniques. Simulation settings and methods are specified and abbreviated as in Table 1.1. The lowest (second and third lowest) RSD values within each setting are again marked in italics (bold).

As to be expected from suitably defined estimators, the MAPE declines with increasing sample size.[9] However, there are drastic differences in the magnitudes of the reductions across estimators. The ND method shows the lowest marginal decrease. For example, it loses the dominant role it has for $n = 21$ in the overall setting (f) when switching to $n = 126$ because the improvement of the historic methods overcompensates the improvement of ND. Furthermore, while the historic methods still enhance from $n = 504$ to $n = 1008$, there is almost no betterment for ND because estimates of mean and variance cannot be noticeably improved anymore (see Moraux, 2011).

In a synopsis of all n, the estimator K2 is characterized by the most significant marginal improvement, followed by POT. Nonetheless, Tables 1.2 and 1.3 advise against exclusively using one of these methods for time horizons of $n = 252$ or smaller. In contrast, for $n = 504$ and $n = 1008$, the error values of POT are close to the historic methods. For these n, it even tends to outperform K1, which it did not for $n = 252$. Unfortunately, we cannot make such a statement for K2. Even for $n = 1008$ it ranks last among the non-parametric methods and is inferior to POT.

[9] For the POT method there are exceptions when raising $n = 21$ to $n = 126$ in settings (b) and (d) which are related to the trade-off between good GPD fit and validity of the limit theorem. In addition, J1 shows some abnormalities between $n = 504$ and $n = 1008$.

A closer look at the differences between the non-parametric methods illustrates some additional aspects. First, we can see that, for $n = 21$ and $n = 126$, J1 is close to H, and often estimates worse than H otherwise. Second, J2 always has higher MAPE and lower RSD than J1. Thus, altogether, the classic historic method tends to dominate the outlier-adjusted methods. Finally, while the kernel-weighted estimator K2 appears to fail categorically, K1 is more reliable. K1 actively competes against H (see also Chen, 2008), but is not able to consistently outperform it in terms of MAPE and RSD. For example, K1 has a better MAPE performance than H in all of our parametrizations for $n = 21$, except (c), whereas K1 performs worse than H in all of our parameterizations for $n = 1008$, except for (a).

As far as the magnitude of the absolute deviations of the best estimators from the true ES values is concerned, we observe about 13 % for $n = 252$ (and 7 % for $n = 1008$). Thus, for a portfolio of 100 million US dollars and true ES of 2.5 % or 2.5 million US dollars, we are on average 0.325 million US dollars (0.175 million US dollars) off. Depending on whether these magnitudes are considered harmful from an economic perspective, larger sample sizes may be preferable in practice.[10]

We complete this section with some sample size-related insights from our performance profile technique. Figure 1.2 plots the profile functions of our estimators for different n in the summary setting (f).[11] For small sample sizes of $n = 21$ and $n = 126$, we see that the outlier-adjusted methods J1 and J2 have the second highest and highest performance functions, respectively, followed by K1. In large samples, other methods become serious competitors. For $n = 504$, K2 has a higher probability $\rho_j(1)$ than K1. However, the function ρ_j for K2 is quickly surpassed by the one for K1. A weaker effect occurs for $n = 1008$, where a higher argument x is required for the performance functions to intersect. Furthermore, we can observe that the POT method successively works its way to the top. In the end ($n = 1008$), it does not have the highest probability $\rho_j(1)$ but shows very high probabilities of belonging to the best estimators. Finally, and in line with our previous results, historic estimators again appear less attractive in the performance profiles than in an assessment based on MAPE and RSD values.

1.4.2.2. Alternative confidence levels

Also from a non-bank perspective, this section analyzes how the ranking of our ES estimators is affected by the choice of confidence level. To this end, we start with an investigation of the MAPE and RSD values for $\gamma = 95\,\%$ and $\gamma = 99\,\%$ in Table 1.4, where the Basel value of $\gamma = 97.5\,\%$ is again supplemented for comparison. To keep our multidimensional set of results tractable, we focus on a sample size of $n = 252$.[12]

In general, our previous assessment of the ES estimators also holds for alternative confidence levels. That is, while the historic methods and K1 tend to have the lowest MAPE values, the RSD of POT and K2 indicate some large estimation errors. J2 has the lowest RSD of all non-parametric methods. As to expect, for all estimators, the magnitude of estimation error tends to increase with the confidence level. MAPE and RSD values of significant size reject K2 for $\gamma = 99\,\%$, while, in the case of $\gamma = 95\,\%$, the error values are closer to the other estimators.

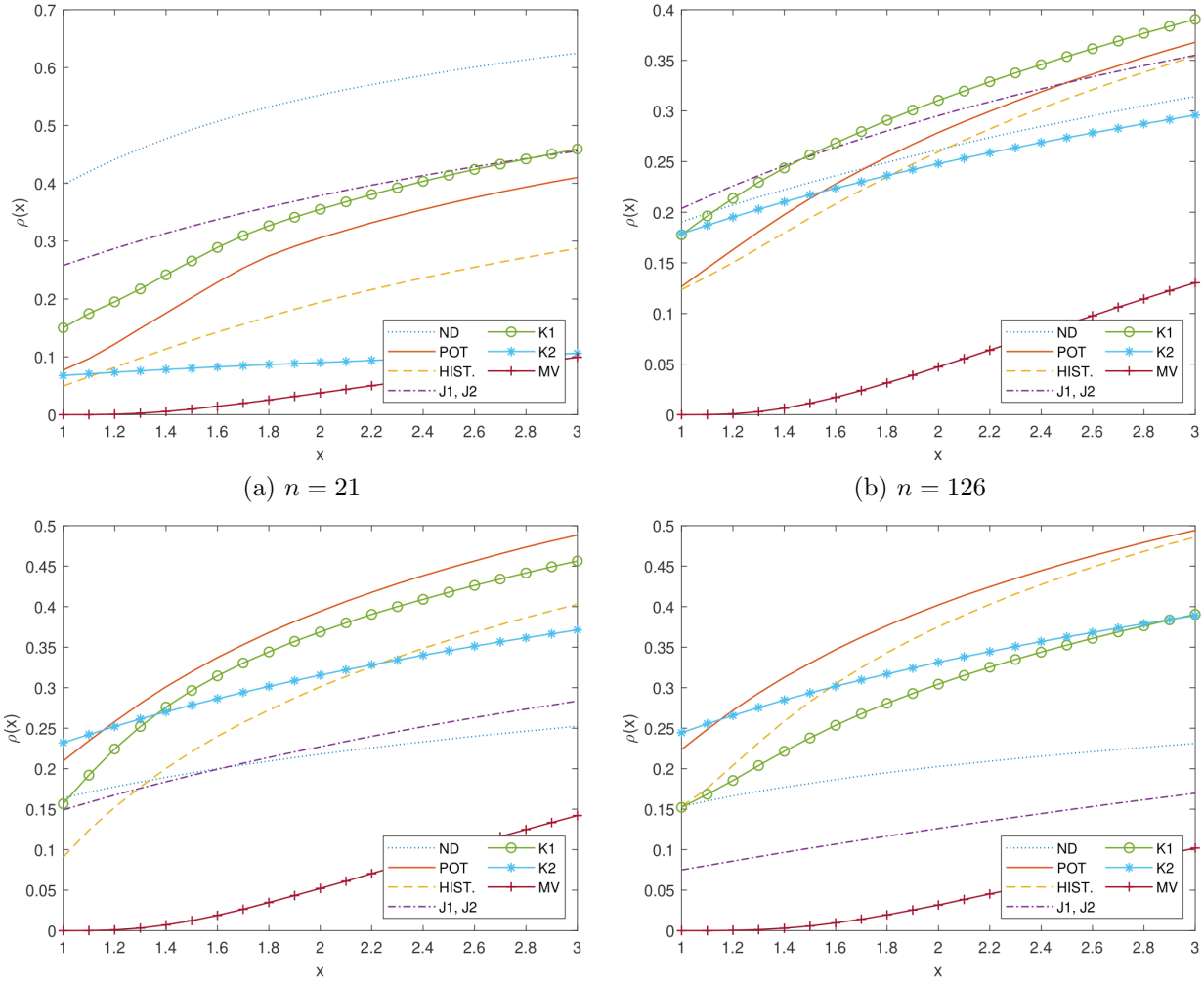
With respect to changing estimator ranks, we find the following. First, H1 systematically outperforms H only if $\gamma = 97.5\,\%$. For $\gamma = 95\,\%$, it performs worse than H in terms of MAPE in settings (c), (e) and (f), and, in the case of $\gamma = 99\,\%$, additionally in setting (a). However, at the same time, the RSD values of H1 are consistently below H. Second, in almost all situations, J2 is dominated by J1 and the historic methods. Also, the relative differences between the MAPE

[10] Determining an accurate level of ES is crucial for several reasons. Probably the most important one is outlined in Furlong and Keeley (1989) and Hugonnier and Morellec (2017). They illustrate that, for a value-maximizing bank or corporation, incentives to decrease asset risk decline as its capital reserves increase.

[11] The profiles for the other distributional settings can be found in Figures A.2 to A.5 of the appendix.

[12] The results for the other sample sizes can be found in Tables A.4 and A.5 of the appendix.

Figure 1.2.: Performance profiles for overall setting (f), $\gamma = 97.5\,\%$ and varied n



(a) $n = 21$ (b) $n = 126$

(c) $n = 504$ (d) $n = 1008$

For a confidence level of $\gamma = 97.5\,\%$ and varied sample size n, this figure plots the performance profiles (defined in Section 1.3.4.2) of our expected shortfall estimators. Each subfigure concentrates on the overall distributional setting (f) (specified in Section 1.3.3). The estimators are abbreviated as in Table 1.1 and grouped as in Figure 1.1.

values of the latter two tend to shrink with increasing γ. Finally, K1 behaves in the opposite fashion. Its mean distance to the MAPE of, for example, H falls with decreasing γ.

Figure 1.3 presents the performance plots for $\gamma = 95\,\%$ and $\gamma = 99\,\%$ in setting (f).[13] It shows that the POT approach plays a very important role along all observed x, directly followed by the kernel method K1 which finally dominates the performance of POT in case of $\gamma = 99\,\%$. The $\rho_j(1)$ rank of the outlier-adjusted methods increases with γ, whereas the rank of K2 decreases. The higher the confidence level, the larger the general probability distance between the outlier-adjusted methods and K2 becomes. For high γ, the adjusted estimators have a higher probability of being the best, but are characterized by a lower efficiency than POT and K1.

[13] The performance plots for the other settings can be found in Figures A.6 and A.7 of the appendix.

Table 1.4.: MAPE and RSD of ES estimates for $n = 252$ and varied γ

	ND	POT	H	H1	H2	H3	J1	J2	K1	K2	MV
MAPE, $\gamma = 95\,\%$											
(a)	20.09	9.22	**8.97**	***8.95***	**8.97**	9.06	10.24	12.76	8.99	12.54	10.98
(b)	18.64	15.72	12.31	**12.11**	**12.30**	12.33	13.66	17.03	***11.91***	14.87	14.09
(c)	30.70	11.28	***5.30***	5.64	***5.30***	**5.37**	6.05	7.55	5.48	14.49	9.72
(d)	***7.58***	13.03	10.58	**10.40**	**10.57**	10.60	11.70	14.59	10.67	14.56	11.43
(e)	***4.32***	**6.02**	**6.05**	6.31	6.06	6.15	6.93	8.68	6.08	12.91	6.95
(f)	16.27	11.05	**8.64**	8.68	**8.64**	8.70	9.72	12.12	***8.63***	13.87	10.63
MAPE, $\gamma = 97.5\,\%$											
(a)	24.28	11.40	**10.71**	***10.22***	10.72	10.82	12.96	15.58	**10.67**	21.23	13.86
(b)	26.23	19.69	**15.55**	***14.93***	15.56	15.61	17.90	21.28	**15.20**	24.90	18.68
(c)	32.77	6.62	**6.60**	***6.30***	**6.60**	6.68	8.03	9.67	6.93	21.21	11.14
(d)	***8.71***	17.49	**13.54**	**12.86**	**13.54**	13.60	15.73	18.74	14.06	21.49	14.98
(e)	***4.17***	**6.82**	7.04	**6.75**	7.04	7.15	8.61	10.44	6.99	19.98	8.50
(f)	19.23	12.40	**10.69**	***10.21***	**10.69**	10.77	12.65	15.14	10.77	21.76	13.43
MAPE, $\gamma = 99\,\%$											
(a)	29.19	15.94	***14.01***	14.70	**14.02**	14.36	14.06	17.38	**14.05**	176.30	32.40
(b)	35.99	30.31	**21.29**	***19.58***	**21.29**	21.40	**21.21**	24.46	21.39	207.67	42.46
(c)	33.41	17.52	***9.08***	13.72	***9.08***	9.34	**9.10**	11.36	10.00	46.13	16.87
(d)	***16.14***	25.80	19.21	**17.69**	19.21	19.34	**19.17**	22.04	20.88	166.85	34.63
(e)	***4.03***	9.09	8.82	13.69	8.82	9.19	**8.78**	11.40	**8.71**	70.83	15.34
(f)	23.75	19.73	**14.48**	15.88	**14.48**	14.73	***14.46***	17.33	15.01	133.55	28.34
RSD, $\gamma = 95\,\%$											
(a)	***8.25***	26.11	11.23	11.11	11.22	11.12	10.95	**10.78**	11.31	21.16	**8.55**
(b)	***12.23***	153.64	15.86	15.61	15.84	15.62	**13.97**	**13.24**	15.04	21.85	19.67
(c)	***5.39***	1164.05	6.58	6.49	6.57	6.50	**6.34**	**6.17**	6.83	36.50	107.61
(d)	***10.62***	192.95	13.53	13.30	13.52	13.31	**11.84**	**11.18**	13.46	25.31	21.88
(e)	***5.41***	7.58	7.51	**7.45**	7.51	7.46	7.50	7.49	7.59	28.73	**6.48**
(f)	***8.38***	308.87	10.94	10.79	10.93	10.80	**10.12**	**9.77**	10.85	26.71	32.84
RSD, $\gamma = 97.5\,\%$											
(a)	***7.97***	45.32	13.11	12.93	13.09	12.96	12.48	**12.12**	13.48	46.11	**11.37**
(b)	***12.02***	97.21	20.02	19.66	20.00	19.73	**16.80**	**15.77**	19.37	50.88	17.96
(c)	***5.42***	8.85	7.96	7.83	7.95	7.85	**7.47**	**7.19**	8.75	54.32	7.74
(d)	***10.47***	310.68	17.23	16.89	17.20	16.95	**14.47**	**13.52**	17.95	43.71	34.84
(e)	***5.22***	8.61	8.45	8.36	8.44	8.38	8.34	**8.24**	8.73	46.07	**8.01**
(f)	***8.22***	94.13	13.35	13.14	13.34	13.17	**11.91**	**11.37**	13.65	48.22	15.98
RSD, $\gamma = 99\,\%$											
(a)	***7.75***	71.35	17.01	**15.96**	16.99	16.21	17.00	**14.76**	17.79	128.38	32.59
(b)	***11.79***	339.34	28.04	**25.89**	28.01	26.42	27.77	**21.33**	27.38	114.17	45.88
(c)	***5.43***	2777.16	10.71	**9.97**	10.70	10.14	10.76	**9.18**	13.08	111.57	240.40
(d)	***10.74***	133.45	25.26	**23.18**	25.23	23.68	24.95	**18.75**	27.30	142.82	37.49
(e)	***5.05***	11.68	10.32	**9.81**	10.32	9.93	10.29	**9.55**	10.74	132.62	22.36
(f)	***8.15***	666.60	18.27	**16.96**	18.25	17.28	18.15	**14.71**	19.26	125.91	75.74

For a sample size of $n = 252$ and varied confidence levels γ, this table presents the mean absolute percentage error (MAPE) and relative standard deviation (RSD) of the expected shortfall (ES) estimates produced by our parametric and non-parametric techniques. Simulation settings and methods are specified and abbreviated as in Table 1.1. The lowest (second and third lowest) MAPE and RSD values within each setting are again marked in italics (bold).

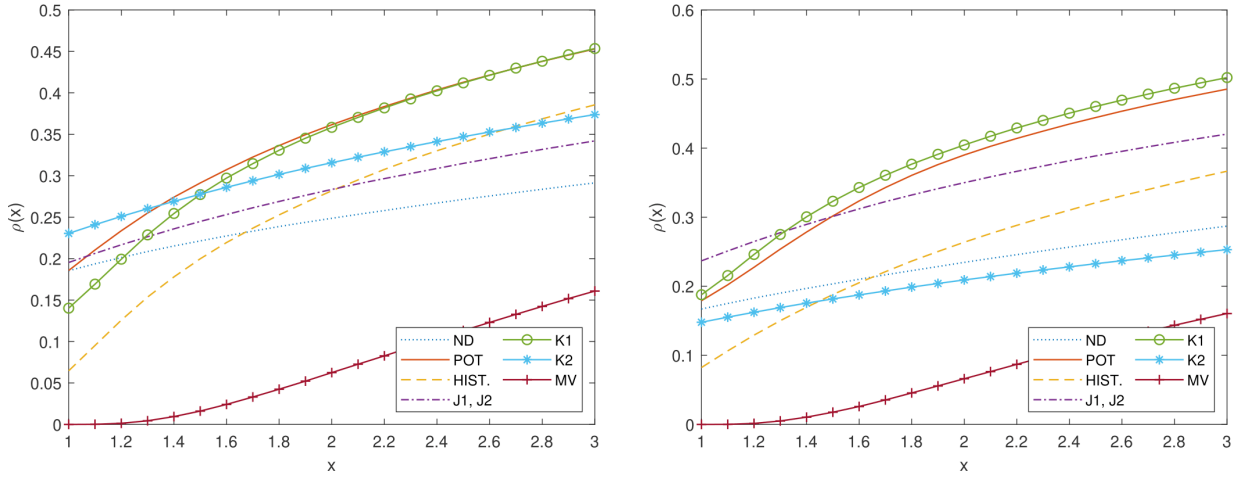
1.4.3. Combined estimates

1.4.3.1. Preliminaries

Before turning to a detailed analysis of combined ES estimators, we take a brief look at the MPE values of the individual estimators, which we document in the appendix (see Tables A.1 to A.3). This is useful because it highlights over-/underestimation tendencies, which might cancel each other out in adequate combinations and thus lead to improved estimates.[14]

[14] Similar to traditional portfolio theory, where asset correlations determine whether portfolio building is beneficial (see Elton et al., 2007, chpt. 5), ES estimators characterized by suitable balancing countermovement should lead to the best combinations.

Figure 1.3.: Performance profiles for overall setting (f), $n = 252$ and varied γ



(a) $\gamma = 95\,\%$ (b) $\gamma = 99\,\%$

For a sample size of $n = 252$ and varied confidence levels γ, this figure plots the performance profiles (defined in Section 1.3.4.2) of our expected shortfall estimators. Each subfigure concentrates on the overall distributional setting (f) (specified in Section 1.3.3). The estimators are abbreviated as in Table 1.1 and grouped as in Figure 1.1.

Regardless of distributional setting, n or γ, almost all techniques tend to underestimate the true ES on average. Overestimation occurs for H1 only in a few situations. In contrast, ND systematically overestimates in setting (c). In addition, the MPE of K2 strongly points to positive when n is small and γ is large. As far as the underestimating techniques are concerned, most of them show rather small MPE values in large samples. While the POT method shines in this respect, the large errors of the outlier-adjusted methods are striking.[15]

1.4.3.2. Alternative combinations

Because Sections 1.4.1 and 1.4.2 show that MV, which pools all available estimators, does not qualify as highly beneficial, we form additional equally weighted combinations based on our previous findings. The combination C1 includes H, H1 and K1 because these methods belong to the top MAPE approaches across our entire set of results. C2 covers all methods except for the high RSD estimators POT and K2. In C3, we combine H1, J1 and K1 to diversify across non-parametric approaches by selecting the most promising (in terms of overall MAPE findings) representative of each sub-class (historic, outlier-adjusted, kernel-based). For C4, we pick POT, J1, J2, K1 and K2 motivated by their outstanding performance profile results. C5 is a combination of H1 and K2, which tend to overestimate in a few and most constellations, respectively. Based on the observation that ND and K2 are characterized by the largest MAPE values in many situations, we combine them to C6. Finally, C7 resembles a summary of C6 and H1, i. e., a combination of the worst and the best methods.

In the following, we concentrate on a sample size of $n = 252$, the confidence levels $\gamma \in \{95\,\%, 97.5\,\%, 99\,\%\}$ and an analysis of MAPE and RSD because this is sufficient to answer the question of whether combinations can be advantageous in the estimation of ES.[16] Table 1.5 presents the results for our alternative combinations and, for a better assessment of their magnitude, also contains our previous findings for H1.

[15] Excluding observations which are valid realizations of the data-generating process is known to lead to bias (see Studenmund, 2017, chpt. 3).

[16] Nonetheless, the results for other sample sizes and the corresponding sets of performance plots are available from the authors upon request.

Table 1.5.: MAPE and RSD of combined ES estimators for $n = 252$ and varied γ

	H1	C1	C2	C3	C4	C5	C6	C7
MAPE, $\gamma = 95\,\%$								
(a)	8.95	8.89	8.12	**6.87**	***6.85***	**7.79**	12.82	9.06
(b)	12.11	11.88	10.58	***9.30***	**9.57**	**9.72**	13.34	10.51
(c)	5.64	**5.25**	**4.09**	***3.99***	6.37	7.92	17.80	12.78
(d)	10.40	10.40	**8.08**	**8.10**	8.47	9.11	8.22	***7.46***
(e)	6.31	6.00	**4.69**	***4.56***	**5.73**	8.79	7.82	6.86
(f)	8.68	8.48	**7.11**	***6.56***	**7.40**	8.66	12.00	9.33
MAPE, $\gamma = 97.5\,\%$								
(a)	10.22	10.45	**10.15**	***8.45***	**9.40**	12.28	17.74	12.68
(b)	14.93	14.91	**14.02**	***12.20***	**13.16**	15.39	20.45	15.66
(c)	**6.30**	6.41	***4.42***	**5.11**	6.93	11.31	20.89	14.46
(d)	12.86	13.29	**11.03**	**10.96**	11.74	13.14	12.43	***10.39***
(e)	**6.75**	6.80	**5.50**	***5.39***	7.31	12.19	10.97	8.89
(f)	10.21	10.37	**9.02**	***8.42***	**9.71**	12.86	16.49	12.42
MAPE, $\gamma = 99\,\%$								
(a)	14.70	**13.09**	**11.79**	***9.94***	38.17	90.44	90.03	60.22
(b)	19.58	**19.55**	**17.64**	***15.02***	46.77	106.59	106.03	71.75
(c)	13.72	**8.55**	***5.70***	**6.60**	13.13	26.10	33.85	26.81
(d)	**17.69**	18.15	**14.80**	***14.05***	39.61	86.19	84.90	58.14
(e)	13.69	**8.33**	***6.01***	**6.34**	16.44	39.24	35.92	26.65
(f)	15.88	**13.53**	**11.19**	***10.39***	30.82	69.71	70.15	48.71
RSD, $\gamma = 95\,\%$								
(a)	11.11	11.20	**8.51**	***8.40***	**9.06**	11.86	12.26	9.70
(b)	15.61	15.13	**11.67**	***11.16***	34.10	13.30	13.04	**11.66**
(c)	**6.49**	6.61	***4.65***	**4.94**	224.54	18.72	16.52	11.82
(d)	13.30	13.24	***9.69***	**9.78**	41.49	14.18	13.49	**10.91**
(e)	**7.45**	7.51	***5.56***	**5.64**	7.89	15.65	15.29	11.26
(f)	10.79	**10.74**	**8.02**	***7.98***	63.41	14.74	14.12	11.07
RSD, $\gamma = 97.5\,\%$								
(a)	**12.93**	13.14	***9.75***	**9.85**	15.28	24.39	26.85	18.35
(b)	19.66	**19.03**	**14.39**	***14.03***	25.48	27.57	29.78	21.30
(c)	**7.83**	8.07	***5.47***	**6.05**	12.82	28.39	24.67	17.63
(d)	16.89	16.97	***12.15***	**12.62**	67.46	23.49	22.93	**16.87**
(e)	**8.36**	8.49	***6.12***	**6.38**	11.50	24.61	24.27	17.07
(f)	**13.14**	**13.14**	***9.58***	**9.79**	26.51	25.69	25.70	18.24
RSD, $\gamma = 99\,\%$								
(a)	**15.96**	16.79	***12.19***	**12.65**	54.11	90.35	100.69	75.77
(b)	25.89	**25.48**	***19.09***	***19.09***	73.60	83.53	93.06	71.81
(c)	**9.97**	10.78	***7.06***	**8.23**	493.16	60.74	55.98	39.52
(d)	**23.18**	24.13	***17.03***	**18.17**	61.94	99.53	105.85	80.17
(e)	**9.81**	10.21	***7.19***	**7.70**	39.98	78.27	81.26	57.27
(f)	**16.96**	17.48	***12.51***	**13.17**	144.56	82.49	87.37	64.91

For a sample size of $n = 252$ and varied confidence levels γ, this table presents the mean absolute percentage error (MAPE) and relative standard deviation (RSD) of several (mean-)combined expected shortfall (ES) estimators. Simulation settings and abbreviations of the individual methods are used as in Table 1.1. The combinations are C1 (H, H1, K1), C2 (all estimators except POT and K2), C3 (H1, J1, K1), C4 (POT, J1, J2, K1, K2), C5 (H1, K2), C6 (ND, K2) and C7 (ND, K2, H1). The lowest (second and third lowest) MAPE and RSD values within each setting are again marked in italics (bold).

The combination of the three best estimators (C1) produces better MAPE values than H1 when $\gamma = 95\,\%$. Unfortunately, this does not hold for most distributional settings when $\gamma = 97.5\,\%$ and one setting when $\gamma = 99\,\%$. However, it should be noted that, for high γ, C1 tends to belong to the top three combinations. Excluding POT and K2 from MV (C2) results in generally better MAPE than H1 and all other non-parametric approaches. This combination also produces the lowest RSD values for almost all covered settings and confidence intervals. The MAPE results of a merger of the best-in-class non-parametric representatives (C3) are the most impressive. C3 usually ranks first among all combinations. The only exceptions are three light-tailed constellations, where C2 leads, and two heavy-tailed cases, where C7 is superior.

C4, C5, C6 and C7 fail in particular for a high confidence level $\gamma = 99\,\%$. C4, the combination of methods with best performance profiles, has its merits in positively skewed data when low confidence levels are used. To a somewhat weaker extent, this is also true for C5, which combines

methods with mixed tendencies to overestimate. While C6 never ranks among the top three combinations, C7 is the dominant combination for negatively skewed and fat-tailed data when γ is not too high.

In summary, we can always find a combination which has a lower MAPE than H1. Put differently, in a comparison of the top estimators of Tables 1.4 and 1.5, we can often identify constellations where combinations outperform each individual estimator. On the one hand, there is no combination that consequently ranks first across all distributional settings and confidence levels. The performance and ranking of a given combination varies just like its components. On the other hand, a best-in-class combination often scores very high and can therefore be considered promising.

1.4.4. Robustness

To check whether our findings are driven by some specifics of our simulation design or the parametrization of our estimators, we performed two sorts of robustness checks.

First, as far as the stability of our simulation results is concerned, we ensured that $m = 10^5$ is sufficient to provide the same estimator rankings when the entire simulation study is repeated.[17] In other words, this m ensures that we almost capture the entire population such that the derived MAPE, RSD and profile probability values are not estimates but population properties. For smaller m, such as $m = 10^4$, this is not guaranteed.

Second, we evaluated the influence of changing some of the parameters in the ES estimation methods. For the POT parameter q, we considered the alternative values 0.05, 0.08, 0.12 and 0.15 because these have also been used in previous research (see Herrera, 2013; Chavez-Demoulin and McGill, 2012).[18] Of course, this changes MAPE and RSD values, but does not influence our overall assessment of the POT method. A similar effect can be found in the estimation of J1 and J2 when changing the parameter a to 0.05 or 0.10. As indicated by the results, our implemented bandwidth selection rule for K1 appears to choose suitable bandwidths h. However, there might be problems in the bandwidth selections for K2. Therefore, we experimented with alternatives available in Cheng and Sun (2006) and Wand and Jones (1995). Unfortunately, this either did not change the results significantly or led to crucial ES outliers. There may be specifications yielding better results. However, the focus of our study is not on optimal bandwidth choice but on the quality of specifications often used in practice.

1.5. Conclusion

Motivated by recent proposals to replace the VaR with the ES in the calculation of capital requirements and other portfolio management applications, we present a rich set of tables and figures which systematically compare the performance of popular non-parametric estimators of ES to each other, well-known parametric benchmarks and combinations of different ES estimators. This material allows researchers and risk managers to quickly look up the most suitable estimator for a specific distributional environment and decision problem.

While no estimator outperformed all others in all specified situations, we can observe some general tendencies. Based on the classic evaluation measures MAPE and RSD, a slightly modified version of the traditional historic estimator often stands out. Outlier-adjusted variants have some merits but should be handled with care because their design artificially ignores risk-relevant large losses. Our extreme value theory method is characterized by some rare but drastic misestimations with crucial impact on its ranking in comparison to other estimators. By chance, we found that,

[17] This is in line with the simulation sizes typically used in the evaluation of risk and performance measures (see, for example, Schuhmacher and Eling, 2011; Schuhmacher and Auer, 2014).

[18] For a review of the literature dealing with the choice of q, see Scarrott and MacDonald (2012). For applications, in which the choice of q appears to be irrelevant, see Auer (2015).

in a setting specified with realistic levels of skewness and kurtosis, even a simple normally distributed estimation approach can be highly valuable. Furthermore, thoughtful combinations of ES estimators have significant potential to reduce estimation error.

In an application of a performance profiling method, where the focus lies on the probability that a given estimator is the best (or belongs to the best) estimator(s) of a selection of estimators, we gained some additional insights. Because this method is less sensitive to extraordinary estimation errors than the traditional evaluation measures, it often favors kernel-based methods and the extreme value method. These techniques often rank first and, if they do not, they still belong to the highest-ranked estimators.

Given these two perspectives, readers of our guide have to decide whether they want an otherwise good estimator to be punished for few failures. If the answer is yes, they should rely on our MAPE and RSD values when selecting the optimal estimator. If the answer is no, they should attend to our performance profiles.

In future research, our work could be extended in several ways. First, it might be interesting to analyze the error introduced by time series filters often used in practice (see McNeil and Frey, 2000; Marimoutou et al., 2009; Auer, 2015). In our iid setting, we know that there is no serial correlation in returns and variances such that applying a time series filter in ES estimation will definitely introduce finite-sample distortions. Their magnitude could be derived by comparing the errors with and without filter. In similar extensions, we could generally quantify the role of error originating from the assumption of potentially inadequate time series models. Second, one might consider additional classes of parametric estimators. While standard theoretical distributions are rather uninteresting in this respect (because we already know how close they will be to our simulated distribution), flexible data-oriented distribution systems are more interesting. Mixtures (of normal, student t or stable distributions), g-and-h settings or the Johnson framework are typical examples of such systems (see Degen et al., 2007; Nadarajah et al., 2014; Novales and Garcia-Jorcano, 2019) and a potential starting point for new research endeavors. Finally, our results may be revisited with a focus on marginal ES which measures the impact of a single company on the tail risk of the market (see Caporin and de Magistris, 2012; Daníelsson et al., 2016). This can provide valuable insights into the role of estimation error in the measurement of systemic risk.

Part II

Motivation

In the last chapter, we discussed several promising ES estimators and analyzed their performances in a well-defined simulated framework that enables us to identify their principal strengths and weaknesses segregated from any distorting real-world influences. In the next chapter, we extend this knowledge by conducting an analysis of ES estimators in commodity futures markets (in other words, move on to real data).

As we can no longer rely on independent and identically distributed (iid) data, we additionally allow for some underlying time-varying processes of mean and volatility. For that, we fit the commodity futures indices to an autoregressive (AR) mean and a generalized autoregressive conditional heteroscedasticity (GARCH) volatility process first (see Box and Jenkins, 1970; Bollerslev, 1986) and, following, concentrate on the resulting mean- and volatility-adjusted data. Naturally, our assumptions of the underlying mean- and volatility processes have an additional effect on the quality of ES estimations.[19] In contrast to the simulation approach of Chapter 1, we are unaware of the correct ES values when considering commodity futures indices, such that assessing estimation quality becomes more complex, too. Thus, for our evaluation of ES estimations that follow from an AR-GARCH specification and the subsequent application of ES estimators, we rely on recently developed backtest procedures of Du and Escanciano (2017) to evaluate the reliability of our estimated results.

Albeit Chapter 1 suggests several appropriate ES estimators for our filtered data[20] (such as peak over threshold, a kernel density approach or some historic variants), we instead focus on estimators based on invertible distribution functions in order to secure reliable backtest results. In other words, we mainly consider parametric estimators (as the peak over threshold method and several others) and just one semi-/ non-parametric approach (a kernel density method) that bases on a smoothed data-dependent distribution.

Altogether, we estimate daily-changing risk levels for more than a quarter-century of commodity futures indices and can identify whether/ in which way past crises and bear markets affected the commodity sectors. Based on the backtest procedures, we provide a ranking of (AR-GARCH adjusted) ES estimators for several classes of commodity futures indices and detect when the estimations remained critical in the past.

[19] It is worth mentioning to note that, for this reason, the term ES estimator does not any longer refer to a method that models the distribution of pure iid losses, but to the package of AR-GARCH filter and distribution specification both, when it is mentioned below in Chapter 2.

[20] In the following analysis, we will apply ES estimators at a 95 % confidence level, which is evaluated in Table A.4. The characteristics of our available commodity futures indices indicate that we have to consider specifications (a) and (b) (according to their definition in Section 1.3.3). Due to the 2500 trading days spanning (time-varying) estimation periods our analysis bases upon, $n = 1008$ remains the most adequate of the evaluated sample sizes.

2. Time-varying dynamics of expected shortfall in commodity futures markets

Abstract: Motivated by the growing interest of investors in commodities and by advances in risk measurement, we present a full-scale analysis of expected shortfall (ES) in commodity futures markets. Besides illustrating the dynamics of historic ES, we evaluate whether popular estimators are suitable for forecasting future ES. By implementing a new backtest, we find that the performance of estimators hinges on market stability. Estimators tend to fail when markets are in turmoil and accurate forecasts are urgently needed. Even though a kernel method performs best on average, our results advise against the use of established estimators for risk (and margin) prediction.

2.1. Introduction

Over the last decades, commodity futures markets have grown significantly because, in addition to hedgers using futures for risk management, investors have discovered the potential of futures in investment products (see Rouwenhorst and Tang, 2012; Cheng and Xiong, 2014). In addition to documenting general properties of commodity futures (see Jagannathan, 1985; Gorton and Rouwenhorst, 2006), the success of long-short trading strategies (see Miffre, 2016) and diversification benefits in multi-asset portfolios (see Daskalaki et al., 2017), researchers have paid significant attention to the risk of commodity futures positions (see Hirshleifer, 2015; Carter et al., 2017).[1]

Following the banking and industry standard, the investment risk of commodity futures has typically been quantified by the value at risk (VaR) and the main objective of most studies has been to identify the most suitable VaR estimation technique (see Aloui and Mabrouk, 2010; Füss et al., 2010; Laporta et al., 2018). This, however, is problematic. While the VaR is a quite intuitive measure, one that captures the loss of a financial instrument that is not exceeded with a certain probability, it is not sub-additive (see Artzner et al., 1997, 1999; Yamai and Yoshiba, 2005). When using the VaR, the risk of a diversified portfolio can therefore be higher than the sum of its components' stand-alone risks.[2] Furthermore, the VaR focuses on the frequency but ignores the magnitude of tail events. The expected shortfall (ES), an alternative risk measure, defined as the loss to be expected when the VaR is exceeded, has no such shortcomings. After many years of neglect, regulators have just recently focused on these problems and now suggest steadily replacing the VaR with the ES in risk management applications (see Basel Committee of Banking Supervision, 2012; Kinateder, 2016). However, apart from a few exceptions, researchers and practitioners in the commodity sector still appear reluctant to follow this lead. Consequently only little is known about ES in commodity futures markets and there is a significant research gap which we intend to fill. Reservations arise mainly because implementing the new measure requires a suitable ES estimator but the empirical evaluation of available alternatives is not as straightforward as it is for VaR estimators. For the VaR, the backtests of Kupiec (1995), Christoffersen (1998)

[1] For other investment vehicles, such as commodity ETFs, see Del Brio et al. (2017).

[2] Daníelsson et al. (2013) discuss the sub-additivity issue in detail and show how the choice of VaR estimator can mitigate this problem which is particularly serious when estimating via historical simulation.

and Berkowitz et al. (2011) have become standard (see Kuester et al., 2006; Basel Committee of Banking Supervision, 2011).[3] The problem with finding a suitable backtesting procedure for ES is that it does not fulfill the property of identifiability (see Nolde and Ziegel, 2017).[4] Nevertheless, various rudimentary backtests have been proposed over the years and occasionally applied in the commodity sector (see, for example, Youssef et al., 2015; Del Brio et al., 2017). In a recent article, Du and Escanciano (2017) make an important contribution to ending the backtesting debate for ES by developing unconditional and conditional coverage tests (with suitable size and power properties), which are easy to implement and to comprehend because they use ideas similar to the established VaR backtests. In many fields, these new tests are receiving significant attention because, after years of theoretical research, they finally allow a judgment of available ES estimators (see Novales and Garcia-Jorcano, 2019; Hoga, 2019; Le, 2020). In the commodity context, there is now no more reason to put research on hold.

With the overdue regime shift in risk measurement and the eventual availability of methods to empirically evaluate ES estimators, several important questions arise for commodity investors. Some of these questions are backward-looking: What is a typical level of ES in the overall commodity futures market and its subsectors? What kind of events have historically had the most significant impact on the ES of commodity futures? Other questions are forward-looking: What is the best estimator of ES when it comes to predicting future ES? Does the relative performance of ES forecasts produced by different estimators vary over time? In other words, should we switch estimators based on the market phase?

To answer these questions, we analyze a quarter-century of daily futures market data. As in Bianchi et al. (2016) and Georgopoulou and Wang (2017), we focus on the futures-based Standard and Poor's Goldman Sachs Commodity Index (S&P GSCI) and its five sub-indices − energy, precious metals, industry metals, agriculture and livestock − to capture the most important commodity contracts. In other words, we consider investors who participate in the market either by rolling the underlying futures themselves or by taking positions in investment products mimicking these well-known commodity benchmarks. By modeling the ES of the index returns, we can capture the risk to which these investors are exposed. Furthermore, because ES forecasts have become a key input variable in the margin calculations of some exchanges, we can simultaneously shed light on whether different ES estimators tend to cause inadequate margin settings by systematic under- or overprediction. Keeping daily margins at proper levels is important because too-low margins are insufficient collateral against default; too-high margins increase traders' transaction costs or force them out of the market (see Brooks et al., 2005; Ho et al., 2008).

To obtain results on past and future ES, we use six estimators which have become quite popular in the stock market literature (and are starting to attract attention in applications with commodity data) but whose performance has not jet been fully verified by adequate backtesting. Five of these estimators are parametric; one is (semi) non-parametric. The commonality of all estimators is that they capture the time-varying dynamics of losses via a generalized autoregressive conditional heteroscedasticity setting which is the workhorse of many commodity market studies (see Marimoutou et al., 2009; Watugala, 2019). They differ in the form they assume for the conditional distribution of losses. In the parametric cases, we consider Hansen's skewed extension of the Student t distribution (see Jondeau and Rockinger, 2003), an extreme value setting based on the generalized Pareto distribution (see Gençay and Selçuk, 2004), the g-and-h distribution (see Degen et al., 2007), the Johnson system of distributions (see Brooks et al., 2005) and a two-component Gaussian mixture (see Kuester et al., 2006).[5] The shapes of these distributions (or distribution systems) are quite flexible, making them promising candidates for matching empirical

[3] For recently proposed extensions, see Ziggel et al. (2014), Wied et al. (2016) and Kratz et al. (2018).

[4] Identifiability allows for sensible forecast evaluation whereas elicitability is required for forecast comparisons (see Gneiting, 2011; Fissler and Ziegel, 2016; Nolde and Ziegel, 2017).

[5] The robustness checks of Section 2.4.3 also document the consequences of simply assuming a normal distribution.

data (see Nadarajah et al., 2014). In the non-parametric case, we apply a kernel density method (see Nadaraya, 1964) which does not impose a specific theoretical distribution but instead derives the distribution of losses by smoothing the empirical distribution with an appropriate kernel function and bandwidth parameter (see Scaillet, 2004; Chen, 2008).[6] For all six methods, we estimate the full set of parameters (time series model and distribution model) by the two-step procedure of McNeil and Frey (2000) and Bhattacharyya et al. (2008), which, in parametric approaches, can reduce misspecification error (in the parameters of the time series model) related to an incorrect distributional choice (see Ergen, 2015) and, in non-parametric ones, is the de facto standard (see Gao and Song, 2008). To characterize past investment risk in commodity futures markets and analyze its behavior in turbulent phases like recessions and stock market downturns, we apply the estimators to our entire range of return data.

To investigate whether our ES estimators can provide accurate forecasts of future ES, we rely on the unconditional and conditional backtests of Du and Escanciano (2017). We do this in a rolling-window setting because the quality of ES predictions may vary over time.[7] In addition, a rolling-window approach captures the empirical practice of regular model updating in VaR and ES estimation (see Hillebrand, 2005; Ardia and Hoogerheide, 2014). In other words, we take the perspective of an investor (or exchange with margin focus) continuously applying our estimators to obtain up-to-date ES forecasts and check how often they failed. What's even more important is that we can also see in which market phases they were unable to support decision making.

The remainder of our study is organized as follows. Section 2.2 introduces our general approach of ES estimation, the above-mentioned distribution models and our backtesting techniques. Section 2.3 reports the key characteristics of our dataset. Section 2.4 presents our results, which we subdivide into an analysis of the historic risk levels in the commodity sector, the evaluation of the forecasting abilities of our ES estimators and a series of robustness checks. Section 2.5 concludes and highlights directions for future research.

2.2. Methodology

2.2.1. General estimation procedure

Our starting point is a strictly stationary time series $(L_t)_{t\in\mathbb{N}}$ representing the losses (i.e., negative log returns) of commodity futures investments. Following McNeil and Frey (2000), we assume that its dynamics are captured by

$$L_t = \mu_t + \sigma_t X_t, \tag{2.1}$$

where the innovations (X_t) are an independent and identically distributed (iid) process with zero mean, unit variance and continuous marginal cumulative distribution function (cdf) F. For each t the time-varying mean μ_t and standard deviation σ_t are assumed to be measurable with respect to Ω_{t-1}, the information about the process up to time $t-1$.

In this context, it can be shown that, under a coverage level of α, such as 5 %, the time-varying VaR and ES of the losses are

$$\mathrm{VaR}_{\alpha,t}(L_t) = \mu_t + \sigma_t \mathrm{VaR}_\alpha(X), \tag{2.2}$$
$$\mathrm{ES}_{\alpha,t}(L_t) = \mu_t + \sigma_t \mathrm{ES}_\alpha(X), \tag{2.3}$$

[6] Section 2.2.2.6 explains why we have to focus on the kernel density method and cannot consider non-parametric alternatives like historical simulation.

[7] This is because the frequency of extremes and consequently the distributional form of losses may change (see Bali, 2007; Bali et al., 2008).

where $\text{VaR}_\alpha(X)$ and $\text{ES}_\alpha(X)$ denote the VaR and ES of the distribution of (X_t), which by assumption does not depend on t (see Hoga, 2018). In accordance with Du and Escanciano (2017), $\text{VaR}_\alpha(X)$ is defined as the $(1-\alpha)$-quantile of F, i.e., $\text{VaR}_\alpha(X) = F^{-1}(1-\alpha)$, which means that, in line with actuarial practice, it is a positive value not exceeded by most realizations of (X_t). Based on that, the ES is defined as[8]

$$\text{ES}_\alpha(X) = \frac{1}{\alpha}\int_{\text{VaR}_\alpha(X)}^{\infty} x f(x)\mathrm{d}x = \frac{1}{\alpha}\int_0^\alpha \text{VaR}_v(X)\mathrm{d}v, \tag{2.4}$$

where f is the continuous marginal probability density function (pdf) of (X_t).[9] It represents the expected value of (X_t) in the worst-case of a tail-event, i.e., when the VaR is exceeded.

To obtain ES estimates based on this framework, McNeil and Frey (2000), Bhattacharyya et al. (2008) and Ergen (2015) propose a parsimonious but effective two-step procedure.[10] We follow its basic idea. Also, note that, especially in our backtests, we split given loss data $(L_t)_{t=1,\dots,T,\dots,T+n}$ into an in-sample period of size T (with indices $1,\dots,T$) for parameter estimation and a subsequent out-of-sample period of size n (with indices $T+1,\dots,T+n$) for evaluation.

In the first step, we have to choose a suitable model for the conditional mean and standard deviation in Equation (2.1), fit it to the available in-sample loss data, estimate (μ_t) and (σ_t), and calculate the implied model residuals (X_t), which will be relevant for the second step. When it comes to volatility modeling, the literature tells us that it is unlikely to find a model that perfectly describes our data (see Bollerslev, 1986) and that, out of hundreds of competing models, the simple GARCH(1,1) model tends to perform best (see Hansen and Lunde, 2005).[11] Therefore, supplemented by the fact that higher-order autocorrelation of losses are typically negligible (see Campbell et al., 1993), researchers and practitioners often prefer AR(1)-GARCH(1,1) settings (see McNeil and Frey, 2000; Auer, 2015; Du and Escanciano, 2017; Le, 2020). We follow this majority approach and use the specification

$$\mu_t = \alpha_0 + \alpha_1 L_{t-1}, \tag{2.5}$$

$$\sigma_t^2 = \beta_0 + \beta_1 \left(\sigma_{t-1}X_{t-1}\right)^2 + \beta_2\sigma_{t-1}^2, \tag{2.6}$$

where α_0,α_1 and β_0,β_1,β_2 are the parameters of the mean and variance equation, respectively. We fit the model making no complex assumption about F. More precisely, we estimate its parameters via quasi maximum likelihood (QML), i.e., by using the normal distribution assumption which we do not necessarily believe. QML has the powerful property that, as long as the mean and variance equations are correctly specified (and some mild additional conditions hold), it delivers consistent and asymptotically normal estimates even if the true distribution is different from normal (see Bollerslev and Wooldridge, 1992; Francq and Zakoïan, 2004, 2012; Fan et al., 2014). While QML avoids the problem of inconsistent ML estimates in the case of a misspecified skewed innovation distribution (see Newey and Steigerwald, 1997), the price to pay is that the estimates may not be fully efficient. However, especially in large samples, this issue is not relevant for forecasting VaR and ES (see Bauwens and Laurent, 2005). With the resulting parameter estimates at hand, we can estimate (μ_t) and (σ_t) via the structures in Equations (2.5) and (2.6) and obtain the model residuals (or standardized losses) via Equation (2.1), i.e., $X_t = (L_t - \mu_t)/\sigma_t$.

[8] Note that this notation of ES corresponds to that of Du and Escanciano (2017) and therefore, slightly deviates from our definition of ES in Chapter 1.

[9] In the case of a continuous distribution, ES is also given by $\text{ES}_\alpha(X) = \mathbb{E}(X|X \geq \text{VaR}_\alpha(X))$.

[10] For general discussions of one-step vs. multi-step estimation techniques, see (Tsay, 2015, chpt. 3.5.3), Bauwens and Laurent (2005), Carnero and Eratalay (2014) and the references therein. Jalal and Rockinger (2008) show that the two-step procedure (with extreme value theory specification) performs well even in non-GARCH data.

[11] In addition, Furió and Climent (2013) and Novales and Garcia-Jorcano (2017) emphasize that the choice of innovation distribution is typically more influential on VaR forecasting than the conditional volatility specification.

In the second step, we fit one of the distribution models outlined in Section 2.2.2 to the residuals of the first step.[12] This enables us to estimate $\text{VaR}_\alpha(X)$ by using the inverse cdf F^{-1} of (X_t) and to obtain $\text{ES}_\alpha(X)$ by utilizing Equation (2.4). Rescaling to the original losses via Equation (2.3) delivers in-sample ES estimates, which we analyze in Section 2.4.1. Furthermore, out-of-sample standardized losses, which will be required for empirical backtesting in Section 2.4.2, can be obtained by successively filling $X_t = (L_t - \mu_t)/\sigma_t$ with out-of-sample data (see Engle, 2001). To avoid confusion, note that, in the remainder of the text, we use the term ES estimator to refer to the package of AR-GARCH model and distribution specification.

In comparison to the one-step estimation of parametric setups, where the parameters of the AR(1)-GARCH(1,1) model and assumed innovation distribution are estimated simultaneously via ML, the properties of QML make the two-step procedure less sensitive to misspecification (see Ergen, 2015). Furthermore, with the two-step approach any difference in estimator performance will be entirely attributable to the choice of distribution in the second step because the AR(1)-GARCH(1,1) parameters obtained in the first step will be identical for all estimators.[13] Finally, one-step estimation can be quite problematic if a distribution (such as the g-and-h distribution) does not have a closed-form cdf. The two-step approach mitigates this issue. This is why two-step estimation is the established technique in non-parametric settings (see Gao and Song, 2008).

2.2.2. Distribution models

2.2.2.1. Hansen's skewed t distribution

One of the best-known methods of forecasting ES is to assume that the data follows the skewed t distribution of Hansen (1994) which, in contrast to a normal distribution, can model the empirically relevant features of asymmetry and heavy tails. In a commodity context, Degiannakis and Potamia (2017) have used it to forecast VaR and ES of COMEX gold, silver, and copper futures and pointed out its theoretical merits in comparison to simple standard approaches.[14]

Hansen's skewed t distribution for $x \in \mathbb{R}$ is characterized by the pdf

$$f(x) = \begin{cases} bc\left(1 + \frac{1}{\nu-2}\left(\frac{bx+a}{1-\lambda}\right)^2\right)^{-\frac{\nu+1}{2}} & \text{if } x < -\frac{a}{b}, \\ bc\left(1 + \frac{1}{\nu-2}\left(\frac{bx+a}{1+\lambda}\right)^2\right)^{-\frac{\nu+1}{2}} & \text{if } x \geq -\frac{a}{b}, \end{cases} \tag{2.7}$$

where $2 < \nu < \infty$, $-1 < \lambda < 1$ and

$$a = 4\lambda c\frac{\nu-2}{\nu-1}, \qquad b^2 = 1 + 3\lambda^2 - a^2, \qquad c = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi(\nu-2)}\Gamma(\nu/2)}. \tag{2.8}$$

By definition, a skewed t random variable x has zero mean and unit variance. The parameters λ and ν control the degree of skewness and kurtosis, respectively. Thus, if $\lambda = 0$, the skewed t distribution reduces to the Student t distribution and, if additionally $\nu \to \infty$, it converges to the standard normal distribution.

To determine the parameter values, for which a theoretical distribution matches given data (in our case model residuals) best, we can choose among fitting methods based on moments, quantiles or maximum likelihood (see Cramér, 1946). For Hansen's skewed t distribution, we use the maximum likelihood method. Afterward we plug the resulting values into the pdf and integrate it to obtain the cdf relevant for estimating and backtesting the ES.

[12] Guided by the general properties of our residuals, we fit all distribution models under the constraints of zero mean and unit variance.

[13] In simultaneous estimation, each distribution would lead to other AR(1)-GARCH(1,1) parameter estimates.

[14] For other kinds of skewed t distributions and an application in commodity research, see Cheng and Hung (2011).

2.2.2.2. Generalized Pareto distribution

Because the ES is a risk measure with natural focus on extreme losses, some researchers recommend its estimation based on extreme value theory (see McNeil and Frey, 2000; Gilli and Këllezi, 2006; Martins-Filho et al., 2018). One of the most popular approaches originating from this field, which has been used by Krehbiel and Adkins (2005) (Marimoutou et al., 2009) in a VaR analysis of NYMEX energy futures (Brent and WTI crude oil spot markets) and will also be the implemented in our study, is the peak over threshold (POT) method. It builds on the limit theorem of Balkema and de Haan (1974) and Pickands (1975), which, in our context, states that, for (almost) any form of loss distribution, the distribution of excesses $Y_t := X_t - u$ over a large threshold u for $t \in \mathbb{N}$ is well approximated by the generalized Pareto distribution (GPD). This result allows us to model the tail of a distribution without having to specify the form of the entire distribution function. In other words, we can derive ES based on the cdf of excesses, given by

$$G(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0, \\ 1 - e^{-\frac{y}{\sigma}} & \text{if } \xi = 0, \end{cases} \tag{2.9}$$

where ξ and $\sigma > 0$ are shape and scale parameters, respectively (see McNeil, 1997). The support of this function is $y \geq 0$ when $\xi \geq 0$ and $0 \leq y \leq -\frac{\sigma}{\xi}$ when $\xi < 0$. If $u = 0$ and $\xi = 0$, the dedicated G is equivalent to the exponential distribution.

Denoting by $q > \alpha$ the percentage of observations in the data exceeding u, the cdf of excesses implies the following cdf for the tail data ($x \geq u$ if $\xi \geq 0$ and $u \leq x \leq u - \frac{\sigma}{\xi}$ otherwise):

$$F(x) = \begin{cases} 1 - q\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0, \\ 1 - qe^{-\frac{x-u}{\sigma}} & \text{if } \xi = 0. \end{cases} \tag{2.10}$$

Consequently, we can apply the POT method by setting the threshold u (or, equivalently, make a choice of q), fitting a GPD via maximum likelihood to the corresponding excesses of our in-sample data and plugging the estimated parameters into Equation (2.10).[15] As discussed in McNeil (1997) and Novales and Garcia-Jorcano (2019), there are several strategies for the choice of threshold u, mostly resulting in q values between 0.08 and 0.15. To simplify our investigation, we follow their approaches and choose $q = 0.1$.

2.2.2.3. g-and-h distribution

The g-and-h distribution of Tukey (1977) is another flexible distribution model, which originates from transformations of standard normal variables and whose properties have been studied by Martinez and Iglewicz (1984) and Headrick et al. (2008). After a series of stock market applications (see, for example, Badrinath and Chatterjee, 1988; Mills, 1995; Jiménez Moscoso and Arunachalam, 2011), it has only recently found its way into the commodity sector, where it has been found to model the VaR of the futures-based Bloomberg Commodity Index better than simpler traditional distributions (see Díaz et al., 2017).

[15] Fitting the GPD in-sample ensures that the cdf $F(x)$ is real-valued for each in-sample argument x. However, when applying the estimated F out-of-sample, which has to be done in our backtests, complex results can arise. This occurs when there are strong differences between in-sample and out-of-sample data that lead to out-of-sample function arguments outside the support of the fitted cdf. In such situations, we set the corresponding $F(x)$ to *NaN* and perform our backtests without the troublesome date. In a total of 6,637 evaluations per index, we had to do this 8 (S&P GSCI), 9 (energy), 1,411 (precious metals), 151 (industry metals), 0 (agriculture) and 58 (livestock) times.

For the g-and-h distribution, a cdf formula in closed form does not exist. However, a g-and-h distributed variable x can be defined with G^{-1}, a part of its quantile function, as satisfying

$$x = G^{-1}(z) := \begin{cases} \frac{e^{gz}-1}{g} e^{\frac{hz^2}{2}} & \text{if } g \neq 0, \\ ze^{\frac{hz^2}{2}} & \text{if } g = 0, \end{cases} \tag{2.11}$$

where z follows a standard normal distribution, g controls the skewness and $h \geq 0$ is related to the kurtosis of the distribution. Consequently, the corresponding cdf can be derived as

$$F(x) = \Phi(G(x)), \tag{2.12}$$

where Φ is the standard normal cdf. If $g = h = 0$, the g-and-h distribution reduces to the standard normal case. It can also capture other commonly used distributional shapes, such as log-normal, Weibull and exponential, most of the Pearson family and even distributions that do not have finite first four moments, such as the Cauchy distribution (see Xu et al., 2014).

To estimate the parameters g and h we follow Headrick et al. (2008) by setting skewness and excess kurtosis of our in-sample residuals to the theoretical third and fourth standardized moments of the g-and-h distribution, respectively, and solving the resulting system of equations. Afterwards, we insert the estimates of g and h into Equation (2.11), solve it numerically for all $z = G(X_t)$ with in-sample t and compute outcomes of the cdf with Equation (2.12).

2.2.2.4. Johnson distribution

Additional distributions, based on and generalizing normal transformations, are contained in the distribution system of Johnson (1949), which has been compared in detail to the g-and-h distribution in Mac Gillivray (1992) and applied by Mögel and Auer (2018) to model the VaR of London Bullion Market gold spot returns. In this system, the associated cdf for $x \in \mathbb{R}$ is

$$F(x) = \Phi\left(\gamma + \delta g\left(\frac{x-\xi}{\lambda}\right)\right), \tag{2.13}$$

where γ and $\delta > 0$ determine the shape of the distribution, ξ is a location and $\lambda > 0$ a scale factor. The function g can take one of the following forms: in the log-normal system $g^{SL}(x) = \ln(x)$ for each $x > 0$, in the unbounded system $g^{SU}(x) = \sinh(x)$ for all $x \in \mathbb{R}$ or in the bounded system $g^{SB}(x) = \ln\left(\frac{x}{1-x}\right)$ for each $x \in (0, 1)$.

To empirically implement the Johnson system, we have to choose the form of g and estimate the parameters γ, δ, ξ and λ. To do that, we follow George and Ramachandran (2011) by using the quantile estimation approach of Wheeler (1980), where we modified the minimum and maximum orders of sample quantiles to be 0.05 and 0.95, respectively, as recommended by Aitchison and Brown (1957).

2.2.2.5. Gaussian mixture distribution

While the unbounded Johnson system allows only unimodal non-normality (see DeBrota et al., 1989), Gaussian mixtures break this limitation (see Broda and Paolella, 2011). According to Kon (1984) daily financial data is better described by Gaussian mixtures than by non-mixed classic distributions. Meade (2010) additionally suggest suitability for crude oil spot returns.

We set up a two-component Gaussian mixture via the weighted sum of two normal distributions with means μ_1, μ_2 and standard deviations σ_1, σ_2, respectively, and a weight λ which is assumed

to range between 0 and 1.[16] The resulting cdf for all $x \in \mathbb{R}$ is denoted as

$$F(x) = \lambda\Phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-\lambda)\Phi\left(\frac{x-\mu_2}{\sigma_2}\right). \tag{2.14}$$

The mixture reduces to a normal distribution if $\lambda = 0$ or $\lambda = 1$. Depending on its parametrization, it can incorporate skewness, kurtosis, unimodality or bi-modality (see Rossi, 2014).

To fit this model, i.e., to find maximum likelihood estimates of the parameters, we use the expectation-maximization algorithm (see Hastie et al., 2001).

2.2.2.6. Smoothed empirical distribution

The popular alternatives to parametric distribution fitting are non-parametric estimation via historical simulation (see Chapter 1), quantile regression (see Taylor, 2008b) and kernel techniques (see Scaillet, 2004). These methods do not assume that a theoretical distribution model holds but instead build on the non-smoothed (historical simulation, quantile regression) or smoothed (kernel techniques) empirical distribution function.

Unfortunately, not all of these techniques can be evaluated in the framework of Du and Escanciano (2017) because a closer look at its derivation reveals that the backtests require *invertible* distribution functions to be applicable (see also Section 2.2.3). Only kernel techniques which transform empirical staircase functions to continuous functions are in line with this premise. For this reason and because its performance tends to be similar to classic historical simulation (see Chen, 2008), we chose a standard kernel density estimator (KDE) to represent the class of non-parametric methods.[17]

Our approach is based on Nadaraya (1964) and requires chosing a kernel function k, which can be a standard (often symmetric) pdf, and a bandwidth parameter $h > 0$. With these components, the pdf f corresponding to empirical data $X_1, \ldots, X_T$ can be estimated via

$$\hat{f}(x) = \frac{1}{hT}\sum_{t=1}^{T} k\left(\frac{x-X_t}{h}\right) \text{ for } x \in \mathbb{R}. \tag{2.15}$$

Consequently, the associated estimated cdf is

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(u)\mathrm{d}u = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\frac{x-X_t}{h}} k(u)\mathrm{d}u \text{ for } x \in \mathbb{R}. \tag{2.16}$$

According to Nadaraya (1965), for uniformly continuous f and a kernel function with bounded variation, an appropriate bandwidth ensures that $\hat{f}$ converges uniformly with probability one to f as the sample size T tends to infinity. While switching the utilized kernel function typically does not have a crucial impact, estimation results are very sensitive to the selected bandwidth (see Wand and Jones, 1995, chpt. 2.7). Because the bandwidth controls the smoothness of $\hat{f}$, an unsuitable choice of h can lead to under- or over-smoothing and thus misleading distribution shapes.

To mitigate bandwidth selection risk, we follow Chen (2008) and Yu et al. (2010) in two steps. First, we use the standard normal pdf ϕ for the kernel k. Hence, our estimated cdf simplifies to $\hat{F}(x) = \frac{1}{T}\sum_{t=1}^{T}\Phi(\frac{x-X_t}{h})$ for $x \in \mathbb{R}$. Second, we implement an effective normal scale rule to select the bandwidth h (see Wand and Jones, 1995, chpt. 3.2.1). That is, we compute

$$h = \left(\frac{8\sqrt{\pi}\int\phi(x)^2\mathrm{d}x}{3T\left(\int x^2\phi(x)\mathrm{d}x\right)^2}\right)^{\frac{1}{5}}\hat{\sigma}, \tag{2.17}$$

[16] Relying on a m-component mixture and letting the data determine the value of m (as in Kuester et al., 2006) does not qualitatively influence our results.

[17] Recent research extends the standard or single kernel idea to double kernels. However, its additional smoothing tends to introduce additional estimation error (see Kato, 2012).

where $\hat{\sigma}$ is an unbiased estimate of the population standard deviation.[18]

2.2.3. Backtests

After the backtests for VaR, proposed by Kupiec (1995), Christoffersen (1998) and Berkowitz et al. (2011) had found their way into the VaR backtesting standards of bank regulators (see Basel Committee of Banking Supervision, 2011),[19] researchers immediately started asking whether there are simple but effective ways to backtest ES. In recent years, research has produced significant results. McNeil and Frey (2000) developed an unconditional ES evaluation procedure, based on a likelihood residual approach, that uses bootstrap simulations to conduct a one-sided test against the alternative hypothesis that ES is systematically underestimated. Berkowitz (2001) also suggested a one-sided test, based on a censored normal likelihood, including a failure tolerance term and the functional delta approach of Kerkhof and Melenberg (2004). Wong (2008) presented a saddle-point technique and some empirical examples that illustrate higher small-sample power than the two latter backtests. However, it is limited by the requirement of normally distributed data. The additional backtests of Righi and Ceretta (2013) and Acerbi and Szekely (2014) are not restricted by specific distributional assumptions but require extensive simulations to obtain critical values.

Constanzino and Curran (2015) finally designed the first unconditional coverage test for ES, which is similar to the unconditional VaR backtest of Kupiec (1995) and thus easy to comprehend and to compute. It was extended in Du and Escanciano (2017), where the authors also present a concise conditional coverage test for ES, which is an ES analogue to the conditional VaR backtest of Christoffersen (1998) and Berkowitz et al. (2011). These new coverage tests for ES are based on distribution-free cumulative violations, which, in contrast to the pure violations of traditional VaR backtests, can consider tail risk magnitude. In comparison to the alternatives, the latest unconditional coverage test evaluates against the alternative of under- and over-estimation, thus also protecting institutions against inefficient use of capital. While the unconditional coverage test looks at under- and over-estimation of risk, the conditional coverage test enables us to check whether both the unconditional coverage property and the independence of cumulative violations are fullfilled simultaneously. This is important because dependent cumulative violations indicate that a given ES estimator neglects available predictive information. We implement both the unconditional and the conditional versions of the new ES backtest by using the following approach and specification.

Because $\mathrm{ES}_\alpha(X)$ covers X_t exceeding the $\mathrm{VaR}_\alpha(X)$, for all out-of-sample $t \in \{T+1, \ldots, T+n\}$, a matching α-violation (or hit) can be defined as

$$h_t(\alpha) = I(X_t \geq \mathrm{VaR}_\alpha(X)), \tag{2.18}$$

where I denotes a mathematical indicator function which maps to 1 if its argument is true and to 0 otherwise. Because Equation (2.4) allows expressing the ES as the definite integral of the VaR and the popular unconditional VaR backtest focuses on the number of α-violations, its extension to the ES looks, for $t \in \{T+1, \ldots, T+n\}$, at the integrated $(h_t(\alpha))$, i.e.

$$H_t(\alpha) = \frac{1}{\alpha} \int_0^\alpha h_t(v) \mathrm{d}v. \tag{2.19}$$

If the VaR model (and, hence, the ES model) is correctly specified, the mean of $(H_t(\alpha))$ has to equal $\frac{\alpha}{2}$. To test whether this requirement is fulfilled, the $(H_t(\alpha))$ can be computed by using

[18] Note that Footnote 12 also applies to the non-parametric case.

[19] The regulatory standard backtest for VaR builds on the requirement that the rate of out-of-sample violations, i.e. losses exceeding a well-estimated VaR_α, should range around α. Compliance with this coverage condition can be statistically evaluated with the binomial distribution.

Equation (2.18) and the fitted invertible $\hat{F}$ as follows:[20]

$$\begin{aligned}\hat{H}_t(\alpha) &= \frac{1}{\alpha}\int_0^{\alpha} I(\hat{F}(X_t) \geq 1-v)\mathrm{d}v \\ &= \frac{1}{\alpha}(\hat{F}(X_t) - 1 + \alpha) I(\hat{F}(X_t) \geq 1-\alpha).\end{aligned} \tag{2.20}$$

With these results at hand, the *unconditional backtest* with null hypothesis $H_0^u : \mathbb{E}(H_t(\alpha) - \frac{\alpha}{2}) = 0$ is conducted via the test statistic

$$U = \frac{\sqrt{n}(\bar{H} - \frac{\alpha}{2})}{\sqrt{\alpha(\frac{1}{3} - \frac{\alpha}{4})}}, \tag{2.21}$$

where $\bar{H} = \frac{1}{n}\sum_{t=T+1}^{T+n} \hat{H}_t(\alpha)$ is the mean of the $(\hat{H}_t(\alpha))$. Du and Escanciano (2017, Corollary 1) show that the statistic U is asymptotically standard normal if the (in-sample) estimation period T is much larger than the (out-of-sample) evaluation period n.[21] Consequently, H_0^u can be rejected at the $\bar{\alpha}$-level if the realized U lies outside the interval $[-\Phi^{-1}(1-\frac{\bar{\alpha}}{2}), \Phi^{-1}(1-\frac{\bar{\alpha}}{2})]$.

The *conditional backtest* evaluates the null hypothesis $H_0^c : \mathbb{E}(H_t(\alpha) - \frac{\alpha}{2}|\Omega_{t-1}) = 0$ via a Box-Pierce test. For its implementation and for $j \in \{0, \ldots, m\}$, the lag-j autocovariances $\gamma_j = \mathrm{Cov}(H_t(\alpha), H_{t-j}(\alpha))$ of $(H_t(\alpha))$ have to be estimated via

$$\hat{\gamma}_j = \frac{1}{n-j}\sum_{t=T+1+j}^{T+n} (\hat{H}_t(\alpha) - \tfrac{\alpha}{2})(\hat{H}_{t-j}(\alpha) - \tfrac{\alpha}{2}). \tag{2.22}$$

We have approximately $\hat{\gamma}_j \approx \frac{1}{n-j}\sum_{t=T+1+j}^{T+n}((\hat{H}_t(\alpha) - \bar{H})(\hat{H}_{t-j}(\alpha) - \bar{H})) + (\bar{H} - \frac{\alpha}{2})^2$. Thus, testing based on these autocovariances brings power against deviations from zero autocovariance (first term) and deviations from H_0^u (second term).[22] The relevant test statistic, which is based on estimates of lag-j autocorrelations $\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0}$, can be expressed as

$$C(m) = n\sum_{j=1}^{m} \hat{\rho}_j^2. \tag{2.23}$$

According to Du and Escanciano (2017, Corollary 2) it asymptotically follows a chi-square distribution with m degrees of freedom, if again $n/T \to 0$.[23] Therefore, H_0^c is rejected if $C(m) > \chi^2_{1-\bar{\alpha},m}$, where $\chi^2_{1-\bar{\alpha},m}$ is the $(1-\bar{\alpha})$-quantile of the chi-square distribution.

2.3. Data

To capture representative commodity investments, we obtain daily data from Thomson Reuters Datastream for the world-production weighted S&P GSCI and its five sector sub-indices: energy (including Brent crude oil, WTI crude oil, gas oil, heating oil, natural gas, unleaded gasoline),

[20] Invertibility, which is satisfied by each strictly monotonically increasing cdf, is necessary to ensure the property $I(X_t \geq \mathrm{VaR}_\alpha(X)) = I(F(X_t) \geq 1-\alpha)$. If F is not strictly monotonic, such as in the case of an empirical distribution function, we have a $\leq$ relation instead of equality. Put differently, in this situation, no unique quantile function exists because of an absence of invertibility.

[21] Otherwise a parameter estimation effect requires using a normal distribution with a mean of zero and a variance depending on the asymptotic relative magnitude of T and n (see Du and Escanciano, 2017, Theorem 1).

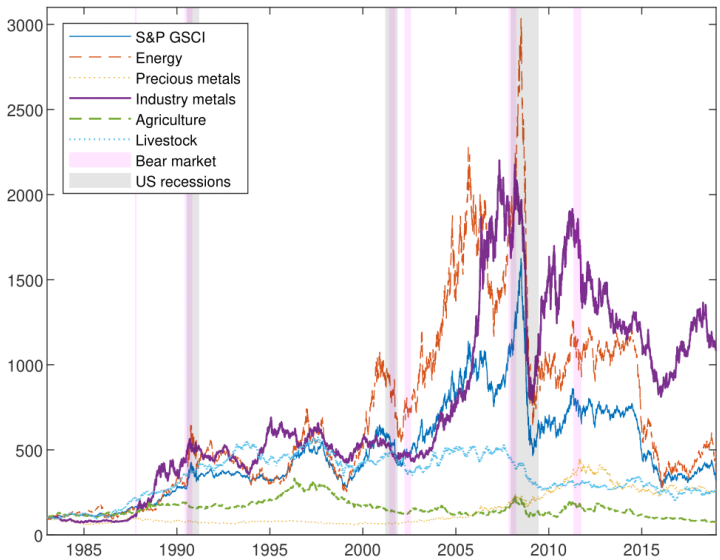
[22] Disentangling both hypotheses delivers a test of lower power (see Du and Escanciano, 2017, Footnote 8).

[23] Similar to footnote 21, violation of this condition leads to another distribution: a weighted chi-square distribution that depends on the model and data generating process (see Du and Escanciano, 2017, Theorem 2).

precious metals (gold, silver), industry metals (aluminum, copper, lead, nickel, zinc), agriculture (cocoa, coffee, corn, cotton, soybeans, sugar, Chicago wheat, Kansas wheat) and livestock (feeder cattle, lean hogs, live cattle). We focus on the total return versions of the indices which capture returns accrued from investing in liquid fully collateralized futures nearest to expiration. Unlike their spot return counterparts, they are directly replicable in practice and comparable to the returns from an investment in the S&P 500 stock market index with dividend reinvestment. For reasons of sample consistency, our investigation period spans from January 7, 1983 to December 31, 2018.

As a first look at the characteristics of our data, Figure 2.1 illustrates the appreciation and depreciation of $ 100 invested into the S&P GSCI and its five sub-indices, respectively, on January 7, 1983. Especially for later discussions, it also includes US recession periods (according to the National Bureau of Economic Research, 2010) and periods of S&P 500 bear markets (as defined in Pagan and Sossounov, 2003).[24] We can nicely identify the effects of the 2000 commodity boom or super cycle (see Erten and Ocampo, 2013). While the total GSCI investment increased to about $ 1,000 until July 2008, the investment in the energy sub-index was the most profitable of all sub-indices with a significant rise to more than $ 3,000. The global financial crisis caused a crash of the commodity market in 2008. Especially oil markets significantly contributed to the GSCI downturn because, in July 2008, the crude oil price peaked at its historic maximum of more than $ 145 per barrel and then plunged to less than $ 35 per barrel in December 2008 (see Lang and Auer, 2020). Furthermore, weakened demand in the ensuing recession caused prices for industrial metals to drop (see Jacobsen et al., 2019). After 2008, the energy investment was outperformed by the industrial metals investment, which until April 2011 had increased almost 20-fold since inception. While the latter scores first with respect to terminal value, precious metals and livestock rank far behind and agriculture even closes with a capital loss.

Figure 2.1.: Performance of investments in the S&P GSCI and its sub-indices



This figure shows the development and final values of $100 invested into the S&P GSCI (total return futures version) and its five sub-indices: energy, precious metals, industry metals, agriculture and livestock. The investment period spans from January 7, 1983 to December 31, 2018. US recession periods (according to the National Bureau of Economic Research, 2010) and S&P 500 bear markets (as defined in Pagan and Sossounov, 2003) are included as shaded areas.

[24] Bear situations are considered present when there are at least four months of negative returns or when there has been a decline greater or equal than 20 % in a single month.

To conduct our ES analysis, we calculate the daily losses of the commodity future index levels (I_t). That is, we obtain negative daily percentage log returns via $L_t = -100(\ln(I_t) - \ln(I_{t-1}))$. Table 2.1 reports some descriptive statistics for these losses: their minimum, maximum, mean, standard deviation, skewness and kurtosis. A look at the mean losses shows that only investments in agriculture suffer on average daily losses whereas all others exhibit on average daily gains, a typical observation in commodity futures markets (see Wang, 2001). Investment risk in terms of standard deviation (and minimum-maximum spread) of losses is highest in the energy sector and lowest in livestock. All time series are positively skewed, indicating that large losses are more likely than large gains. This is also reflected by the fact that maximum losses are (in absolute terms) larger than the minimum losses. Kurtosis strongly deviates from normality, with the exception of agriculture, which is close to normal.[25]

Table 2.1.: Descriptive statistics

	Min	Max	Mean	Std	Skewness	Kurtosis
S&P GSCI	-7.600	18.431	-0.013	1.233	0.563	12.143
Energy	-12.938	34.828	-0.014	1.923	0.699	17.907
Precious metals	-8.763	10.105	-0.010	1.069	0.308	9.770
Industry metals	-8.420	12.495	-0.025	1.348	0.214	7.554
Agriculture	-7.157	7.475	0.003	1.067	0.075	6.445
Livestock	-3.254	4.248	-0.010	0.871	0.109	3.832

For the period from January 10, 1983 to December 31, 2018, this table reports minimum, maximum, mean, standard deviation, skewness and kurtosis of daily losses (i.e. negative daily percentage log returns) incurred from investing in the S&P GSCI and its five sector sub-indices.

2.4. Results

2.4.1. Historic risk characterization

We start our analysis of commodity futures ES with a documentation of its past levels over time and across sectors. That is, for the 26-year period from 1983 to 2018, we apply the estimators given by Sections 2.2.1 and 2.2.2 to the S&P GSCI and its sub-indices and inspect the in-sample estimates. We limit ourselves to a presentation of results for a common coverage level of $\alpha = 0.05$.[26]

Table 2.2 reports the estimates of $\mathrm{ES}_\alpha(X)$, i.e. the ES of the AR(1)-GARCH(1,1) standardized losses, resulting under our different distributional assumptions. Note that, apart from the Johnson and g-and-h distributions, standard diagnostic checking does not reject our estimated models.[27] Across all indices, using the Johnson system delivers the smallest ES estimates, whereas the g-and-h distribution consistently yields the highest. As far as the other estimators are concerned, an ordering of estimators by ES magnitude differs from index to index. For example, for the energy sub-index the ES levels ascend when moving from the Gaussian mixture over Hansen's t to POT,

[25] Nonetheless, a Jarque-Bera test conducted at conventional significance levels rejects the null of normally distributed losses for all six time series.

[26] Historic estimates for $\alpha = 0.01$ are given in Table B.1 and Figure B.1 of the appendix. We do not conduct the backtests of Section 2.2.3 for $\alpha = 0.01$ because Du and Escanciano (2017) and Hoga (2019) advise against it. The reason for this is that, in the case of very low α, the validity of the asymptotic test distributions may break down.

[27] Ljung-Box tests for the standardized and squared standardized losses indicate that our AR(1)-GARCH(1,1) filter is suitable. Exemplary findings for the S&P GSCI are presented in Table B.2 of the appendix. Figure B.2 visualizes the corresponding fit of our distribution models. Results for the sub-indices and the outcomes of Kolmogorov-Smirnov goodness-of-fit tests (as specificed in Stavroyiannis, 2018) are available upon request.

and descend for agriculture. Kernel density estimates are often either similar to or slightly higher than these three methods.

Table 2.2.: Estimated ES of standardized losses for $\alpha = 0.05$

	Hansen	POT	g-and-h	Johnson	GM	KDE
S&P GSCI	2.240	2.264	**2.604**	***2.110***	2.278	2.278
Energy	2.243	2.288	**2.627**	***2.117***	2.241	2.297
Precious metals	2.295	2.360	**2.963**	***2.098***	2.342	2.375
Industry metals	2.214	2.249	**2.689**	***2.075***	2.239	2.270
Agriculture	2.155	2.149	**2.483**	***2.065***	2.172	2.166
Livestock	2.166	2.219	**2.279**	***2.116***	2.196	2.243

This table reports the estimated ES of standardized losses related to investments in the S&P GSCI and its sector sub-indices. The estimates are obtained using the methodology of Sections 2.2.1 and 2.2.2. That is, under an AR(1)-GARCH(1,1) model for daily losses between January 10, 1983 to December 31, 2018, the ES with coverage level $\alpha = 0.05$ is obtained by applying different types of innovation distribution models: Hansen's skewed t, peak over threshold (POT), g-and-h, Johnson system, Gaussian mixture (GM) and kernel density estimation (KDE). The lowest ES estimates for each index are marked in italics and the highest in bold.

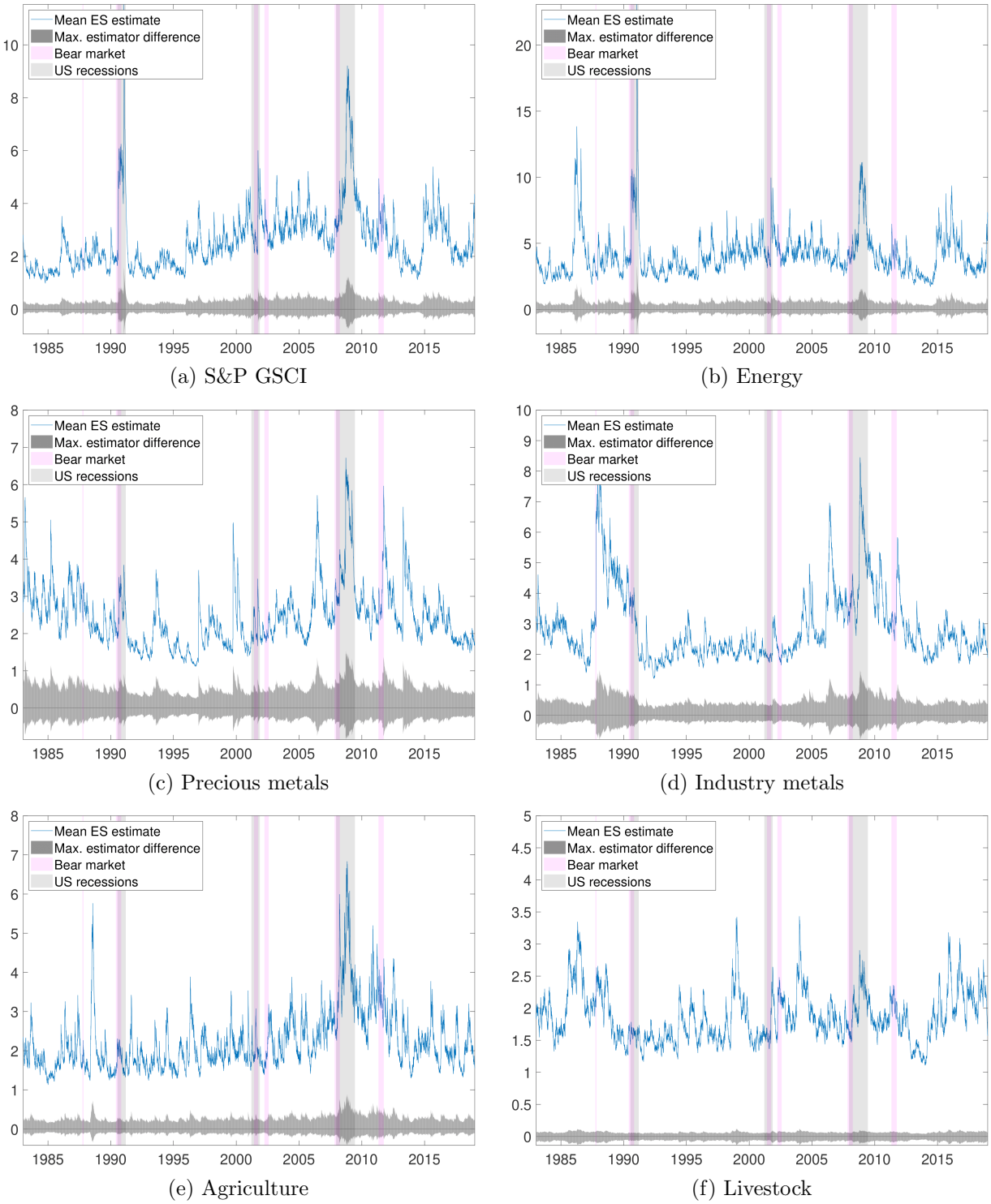
Time-varying risk levels are obtained by daily rescaling the estimates in Table 2.2 via Equation (2.3) using AR(1)-GARCH(1,1) estimates of the time-dependent mean (μ_t) and standard deviation (σ_t). Because, with six indices and six distributional settings, this generates 36 ES time series, and to allow a discussion, which is not driven by estimator-specific results, we opt for a compact form of result visualization. Using the fact that weighted averages of estimates derived from estimators with different performance are typically superior to estimates from a single estimator (see Timmermann, 2006; Wang et al., 2009, 2016; Baumeister and Kilian, 2015), we calculate the simple arithmetic mean of the estimates produced by our six distribution models. For each index, Figure 2.2 presents the resulting time-varying ES averages. To indicate the range of estimates delivered by our different estimators, we calculate their deviations from this average and plot the smallest and largest deviation in the form of shaded bars around the zero line. Consequently, the value of the largest (smallest) ES estimate can be obtained by adding the plotted positive (negative) deviation to the average estimate.[28] Similar to Figure 2.1, we extend Figure 2.2 by US recessions and S&P 500 bear sequences.

Figure 2.2 shows that our estimators are close to each other when the general risk level is low and are farther apart when risk is high. This is not surprising because higher volatility σ_t in Equation (2.3) has a magnifying effect. When taking a closer look at the overall picture, we can see how spikes in ES, which are of crucial interest from an investment perspective, relate to recession periods and bear phases in the stock market. Previous research indicates that commodity markets and the economy are linked by supply and demand effects (see Sockin and Xiong, 2015; Clayton, 2016) and that, in recent years, shocks in the stock market often tend to transmit to the commodity market (and vice versa) because, in a process of financialization, many futures contracts have become investment vehicles (see Cheng and Xiong, 2014; Adams and Glück, 2015).[29] In both cases, the strength of transmission depends on the type of commodity.

The first relevant period in our sample is a bear market at the end of 1987. It followed Black Monday (October 19, 1987), when a sudden stock market crash shocked markets all over the world (see Wang, 2001). While, in this less financialized period, we observe only little impact on most commodity sectors, the ES of industry metal futures rose significantly.

[28] Note that, in our case, the plotted deviations relate exclusively to the Johnson and the g-and-h estimates.
[29] The impact of speculation on futures markets is still controversial and hard to measure (see Haase et al., 2016).

Figure 2.2.: Time-varying ES estimates, $\alpha = 0.05$



(a) S&P GSCI (b) Energy

(c) Precious metals (d) Industry metals

(e) Agriculture (f) Livestock

This figure illustrates the behavior of conditional ES estimates (with $\alpha = 0.05$) for the S&P GSCI index and its five sector sub-indices from January 10, 1983 to December 31, 2018. Following Sections 2.2.1 and 2.2.2, in-sample ES estimates are obtained by applying an AR(1)-GARCH(1,1) model to daily losses and using six alternative ways to approximate the distribution of the model innovations. In other words, they are generated by daily AR(1)-GARCH(1,1) rescaling of the ES values in Table 2.2. We do not plot the ES time series for all individual estimators but their average. In addition, we present the maximum (and minimum) of the differences between the single estimators and the reported average in the form of bars around the zero line. US recessions and S&P 500 bear markets are highlighted similar to Figure 2.1.

The next period of interest is the recession from July 1990 to March 1991, and the corresponding bear market, which originated in the 1990 oil price shock, the end of the Cold War and the savings and loan crisis (see Fintzen and Stekler, 1999). Here we detect a significant rise of investment risk in the energy and precious metals sectors (and consequently in the total commodity index). This is linked to supply-side problems in the oil sector (caused by Iraq's invasion of Kuwait in August 1990 and the ensuing First Gulf War) and safe haven capital flows in gold and silver markets (see Baur and McDermott, 2010; Lucey and Li, 2015). Agriculture and livestock are almost unaffected, which is in line with earlier evidence indicating a weaker link to the stock market than can be observed for other commodities (see Nguyen et al., 2015). It also partially explains why livestock (energy) exhibits the lowest (highest) ES in many time frames.

The recession from March 2001 to November 2001, fueled by the dot-com crash in 2000 and the September 11, 2001 terrorist attacks (see Mostaghimi, 2004), had a less significant effect on commodity markets. Because the stock market shock primarily hit internet-based companies, no crucial transmission occurred. That is, over all indices, risk spikes do not stand out in comparison to the pre-recession months. A more detailed look reveals that the main ES peaks come up directly after the 9/11 attacks (and at the beginning of the bear market that accompanied the recession), which were followed by price increases in the energy sector. The additional bear market shortly after the recession apparently did not have a crucial effect on commodity futures markets.

In contrast to the millennium crash, the global financial crisis, ushering in a bear market in winter 2007/2008 and a recession from December 2007 to June 2009, caused a risk surge in all commodity sectors. The core of this impact is often considered to be a set of contemporaneous supply and demand surprises that coincided with low inventories and that were magnified by macroeconomic shocks and policy responses (see Carter et al., 2011). Furthermore, one may also argue that shock transmission from the stock market was strong because, induced by financialization, correlations between stock and commodity markets, which have been low (or negative) historically, had significantly increased (see Silvennoinen and Thorp, 2013).[30] To illustrate this, Figure 2.3 plots the time-varying correlations between the returns of our commodity indices and the S&P 500 index. Between July 2008 and 2009, correlations rose from about -0.2 to 0.6. Thus, considering empirical evidence on causality of stock market movements for commodity futures fluctuations in this period (see Nguyen et al., 2015), shocks in the stock market could easily spill over to the commodity futures market.[31]

Finally, as a last period of interest, we have the August 2011 bear market, triggered by the European debt crisis (see Majewska and Olbrys, 2017). Similar to the recessions before the global financial crisis, we can detect mild spikes in the ES of several sectors; however, they are not comparable to the magnitude of the 2007-2009 increases in investment risk.

To extend our discussion beyond the impact of general shocks on commodity markets, the appendix provides additional information on commodity-specific events and their consequences for the risk levels in commodity subsectors.

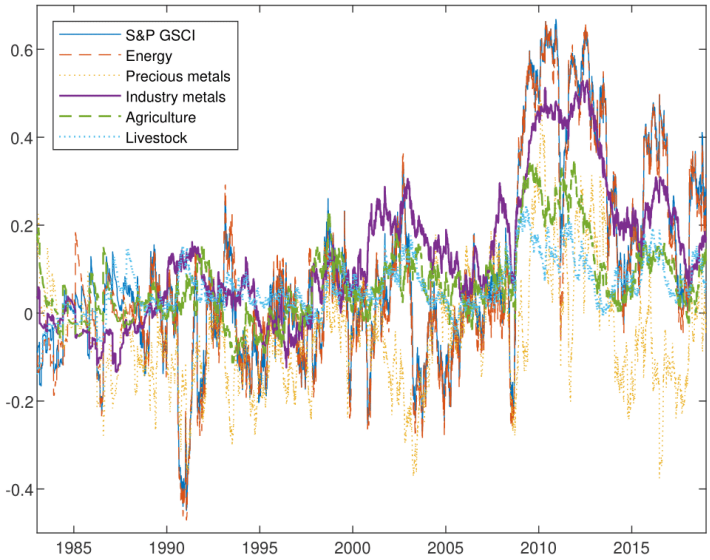
2.4.2. Risk prediction accuracy

While our previous results indicate that many of our estimators can provide useful in-sample ES estimates, which may be the input for mean-risk portfolio models comparing past risk-adjusted investment performance (see Schuhmacher and Auer, 2014), investors are often concerned with an even more important issue: the prediction of future ES. Since we know from basic econometrics

[30] In addition, transmissions between commodity sectors have intensified (see Kang et al., 2017).

[31] Table B.3 of the appendix shows that, in the recession, hedge and safe haven properties (as defined in Baur and Lucey, 2010) of commodities for stocks are largely illusory. Only precious metals may be considered a hedge (negative correlation on average) and a safe haven (negative correlation in times of stress). However, their correlation magnitudes have rather weak economic relevance.

Figure 2.3.: Correlations between commodity futures and stock returns



This figure shows the time-varying correlations between the returns on the S&P GSCI (and its sub-indices) and the S&P 500 index. Following Joy (2011) and Manera et al. (2013) correlations are determined via a DCC(1,1)-GARCH(1,1) model of the Engle (2002) type, which is estimated via multi-step QML.

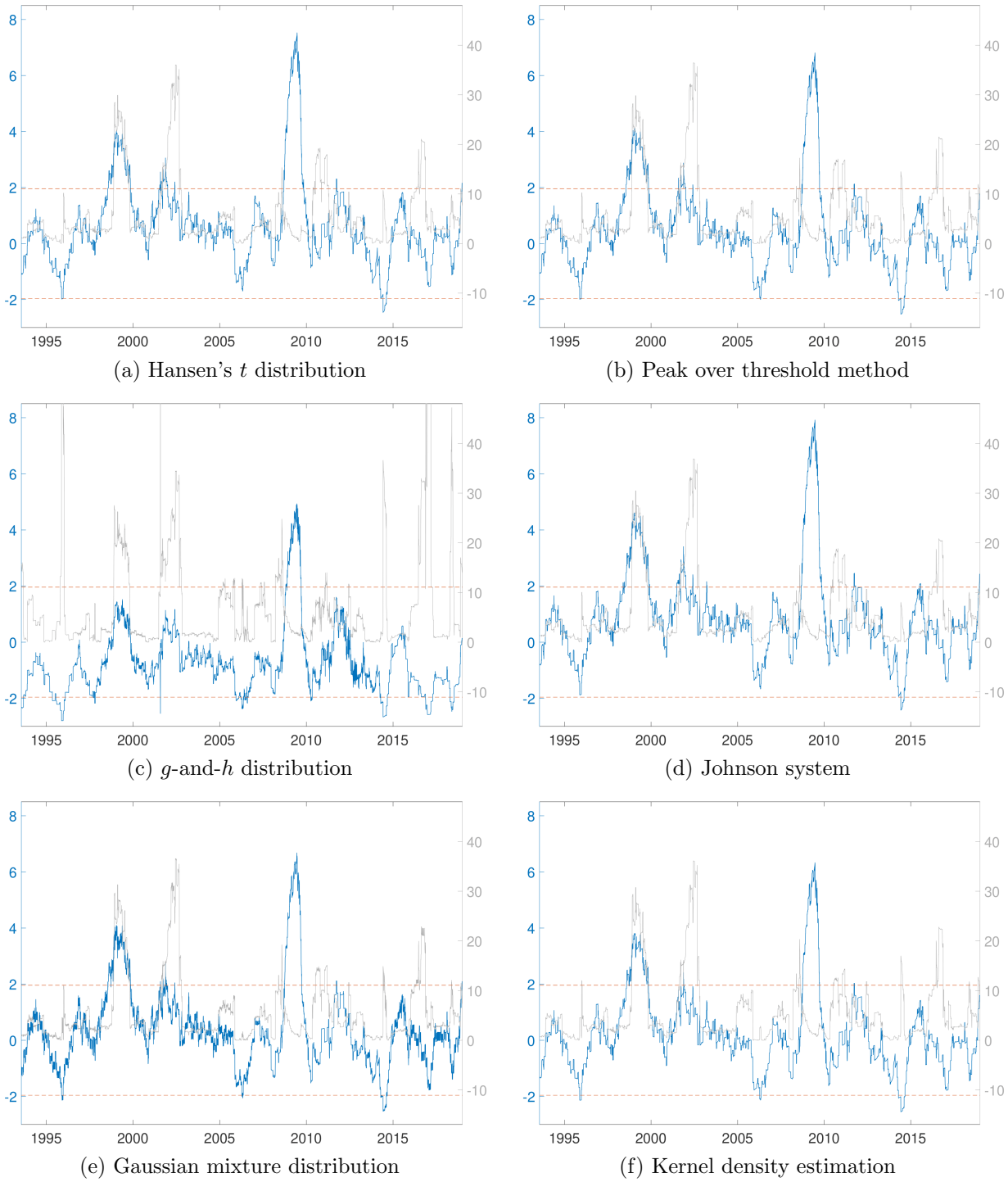
that models with adequate in-sample fit do not necessarily provide useful out-of-sample results (see Pindyck and Rubinfeld, 1998), we dedicate this section to backtesting our ES estimators.

Because we are interested in whether the predictive ability of our ES estimators changes over time, we perform a rolling-window analysis, where a time window is moved in steps of one day from the beginning to the end of our sample. This window consists of an in-sample period of size $T = 2,500$ and an out-of-sample period of size $n = 250$, which is in line with the backtest sample size recommendation of Du and Escanciano (2017). It is also consistent with the minimum requirements for reliable GARCH estimation (see Hwang and Valls Pereira, 2006) and the evaluation periods typically set by regulators (see Argyropoulos and Panopoulou, 2019). Hence, our first out-of-sample period spans from August 10, 1992 to July 23, 1993 and its backtest results are assigned to the date July 23, 1993. For the following windows, we use similar mapping. Within the in-sample period of each window, we re-estimate the models of Sections 2.2.1 and 2.2.2, i.e., the parameters of our AR(1)-GARCH(1,1) model and the parameters of our different distribution models.[32] After this, we derive the standardized losses of the out-of-sample-period and use them in conjunction with the in-sample estimated cdfs to conduct the backtests of Section 2.2.3. Following Du and Escanciano (2017), we document the results of the unconditional and conditional backtests for a significance level $\bar{\alpha} = 0.05$ and $m = 5$ lags.

Figure 2.4, which captures a total of 6,637 out-of-sample periods between August 10, 1992 and December 31, 2018, presents the backtest statistics for the S&P GSCI. It is also supplemented by dashed lines representing the rejection regions for $\bar{\alpha} = 0.05$. The conditional backtest rejects an ES estimator if its test statistic is above $\chi^2_{1-\bar{\alpha},5} \approx 12.07$, whereas the unconditional backtest rejects if its test statistic falls outside $[-\Phi^{-1}(1-\frac{\bar{\alpha}}{2}), \Phi^{-1}(1-\frac{\bar{\alpha}}{2})] \approx [-1.96, 1.96]$. A large unconditional statistic above (below) the upper (lower) rejection boundary indicates that an estimation technique tends to underpredict (overpredict) the ES in the corresponding out-of-sample period.

[32] With a focus on the S&P GSCI, Figure B.3 plots the Ljung-Box statistics of the standardized residuals and the squared standardized residuals over time. The results for the other indices are available upon request. In general, we detect few instances of rejection, suggesting only occasional switching to models of higher order.

Figure 2.4.: Backtest statistics for S&P GSCI, $\alpha = 0.05$



(a) Hansen's t distribution

(b) Peak over threshold method

(c) g-and-h distribution

(d) Johnson system

(e) Gaussian mixture distribution

(f) Kernel density estimation

For the S&P GSCI and our different ES estimation methods, this figure plots the test statistics of the Du and Escanciano (2017) unconditional (blue, left axis) and conditional (gray, right axis) backtests resulting in a rolling-window approach with in-sample size $T = 2,500$, out-of-sample size of $n = 250$ and step size of one day. The coverage level is $\alpha = 0.05$ and the number of lags in the conditional test is set to $m = 5$. The test statistics, which are assigned to the end date of each out-of-sample period, span from July 23, 1993 to December 31, 2018. Dashed horizonal lines mark the backtest rejection areas for a significance level of $\bar{\alpha} = 0.05$. While, for the unconditional backtest, the normal quantiles of the upper and lower line are relevant, the conditional backtest can be conducted with the upper line, which, with reference to the right axis, also resembles the required critical chi-square value.

The statistics in Figure 2.4 reveal two important findings. First, none of our popular estimators was capable of delivering adequate ES predictions during the global financial crisis. The unconditional backtest rejects all distributional specifications. Even a continuous updating of the estimators with fresh data, which is implemented in our rolling-window approach and standard in practice (see Ardia and Hoogerheide, 2014), does not prevent this failure. This result is striking because it calls for abandoning the established standard methods and searching for risk forecasting techniques more effective in extremely turbulent market phases. Second, reasons for estimator rejection are quite different. For example, in the one-year period after September 11, 2001, the conditional null hypothesis $\mathbb{E}(H_t(\alpha) - \frac{\alpha}{2}|\Omega_{t-1}) = 0$ is rejected for all distributions, while rejections of the unconditional null hypothesis $\mathbb{E}(H_t(\alpha) - \frac{\alpha}{2}) = 0$ occur less frequently or, as in the case of the g-and-h and kernel density settings, (almost) do not arise at all. This indicates that the models are inadequate mainly because of autocorrelated cumulative violations. Similar situations can be observed between June 2010 and April 2011 as well as from August to November 2016. In contrast, from October 2008 to September 2009, the conditional backtest does not reject, but the unconditional one does.[33] In this case, the mean of cumulated violations tends to be the main reason for rejection.

Turning to the differences between ES estimators, we can see that, on the one hand, the unconditional test statistics for the g-and-h distribution are generally lower than for the other distributions. Consequently, when rejections of this distribution model occur, they are often related to an overestimation of risk, whereas other models tend to be rejected for underestimation. On the other hand, the conditional test statistics deliver higher amplitudes for g-and-h than for the other estimators, with the effect that it rejects the model more frequently than the others.

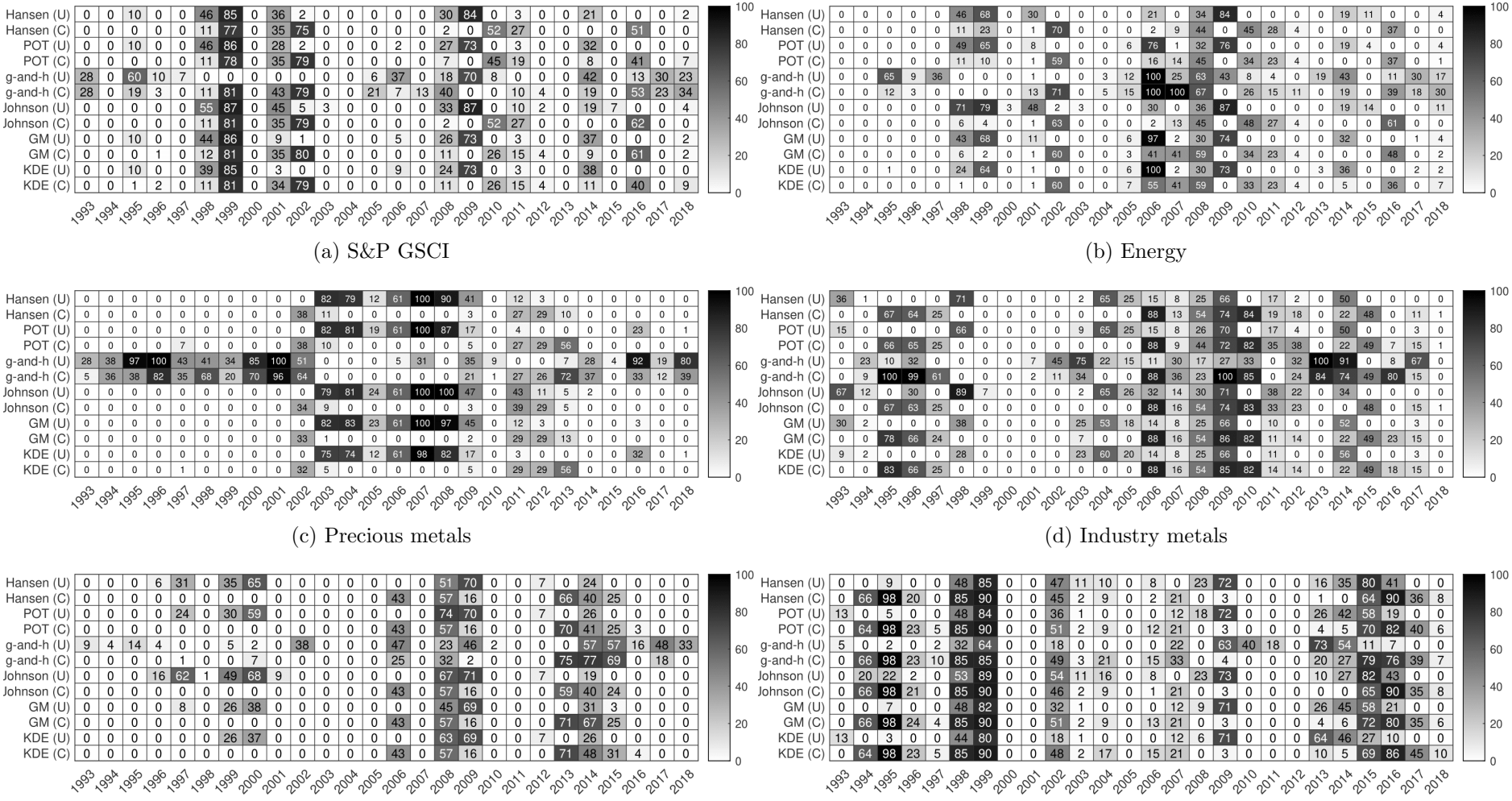
While the g-and-h distribution is conspicuously poor, the test statistics for our alternative distributions appear quite similar in the plots of Figure 2.4, which complicates a detailed comparison. Therefore, we switch to a compact form of visualization that focuses on the resulting test decisions. In this respect, Figure 2.5(a) depicts the percentages of trading days on which the tests of Figure 2.4 reject their null hypotheses each year. In a comparison of Figures 2.2(a) and 2.5(a), we see that all of our ES estimators inaccurately predict the future ES of the S&P GSCI especially when risk temporarily deviates from its long-run level. This holds not only for the global financial crisis but also for smaller shocks. In other words, the estimators fail in turbulent times when accurate forecasts are most needed. In contrast, they all appear reasonable in calm market phases. In addition, Figure 2.5(a) illustrates that a general estimator (the kernel method) is not necessarily always superior to a more restrictive one (the Hansen distribution). For example, shortly after the millennium, more unconditional rejections occur for the latter than for the former. In contrast, the conditional backtests of 2014 show no discrediting evidence against the latter but against the former.

After our focus on the backtest results for the overall S&P GSCI, we now look at its sub-indices. To this end, we again opt for compact summaries of annual rejection percentages and present them in Figures 2.5(b) to 2.5(f).[34] Again, we detect that, regardless of the choice of distribution and during several phases of the global financial crisis 2008-2010, risk is seriously mispredicted in all commodity sectors. Furthermore, extending our previous discussion, the g-and-h distribution attracts attention by overestimation in many sectors. While this is particularly evident for precious metals (see Figure 2.5(c)), the performance of the g-and-h model is closer to the other distributions when focusing attention on livestock (see Figure 2.5(f)). Interestingly, in the precious metals sector, the g-and-h model appears to pass our backtests in several phases where the other estimators fail

[33] Because the conditional test statistic has to pay attention to both autocorrelation and mean of $(H_t(\alpha))$, it does not necessarily bring the same sensitivity to the latter component as the unconditional test which only focuses on the mean of $(H_t(\alpha))$. This may result in a high unconditional statistic, while the conditional statistic is simultaneously damped by an absence of autocorrelation.

[34] The underlying time series of test statistics are available from the authors upon request.

Figure 2.5.: Backtest rejections per year



(a) S&P GSCI

(b) Energy

(c) Precious metals

(d) Industry metals

(e) Agriculture

(f) Livestock

Based on the Du and Escanciano (2017) statistics (and critical values) of Figure 2.4, which evaluate different ES estimators for the S&P GSCI, Subfigure 2.5(a) of this figure uses a heat map design to capture the (rounded) annual percentages of trading days on which the unconditional (U) and conditional (C) backtests reject their null hypotheses. Abbreviations of distribution models are used as in Table 2.2. Subfigures 2.5(b) to 2.5(f) present similar visualizations for the sub-indices of the S&P GSCI. Figure B.4 of the appendix illustrates how rejections cluster within years and whether unconditional rejections are associated with overestimation or underestimation.

and vice versa. Thus, especially in this case, a combination of forecasts may have its merits (see Timmermann, 2006).[35] Finally, another observation is noteworthy. Although livestock can be characterized as the least volatile sector, we see a significant number of rejections from 1998 to 2000. In comparison, the more volatile agriculture sector (see Figure 2.5(e)) shows fewer rejections. Here, rejection periods are mainly around 2000 and from August 2008 to September 2009 (unconditional test) as well as in summer 2006 and from May 2013 to September 2014 (conditional test).

Even though our results show that all estimators are disappointing, it may be instructive to document which one is the least likely to fail and which sectors have the best forecasts. To answer these questions, we follow Brandolini and Colucci (2012) by counting the total number of backtest rejections and present them in Table 2.3.[36] Starting with the latter question, we find that the lowest (highest) numbers of unconditional rejections can be found for agriculture (precious metals). In contrast, conditional rejections are lowest (highest) for precious metals (livestock). With respect to a ranking (from best to worst) of our estimators for given sectors, the unconditional test results of Table 2.3 tell us that the kernel density estimation is the best method for the S&P GSCI and the subsectors precious metals, industry metals and livestock. Hansen's skewed t model and the Gaussian mixture rank first for energy and agriculture, respectively. In contrast, the g-and-h distribution scores last for all indices but the S&P GSCI and livestock. For the latter two, the Johnson system has the highest number of rejections. In the conditional tests, the g-and-h model consistently ranks last. The kernel method (Hansen's skewed t) shines for the S&P GSCI (industry metals and livestock). Finally, the Gaussian mixture, the Johnson system and the POT method rank highest for precious metals, agriculture and energy, respectively.

Trusting both backtests, we can derive an aggregate estimator ranking which looks for moderate rejection numbers in both tests and across commodity types. In this respect, with the lowest mean of summarized rejection numbers, the kernel density method comes first. It is followed by Hansen's skewed t distribution, which delivers quite similar average rejection numbers. The Gaussian mixture and the POT method rank in the middle. The former slightly outperforms the latter because it compensates for its worse conditional backtest performance via fewer significant risk underestimations. The Johnson distribution comes in fifth, whereas g-and-h distributed innovations are the least reliable assumption.

2.4.3. Robustness checks

To investigate whether some specific choices in our research design influence our overall conclusions, we performed several robustness checks with respect to the used time series (or filter) model, the sample sizes, the backtest specification and looked at another often-mentioned estimator.[37]

Filtering In accordance with the literature, our main approach used an AR(1)-GARCH(1,1) model whose parameters were estimated via QML with normal distribution assumption. We modify this in two ways. First, because QML preserves its key features under other simple distributions – provided that they are unimodal and symmetric around the origin (see Newey and Steigerwald, 1997; Klar et al., 2012) – we additionally perform QML with a Student t distribution. However, we observe no notable differences in the rejection behavior of our test statistics. On average, there is a slight decrease of unconditional backtest rejections and a small increase of conditional backtest

[35] Unfortunately, our backtests cannot straightforwardly check this statement. In contrast to ES estimation, which is based on inverse distribution functions, the tests require the specification of a cdf. Because mapping a cdf onto its inverse function is a nonlinear transformation, backtest computations based on a mean of several cdfs would not generally correspond to the mean ES of the individual estimators.

[36] We could also use loss functions (see Le, 2020) or comparative backtests (see Nolde and Ziegel, 2017). However, this would introduce arbitrariness and lead to an inconsistent abandonment of our chosen backtesting framework.

[37] Because of space considerations, we largely concentrate on describing their main idea and outcome. Detailed results are available upon request.

Table 2.3.: Total number of backtest rejections

	Hansen's t		POT		g-and-h		Johnson		GM		KDE	
Panel A: Unconditional backtest												
S&P GSCI	$^{751}_{80}$	831	$^{693}_{112}$	805	$^{230}_{644}$	874	$^{882}_{50}$	**932**	$^{640}_{133}$	773	$^{591}_{147}$	***738***
Energy	$^{688}_{140}$	***828***	$^{566}_{319}$	885	$^{113}_{1165}$	**1278**	$^{888}_{162}$	1050	$^{547}_{418}$	965	$^{454}_{444}$	898
Prec. metals	$^{1254}_{0}$	1254	$^{1179}_{64}$	1243	$^{92}_{2281}$	**2373**	$^{1444}_{0}$	1444	$^{1323}_{8}$	1331	$^{1099}_{86}$	***1185***
Ind. metals	$^{811}_{135}$	946	$^{787}_{160}$	947	$^{89}_{1601}$	**1690**	$^{1211}_{93}$	1304	$^{647}_{210}$	857	$^{614}_{217}$	***831***
Agriculture	$^{692}_{62}$	754	$^{688}_{68}$	756	$^{194}_{851}$	**1045**	$^{914}_{49}$	963	$^{485}_{90}$	***575***	$^{527}_{68}$	595
Livestock	$^{1132}_{134}$	1266	$^{891}_{224}$	1115	$^{514}_{551}$	1065	$^{1294}_{97}$	**1391**	$^{858}_{218}$	1076	$^{683}_{332}$	***1015***
Panel B: Conditional backtest												
S&P GSCI		863		859		**1232**		915		878		***847***
Energy		717		***669***		**1422**		714		844		867
Prec. metals		308		451		**2033**		310		***282***		413
Ind. metals		***1530***		1610		**2538**		1538		1654		1648
Agriculture		643		666		**806**		***624***		729		706
Livestock		***1665***		1746		**1925**		1667		1747		1819
Panel C: Backtest summary												
Mean		967		979		**1523**		1071		976		***964***
Std		390		***388***		**582**		408		424		410

Row-wise summarizing the results of Figures 2.5 and B.4, for each commodity index and ES estimation method, Panel A of this table reports the total number of unconditional backtest rejections (regular font). Additionally, these numbers are split into rejections related to underestimation (small upper-case number) and overestimation (small lower-case number) of risk. Panel B presents the rejection numbers for the conditional backtest. Panel C aggregates Panels A and B by reporting the (rounded) mean and standard deviation of the rejections across test types and commodity indices. Abbreviations are used as in Table 2.2. For each index, the lowest total rejection number is marked in italics and the highest in bold.

numbers. Second, we implement some GARCH alternatives. Even though our specification tests indicate no misspecified volatility dynamics in the plain vanilla GARCH setting, we also use a TGARCH (or GJR-GARCH) model (as in Hammoudeh et al., 2014) and an EGARCH model (as in Del Brio et al., 2017) because such models have been found to do quite well in predicting extreme events, as required for accurate ES forecasts (see Trapin, 2018). However, we do not detect any forecast improvements or changes in estimator ranking. On the contrary, the former model led to a lot of additional backtest rejections between 2015 and 2018.[38]

Sample sizes While our in-sample size $T = 2,500$ followed the recommendation of Du and Escanciano (2017), we also look at $T = 2,000$ and $T = 3,000$. In general, we observe lower peaks of the conditional backtest statistics for larger sample sizes (especially during the global financial crisis). This demonstrates that, in our application, a more extensive sample for parameter estimation should be preferred to a small sample that places more emphasis on recent information. However, it does not change our finding that all estimators perform inadequately. Furthermore, we detect almost no changes in estimator rankings based on the unconditional backtest and find that Hansen's skewed t distribution improves more significantly with rising sample size than other

[38] This can be related to the fact that TGARCH residuals typically exhibit a non-iid dependence structure that can best be described either non-parametrically or by means of copulae (see Beckers et al., 2017).

parametric models. In contrast, rankings derived from the conditional backtests are slightly affected. In the case of $T = 3,000$, we observe quite similar rejections numbers. For example, when looking at the overall S&P GSCI, most distributions exhibit rejection numbers ranging within a small interval with absolute width of 45. Consequently, changes in ranking do not necessarily indicate remarkable differences in predictive power. When $T = 2,000$ is chosen, there are no red flags.

Backtest specification Since a conditional backtest decision depends on the number m of included autocorrelation coefficients, our evaluation may also be influenced by our choice of $m = 5$. Therefore, we follow Escanciano and Lobato (2009) and Novales and Garcia-Jorcano (2019) by also considering $m = 1$ and $m = 10$. While in the latter case the Johnson system can improve its positions in our estimator rankings, all other techniques hold their relative order in ranking.

Normal distribution The assumption of conditional normality leads to a very simple cdf and a quite compact ES estimation formula (see Frey and McNeil, 2002, chpt. 2.2.4). Thus, risk managers may be tempted to use it in practice. To elaborate on the consequences of such an action, we also analyzed this estimator. As far as the ES of standardized losses (extending Table 2.2) is concerned, we obtain around 2 %. That is, not surprisingly, assuming a normal distribution leads to the lowest risk estimates across all approaches. The unconditional backtest (extending Panel A of Table 2.3) rejects the model in 1093 (1046, 47), 1207 (1045, 162), 1395 (1395, 0), 1203 (1123, 80), 981 (924, 57) and 1668 (1606, 62) cases. The conditional procedure (extending Panel B of Table 2.3) does so 913, 740, 311, 1548, 640 and 1688 times. Thus, even though it can compete with some distributions for certain subindices, it is inferior to the highscore models of our main analysis.

2.5. Conclusion

Motivated by (i) increased trading activity in commodity futures markets, (ii) emerging efforts to replace the established risk measure VaR by ES and (iii) a breakthrough with respect to backtesting procedures for the evaluation of competing ES estimators, we dedicated this study to a full-scale analysis of the dynamic features of ES in commodity futures markets.

Focusing on six parametric and non-parametric estimators (with the same time series setup but that differ with respect to the continuous loss distribution), which are particularly popular in academia and practice, we started with a documentation of historic risk levels in the overall commodity market and its most important subsectors. We identified and discussed risk peaks in several recessions, bear stock markets and commodity-specific shock environments. Furthermore, we showed that investment risks tend to be highest in the energy sector and lowest in livestock.

In an out-of-sample evaluation of the predictive accuracy for future ES, the unconditional and conditional backtests of Du and Escanciano (2017) revealed serious limitations of the estimators. While most estimators tend to underpredict investment risk, the estimator using the flexible g-and-h distribution seriously overshoots. Hansen's skewed t, the POT extreme value approach and the Gaussian mixture mispredict less frequently than the g-and-h model. The Johnson system performs somewhere between the former and the latter group of estimators and a kernel density approach emerges as the most useful choice on average. Across all commodity sectors, rejection numbers are lowest for agriculture. Nonetheless, even the better distribution models fail with respect to unconditional and/or conditional coverage when it comes to forecasting the ES in turbulent market phases. In summary, all of them can mislead investors in their evaluation of risky futures positions and cause exchanges to set inadequate margin levels.

Where do these results leave us? First, they advise against trusting the standard estimators currently used in many studies unless we can be certain that smooth markets lie ahead. Second,

they call for additional research because our study naturally cannot cover all potential alternative estimation models. There is a variety of additional techniques which do not require a full specification of the conditional distribution of the data and may thus reduce misestimation risk (rejections) below the levels observed for our best models (see Cotter and Dowd, 2010; Escanciano and Mayoral, 2009; Le, 2020). Unfortunately, before such alternatives can be tested within a framework of the Du and Escanciano (2017) type, it needs to be enhanced beyond a setup bound to invertible distribution functions. Furthermore, it may be insightful to extend or abandon the established simple AR-GARCH time series setting because, even though it cannot be rejected in standard diagnostic checking, other models might provide better forecasts of future means and volatilities. In this context, we suggest three possible endeavors. (a) We might incorporate time-varying skewness and kurtosis, which has recently been tried in asset allocation applications involving commodity-based portfolio components (see Gao and Nardari, 2018). Even though some stock market studies have found that these moments do not necessarily vary significantly over time (see Bali et al., 2008; Ergün and Jun, 2010), preliminary results in commodity futures markets suggest otherwise (see Fernandez-Perez et al., 2018a). (b) We could turn our back on the GARCH world and use alternative processes (such as diffusion and stochastic volatility models) instead (see Chen and Tang, 2005). (c) Leaving the narrative of pure time series models by adding exogenous predictors (such as recession or stock market indicators or inventory variables) might also be fruitful (see Ye et al., 2005; Peracchi and Tanase, 2008). However, because of a wide variety of potential forecasting variables, the curse of dimensionality has to be faced in such considerations (see Guidolin and Pedio, 2020).

Part III

Motivation

In Chapter 2, we analyzed the risks that several passive commodity futures investments have offered to investors in recent decades. In this context, we have seen that the common distribution models that we used for ES forecasting tend to work insufficiently when markets are in turmoil. Thus, market participants that invest in passive commodity futures strategies risk overlooking or misjudging chancy market phases in spite of constantly assessing risks with established mean- and volatility-adjusted ES estimators. This motivates us to study whether active investments strategies in commodity futures markets also suffer from the same problems as passive.

In the following chapter, we turn to the analysis of several active commodity investment strategies, to which we attach importance to their risks and returns, especially. First, we investigate the past performance of traditional cross-sectional momentum strategies, which suggest to buy (sell) the recently best- (worst-)performing commodity futures indices. Next, we study whether/ to what extent an additional incorporation of autocorrelation measures helps to deal with sways in the market. For this, we design memory-enhanced momentum strategies to construct portfolios that continue investing with established past-focusing momentum strategies when stable market phases are indicated by a memory measure and stop it otherwise. In this context, we concentrate on measures of both, short and long memory by first incorporating variances ratios (according to Lo and MacKinlay, 1988) and second, Hurst coefficients (based on Hurst, 1951). Besides, we analyze the extent to which strategy performances can be improved for volatile market phases by allowing for (short-term) reversal.

3. Memory-enhanced momentum in commodity futures markets

Abstract: Motivated by the deteriorating performance of traditional cross-sectional momentum strategies in commodity futures markets, we propose to resurrect momentum by incorporating autocorrelation information into the asset selection process. Put differently, we introduce measures of short and long memory (variance ratios and Hurst coefficients, respectively) telling us whether past winners and losers are likely to persist or not. Our empirical findings suggest that a memory-enhanced momentum strategy based on variance ratios significantly outperforms traditional momentum in terms of reward and risk, effectively prevents momentum crashes and is not bound to the movement of the overall commodity market. Furthermore, strategy returns cannot be explained by typical factor portfolios and macroeconomic variables and are robust to various parametrization choices, alternative data sets, transaction costs and data snooping. Finally and in contrast to a newly emerging strand of literature promoting the benefits of long memory measures in portfolio management, we show that Hurst coefficients do not carry investment-relevant information in a commodity momentum context.

3.1. Introduction

After the seminal work of Jegadeesh and Titman (1993, 2001), showing that stocks tend to continue their past performance in the near future, the striking effectiveness of cross-sectional momentum investment strategies has been verified in equity markets all over the world (see Fama and French, 2012; Asness et al., 2013). In addition, research activity has quickly spilled over to commodity futures markets because, here, easier shorting, high liquidity, negligible transaction costs and an overseeable asset universe make the strategies particularly successful (see Miffre and Rallis, 2007; Shen et al., 2007; Szakmary et al., 2010; Fuertes et al., 2010, 2015; Bianchi et al., 2015, 2016).[1]

Unfortunately, some recent studies cast doubts on the future usefulness of momentum strategies by indicating that stock momentum profits have significantly declined or even completely vanished in the last decade (see Chordia et al., 2014; Mao and Wei, 2014; Hwang and Rubesam, 2015). In a preliminary examination, we show that similar results can be obtained for commodity momentum. That is, the significant multi-factor alphas of previous studies almost completely disappear using an up-to-date sample. But does this mean that the momentum effect in commodities should be declared dead, as has been done by many for the size effect in equities (see van Dijk, 2011)?

Asness et al. (2018) emphasize that the size effect can be resurrected if size is adjusted for firm quality (as captured by the junk metric of Asness et al., 2019). In a similar vein, Asness and Frazzini (2013) show that value portfolios can be improved by using a more adequate measure of value (a book-to-market ratio which, in contrast to the traditional one, uses more timely prices). Motivated by these approaches, we propose a refined momentum selection evaluating potential performance continuation in a more sophisticated way than traditional momentum strategies. Our

[1] Miffre (2016) provides an excellent review of long-short commodity investing. Bond and foreign exchange markets have been excessively studied as well (see Gebhardt et al., 2005; Serban, 2010).

proposal originates from a closer look at the sources of momentum profits. In theoretical models, empirical setups and simulation studies, Lewellen (2002), Pan et al. (2004) and Hong and Satchell (2015) illustrate that, ceteris paribus, cross-sectional (and time series) momentum is particularly strong if asset returns exhibit significant positive autocorrelation.[2] Without such autocorrelation, momentum profits tend to be lower or nonexistent. Therefore, we argue that it can be suboptimal to base commodity momentum strategies on past performance (i.e., the cumulative returns of the most recent months) only.[3] Traditionally, a momentum strategy buys past winners and sells past losers. If they are positively autocorrelated (persistent), it is likely that they continue their past performance in the near future and the strategy succeeds. However, if there are phases where they are uncorrelated or negatively autocorrelated (anti-persistent), performance is less likely to continue or might even reverse such that the strategy is in danger of failure. Consequently, an effective trading strategy should dynamically adjust to time-varying autocorrelation conditions in the market (see DeMiguel et al., 2014) instead of just clinging to an established selection rule which, depending on the current level of autocorrelation, may be misleading.

To take into account the empirical evidence on serial dependence in commodity futures markets (see Kamara, 1984; Kristoufek and Vosvrda, 2014) and to limit the risk of entering disadvantageous investment positions, we suggest using both past performance and autocorrelation to determine the long and short positions in commodity momentum strategies. To compactly capture the latter, we use the well-known variance ratio (VR) of Lo and MacKinlay (1988).[4] Because, if adequately specified, it represents a linear combination of low-order autocorrelation coefficients, the VR measures the intensity of short-term autocorrelation or short memory and, in contrast to other aggregates, like the Ljung and Box (1978) statistic with quadratic combination, allows a differentiation between persistence ($VR > 1$) and anti-persistence ($VR < 1$). Using past returns and variance ratios, our bivariate strategy consists of taking long positions into persistent winners and anti-persistent losers as well as short positions into persistent losers and anti-persistent winners. Variance ratio significance is evaluated via established statistical procedures such that we do not trade based on potentially random signals but only on significant ones. Furthermore, even though earlier studies indicate that (short-term) reversal is less relevant in commodity futures markets (see Miffre and Rallis, 2007; Shen et al., 2007), our strategy design does not generally rule out its existence because, especially in commodity markets, return behavior is subject to significant variation (see Adams and Glück, 2015). In other words, we not only enhance traditional momentum by refining the signals for momentum trades but also allow the momentum strategy to consider (short-term) reversals if significant negative autocorrelation suggests doing so.

We find that tactically allocating wealth based on our strategy, which we call memory-enhanced momentum (MEM), generates economically and statistically significant profits. It outperforms traditional momentum in various ways. First, the best specifications (resulting from short holding and ranking periods combined with first-order autocorrelation) earn alphas of about 2 % per month and exhibit notably lower tail risk. The latter observation is related to the fact that, especially in phases where traditional momentum suffers its worst losses, MEM enters more suitable positions. Second, strategy returns are not significantly linked to the overall commodity market. That is, the profits are not just the result of strong upward trends observable in the recent history of some commodity market sectors. Finally, while the short leg of traditional momentum significantly weakens its investment outcome, the short leg of MEM is particularly strong.

Overall, MEM performs well in a wide variety of settings. Even without the short leg and the trades related to detected anti-persistence, it is better than traditional momentum. Furthermore, it

[2] For earlier work in this area, see Lo and MacKinlay (1990) and Conrad and Kaul (1998).

[3] Rachev et al. (2007), Zaremba et al. (2021) and Chen et al. (2021) alternatively capture past performance via reward-to-risk ratios, alphas as well as outlier-robust rank and sign measures, respectively.

[4] We might alternatively think of using autoregressive models (see Gaunt and Gray, 2003; DeMiguel et al., 2014). However, our intention is to make the practical implementation of our strategy as simple as possible.

is robust to using alternative futures data sets and survives both transaction costs and established data mining tests. Last, but not least, the returns of MEM cannot be explained by a large set of (stock, bond and commodity) factor portfolios and classic macroeconomic variables. This stability and independence makes the strategy a valuable tool for commodity market investors.

Besides using variance ratios, we investigate the potential of another aggregate measure of autocorrelation: the Hurst coefficient (HC). It captures the level and behavior of low- and high-order autocorrelation and is thus typically referred to as a measure of long memory. In recent research, estimates of the HC have been suggested to contain valuable information for portfolio managers (see De Souza and Gokcan, 2004; Clark, 2005; Batten et al., 2013; Ramos-Requena et al., 2017; López-García et al., 2021). For example, while De Souza and Gokcan (2004) and Batten et al. (2013) point out their worth for hedge fund selection and gold market timing, respectively, López-García et al. (2021) indicate that they can be used for the construction of profitable long-short portfolios (and asset pricing factors) in equity markets. We extend this literature by investigating whether using the HC can be beneficial within a memory-enhanced commodity momentum strategy. Again, the measure allows an identification of persistence ($HC > 0.5$) and anti-persistence ($HC < 0.5$). We obtain the HC by averaging the popular estimators of Hurst (1951), Higuchi (1988), Barabási and Vicsek (1991) and Peng et al. (1994) and, because the HC literature is not yet as advanced as the VR literature, use a heuristic decision rule do determine relevant levels of persistence or anti-persistence. The general design of the HC investment strategy then follows the same principles as the proposed VR strategy.

Interestingly, we find that MEM investing based on the HC does not perform particularly well. This also holds when using an extended set of HC estimators and when modifying the employed decision rule. Even if we explicitly allowed for data snooping, we would not be able to generate persuasive outcomes. Consequently, we have to conclude that the HC does not contain information valuable for improving momentum strategies in commodity futures markets. Despite the growing popularity of the HC, its investment worth is far behind the one of the VR.

Our study is organized as follows. Section 3.2 briefly introduces our commodity data set. Section 3.3 describes our approach to portfolio construction. Section 3.4 presents our empirical results which we subdivide into traditional (univariate) momentum, memory-enhanced (bivariate) momentum and robustness checks. Section 3.5 concludes and outlines directions for future research.

3.2. Data

Following Bianchi et al. (2015), our study captures investments in individual commodity futures via the subindices of the well-known Standard and Poor's Goldman Sachs Commodity Index (S&P GSCI) and additional indices published by S&P (but not included in the S&P GSCI). Using the continuous price series of S&P has several advantages over self-compiled series based on raw futures contracts. First, the indices are much more accessible because they are established commodity market benchmarks reflecting the real returns (from investing in fully collateralized futures nearest to expiration) available to large market participants.[5] They simultaneously generate a focus on the most liquid futures. Second, while the immediate rollover implemented in many studies (see Miffre and Rallis, 2007; Shen et al., 2007; Fuertes et al., 2010) is impractical for investors with large positions because it would result in adverse price impact, the gradual rollover used in the indices absorbs such an impact. Finally, individual futures contracts are difficult to manage because they are traded across different exchanges. The pre-compiled indices are uniform and additionally make our results easier to replicate.

[5] For a discussion of the implications of full collateralization, see Gorton and Rouwenhorst (2006), Fuertes et al. (2010), Bianchi et al. (2016) and the review of Woodard et al. (2011).

3.2. Data

Our sample consists of commodities from six sectors: energy (Brent and WTI crude oil, gas oil, heating oil, natural gas, petroleum, unleaded gasoline), precious metals (gold, platinum, silver), industry metals (aluminum, copper, lead, nickel, tin, zinc), agriculture (cocoa, coffee, corn, cotton, soybeans, soybean oil, sugar, Chicago and Kansas wheat) and livestock (feeder cattle, lean hogs, life cattle).[6] The corresponding price data, which we obtained from Thomson Reuters Datastream, spans from December 1970 to December 2019. As can be seen in Table 3.1, in the early years of our sample period, investors only had access to four traded commodity indices (corn, soybeans, Chicago wheat, live cattle). Over time, the investment opportunity set then successively increased to 28 indices with the latest addition (tin) in April 2007.

Table 3.1.: Descriptive statistics

	Mean	Volatility	SR	VaR	ES	Min	Max	Pos. mths	Inception
S&P GSCI	0.54	5.72	0.03	9.25	12.41	-28.20	23.83	56.19	1970
Energy									
Crude oil (Brent)	1.14	8.96	0.11	14.23	19.03	-33.75	36.56	57.77	1999
Crude oil (WTI)	0.85	9.48	0.06	13.38	18.97	-32.43	48.89	54.68	1987
Gas oil	1.13	9.04	0.11	13.76	19.01	-30.93	31.19	54.58	1999
Heating oil	0.91	9.02	0.07	13.65	17.74	-28.86	37.60	54.18	1983
Natural gas	-1.17	14.22	-0.10	24.46	29.31	-37.63	53.08	44.69	1994
Petroleum	0.95	8.90	0.07	12.64	17.72	-32.75	37.73	53.83	1983
Unleaded gasoline	1.34	9.75	0.11	13.09	18.64	-39.52	49.46	57.44	1988
Precious metals									
Gold	0.56	5.46	0.04	6.73	10.57	-20.41	28.23	50.40	1978
Platinum	0.66	6.34	0.06	9.04	13.47	-31.24	34.64	53.94	1984
Silver	0.68	9.28	0.03	11.46	18.30	-46.87	55.91	49.01	1973
Industry metals									
Aluminum	-0.03	5.42	-0.04	8.01	11.25	-16.76	15.92	44.67	1991
Copper	0.98	7.50	0.08	9.30	14.56	-35.55	38.43	52.23	1977
Lead	0.71	7.99	0.07	10.78	16.85	-27.43	27.03	53.85	1995
Nickel	0.89	9.82	0.07	13.18	18.91	-27.48	35.16	50.46	1993
Tin	0.55	7.47	0.07	12.80	14.70	-21.55	26.75	50.33	2007
Zinc	0.33	7.07	0.02	9.53	14.68	-34.17	28.06	48.13	1991
Agriculture									
Cocoa	0.05	8.14	-0.03	12.14	15.20	-24.94	35.22	47.56	1984
Coffee	0.34	10.48	0.00	13.32	18.19	-30.89	54.24	46.04	1981
Corn	0.01	7.29	-0.05	11.63	14.83	-22.80	46.55	45.78	1970
Cotton	0.43	6.92	0.01	10.52	14.27	-22.58	27.52	52.82	1977
Soybeans	0.55	7.42	0.02	10.94	15.07	-21.98	56.64	50.81	1970
Soybean oil	0.18	6.80	0.01	10.36	15.43	-25.10	26.68	46.93	2005
Sugar	0.54	11.37	0.01	15.27	19.54	-29.69	68.63	49.37	1973
Wheat (Chicago)	0.01	7.84	-0.05	11.63	15.59	-25.27	42.40	48.65	1970
Wheat (Kansas)	-0.28	8.32	-0.05	12.43	16.38	-23.72	36.01	44.62	1999
Livestock									
Feeder cattle	0.28	4.74	0.04	7.95	10.47	-16.19	15.06	53.95	2002
Lean hogs	0.30	7.35	-0.01	11.35	14.77	-25.87	24.84	51.80	1976
Live cattle	0.58	4.97	0.04	7.93	10.75	-21.02	22.24	54.40	1970

Covering our strategy evaluation period from August 1973 to December 2019, this table presents some descriptive statistics for the monthly percentage returns of S&P indices designed for the aggregate commodity market (S&P GSCI) and individual commodities (grouped by sector). Besides mean, standard deviation (volatility) and Sharpe ratio (SR), we report the empirical 95 % value at risk (VaR) and expected shortfall (ES), the minimum and maximum of returns as well as the percentage of positive months. We also state the inception year of each index.

In addition to data availability information, Table 3.1 reports some descriptive statistics for the monthly percentage returns of the S&P GSCI and the individual commodity indices. To be consistent with our main strategy evaluation, these statistics concentrate on the period from August 1973 to December 2019. Besides mean, standard deviation (volatility) and Sharpe ratio (SR), we compute value at risk (VaR) and expected shortfall (ES), the minimum and maximum

[6] RBOB gasoline replaced unleaded gasoline in October 2006 (see https://www.goldmansachs.com/what-we-do/global-markets/business-groups/sts-folder/gsci/components-weights-index-levels.html). The commodities petroleum, platinum, tin and soybean oil are currently not constituents of the S&P GSCI.

of returns as well as the percentage of positive months. SR is the ratio of mean excess returns (over the risk-free rate proxied by the 1-month TBill rate) to volatility (see Sharpe, 1994).[7] The estimation of VaR and ES is based on the empirical 95 % quantile (see Frey and McNeil, 2002).

Most commodities earn positive mean returns. In the energy sector, i.e., for unleaded gasoline, Brent crude oil and gas oil, we can even observe mean returns greater than 1 % per month. These three commodities simultaneously exhibit the highest Sharpe ratios and are thus the most interesting standalone investments. In contrast, many instances of negative Sharpe ratios make the agricultural sector the least relevant. Turning to our downside risk measures VaR and ES, natural gas shows the highest values of about 25 % and 30 %, respectively, across all commodities. Interestingly, the lowest risk can be detected for gold (followed by feeder cattle and live cattle). Finally, for many individual indices, we see that the maximum returns are (in absolute terms) larger than the minimum returns. Furthermore, most sectors exhibit more positive than negative returns. In this respect, agriculture appears to be an outstanding exception.

3.3. Methodology

3.3.1. Traditional momentum

Our construction of traditional commodity momentum portfolios is guided by Miffre and Rallis (2007). That is, at the end of each month and based on the cumulative returns of the previous R months (ranking period), we sort our commodities into quintiles.[8] We then form equally weighted winner and loser portfolios using the commodities in the top and bottom quintiles, respectively, and monitor their returns over the subsequent H months (holding period). The corresponding R-H momentum strategy involves buying the winner and selling the loser portfolios.

For the ranking and holding period lengths, we consider 1, 3, 6, 9 and 12 month(s). This leads to a total of 25 strategies.[9] If, for a given R, we have $H > 1$, the holding periods of constructed portfolios naturally overlap. In these cases, we follow Fuertes et al. (2010) by computing the monthly portfolio return as the return average of the portfolios formed in the recent H months. Such averaging ensures that neither the strategy initiation month nor omitted momentum updating in the holding period drive portfolio performance.

3.3.2. Memory-enhanced momentum

3.3.2.1. Variance ratios

Our first kind of MEM captures the strength of short-term memory in commodity returns. As straightforward measures for the latter, we might think of using autocorrelation coefficients $\rho_k \in [-1, 1]$ of order $k \in \{1, 2, \ldots, T-1\}$, which can be estimated via

$$\hat{\rho}_k = \frac{1}{(T-1)\hat{\sigma}^2} \sum_{t=k+1}^{T} \left((r_t - \hat{\mu})(r_{t-k} - \hat{\mu})\right), \tag{3.1}$$

where r_t is the return in month t, T the sample size, $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} r_t$ and $\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}(r_t - \hat{\mu})^2$ (see Campbell et al., 1997).[10] They describe the linear dependence between the returns in t and

[7] The risk-free rate is available in Kenneth French's Data Library: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[8] Switching to quartiles or terciles does not change our overall picture.

[9] While momentum studies in equity markets skip one month between ranking and actual investment (see Jegadeesh and Titman, 1993; Asness et al., 2013), this is often not done in commodity markets because it has been shown to diminish strategy returns (see Miffre and Rallis, 2007; Fuertes et al., 2010; Bianchi et al., 2015).

[10] In contrast to our portfolio evaluations, which use simple returns capturing actual investment outcomes, variance ratios and Hurst coefficients require log returns (see Peters, 1992; Charles and Darné, 2009).

$t-k$. If $\hat{\rho}_k > 0$, we expect a return $r_{t-k} > \hat{\mu}$ ($r_{t-k} < \hat{\mu}$) to go along with $r_t > \hat{\mu}$ ($r_t < \hat{\mu}$). In contrast, we would predict that $r_{t-k} - \hat{\mu}$ and $r_t - \hat{\mu}$ have opposite signs when $\hat{\rho}_k < 0$.

Because different lags k may lead to different expectations regarding future return directions and because we are interested in a clear trading rule, we are looking for an aggregate autocorrelation measure. As shown by Lo and MacKinlay (1988), variance ratios can serve this purpose because $VR(q)$ is not only the ratio of the variance of q-period returns and q times the variance of one-period returns but, for $q \geq 2$, can also estimated as

$$\widehat{VR}(q) = 1 + 2\sum_{k=1}^{q-1}\left(1 - \frac{k}{q}\right)\hat{\rho}_k. \tag{3.2}$$

Put differently, it is a linear combination of the first $q-1$ autocorrelation coefficients $\hat{\rho}_k$ which receive weights decreasing with lag size k. If returns are not autocorrelated, we expect to observe $\widehat{VR}(q) \approx 1$. However, if positive (negative) autocorrelation dominates the statistic, we will have $\widehat{VR}(q) > 1$ ($\widehat{VR}(q) < 1$).

To determine whether a variance ratio is significantly different from 1, Lo and MacKinlay (1988, 1989) have developed several statistical tests and intensively studied their small-sample properties. We use their heteroscedasticity-robust version. That is, we base our decisions on the asymptotically standard normal statistic $(\widehat{VR}(q) - 1)/\hat{V}(q)$, where

$$\hat{V}(q) = \sqrt{\sum_{k=1}^{q-1}\left(\left(\frac{2(q-k)}{q}\right)^2 \frac{\sum_{t=k+1}^{T}(r_t - \hat{\mu})^2 (r_{t-k} - \hat{\mu})^2}{\left(\sum_{t=1}^{T}(r_t - \hat{\mu})^2\right)^2}\right)} \tag{3.3}$$

is the estimated asymptotic standard deviation of $\widehat{VR}(q)$.

3.3.2.2. Short memory strategy

The MEM strategy based on variance ratios supplements past performance by additional information indicating whether it can be expected to persist or not. The core of the strategy can be summarized by the following matrix:

$$\begin{array}{c} \\ Winner \\ Loser \end{array} \begin{array}{c} \begin{array}{cc} Persistent & Anti\text{-}persistent \end{array} \\ \begin{bmatrix} Long & Short \\ Short & Long \end{bmatrix} \end{array} \tag{3.4}$$

A winner (loser) is bought (sold) only if our variance ratios tell us that it is persistent. If they indicate anti-persistence, we sell (buy) past winners (losers). Besides the momentum-typical persistence focus, we explicitly include the anti-persistence side because previously documented marginality of (short-term) reversal (see Miffre and Rallis, 2007; Shen et al., 2007) may no longer hold today. If there were phases of anti-persistence, it would not be wise to ignore this investment-relevant information by using a outdated rule suggesting that there is never any reversal.

For the exact strategy specification, several aspects have to be noted. First, in contrast to a focus on relative strength, where a commodity with negative past performance may show up on the winner side because it is less negative than others, we follow Erb and Harvey (2006) by using positive and negative returns to allow a clearer distinction between winners and losers. In this setup, a combination with autocorrelation information is more meaningful. Second, we link q to the ranking period R. That is, we concentrate our analysis on $q \in \{2, 4, 7, 13\}$ and, for each R, set $q \leq R + 1$.[11] With higher levels of q, we would leave our short memory focus. Third, because

[11] Recall from Section 3.3.2.1 that this choice of q-values considers autocorrelations up to 1, 3, 6 and 12 lags.

variance ratios above or below 1 may just be random encounters, we label a commodity persistent or anti-persistent only if the test of Section 3.3.2.1 suggests significant deviations from 1 at a 10 % level. The test is performed based on data covering 32 months. Finally, because there is also some empirical evidence on time-dependent random walk behavior in commodity futures markets (see Sensoy and Hacihasanoglu, 2014), there can be phases where our strategy does not suggest taking a commodity position at all. In these cases, the futures collateral is assumed to be invested risk-free earning the 1-month TBill rate.

3.3.2.3. Hurst coefficients

Originating from the empirical hydrology research of Hurst (1951) and theoretically incorporated into a stationary stochastic process (the fractional Brownian motion, FBM) by Mandelbrot and Van Ness (1968), the Hurst coefficient is one of the best-known measures of long memory. Its interpretation can be nicely illustrated in the context of the latter because, here, $HC = 0.5$ implies that the FBM reduces to a classic Brownian motion with independent increments. In contrast, $HC > 0.5$ introduces a strictly positive autocorrelation function, where the limit sum of the autocorrelation coefficients goes to infinity. For $HC < 0.5$, the autocorrelation is negative for arbitrary chosen time intervals and its coefficient sum tends to zero. Overall, the larger the distance to 0.5, the higher is the absolute autocorrelation for small and large lags.[12]

Various methods have been proposed to estimate HC based on empirical data. Unfortunately, many estimators lack a supplementary distribution theory such that testing hypotheses is often impossible. In addition, Taqqu et al. (1995) and Rea et al. (2013) emphasize that there is no consensus on the most suitable estimator and that the results of available estimators can differ considerably. To take these issues into account, i.e., to limit estimation risk, we do not rely on a single estimator but, in spirit of Liu (2015), use an equally weighted average of four particularly promising ones (see below). To evaluate whether an estimate has relevant magnitude, we derive a simple heuristic decision rule which proxies typical practitioner behavior (see below).

Rescaled range analysis Established by Mandelbrot and Wallis (1968, 1969) and Mandelbrot (1971), rescaled range analysis (RSA) is the most frequently used estimation method. For various τ, it begins with dividing the return sample into $n_\tau = \frac{T}{\tau}$ adjacent subsamples of length τ and forming partial sums $(X_{j,t})_{t=1,\dots,\tau} = \sum_{t=(j-1)\tau+1}^{j\tau} r_t$ for $j = 1, \dots, n_\tau$. For each subsample j, the statistical range of mean-adjusted partial sums is then rescaled as

$$RS_j(\tau) = \frac{1}{\hat{\sigma}_{j,\tau}} \left(\max_{1 \le t \le \tau} \left(X_{j,t} - \hat{\mu}_{j,\tau} \cdot t \right) - \min_{1 \le t \le \tau} \left(X_{j,t} - \hat{\mu}_{j,\tau} \cdot t \right) \right), \tag{3.5}$$

where $\hat{\mu}_{j,\tau} = \frac{1}{\tau} \sum_{t=(j-1)\tau+1}^{j\tau} r_t$ and $\hat{\sigma}_{j,\tau} = \sqrt{\frac{1}{\tau} \sum_{t=(j-1)\tau+1}^{j\tau} \left(r_t - \hat{\mu}_{j,\tau} \right)^2}$ are the maximum likelihood estimators of the return mean and standard deviation in a subsample. To deal with the fact that, for small τ, RSA can deliver HC values different from 0.5 even for independent data, we apply the Anis and Lloyd (1976) correction to obtain adjusted RS statistics $RS_j^*(\tau)$.[13,14] Next, for each τ, we

[12] For a nice illustration of autocorrelation behavior in typical long memory models, see Granger and Ding (1996) and Campbell et al. (1997, chpt. 2.6).

[13] Formally, we have $RS_j^*(\tau) = RS_j(\tau) - \left(\mathbb{E}(RS(\tau)) + \sqrt{\frac{1}{2}\pi\tau} \right)$, where $\mathbb{E}\left(RS(\tau)\right) = \frac{\Gamma\left(\frac{\tau-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\tau}{2}\right)} \sum_{t=1}^{\tau-1} \sqrt{\frac{\tau-t}{t}}$ is the expected value of the rescaled range in independent Gaussian data and whose logarithm asymptotically converges to $\frac{1}{2} \log\left(\frac{1}{2}\pi\tau\right)$. We do not apply the additional 'correction' of Peters (1994) because Couillard and Davison (2005) and Ellis (2006) show that it introduces further bias.

[14] Another well-known modification, which scales via an autocovariance-adjusted variance (to avoid distorting effects of short memory), has been proposed by Lo (1991). However, Teverovsky et al. (1999) and Kristoufek (2012) show that it is biased toward independence primarily because the relevant autocovariance lag size is unknown.

calculate the arithmetic mean $\overline{RS}^*(\tau)$ of adjusted RS statistics over all $j = 1, \ldots, n_\tau$ subsamples of length τ. This mean scales as $\overline{RS}^*(\tau) \approx c\tau^{HC}$ with a finite constant c independent of τ. Thus, we can estimate HC with the regression $\log\left(\overline{RS}^*(\tau)\right) = \log(c) + HC\log(\tau)$.

Detrended fluctuation analysis As shown by Weron (2002), detrended fluctuation analysis (DFA), proposed by Peng et al. (1994) and further developed by Kantelhardt et al. (2002), can outperform RSA in small samples. To implement this method, a sample again has to be subsequently subdivided into $n_\tau = \frac{T}{\tau}$ adjacent subsamples with cumulative returns $(X_{j,t})_{t=1,\ldots,\tau}$. The $(X_{j,t})$ are then detrended by subtracting the prediction $a_j + b_j \cdot t$ of a linear regression of $(X_{j,t})$ on time t. The results deliver mean square detrended variables

$$F_{j,s}(\tau) = \left(\frac{1}{\tau}\sum_{t=1}^{\tau}(X_{j,t} - a_j - b_j \cdot t)^2\right)^{s/2} \tag{3.6}$$

for all $j = 1, \ldots, n_\tau$. Next, for each τ, the sth root of the arithmetic mean $\overline{F_s(\tau)}$ of $F_{j,s}(\tau)$ over all n_τ subsamples is denoted as sth order fluctuation. It also scales as $\left(\overline{F_s(\tau)}\right)^{1/s} \approx c\tau^{HC}$ such that HC can again be estimated via logarithmic regression. Guided by Weron (2002), we restrict τ to the interval $[50, T]$ (in both DFA and RSA) and concentrate on applying DFA with $s = 1$.

Higuchi approach According to Montanari et al. (1999) and Taqqu and Teverovsky (1998), the estimator developed by Higuchi (1988) is particularly valuable when time series exhibit seasonality (or are rather short). The key difference from RSA and DFA is that Higuchi's approach is based on sliding windows which, in contrast to using non-intersecting blocks, is more computationally intensive. Specifically, we accumulate returns as $X_t = \sum_{i=1}^{t} r_t$ for $t = 1, \ldots, T$. With the floor function $\lfloor\cdot\rfloor$ and an independent constant c, we then define

$$L(\tau) = \frac{T-1}{\tau^3}\sum_{t=1}^{\tau}\left\lfloor\frac{T-t}{\tau}\right\rfloor^{-1}\sum_{i=1}^{\left\lfloor\frac{T-t}{\tau}\right\rfloor}|X_{t+i\tau} - X_{t+(i-1)\tau}| \tag{3.7}$$

and estimate HC by log-linearizing $L(\tau) \approx c\tau^{HC-2}$ and subsequent regression. Similar to Boutahar et al. (2007), τ is chosen based on sample size, which yields $\tau \in [4, 19]$.

Generalized Hurst exponent approach Finally, screening Barunik and Kristoufek (2010) for an estimator which is well-behaved in non-Gaussian data, we discover the generalized Hurst exponent (GHE) approach of Barabási and Vicsek (1991).[15] It is based on the computation of w-order moments of financial increment processes, such that we again require interval returns $X_t = \sum_{i=1}^{t} r_i$. For increasing lags τ, we compute

$$K_w(\tau) = \sum_{t=1}^{T-\tau}\left(|X_{t+\tau} - X_t|^w\right) / \sum_{t=1}^{T-\tau}\left(|X_t|^w\right) \tag{3.8}$$

statistics which approximately scale as $K_w(\tau) \approx c\tau^{wHC}$. Following Barunik and Kristoufek (2010), we regress with $w = 2$ such that K_w is proportional to the autocorrelation function of the interval return processes. As in Di Matteo et al. (2005), we set τ to range within $[1, 19]$.

[15] Barunik and Kristoufek (2010) simulate time series with different lengths following a family of stable distributions. This family is based on a characteristic exponent $\alpha \leq 2$ which, fitted to our data, ranges from 1.676 to 1.974. Hence, we consult their simulation results for α-segments from 1.7 to 2.

3.3.2.4. Long memory strategy

Our MEM strategy incorporating Hurst coefficients follows the basic design of matrix (3.4). However, several points require clarification. First, while many studies document the properties of HC estimators for very large sample sizes (such as 10.000 or more observations; see Taqqu and Teverovsky, 1998; Kristoufek, 2012), which are clearly not feasible for most financial applications, Weron (2002) and Chamoli et al. (2007) point out that, depending on the method, smaller samples (fewer than 500 observations) may be used to obtain reasonable estimates. Batten et al. (2013) go even further and argue that HC measurements based on just 22 or 66 returns deliver exploitable information. Taking these findings into account, as a compromise, we fix our rolling window size for HC estimation to 4 years of daily data.[16] Second, we combine two sources to obtain a decision rule for judging whether a deviation of our estimator average from 0.5 can be considered relevant. One source is the simulation study of Weron (2002) providing approximate confidence intervals for RSA and DFA. In our setting, the rounded 90 % intervals would be (0.37; 0.62) and (0.41; 0.58), respectively. The other source are rule-of-thumb boundaries popular in the applied literature. Screening Couillard and Davison (2005), Batten et al. (2013), Hull and McGroarty (2014) and others, persistence (anti-persistence) is typically declared for estimates within [0.55; 0.65] ([0.35; 0.45]) and higher (lower). For the sake of brevity and because we find that HC strategies cannot compete with VR strategies, we concentrate on the strategies resulting from a (0.38; 0.62) insignificance interval, which is a typical choice of investors following the mainstream literature.

3.4. Empirical results

3.4.1. Traditional momentum

We start our empirical analysis with the performance of traditional momentum. To conserve space, we follow Miffre and Rallis (2007), Fuertes et al. (2010) and Bianchi et al. (2015) by focusing on the dynamics of the strategies where the mean returns are statistically significant at the 10 % level. That is, we document 11 of 25 strategies.[17] Specifically, we have five strategies with a 1-month ranking period (R1-H1, R1-H3, R1-H6, R1-H9, R1-H12), four strategies with a 3-month ranking period (R3-H1, R3-H3, R3-H9, R3-H12), and one strategy with a 6-month (R6-H1) and a 12-month (R12-H1) ranking period each.

Table 3.2 reports basic descriptive statistics for the portfolio returns subdivided into the long (winner) leg, the short (loser) leg and the long-short (momentum) combination. For comparison, it also includes a simple equally weighted benchmark portfolio of all 28 commodities. Because many studies evaluating momentum performance focus on standard statistical measures summarizing raw returns, Bianchi et al. (2015) suggest the additional calculation of metrics popular in investment practice. We follow this lead by using the set of measures (especially tail risk metrics) already introduced in Section 3.2, which will be relevant for the comparison of traditional momentum and our new strategies.[18]

In the overall perspective, we can confirm a typical finding for commodity momentum (see Miffre, 2016; Shen et al., 2007): mean returns of long-short momentum portfolios are dominated by the returns of the long legs. In contrast to previous work, we see that, in recent data, the short leg contributes negatively, not positively, to the long-short portfolio. The best long-short

[16] While using daily instead of monthly data may appear to be a break in consistency, it is a standard approach to improve estimation (and prediction) quality in monthly portfolio settings (such as, for example, beta estimation for monthly cross-sectional regressions; see Frazzini and Pedersen, 2014; Jegadeesh et al., 2019).

[17] The detailed results for the 14 remaining strategies are available from the authors upon request.

[18] Because Ljung and Box (1978) and Engle (1982) tests (with up to 12 lags) indicate autocorrelation and heteroscedasticity in momentum returns, we base all t-tests in our study on Newey and West (1987) standard errors.

strategies are the ones with short holding periods. The leading strategy R1-H1 earns mean returns of 1.17 % (= 1.23 % − 0.06 %) per month, which is nearly twice as much as the benchmark.[19] On an annual basis this implies a mean return of 14.04 % (= 14.76 % − 0.72 %). In comparison, for a R1-H1 strategy, the earlier study of Fuertes et al. (2010) documents mean returns of 17.69% (= 12.39 % − (−5.30 %)) per year.[20] On a risk-adjusted basis, the monthly (annualized) Sharpe ratios of our best long-short strategy is, with a value of 0.09 (0.31), also larger than the one of the benchmark. Again, a comparison to Fuertes et al. (2010), who report a Sharpe ratio of 0.66 per year, indicates a decline in performance.[21]

Table 3.3 reports the estimates (and corresponding t-statistics) of the coefficients α, β_S, β_B and β_C in the multi-factor model $r_{P,t} = \alpha + \beta_S r_{S,t} + \beta_B r_{B,t} + \beta_C r_{C,t} + \varepsilon_t$, where $r_{P,t}$ reflects our strategy portfolio returns and $r_{S,t}$, $r_{B,t}$ and $r_{C,t}$ denote the returns of the S&P 500 Composite Stock Market Index, the S&P 10-Year US Government Bond Index and the S&P GSCI, respectively. Furthermore, it presents idiosyncratic volatility (i.e., the standard deviation of model residuals), the multi-factor version of the information ratio (i.e., the ratio of alpha to idiosyncratic volatility; see Goodwin, 1998) and the adjusted coefficient of determination. Our model choice is guided by Miffre and Rallis (2007), Fuertes et al. (2010) and Bianchi et al. (2015) because we wish to compare our results to the previous literature and our goal is to evaluate *investment performance* which requires that the chosen factors are the returns of portfolios which can be realized by investors.[22] Consequently, α measures the worth momentum strategies generate in excess of index investments in the three major asset classes stocks, bonds and commodities. Note that, when it comes to the analysis of (potentially non-tradeable) factors *explaining* strategy returns (see Section 3.4.4.5), we use a significantly extended model.

Looking at the model betas, we detect that the long legs are significantly positively connected to the movements of the entire commodity market. The short legs show additional positive (negative) linkage to the stock (bond) market. In contrast, evidence on a relation of long-short momentum and market index returns is strongly limited. Interestingly, and in comparison to previous studies (see Fuertes et al., 2010; Bianchi et al., 2015), none of the long-short strategies generates a significantly positive alpha.[23] While this suggests that traditional momentum has lost its merits, focusing on the long legs may still be considered beneficial. For example, the long side of R3-H1 not only earns a high and statistically significant monthly alpha of 0.61 % but also exhibits the highest monthly information ratio of 0.11 which captures the compensation for bearing one unit of idiosyncratic risk generated by turning away from passive (diversified) market investments and toward an active (less diversified) momentum strategy.

3.4.2. Short memory momentum

3.4.2.1. Baseline results

After observing that traditional momentum has significantly weakened, we now turn to our first variant of MEM which, as discussed in Section 3.3.2.2, focuses on short-term autocorrelation captured via variance ratios. Table 3.4 reports the basic return characteristics of several short memory strategies *R-q* with a focus on a 1-month holding period because, similar to Section 3.3.1,

[19] Figure C.1 of the appendix illustrates the portfolio compositions suggested by R1-H1 over time.

[20] Note that a comparison of results in the momentum literature requires special care because their presentation varies significantly. For example, some studies present monthly returns, others annualize them. Furthermore, there are articles reporting non-percentage returns instead of percentage values.

[21] A subsample analysis performed in Bianchi et al. (2015) further supports this finding.

[22] Because the bond index launched in September 2013, we approximate the years before by the (highly correlated) maturity-congruent Datastream Bond Index.

[23] Very low and partially negative values of $\bar{R}^2$ are consistent with the observations of Fuertes et al. (2010) and Bianchi et al. (2015).

Table 3.2.: Basic characteristics of traditional momentum strategies

	R1-H1	R1-H3	R1-H6	R1-H9	R1-H12	R3-H1	R3-H3	R3-H9	R3-H12	R6-H1	R12-H1	Benchmark
Panel A: Long												
Mean	**1.23**	**1.07**	**0.98**	**0.83**	**0.79**	**1.41**	**1.07**	**0.92**	**0.83**	**1.28**	**1.01**	**0.60**
t-Statistics	3.35	3.49	3.43	3.02	2.89	3.58	3.06	2.95	2.75	3.40	2.52	2.76
Volatility	7.53	5.97	5.50	5.34	5.24	7.95	6.75	5.95	5.83	7.58	8.04	4.35
SR	0.11	0.12	0.11	0.09	0.08	0.13	0.10	0.09	0.08	0.12	0.08	0.05
VaR	9.87	8.18	7.43	7.38	6.90	9.83	9.26	8.23	7.77	10.68	11.66	6.39
ES	13.27	11.30	10.83	10.68	10.60	13.89	13.07	11.77	12.12	14.96	17.00	9.20
Min	-20.29	-21.30	-21.95	-22.51	-22.17	-33.31	-33.42	-28.50	-29.34	-33.31	-33.31	-21.74
Max	51.20	35.56	29.20	24.48	25.15	51.20	35.56	36.44	34.83	41.87	41.87	25.81
Pos. mths	55.30	56.73	58.17	57.27	55.12	55.66	55.30	56.73	55.12	54.94	54.94	55.48
Panel B: Short												
Mean	0.06	0.28	0.37	**0.49**	0.42	0.26	0.36	**0.47**	0.44	0.10	0.02	**0.60**
t-Statistics	0.19	1.07	1.59	2.20	1.87	0.89	1.47	1.96	1.90	0.35	0.06	2.76
Volatility	6.67	5.40	5.14	4.95	4.82	6.29	5.70	5.20	4.96	6.73	6.70	4.35
SR	-0.05	-0.02	-0.00	0.02	0.01	-0.02	-0.00	0.02	0.01	-0.04	-0.05	0.05
VaR	10.21	8.06	7.25	6.81	6.89	10.26	9.03	7.27	6.94	10.57	9.89	6.39
ES	15.91	12.48	10.96	10.12	9.88	14.41	12.16	10.15	9.58	14.67	14.36	9.20
Min	-24.34	-20.75	-20.52	-20.64	-20.97	-20.75	-21.54	-20.50	-20.14	-20.40	-22.34	-21.74
Max	25.00	27.95	22.71	30.14	31.18	22.71	22.71	25.49	24.99	32.20	33.77	25.81
Pos. mths	52.24	52.96	53.68	55.48	54.40	50.09	52.24	53.68	54.40	49.91	49.01	55.48
Panel C: Long-short												
Mean	**1.17**	**0.79**	**0.61**	**0.35**	**0.37**	**1.15**	**0.71**	0.45	0.39	**1.17**	**1.00**	**0.60**
t-Statistics	2.90	2.83	3.00	2.18	2.62	2.54	2.07	1.89	1.84	2.87	2.22	2.76
Volatility	8.94	5.63	4.41	3.87	3.44	9.05	7.29	5.26	4.68	9.15	9.58	4.35
SR	0.09	0.07	0.05	-0.01	-0.00	0.09	0.05	0.01	0.00	0.09	0.06	0.05
VaR	12.60	7.02	5.36	4.93	4.31	11.54	8.98	7.43	6.08	11.78	14.29	6.39
ES	17.51	10.73	8.15	7.41	6.50	16.40	14.00	10.83	10.36	17.57	19.38	9.20
Min	-30.79	-30.79	-20.76	-16.48	-13.69	-30.79	-30.79	-17.91	-19.58	-30.79	-37.26	-21.74
Max	47.01	26.31	25.87	28.67	30.78	40.61	33.10	42.54	32.24	52.67	52.67	25.81
Pos. mths	56.01	54.22	54.94	52.78	54.76	53.68	53.68	52.06	51.89	53.14	53.50	55.48

Based on data from August 1973 to December 2019, this table presents descriptive statistics for the *monthly* percentage returns of traditional long, short and long-short momentum portfolios (constructed as outlined in Section 3.3.1). R and H refer to the ranking and holding periods, respectively. The benchmark is an equally weighted portfolio of all commodity futures. Further abbreviations are used as in Table 3.1. t-statistics are based on Newey-West standard errors. Significant means at a 5 % level are marked in in bold.

Table 3.3.: Alphas and betas of traditional momentum strategies

	R1-H1	R1-H3	R1-H6	R1-H9	R1-H12	R3-H1	R3-H3	R3-H9	R3-H12	R6-H1	R12-H1
Panel A: Long											
α	0.52	0.37	**0.33**	0.25	0.20	**0.61**	0.38	**0.33**	0.23	0.35	0.24
t_α	1.74	1.88	2.25	1.88	1.48	1.98	1.77	2.33	1.68	1.50	1.02
σ_ε	5.60	3.68	3.12	2.85	2.71	5.53	4.26	3.16	2.97	5.23	5.27
IR	0.09	0.10	0.11	0.09	0.07	0.11	0.09	0.10	0.08	0.07	0.05
β_S	0.06	0.07	**0.10**	**0.10**	**0.11**	0.07	0.09	**0.11**	**0.11**	0.06	0.07
t_S	0.82	1.53	2.71	2.88	3.51	0.92	1.77	3.13	3.07	1.00	1.15
β_B	0.02	0.00	-0.04	-0.07	-0.05	0.04	-0.10	-0.10	-0.09	0.06	0.07
t_B	0.21	0.01	-0.53	-1.03	-0.86	0.28	-0.94	-1.40	-1.30	0.49	0.60
β_C	**0.68**	**0.69**	**0.64**	**0.64**	**0.63**	**0.73**	**0.69**	**0.67**	**0.70**	**0.71**	**0.79**
t_C	7.81	15.76	19.78	20.61	21.59	9.05	14.46	17.04	17.48	11.64	10.35
$\bar{R}^2$	0.32	0.52	0.58	0.62	0.64	0.35	0.46	0.60	0.64	0.37	0.41
Panel B: Short											
α	-0.14	0.07	0.13	0.15	0.06	-0.05	0.08	0.00	0.05	-0.10	-0.16
t_α	-0.59	0.46	0.82	0.98	0.41	-0.21	0.37	0.02	0.29	-0.37	-0.55
σ_ε	5.06	3.46	3.07	2.93	2.80	4.97	4.19	3.50	3.39	5.28	5.44
IR	-0.03	0.02	0.04	0.05	0.02	-0.01	0.02	0.00	0.01	-0.02	-0.03
β_S	**0.15**	**0.14**	**0.11**	**0.11**	**0.10**	**0.17**	**0.13**	**0.13**	**0.12**	**0.20**	0.13
t_S	2.58	3.43	3.14	3.11	3.17	2.86	2.43	3.08	3.09	3.46	1.78
β_B	**-0.25**	**-0.16**	**-0.14**	**-0.12**	**-0.12**	-0.17	-0.14	-0.12	**-0.14**	-0.13	-0.18
t_B	-2.24	-2.35	-2.31	-1.97	-2.12	-1.68	-1.80	-1.61	-2.03	-1.21	-1.54
β_C	**0.61**	**0.55**	**0.59**	**0.57**	**0.56**	**0.51**	**0.56**	**0.56**	**0.54**	**0.56**	**0.55**
t_C	10.18	14.95	16.38	17.67	19.01	9.22	11.19	12.74	13.67	9.06	6.58
$\bar{R}^2$	0.34	0.47	0.56	0.56	0.57	0.28	0.38	0.47	0.47	0.29	0.26
Panel C: Long-short											
α	0.66	0.30	0.21	0.10	0.14	0.66	0.30	0.32	0.18	0.45	0.40
t_α	1.50	1.17	1.29	0.78	1.14	1.52	1.00	1.65	0.98	1.28	1.01
σ_ε	8.25	4.74	3.48	2.94	2.57	7.95	6.16	4.28	3.86	7.84	8.18
IR	0.08	0.06	0.06	0.03	0.05	0.08	0.05	0.08	0.05	0.06	0.05
β_S	-0.09	-0.07	-0.01	-0.01	0.01	-0.09	-0.04	-0.02	-0.01	-0.14	-0.06
t_S	-0.75	-1.10	-0.26	-0.34	0.31	-0.77	-0.47	-0.51	-0.26	-1.63	-0.48
β_B	0.27	0.16	0.10	0.05	0.07	0.20	0.04	0.02	0.05	0.19	0.25
t_B	1.69	1.62	1.36	0.87	1.30	1.12	0.32	0.24	0.67	1.13	1.45
β_C	0.07	**0.14**	0.05	0.07	**0.07**	0.21	0.14	0.12	**0.16**	0.15	0.24
t_C	0.55	2.35	1.14	1.64	2.07	1.76	1.68	1.78	2.52	1.46	1.64
$\bar{R}^2$	0.00	0.02	0.00	0.01	0.02	0.02	0.01	0.01	0.04	0.01	0.02

For the strategies of Table 3.2, this table presents estimates for the coefficients α, β_S, β_B and β_C of a linear regression of *monthly* percentage strategy returns on investable stock (S&P 500 Composite Stock Market Index), bond (S&P 10-Year US Government Bond Index) and commodity (S&P Goldman Sachs Commodity Index) indices. Their dedicated t-statistics are calculated with Newey-West adjustment. σ_ε, IR and $\bar{R}^2$ refer to the idiosyncratic volatility, the information ratio and the adjusted coefficient of determination, respectively. Significant coefficients at a 5 % level are marked in bold.

Table 3.4.: Basic characteristics of short memory strategies

	R1-q2	R3-q2	R3-q4	R6-q2	R6-q4	R6-q7	R12-q2	R12-q4	R12-q7	R12-q13	Benchmark
Panel A: Long											
Mean	**1.41**	**1.06**	**1.01**	**0.91**	**1.07**	**0.49**	**0.77**	**0.76**	0.34	0.28	**0.60**
t-Statistics	7.04	5.92	4.72	6.63	5.75	2.34	4.62	3.97	1.34	1.17	2.76
Volatility	4.30	4.14	4.95	3.62	4.32	5.00	3.98	4.47	5.07	4.83	4.35
SR	0.24	0.17	0.13	0.15	0.16	0.02	0.10	0.09	-0.01	-0.02	0.05
VaR	4.13	4.18	7.03	4.33	4.85	6.18	5.62	7.09	9.22	8.23	6.39
ES	7.24	8.52	10.75	7.38	8.52	12.12	9.85	10.41	13.59	14.23	9.20
Min	-22.08	-16.84	-21.46	-16.63	-17.29	-32.43	-16.84	-17.29	-31.74	-31.74	-21.74
Max	25.09	23.90	36.37	25.09	28.88	42.91	23.90	21.27	42.91	23.05	25.81
Pos. mths	83.84	83.12	82.05	82.59	81.87	79.53	82.59	80.43	77.38	78.64	55.48
Panel B: Short											
Mean	**-0.66**	-0.28	-0.12	-0.32	-0.23	-0.36	-0.39	-0.22	**-0.51**	**-0.63**	**0.60**
t-Statistics	-2.93	-1.28	-0.68	-1.17	-1.02	-1.32	-1.61	-0.99	-2.11	-2.66	2.76
Volatility	4.82	4.79	5.04	5.38	5.29	5.58	5.18	5.26	5.28	5.02	4.35
SR	-0.21	-0.14	-0.10	-0.13	-0.11	-0.13	-0.15	-0.11	-0.17	-0.20	0.05
VaR	8.80	8.07	8.59	8.53	8.86	9.06	8.52	8.86	10.51	9.94	6.39
ES	13.52	12.87	13.32	14.17	14.26	16.43	13.98	14.35	16.02	15.76	9.20
Min	-23.46	-23.46	-24.62	-23.46	-24.62	-27.67	-23.46	-24.07	-24.78	-24.78	-21.74
Max	44.05	44.05	36.13	38.43	36.13	36.13	44.05	36.13	36.13	24.71	25.81
Pos. mths	14.18	16.34	15.98	17.95	16.88	15.08	16.70	16.88	16.16	13.29	55.48
Panel C: Long-short											
Mean	**2.07**	**1.35**	**1.14**	**1.22**	**1.30**	**0.85**	**1.16**	**0.98**	**0.85**	**0.91**	**0.60**
t-Statistics	7.52	5.02	4.34	4.28	4.90	3.05	4.35	3.74	3.04	3.38	2.76
Volatility	6.09	5.97	6.93	6.00	6.67	7.10	6.26	6.81	6.95	6.57	4.35
SR	0.28	0.16	0.11	0.14	0.14	0.07	0.12	0.09	0.07	0.08	0.05
VaR	6.42	7.76	10.16	7.78	9.98	9.60	8.65	10.47	10.54	10.28	6.39
ES	10.57	12.51	14.41	13.04	14.12	15.80	14.15	14.65	15.58	15.38	9.20
Min	-42.24	-42.24	-35.70	-38.08	-39.28	-41.52	-43.40	-39.28	-39.28	-24.71	-21.74
Max	30.91	23.90	36.37	25.09	29.89	42.91	23.90	24.15	42.91	29.07	25.81
Pos. mths	78.10	74.69	74.87	74.15	75.76	72.71	74.87	72.89	70.56	72.71	55.48

From August 1973 to December 2019, this table presents descriptive statistics for the *monthly* percentage returns of long, short and long-short memory-enhanced momentum portfolios (constructed as outlined in Section 3.3.2.2). R refers to the period used to rank the commodities, $q-1$ to the number of lags used in the calculation of variance ratios. All strategies use a holding period of 1 month. The benchmark is an equally weighted portfolio of all commodity futures. Further abbreviations are used as in Table 3.1. *t*-statistics are based on Newey-West standard errors. Significant means at a 5 % level are marked in bold.

Table 3.5.: Alphas and betas of short memory strategies

	R1-q2	R3-q2	R3-q4	R6-q2	R6-q4	R6-q7	R12-q2	R12-q4	R12-q7	R12-q13
Panel A: Long										
α	**1.47**	**1.13**	**1.19**	**0.96**	**1.17**	0.15	**0.77**	**0.77**	-0.02	0.04
t_α	6.27	5.63	4.30	5.60	5.19	0.55	4.29	3.54	-0.06	0.15
σ_ε	4.32	4.14	5.12	3.71	4.51	5.12	4.09	4.67	5.20	4.82
IR	0.34	0.27	0.23	0.26	0.26	0.03	0.19	0.16	-0.00	0.01
β_S	0.05	0.04	0.02	0.05	0.07	**0.14**	**0.13**	**0.12**	**0.19**	**0.16**
t_S	0.87	0.83	0.40	1.08	1.17	2.16	3.05	2.19	2.62	2.25
β_B	-0.18	**-0.17**	**-0.26**	-0.14	-0.17	0.16	**-0.20**	**-0.18**	0.03	-0.14
t_B	-1.91	-2.16	-2.35	-1.63	-1.83	1.44	-2.27	-2.07	0.32	-1.72
β_C	**0.18**	**0.18**	**0.19**	**0.16**	**0.13**	**0.24**	**0.12**	**0.11**	**0.24**	**0.24**
t_C	3.83	3.48	3.81	5.09	3.33	3.43	3.05	2.90	3.03	2.99
$\bar{R}^2$	0.07	0.06	0.05	0.07	0.04	0.08	0.06	0.04	0.09	0.10
Panel B: Short										
α	**-1.00**	**-0.53**	-0.13	-0.42	-0.35	-0.49	**-0.61**	-0.31	**-0.76**	**-0.89**
t_α	-4.14	-2.22	-0.69	-1.33	-1.46	-1.56	-2.27	-1.32	-2.75	-3.44
σ_ε	4.67	4.60	4.97	4.92	4.92	5.48	4.77	4.84	5.23	4.86
IR	-0.21	-0.12	-0.03	-0.09	-0.07	-0.09	-0.13	-0.06	-0.15	-0.18
β_S	**0.18**	**0.14**	0.02	0.09	**0.13**	0.08	**0.18**	0.10	0.14	**0.17**
t_S	3.38	2.60	0.52	0.99	2.52	1.12	2.49	1.58	1.84	2.08
β_B	-0.10	-0.07	-0.15	-0.18	-0.09	-0.02	-0.09	-0.09	0.06	-0.04
t_B	-0.92	-0.75	-1.95	-1.82	-1.10	-0.14	-0.83	-1.13	0.59	-0.33
β_C	**0.27**	**0.29**	**0.25**	**0.38**	**0.34**	**0.32**	**0.26**	**0.33**	**0.22**	**0.25**
t_C	4.50	5.33	4.25	5.99	5.28	4.04	4.66	4.85	3.38	3.43
$\bar{R}^2$	0.13	0.13	0.08	0.18	0.15	0.10	0.12	0.14	0.07	0.10
Panel C: Long-short										
α	**2.47**	**1.66**	**1.32**	**1.38**	**1.52**	0.64	**1.38**	**1.08**	**0.74**	**0.93**
t_α	7.64	5.19	4.09	3.74	4.48	1.76	4.24	3.36	2.26	2.87
σ_ε	6.26	6.12	7.25	5.99	6.79	7.45	6.26	6.88	7.28	6.79
IR	0.39	0.27	0.18	0.23	0.22	0.09	0.22	0.16	0.10	0.14
β_S	-0.12	-0.11	0.00	-0.04	-0.07	0.06	-0.05	0.02	0.05	-0.01
t_S	-1.92	-1.54	0.00	-0.39	-0.81	0.71	-0.53	0.26	0.66	-0.06
β_B	-0.08	-0.09	-0.11	0.04	-0.08	0.18	-0.12	-0.09	-0.02	-0.11
t_B	-0.55	-0.81	-0.78	0.33	-0.68	1.17	-0.85	-0.79	-0.17	-0.76
β_C	-0.09	-0.11	-0.06	**-0.22**	**-0.20**	-0.08	-0.14	**-0.22**	0.01	-0.01
t_C	-1.31	-1.72	-0.91	-3.48	-2.69	-0.86	-1.94	-2.59	0.12	-0.09
$\bar{R}^2$	0.01	0.01	-0.00	0.04	0.03	0.00	0.01	0.02	-0.01	-0.00

For the strategies of Table 3.4, this table presents estimates for the coefficients α, β_S, β_B and β_C of a linear regression of monthly percentage strategy returns on investable stock (S&P 500 Composite Stock Market Index), bond (S&P 10-Year US Government Bond Index) and commodity (S&P Goldman Sachs Commodity Index) indices. Their dedicated t-statistics are calculated with Newey-West adjustment. σ_ε, IR and $\bar{R}^2$ refer to the idiosyncratic volatility, the information ratio and the adjusted coefficient of determination, respectively. Significant coefficients at a 5 % level are marked in bold.

they perform better than specifications with larger H.[24] For the same strategies, Table 3.5 presents the results of our multi-factor performance regressions.

We observe that mean returns and Sharpe ratios of our long-short strategies are systematically above their traditional counterparts in Table 3.2. Furthermore, by considering autocorrelation information, the short legs of the strategies now deliver the negative returns we would expect from a working strategy (see Jegadeesh and Titman, 1993). Finally, combinations of small ranking periods and variance ratios of low order tend to perform better than strategies with high R and q. Especially the first-order autocorrelation is key to performance improvement.

The best strategy turns out to be R1-q2. It earns a monthly (annualized) mean return of 2.07 % (24.84 %) with a Sharpe ratio of 0.28 (0.97). This not only exceeds the benchmark but also all traditional momentum specifications.[25] In addition, and particularly noteworthy, it nearly halves tail risk magnitude (i.e., VaR and ES) in comparison to the traditional R1-H1 momentum strategy and brings risk to levels close to the benchmark. Finally, R1-q1 (and many other specifications) deliver significant alphas.[26] Specifically, we have a monthly (annualized) value of 2.47 % (29.64 %). The long and short legs of the strategy also earn impressive alphas, which, on their own, outperform traditional momentum strategies.

Figure 3.1 provides further information about the R1-q2 strategy. Figure 3.1(a) compares the evolution and terminal value of US$1 million invested into the long and long-short portfolios of R1-q2 and R1-H1 as well as the equally weighted benchmark. Not surprisingly, the benchmark ranks last with respect to final wealth. Once more, we can observe that removing the short leg from a traditional momentum strategy leads to a higher investment outcome. Interestingly, we can also see that this was not true before the year 2000. Apparently, the short leg turned to a performance killer after 2000, which indicates commodity momentum crashes similar to the crashes documented for stock market momentum (see Daniel and Moskowitz, 2016). Given that the year 2000 coincides with the Commodity Futures Modernization Act (CFMA), which increased media coverage and the speed of information diffusion in commodity futures markets, this (and stagnating long-only performance) is also qualitatively consistent with the behavioral theory of Hong and Stein (1999) which predicts weaker momentum returns under such circumstances.[27] In contrast, the short leg is a valuable addition in the memory-enhanced strategy. However, even ignoring the short side (for example, because of potential short selling limitations; see Alexander, 2000) leads to a valuable strategy. Altogether, MEM is characterized by an almost steady upward movement, effectively avoiding the performance break of traditional momentum around 2000. Furthermore, the decline of traditional momentum performance in the most recent years of our sample is not evident for MEM.

Because MEM is a sophisticated combination of momentum and (short-term) reversal, the question arises on how important the trades based on anti-persistence signals are for the performance

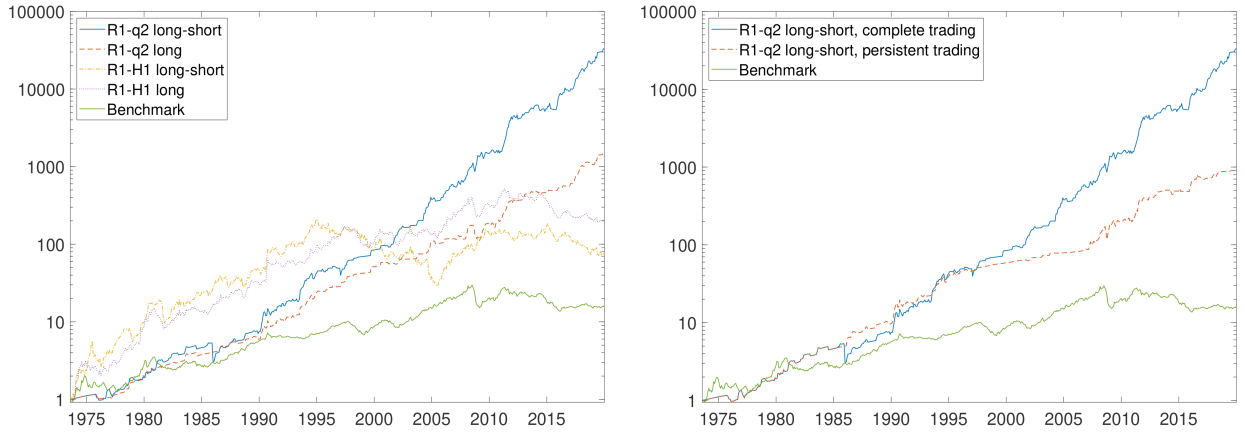
[24] Again, the full set of strategy results is available upon request.

[25] For a statistical verification of this claim, see Table C.1 of the appendix. It presents a performance comparison based on the Ledoit and Wolf (2008) bootstrap test, which does not rely on bivariate normal iid return data (thus improving over Jobson and Korkie, 1981; Memmel, 2003) and performs well in small samples (thus being superior to heteroscedasticity and autocorrelation robust kernel density estimation outlined in Andrews, 1991; Andrews and Monahan, 1992; Romano and Wolf, 2006).

[26] Table C.2 of the appendix, which allows alphas to vary with business cycle proxies (term spread, default spread), shows that alphas can be considered stable over time (see models 1 and 2). Furthermore, alphas remain highly significant if we take into account potential time-variation in betas (see model 3). A ranking of the strategies based on alphas, idiosyncratic volatilities and information ratios leads to similar results as in our main analysis. Consequently, the performance of our strategies is not a compensation for time-varying market exposure. Similar results can be obtained when allowing alphas and betas to vary with lagged strategy and market returns, respectively.

[27] In contrast, the theory of Daniel et al. (1998) would predict a significantly stronger momentum effect because the CFMA facilitated the entry of less-sophisticated investors which are likely to exhibit overconfidence and self-attribution biases (see Goetzmann and Huang, 2018).

Figure 3.1.: Wealth development



(a) Short-selling contribution (b) Anti-persistence contribution

Subfigure (a) presents the evolution and terminal value of US$1 million investments into the long and long-short portfolios of the memory-enhanced R1-q2 strategy and the traditional R1-H1 strategy as well as an equally weighted benchmark portfolio. While this illustrates the contribution of the short legs to the performance of R1-q2 and R1-H1, Subfigure (b) sheds light on the role of trades initiated by anti-persistence signals in R1-q2. That is, it separates persistence-based trades from the complete trading activity. The investment period spans from August 1973 to December 2019.

of R1-q2. To shed light on this issue, Figure 3.1(b) isolates the persistence-based trades from the total trading activity. As we can see, implementing a persistence-only strategy, which simply enhances the traditional momentum signal by autocorrelation testing, is already well-behaved. However, we also detect that, starting around 2000, the anti-persistence trades strongly contribute to the overall strategy performance. This shows that a (short-term) reversal effect can no longer be considered absent in commodity futures markets and should be dynamically considered in the implementation of trend-following strategies.

To conclude this section, we put the performance of R1-q2 into a more general perspective, meaning that we compare it to other bivariate strategies. Among the most prominent ones, we have the term structure extension (TSE) of Fuertes et al. (2010) and the (long-term) reversal inclusion (RI) of Bianchi et al. (2015). On an annual basis, the best specification of TSE exhibits mean returns and alphas of 23.55 % and 23.66 %, respectively. However, especially its tail risk is almost four times higher than that of R1-q2 and the strategy returns are significantly bound to the performance of the overall commodity market. That is, returns tend to decline in falling markets. Table 3.5 indicates no such behavior for our strategy. RI shows slightly different features. For example, it is not bound to the market. Furthermore, while it exhibits return levels similar to TSI, it generates even larger tail risk. However, these observations should not discourage the use of the strategies because both have a solid methodological foundation and their overall performance (relative to passive investments and traditional momentum) makes them highly relevant in practice.

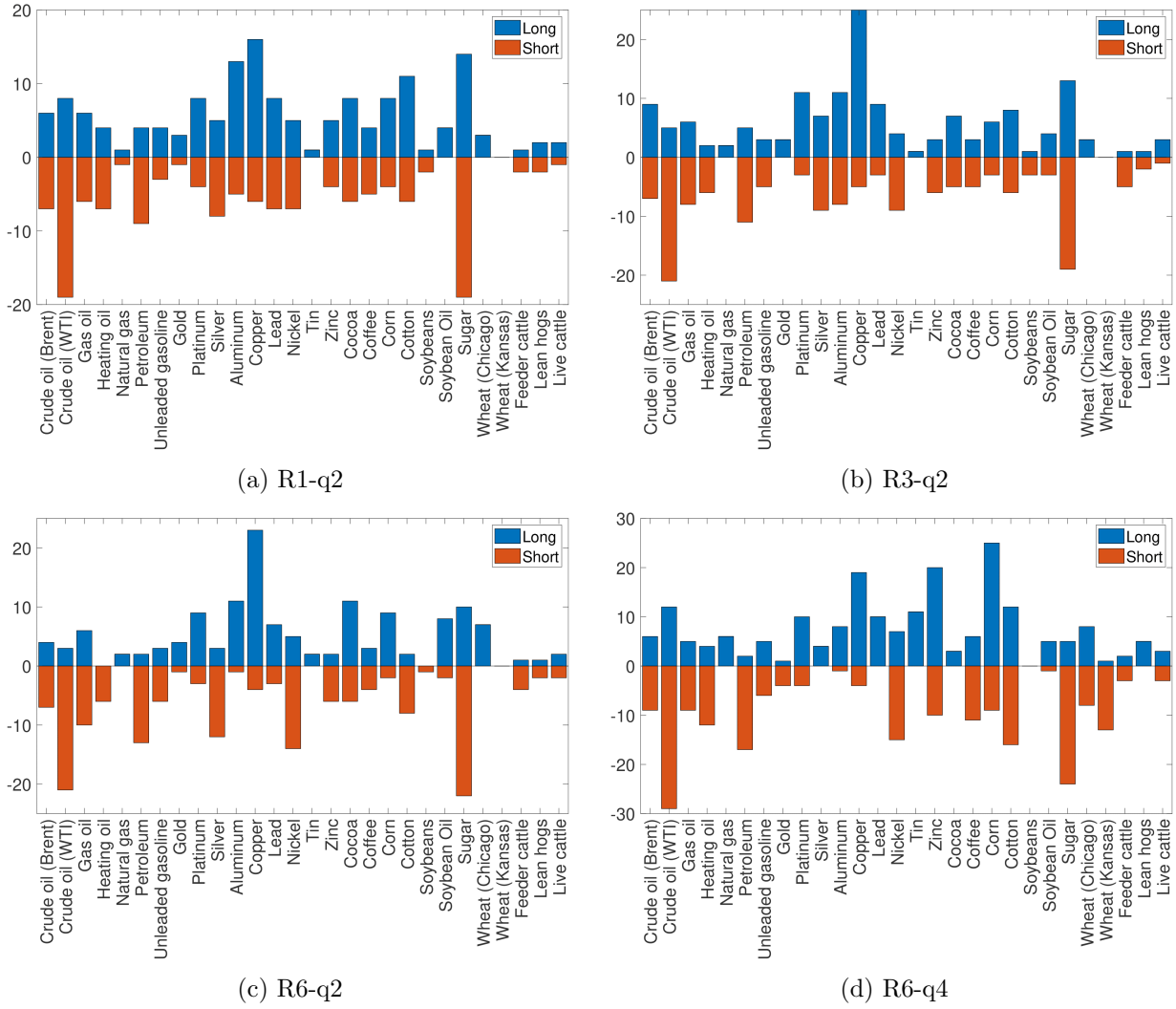
3.4.2.2. Dissecting strategy behavior

To learn more about our short memory strategies, this section has a detailed look at its commodity selection behavior. Figure 3.2 starts by identifying which commodities have been most frequently included in our strategy portfolios. For the sake of brevity, we concentrate on the strategies R1-q2, R3-q2, R6-q2 and R6-q4 outstanding in terms of Sharpe and information ratios.

Our strategies obviously do not favor specific commodity sectors but invest across all of them. However, because the commodities show different autocorrelation dynamics, some are traded more often than others.[28] In particular, WTI crude oil, copper and sugar stand out when considering

[28] Figure C.2 of the appendix illustrates the timely evolution of variance ratios for selected commodities.

Figure 3.2.: Strategy constituents



(a) R1-q2

(b) R3-q2

(c) R6-q2

(d) R6-q4

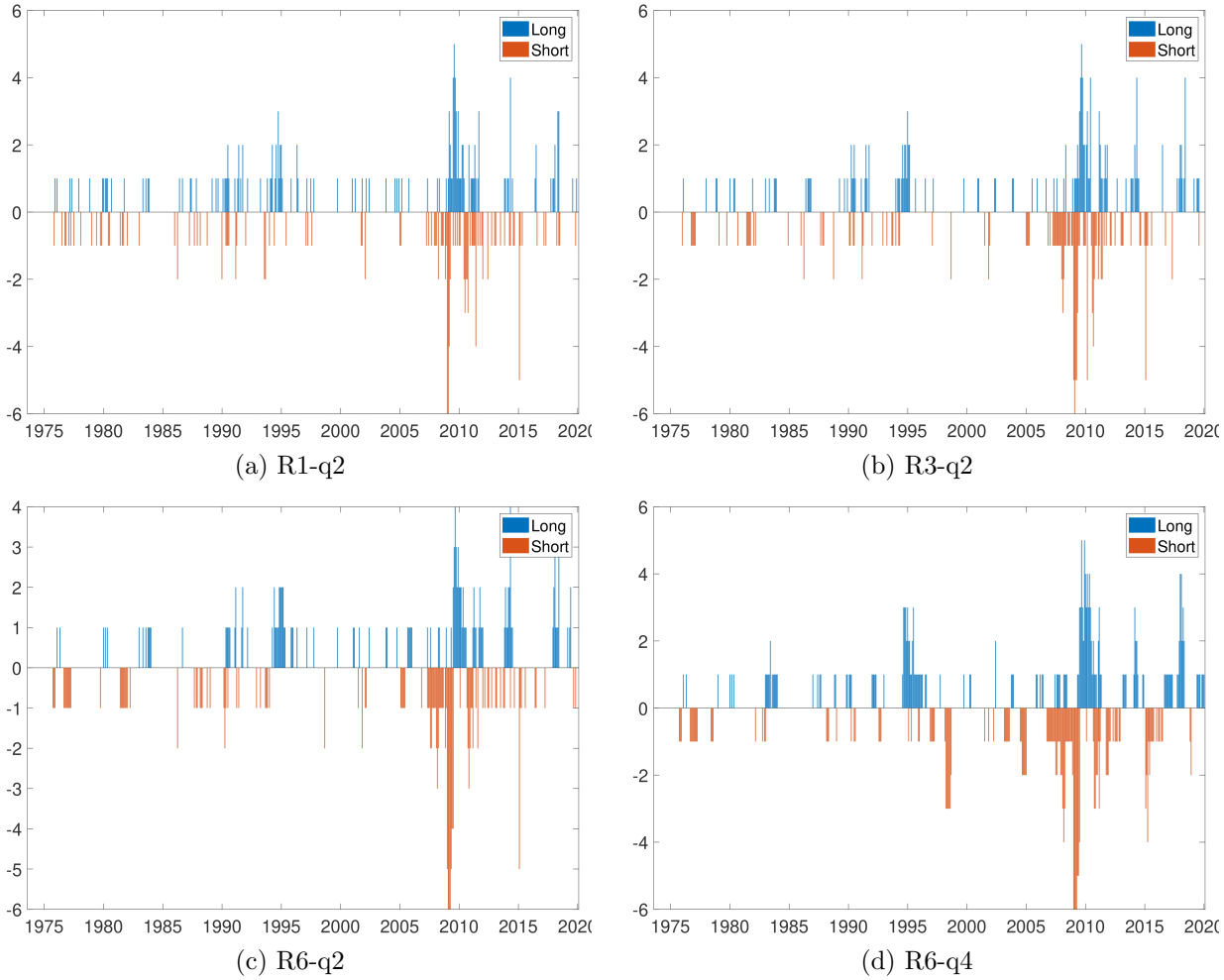
This figure plots the number of months in which a given commodity has been considered by the long and short sides of selected memory-enhanced momentum strategies presented in Tables 3.4 and 3.5.

only first-order autocorrelation. In addition, with a focus on autocorrelation up to three lags, zinc and corn receive special attention. As far as the least relevant commodities are concerned, Kansas wheat (soybeans) is never included in the presented q2 (q4) strategies. Gold, which typically receives significant attention in the financial industry, plays a negligible role in our strategies.

Turning to the positions entered over time, Figure 3.3 illustrates some additional features of our strategies. In the early sample years, they often switch from a long-only to a short-only investment with just one commodity. Because the strength of autocorrelation is time-varying, they also frequently exit the commodity market completely. Thus, despite originating from a cross-sectional setup, the behavior of the strategies is somewhat comparable to technical single-commodity time series strategies (see Marshall et al., 2008) whose timing performance, however, often crucially depends on the choice of commodity (see Han et al., 2016; Rapalias et al., 2021). The entry-exit dynamic becomes particularly valuable between 1995 and 2010, where traditional momentum suffered some of its heaviest losses.[29] While traditional momentum is forced to invest in a balanced long-short portfolio of several commodities, even when statistical testing indicates random walk

[29] Table C.3 of the appendix presents the worst losses of R1-H1 and the corresponding returns of R1-q2.

Figure 3.3.: Strategy positions over time



(a) R1-q2 (b) R3-q2

(c) R6-q2 (d) R6-q4

This figure plots the monthly number of commodities in the long and short legs of selected memory-enhanced momentum strategies presented in Tables 3.4 and 3.5.

behavior, our strategy only invests when there is autocorrelation evidence. Around the year 2000 this often led to a full risk-free investment (supplemented by occasional autocorrelation-indicated bets) and thereby generated no crucial losses.

One may argue that our strategies are not sufficiently diversified (in terms of the number of included futures contracts). However, for three reasons, this should not be a serious concern. First, we have to realize that, especially in the early years of our sample, where the number of commodities is limited, traditional momentum cannot be considered diversified either. Second, diversification is apparently not the key to momentum success. Even though traditional momentum tends to include more commodities, it produces lower returns and higher risk than our approach. Finally, our strategy is not generally an one-asset portfolio. Especially around 2010, where the autocorrelation of many commodities intensified, the position sizes of our strategy significantly increased and simultaneously included long and short investments. These positions have been quite successful. For example, crude oil (Brent and WTI), petroleum, platinum and cocoa on average contributed more than 4 % per traded month to the performance of R1-q2. With the exception of coffee, the other commodities also provided non-negative contributions.[30]

[30] In comparison, R6-q4 is pulled down by negative means of WTI crude oil, platinum, cocoa and lean hogs.

3.4.3. Long memory momentum

After highlighting the merits of incorporating variance ratio information into the selection process, we now turn to the potential of Hurst coefficients which aggregate information on short- and long-term autocorrelation. In a momentum context, we are not interested in exploiting long memory but rather in the information the long memory measure can deliver on the level of short-term autocorrelation.[31] Tables 3.6 and 3.7 report basic return characteristics and multi-factor regression results for the 'long memory' strategy outlined in Section 3.3.2.4, respectively. Similar to our presentation of short memory results, we focus on the best strategies with a 1-month holding period.

Looking at the long-short performance, for $R < 12$, we have significantly positive mean returns around 0.50 % per month accompanied by Sharpe ratios around 0.05. This clearly cannot beat the benchmark. Furthermore, the strategies earn comparably low alphas and thus cannot compete with our short memory strategies. The reasons for this poor performance can be revealed by studying their investment behavior. While the Hurst strategies were more convincing before 1984 and during the global finance crisis from October 2008 to April 2011, they have hardly suggested active trades after these periods. Combined with a historically low (almost negligible) risk-free rate in the recent decade, this crucially attenuates mean returns. Furthermore, in the active phases of the strategies, their long and short sides have focused on only few commodities. In the long legs, silver is most actively traded, followed by platinum, copper, corn, soybeans and several energy indices. Consequently, natural gas, gold, most of the industry metals and agricultural indices as well as the entire livestock sector have never been included.[32] In the short legs, we observe a similarly selective behavior.

Because no specification whatsoever can save the long memory strategies (see Section 3.4.4.1), our results indicate that Hurst coefficients are of low economic relevance. Even though several studies document that they can model the autocorrelation structure in commodity futures returns (see Barkoulas et al., 1999; Coakley et al., 2016),[33] their problem in investment applications might be that they enforce a very specific behavior of the autocorrelation function. Variance ratios allow autocorrelations at different lags to vary widely with respect to sign and magnitude. In contrast, by working with Hurst coefficients investors assume that, for example, the sign is the same for all lags. The latter may reach a better overall fit in the description of the short- and long-term autocorrelation of a time series but at the cost of potentially distorting the levels of short-term autocorrelation. While this is apparently not problematic for other assets like hedge funds (see Clark, 2005), it appears to diminish the usefulness of Hurst coefficients in a commodity momentum context.[34] Put differently, compared to variance ratios they may be a too compact (and thus less informative) measure.

3.4.4. Robustness

To ensure that our results are not driven by some specifics of our research design, we conduct a variety of robustness checks. Besides general issues concerning estimation, data and transaction costs,[35] we discuss the impact of data mining and study whether popular (portfolio-based and macroeconomic) factor variables can explain the returns of our strategies.

[31] This is similar to the focus of previous investment applications of the Hurst coefficient (see Section 3.1).
[32] Figure C.3 of the appendix illustrates the typical evolution of Hurst coefficients for selected actively traded and untraded commodities.
[33] For stock markets, the evidence is rather controversial (see, for example, Willinger et al., 1999; Weron, 2002).
[34] A detailed analysis of this proposition is beyond the scope of our study and thus left for future research.
[35] For the sake of brevity, this first part of our sensitivity checks concentrates on a verbal summary of results. Details are available upon request.

Table 3.6.: Basic characteristics of long memory strategies

	R1-Hurst			R3-Hurst			R6-Hurst			R12-Hurst			Benchmark
	Long	*Short*	*Long-Short*	*Long*	*Short*	*Long-Short*	*Long*	*Short*	*Long-Short*	*Long*	*Short*	*Long-Short*	
Mean	**0.48**	-0.10	**0.58**	**0.37**	-0.10	**0.47**	**0.41**	0.00	**0.41**	0.23	-0.21	0.43	**0.60**
t-Statistics	3.73	-0.88	3.55	2.77	-0.75	2.69	3.48	0.03	2.01	1.44	-1.37	1.90	2.76
Volatility	2.79	3.17	4.16	3.27	3.34	4.61	3.20	3.33	4.56	3.16	3.03	4.41	4.35
SR	0.04	-0.15	0.05	0.00	-0.14	0.02	0.01	-0.11	0.01	-0.04	-0.19	0.01	0.05
VaR	0.03	2.68	4.80	0.00	2.50	4.80	0.00	3.44	5.44	0.00	4.10	2.52	6.39
ES	5.28	8.71	9.67	2.39	9.23	11.36	2.06	8.33	12.38	2.32	9.07	11.99	9.20
Min	-17.29	-33.09	-17.29	-46.87	-33.09	-46.87	-46.87	-20.33	-46.87	-46.87	-20.33	-46.87	-21.74
Max	28.88	23.25	34.35	28.88	22.01	34.35	28.88	27.35	28.88	16.33	24.71	21.68	25.81
Pos. mths	85.37	4.81	82.41	86.85	4.26	83.89	86.30	4.81	82.96	85.93	3.70	84.26	55.48

From January 1975 to December 2019, this table presents descriptive statistics for the *monthly* percentage returns of long, short and long-short memory-enhanced momentum portfolios (constructed as outlined in Section 3.3.2.4). R refers to the period used to rank the commodities. Hurst coefficients are obtained via estimator averaging. All strategies use a holding period of 1 month. The benchmark is an equally weighted portfolio of all commodity futures. Further abbreviations are used as in Table 3.1. t-statistics are based on Newey-West standard errors. Significant means at a 5 % level are marked in bold.

Table 3.7.: Alphas and betas of long memory strategies

	R1-Hurst			R3-Hurst			R6-Hurst			R12-Hurst		
	Long	*Short*	*Long-Short*	*Long*	*Short*	*Long-Short*	*Long*	*Short*	*Long-Short*	*Long*	*Short*	*Long-Short*
α	**0.36**	-0.03	**0.39**	0.25	-0.08	0.32	0.28	0.02	0.26	0.25	-0.23	**0.48**
t_α	3.60	-0.24	2.51	1.26	-0.47	1.15	1.52	0.16	1.02	1.60	-1.37	1.99
σ_ε	2.68	3.18	4.10	3.27	3.37	4.70	3.16	3.06	4.39	3.21	2.85	4.42
IR	0.13	-0.01	0.09	0.08	-0.02	0.07	0.09	0.01	0.06	0.08	-0.08	0.11
β_S	0.09	**0.06**	0.04	0.09	0.04	0.04	0.10	0.03	0.06	0.08	0.04	0.03
t_S	1.67	1.96	0.66	1.39	1.57	0.59	1.81	1.15	1.08	1.60	1.45	0.56
β_B	0.03	-0.19	0.21	-0.06	-0.11	0.05	-0.06	0.02	-0.09	-0.23	-0.04	-0.19
t_B	0.35	-1.38	1.30	-0.97	-0.87	0.36	-1.48	0.28	-0.95	-1.61	-0.80	-1.20
β_C	0.02	0.07	-0.04	0.06	**0.10**	-0.03	0.07	0.08	-0.01	0.04	**0.12**	-0.07
t_C	1.29	1.64	-1.06	1.45	2.43	-0.56	1.69	1.96	-0.22	1.40	2.51	-1.22
$\bar{R}^2$	0.02	0.04	0.01	0.03	0.03	-0.00	0.03	0.02	-0.00	0.04	0.06	0.01

For the strategies of Table 3.6, this table presents estimates for the coefficients α, β_S, β_B and β_C of a linear regression of monthly percentage strategy returns on investable stock (S&P 500 Composite Stock Market Index), bond (S&P 10-Year US Government Bond Index) and commodity (S&P Goldman Sachs Commodity Index) indices. Their dedicated t-statistics are calculated with Newey-West adjustment. σ_ε, IR and $\bar{R}^2$ refer to the idiosyncratic volatility, the information ratio and the adjusted coefficient of determination, respectively. Significant coefficients at a 5 % level are marked in bold.

3.4.4.1. Estimation settings

We start with varying some of the settings in our short memory strategy. First, we enlarge the estimation window for variance ratios to 128 months. This only slightly increases the performance of R1-q2 by improving the short leg of the strategy. Second, we switch to daily and weekly data (and also consider smoothing this data) for variance ratio estimation (as in, for example, Lo and MacKinlay, 1988). However, no crucial impact on strategy performance can be observed. Finally, we follow Marcjasz et al. (2018) by working with averaged estimates across different window sizes. Again, our results turn out to be robust.

Turning to the long memory strategy, we extend the set of Hurst coefficient estimators used in our main analysis by additionally implementing the periodogram regression method, the averaged wavelet estimator and the detrended moving average approach of Geweke and Porter-Hudak (1983), Simonsen et al. (1998) and Alessio et al. (2002), respectively.[36] Furthermore, we follow Batten et al. (2013) who argue that Hurst coefficient estimation can be improved by applying estimators to filtered returns instead of raw returns.[37] However, regardless of the choice of estimation procedure, we find that the long memory strategies do not perform well. Finally, a quite important sensitivity check deals with our heuristic decision rule. That is, we narrow and expand the interval of our main analysis in various ways and find that, regardless of the chosen boundaries, long memory selection does not become successful. Put differently, the lack of statistical tests for Hurst coefficients is not a limitation because their potential decisions are captured by our robustness check.

3.4.4.2. Alternative data set

To independently test and validate the performance of our strategies, we apply them to the futures subindices provided in the context of the Bloomberg Commodity Index (formerly the Dow Jones UBS Commodity Index). They differ from our indices in aspects such as the underlying futures contracts, the rollover period (6th to 10th trading day of a month instead of 5th to 9th) and data availability (starting January 1991 instead of December 1970). We find that, in this data set, traditional momentum barely offers significant mean returns. In comparison, our short memory strategy shows even higher mean returns, Sharpe ratios and alphas than in our main analysis. Consequently, our main conclusions are not driven by the choice of data set.

3.4.4.3. Transaction costs

Even though commodity futures markets are known for their low transaction costs, it has become standard in the momentum literature to discuss their impact on strategy returns. The basis of such evaluations typically is the general futures market documentation of Locke and Venkatesh (1997), largely confirmed for commodity futures by Ferguson and Mann (2001) and Marshall et al. (2012), according to which trading costs typically range from 0.0004 % to 0.033 %.

The study of Fuertes et al. (2010) can be used to provide a particularly interesting conservative proxy for transaction costs. Across their bivariate momentum strategies based on 37 commodity futures, they compute the maximum average annual portfolio turnover (which yields 10.38) and, assuming costs of 0.033 % per trade, estimate maximum costs of just 0.69 % per annum. Given that our leading bivariate strategies trade less frequently (see Figure 3.3) and earn monthly returns above 1 % or even 2 % (see Table 3.4), transaction costs can not compensate for their outstanding performance.[38]

[36] We also consider a modification of DFA using $s = 2$ (as in Peng et al., 1994; Barunik and Kristoufek, 2010).
[37] They estimate the Hurst coefficient based on the residuals of AR(1), AR(2) and ARMA(2,1) models.
[38] Considering brokerage fees in magnitudes suggested by Paschke et al. (2020) does not change this picture.

3.4.4.4. Data mining

To rule out data mining biases, we conduct the White (2000) reality check (RC) and the Hansen (2005) test for superior predictive ability (SPA). Both tests evaluate the null hypothesis that even the best alternative within a set of given active trading strategies does not outperform a given benchmark in terms of expected losses.[39] In contrast to RC, SPA involves a studentized test statistic and a sample-dependent distribution under the null hypothesis. These features make the latter more powerful and less sensitive to the inclusion of poor and irrelevant alternatives.

Similar Fuertes et al. (2010) and Bianchi et al. (2015), we use the equally weighted commodity portfolio of our main analysis as the benchmark and define the set of alternative strategies to contain our traditional momentum and short memory strategies.[40] To ensure robustness, we implement the tests using various (circular and stationary) bootstrap methods and block lengths, and report their consistent p-values in Table 3.8.[41] All tests confirm that the superiority of our outstanding active strategy is not due to data mining.

Table 3.8.: Data mining tests

	Circular				Stationary			
	$b = 0.5$	$b = 0.2$	$b = 0.1$	$b = 0.05$	$b = 0.5$	$b = 0.2$	$b = 0.1$	$b = 0.05$
Long								
RC	**0.006**	**0.011**	**0.011**	**0.008**	**0.006**	**0.006**	**0.008**	**0.009**
SPA	**0.004**	**0.008**	**0.009**	**0.010**	**0.003**	**0.007**	**0.008**	**0.007**
Long-Short								
RC	**0.001**	**0.002**	**0.003**	**0.003**	**0.001**	**0.002**	**0.002**	**0.002**
SPA	**0.000**	**0.001**	**0.000**	**0.001**	**0.000**	**0.001**	**0.000**	**0.000**

This table reports the consistent p-values of the White (2000) reality check (RC) and the Hansen (2005) superior predictive ability (SPA) test. Both procedures are implemented with either circular (see Politis and Romano, 1992) or stationary (see Politis and Romano, 1994) bootstrapping. $\frac{1}{b} \in \{2, 5, 10, 20\}$ describes the fixed block sizes (expected values of the geometrically distributed block lengths) in the former (latter) bootstrapping scheme. For each test, the bootstrap is replicated 10,000 times. The benchmark strategy and the alternative (long and long-short) strategies are collected from Tables 3.2 and 3.4. Significance at a 5 % level is marked in bold.

3.4.4.5. Factor exposures

To study whether systematic or macroeconomic risk factors may explain the variation of our strategy returns, we collect factor data and perform several multivariate regressions. While, in the stock momentum literature, the Fama and French (1993, 2015) three- and five-factor models have become the de facto standards (see Chen et al., 2021), there is currently no widely accepted model for use in commodity research. Fuertes et al. (2010) use passive market indices, exchange rates and inflation as explanatory variables. Moskowitz et al. (2012) opt for passive indices and the classic (stock market) size, value and momentum factors. Bianchi et al. (2015) extend this variable universe by the TED spread (as a measure of global funding liquidity), the VIX volatility index and investor sentiment. Finally, Paschke et al. (2020) have a supplementary look at selected macroeconomic variables (such as industrial production, term spreads and default spreads) and commodity-specific factor portfolios (based on, for example, carry strategies).

We suggest using a model with factor portfolio returns from a particularly interesting source. In a recent study, Ilmanen et al. (2021) construct a data set spanning over a century and containing the most prominent factors that have a strong in- and out-of-sample support in many markets, are commonly employed by quantitative investors and are at the center of most academic and

[39] As in White (2000) and Bianchi et al. (2015), we express losses via negative strategy returns. Some authors, such as Fuertes et al. (2010), apply different loss functions.

[40] Including the long memory strategies does not qualitatively change the test outcomes.

[41] For a comparison of the pros and cons of the bootstrap methods, see Lahiri (1999).

practitioner asset pricing research. Besides representative long-only portfolios for the entire stock, bond and commodity market, they supply the returns of long-short portfolios implementing value, momentum, carry and defensive strategies within these markets.[42] In our context, this selection of factors has the advantages that it captures the most relevant asset classes and includes both passive and active factors. Furthermore, with regard to future research, the factors are freely available and constructed based on a consistent methodology.[43]

Table 3.9 reports the coefficient estimates (and t-statistics) of regressions of the long and long-short returns of R1-q2 on the returns the aforementioned factor portfolios.[44] We find that, while the long-only strategy returns load positively with the overall commodity market, the long-short returns do not load significantly with any of the factors. Thus, supplemented by the observation of a negligibly small coefficient of determination, our strategies cannot be explained by prominent systematic risk factors.

Because Ilmanen et al. (2021) use a rich set of macroeconomic variables to study the characteristics of their factors, they also guide us in our selection of non-tradeable factors.[45] Specifically, we use the Pastor and Stambaugh (2003) aggregate stock market liquidity measure, the Baker and Wurgler (2006) investor sentiment index, the CBOE volatility index, GDP growth, CPI inflation, a geopolitical risk index, the ICE US dollar index as well as the term spread (difference between the 30-year US TBond yield and the 3-month US TBill rate), default spread (yield difference between Moody's seasoned Baa and Aaa US corporate bonds) and TED spread (3-month LIBOR rate minus 3-month TBill rate).[46]

As we can see in Table 3.9, with the exception of inflation in the long case, none of the macroeconomic variables is significantly linked to our strategy returns.[47] Since most traditional investments decline in value during extreme liquidity events (see Pastor and Stambaugh, 2003; Asness et al., 2013) or are strongly driven by investor sentiment (see Jacobs, 2015), our strategy generates returns 'independent' of the overall state of the market and the economy and appears to be largely uninfluenced by investors' mood.

3.5. Conclusion

We have proposed a new kind of bivariate commodity futures trading strategy exploiting significant autocorrelation in futures returns. In contrast to other bivariate strategies brought forth in recent research, we do not combine traditional momentum with other independently profitable strategies (based on cross-sectional selection via, for example, term structure signals, idiosyncratic volatility or skewness; see Fuertes et al., 2010, 2015; Fernandez-Perez et al., 2018b) but instead go back to the roots of momentum by refining the quality of its buy and sell signals. That is, we only enter positions in the case of significant autocorrelation. Furthermore, we allow the strategy to dynamically adjust to the possibility of (short-term) reversal to anticipate momentum crashes.

[42] The term defensive relates to asset selection via market betas (as suggested by Frazzini and Pedersen, 2014).

[43] The data set and a detailed description of its variables can be found in the ARQ Capital Management Data Library: `https://www.aqr.com/Insights/Datasets/Century-of-Factor-Premia-Monthly`.

[44] Ilmanen et al. (2021) do not implement a carry (defensive) strategy in stock (commodity) markets because there are no futures on individual stocks (methodological issues regarding a reference portfolio for beta estimation).

[45] Note that some of the factor portfolios may also be considered non-tradeable because, for example, the stock market factors are based on hundreds of stocks and do not consider transaction costs.

[46] While the majority of these variables can be obtained from Thomson Reuters Datastream, aggregate liquidity, investor sentiment (incorporating information such as industrial production and unemployment) and the geopolitical risk index are available via `https://faculty.chicagobooth.edu/lubos-pastor/data`, `http://people.stern.nyu.edu/jwurgler` and `http://www.policyuncertainty.com/gpr.html`, respectively.

[47] This finding does not qualitatively change when using lagged explanatory variables (as additionally considered in Ilmanen et al., 2021) or performing univariate regressions with only the 20 % most extreme realizations of the explanatory variables (as in Bianchi et al., 2015).

Table 3.9.: Explanatory regressions

Factor portfolios	Long	Long-short	Macroeconomic variables	Long	Long-short
Stock passive	0.05	-0.09	Pastor-Stambaugh liquidity	3.15	-2.52
	0.79	-1.39		0.72	-0.51
Stock value	0.02	0.01	Baker-Wurgler sentiment	0.22	1.32
	0.26	0.10		0.42	1.63
Stock momentum	0.01	0.07	CBOE volatility index	-0.08	-0.04
	0.11	0.72		-1.56	-0.45
Stock defensive	0.05	-0.11	GDP growth	6.59	4.51
	1.00	-1.24		1.60	1.10
Bond passive	-0.23	0.21	CPI inflation	**1.89**	-1.48
	-1.38	0.84		2.36	-0.96
Bond value	-0.14	-0.08	Geopolitical risk index	-0.00	0.01
	-0.77	-0.34		-1.01	1.89
Bond momentum	0.01	0.23	ICE US dollar index	0.00	-0.02
	0.06	1.32		0.06	-0.40
Bond carry	0.16	-0.30	Term spread	0.03	0.13
	0.89	-1.29		0.11	0.40
Bond defensive	0.03	0.15	Default spread	2.03	2.42
	0.22	0.72		1.39	1.44
Commodity passive	**0.30**	-0.03	TED spread	-0.66	0.63
	4.52	-0.45		-0.45	0.47
Commodity value	0.03	0.11			
	0.76	1.56			
Commodity momentum	-0.00	0.07			
	-0.08	1.19			
Commodity carry	-0.01	0.02			
	-0.15	0.42			
Constant	**1.28**	**2.05**	Constant	1.14	1.54
	5.54	6.07		0.42	0.32
$\bar{R}^2$	0.08	0.01	$\bar{R}^2$	0.02	0.00

This table reports the coefficient estimates (and below them the corresponding Newey-West-adjusted t-statistics) of linear regressions explaining the long and long-short returns of our memory-enhanced R1-q2 strategy with a set of factor portfolio returns (left panel) and a set of macroeconomic variables (right panel). The choice of variables is guided by Ilmanen et al. (2021). While the left panel is based on data from August 1973 to December 2019, issues with data availability in the right panel restrict the regression sample to the period from January 1990 to December 2019. $\bar{R}^2$ is the adjusted coefficient of determination. Significance at a 5 % level is marked in bold.

When capturing short-term autocorrelation via variance ratio statistics, we can show that our strategy significantly outperforms traditional momentum in terms of reward (mean returns and alphas) and risk (volatility, value-at-risk, expected shortfall). It has the appealing features of being highly liquid and inexpensive to implement. In addition, it does not require information other than the returns of commodity futures and is not significantly related to a rich set of systematic risk factors and macroeconomic variables. This and its robustness to specification changes, data set variations and other typical influences, make the strategy a particularly valuable addition to the toolkit of quantitative commodity market investors.

Our results offer plenty of scope for future research. For example, because there is empirical evidence that industry momentum in stock markets is significantly driven by return autocorrelation (see Pan et al., 2004), it may be instructive to transfer our strategy to stock markets or apply it across asset classes (see Asness et al., 2013). To preserve the liquid nature of our strategy, we suggest doing this in a small stock universe such as the Dow Jones Industrial Average (or the S&P 500), whose constituents have been shown to exhibit exploitable autocorrelation and momentum patterns (see Fama, 1965; Figelman, 2007; Taylor, 2014). Furthermore, analyzing whether variance ratio information can stimulate new forms of momentum (such as, for example, curve momentum; see Paschke et al., 2020) could also be a fruitful endeavor.

Summary

In this thesis, we explored research questions concerning risk and return of passive and active investment strategies in commodity futures markets. To analyze the former, we concentrate on the risk measure ES; for the latter we deal with momentum-based investment strategies. The thesis consists of three parts, whose main results can be summarized as follows.

First, we make a structured comparison of non-parametric ES estimators, further parametric benchmarks and several combinations of different ES estimators with respect to common error or risk measures and performance profiles in a multidimensional simulation setup (that is naturally not limited to the commodity sector). Although we find that no estimator constantly outperforms all others, the variety of our presented results allows identification of the most suitable estimator situational for certain distributional settings, sample sizes and confidence levels in particular. Risk managers just have to decide whether they search for estimators with the lowest averaged error, fewest variability characteristics or best general performance. They can then consult our study to find the most appropriate ES estimator, depending on their market situation.

Second, we study the behavior of historic risk levels and time-variable risk predictions due to several appropriate ES estimators in commodity futures markets. Relying on backtest procedures of Du and Escanciano (2017), we detected that investment risks in terms of ES tend to be highest in energy, lowest in livestock sector and, on average, can be estimated best-possible with a non-parametric kernel density approach. In total, we find that ES predictions remain insufficient when markets are in turmoil (and accurate risk forecasts are needed most) and therefore, advise risk managers against the use of the established ES estimators for active risk prediction.

Finally, we investigate the performance of traditional cross-sectional momentum and some new memory-enhanced strategies in commodity futures markets. Our memory-enhancements allow us to pause or reverse momentum strategies when turbulent market phases are detected by measures of autocorrelation. We see that incorporation of Hurst coefficients, which belong to the measures of long memory, cannot improve the traditional investment strategies. However, introducing a measure of short memory, variance ratios, significantly improves the performance of traditional momentum strategies in several terms of reward and risk. Furthermore, the strategy returns are not significantly related to several known market indices, risk factors or macroeconomic variables and the results are also robust to a variety of typical influences such as specification or data changes, transaction costs and data mining. Thus, our enhanced momentum strategies that capture short-term autocorrelation via variance ratios can be a lucrative tool for investors in commodity markets.

This work could be extended in several ways. First, future researchers may further relate our general simulation-based evaluation of ES estimators to a potential application within the commodity context by considering additional parametric approaches as g-and-h settings, Johnson's system or distribution mixtures within the simulated framework. Additionally, researchers may look for advanced backtest enhancements that are not limited to ES estimation with invertible distribution functions and therefore, enable evaluation of the performance of general non-parametric estimators, as historic variants, in the commodity futures context. Besides, further modifications or replacements of the established AR-GARCH time series setting of ES estimators in commodity futures markets might be fruitful. It also might be instructive to see how well our results hold for other sectors, like stock markets – especially, whether the active investment strategies can preserve their strong performance.

Acknowledgments

First and foremost, I would like to thank my supervisor Professor Benjamin R. Auer for his mentorship and continuous guidance throughout the last three years. He always gave ear to my problems and also supported me with his constant optimism in respect of outcomes and his ability to place even unexpected results in a positive light. I have benefited from his thorough knowledge of the literature and, thus, his countless ideas on how to improve the papers. His final revisions have made the articles of my thesis what they are today.

I also thank Professor Frank Schuhmacher of the University of Leipzig for his willingness to render an expert opinion on my thesis.

Furthermore, I am deeply grateful to my colleagues at the BTU Cottbus–Senftenberg for creating a good working environment – especially, I enjoyed lots of inspiring and bracing talks with Kerstin Lamert and Anja Vinzelberg during the last years.

Finally, special thanks go to Patrick. Having to work from home with me for more than a year because of the pandemic, he often was the first (and sometimes the only) person to listen to my problems, which he did unfailingly and without any single complaint. Thank you for your emotional support and always being there for me.

Cottbus, November 19, 2021

Julia Mehlitz

Bibliography

Abad, P., Benito, S., López, C., 2014. A comprehensive review of value at risk methodologies. Spanish Review of Financial Economics 12, 15–32.

Acerbi, C., Szekely, B., 2014. Backtesting expected shortfall. RISK Magazine December.

Acerbi, C., Tasche, D., 2002a. Expected shortfall: A natural coherent alternative to value at risk. Economic Notes 31, 379–388.

Acerbi, C., Tasche, D., 2002b. On the coherence of expected shortfall. Journal of Banking and Finance 26, 1487–1503.

Adams, Z., Glück, T., 2015. Financialization in commodity markets: A passing trend or the new normal? Journal of Banking and Finance 60, 93–111.

Aitchison, J., Brown, J.A.C., 1957. The lognormal distribution with special reference to its uses in economics. Cambridge University Press, London.

Alessio, E., Carbone, A., Castelli, G., Frappietro, V., 2002. Second-order moving average and scaling of stochastic time series. European Physical Journal B: Condensed Matter and Complex Systems 27, 197–200.

Alexander, G.J., 2000. On back-testing 'zero-investment' strategies. Journal of Business 73, 255–278.

Aloui, C., Mabrouk, S., 2010. Value-at-risk estimations of energy commodities via long-memory, asymmetry and fat-tailed GARCH models. Energy Policy 38, 2326–2339.

Andrews, D.W., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

Andrews, D.W., Monahan, J.C., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. Econometrica 40, 953–966.

Anis, A.A., Lloyd, E.H., 1976. The expected value of the adjusted rescaled Hurst range of independent normal summands. Biometrika 63, 111–116.

Ardia, D., Hoogerheide, L., 2014. GARCH models for daily stock returns: Impact of estimation frequency on value-at-risk and expected shortfall forecasts. Economics Letters 123, 187–190.

Argyropoulos, C., Panopoulou, E., 2019. Backtesting VaR and ES under the magnifying glass. International Review of Financial Analysis 64, 22–37.

Artzner, P., Delbaen, F., Eber, J.M., Heath, D., 1997. Thinking coherently. Risk 10, 68–71.

Artzner, P., Delbaen, F., Eber, J.M., Heath, D., 1999. Coherent measures of risk. Mathematical Finance 9, 203–228.

Asness, C.S., Frazzini, A., 2013. The devil in HML's details. Journal of Portfolio Management 39, 49–68.

Asness, C.S., Frazzini, A., Israel, R., Moskowitz, T.J., Pedersen, L.H., 2018. Size matters, if you control your junk. Journal of Financial Economics 129, 479–509.

Asness, C.S., Frazzini, A., Pedersen, L.H., 2019. Quality minus junk. Review of Accounting Studies 24, 34–112.

Asness, C.S., Moskowitz, T.J., Pedersen, L.H., 2013. Value and momentum everywhere. Journal of Finance 68, 929–985.

Auer, B., 2015. Extreme value theory, asset ranking, and threshold choice: A practical note on VaR estimation. Journal of Risk 18, 27–44.

Badrinath, S.G., Chatterjee, S., 1988. On measuring skewness and elongation in common stock return distributions: The case of the market index. Journal of Business 61, 451–472.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. Journal of Finance 61, 1645–1680.

Bali, T., 2007. A generalized extreme value approach to financial risk measurement. Journal of Money, Credit and Banking 39, 1613–1649.

Bali, T., Mo, H., Tang, Y., 2008. The role of autoregressive conditional skewness and kurtosis in the estimation of conditional VaR. Journal of Banking and Finance 32, 269–282.

Balkema, A.A., de Haan, L., 1974. Residual life time at great age. Annals of Probability 2, 792–804.

Barabási, A.L., Vicsek, T., 1991. Multifractality of self-affine fractals. Physical Review A 44, 2730.

Barkoulas, J.T., Labys, W.C., Onochie, J.I., 1999. Long memory in futures prices. Financial Review 34, 91–100.

Barunik, J., Kristoufek, L., 2010. On Hurst exponent estimation under heavy-tailed distributions. Physica A: Statistical Mechanics and its Applications 389, 3844–3855.

Basak, S., Shapiro, A., 2001. Value-at-risk based risk management: Optimal policies and asset prices. Review of Financial Studies 14, 371–405.

Basel Committee of Banking Supervision, 1996. Amendment to the capital accord to incorporate market risks. https://www.bis.org/publ/bcbs24.pdf.

Basel Committee of Banking Supervision, 2004. International convergence of capital measurement and capital standards. https://www.bis.org/publ/bcbs107.pdf.

Basel Committee of Banking Supervision, 2011. Messages from the academic literature on risk measurement for the trading book. https://www.bis.org/publ/bcbs_wp19.pdf.

Basel Committee of Banking Supervision, 2012. Fundamental review of the trading book. https://www.bis.org/publ/bcbs219.pdf.

Batten, J.A., Ciner, C., Lucey, B.M., Szilagyi, P.G., 2013. The structure of gold and silver spread returns. Quantitative Finance 13, 561–570.

Baumeister, C., Kilian, L., 2015. Forecasting the real price of oil in a changing world: A forecast combination approach. Journal of Business and Economic Statistics 33, 338–351.

Baur, D., Lucey, B., 2010. Is gold a hedge or a safe haven? An analysis of stocks, bonds and gold. Financial Review 45, 217–229.

Baur, D., McDermott, T., 2010. Is gold a safe haven? International evidence. Journal of Banking and Finance 34, 1886–1898.

Bauwens, L., Laurent, S., 2005. A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. Journal of Business and Economic Statistics 23, 346–354.

Beckers, B., Herwartz, H., Seidel, M., 2017. Risk forecasting in (T)GARCH models with uncorrelated dependent innovations. Quantitative Finance 17, 121–137.

Berkowitz, J., 2001. Testing density forecasts, with applications to risk management. Journal of Business and Economic Statistics 19, 465–474.

Berkowitz, J., Christoffersen, P., Pelletier, D., 2011. Evaluating value-at-risk models with desk-level data. Management Science 57, 2213–2227.

Bhardwaj, G., Gorton, G., Rouwenhorst, G., 2015. Facts and fantasies about commodity futures ten years later. Technical Report. National Bureau of Economic Research.

Bhattacharyya, M., Chaudhary, A., Yadav, G., 2008. Conditional VaR estimation using Pearson's type IV distribution. European Journal of Operational Research 191, 386–397.

Bianchi, R., Drew, M., Fan, J., 2016. Commodities momentum: A behavioral perspective. Journal of Banking and Finance 72, 133–150.

Bianchi, R.J., Drew, M.E., Fan, J.H., 2015. Combining momentum with reversal in commodity futures. Journal of Banking and Finance 59, 423–444.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., Chou, R., Kroner, K., 1992. ARCH-modeling in finance: A review of the theory and empirical evidence. Journal of Econometrics 52, 5–59.

Bollerslev, T., Wooldridge, J.M., 1992. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. Econometric Reviews 11, 143–172.

Boutahar, M., Marimoutou, V., Nouira, L., 2007. Estimation methods of the long memory parameter: Monte Carlo analysis and application. Journal of Applied Statistics 34, 261–301.

Bowman, A., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis. Oxford University Press, Oxford.

Bowman, A., Hall, P., Prvan, T., 1998. Bandwidth selection for the smoothing of distribution functions. Biometrika 4, 799–808.

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Brandolini, D., Colucci, S., 2012. Backtesting value-at-risk: A comparison between filtered bootstrap and historical simulation. Journal of Risk Model Validation 6, 3–16.

Broda, S., Paolella, M., 2011. Expected shortfall for distributions in finance, in: Cizek, P., Haerdle, W., Weron, R. (Eds.), Statistical tools for finance and insurance. Springer, Berlin, pp. 57–99.

Brooks, C., Clare, A.D., Dalle Molle, J.W., Persand, G., 2005. A comparison of extreme value theory approaches for determining value at risk. Journal of Empirical Finance 12, 339–352.

Buchanan, N., 2013. The debt ceiling disasters. Carolina Academic Press, Durham.

Campbell, J.Y., Grossman, S.J., Wang, J., 1993. Trading volume and serial correlation in stock returns. Quarterly Journal of Economics 108, 905–939.

Campbell, J.Y., Lo, A.W., MacKinlay, A.C., 1997. The econometrics of financial markets. Princeton University Press, Princeton.

Caporin, M., de Magistris, P.S., 2012. On the evaluation of marginal expected shortfall. Applied Economics Letters 19, 175–179.

Carnero, M., Eratalay, M., 2014. Estimating VAR-MGARCH models in multiple steps. Studies in Nonlinear Dynamics and Econometrics 18, 339–365.

Carter, C., Rausser, G., Smith, A., 2011. Commodity booms and busts. Annual Review of Resource Economics 3, 87–118.

Carter, D.A., Rogers, D.A., Simkins, B.J., Treanor, S.D., 2017. A review of the literature on commodity risk management. Journal of Commodity Markets 8, 1–17.

Chamoli, A., Bansal, A.R., Dimri, V.P., 2007. Wavelet and rescaled range approach for the Hurst coefficient for short and long time series. Computers and Geosciences 33, 83–93.

Chan, E., 2013. Algoritmic Trading: Winning Strategies and their Rationale. John Wiley and Sons, Hoboken.

Charles, A., Darné, O., 2009. Variance ratio tests of random walk: An overview. Journal of Economic Surveys 23, 503–527.

Chavez-Demoulin, V., McGill, J., 2012. High-frequency financial data modeling using Hawkes processes. Journal of Banking and Finance 36, 3415–3426.

Chen, S., Tang, S., 2005. Nonparametric inference of value-at-risk for dependent financial returns. Journal of Financial Econometrics 3, 227–255.

Chen, S.X., 2008. Nonparametric estimation of expected shortfall. Journal of Financial Econometrics 6, 87–107.

Chen, T., Chou, P., Ko, K., Rhee, S., 2021. Non-parametric momentum based on ranks and signs. Journal of Empirical Finance 60, 94–109.

Cheng, I., Xiong, W., 2014. Financialization of commodity markets. Annual Review of Financial Economics 6, 419–441.

Cheng, M.Y., Sun, S., 2006. Bandwidth selection for kernel quantile estimation. Journal of the Chinese Statistical Association 44.

Cheng, W., Hung, J., 2011. Skewness and leptokurtosis in GARCH-typed VaR estimation of petroleum and metal asset returns. Journal of Empirical Finance 18, 160–173.

Chordia, T., Subrahmanyam, A., Tong, Q., 2014. Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? Journal of Accounting and Economics 58, 41–58.

Christoffersen, P., 1998. Evaluating interval forecasts. International Economic Review 39, 841–862.

Christopherson, J.A., Ferson, W.E., Glassman, D.A., 1998. Conditioning manager alphas on economic information: Another look at the persistence of performance. Review of Financial Studies 11, 111–142.

Claeskens, G., Magnus, J., Vasnev, A., Wang, W., 2016. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting 32, 754–762.

Clark, A., 2005. The use of Hurst and effective return in investing. Quantitative Finance 5, 1–8.

Clayton, B., 2016. Commodity markets and the global economy. Cambridge University Press, New York.

Coakley, J., Kellard, N., Wang, J., 2016. Commodity futures returns: More memory than you might think! European Journal of Finance 22, 1457–1483.

Conrad, J., Kaul, G., 1998. An anatomy of trading strategies. Review of Financial Studies 11, 489–519.

Constanzino, N., Curran, M., 2015. Backtesting general spectral risk measures with application to expected shortfall. https://ssrn.com/abstract=2514403.

Cont, R., 2001. Empirical properties of asset returns: Stylized facts and statistical issues. Quantitative Finance 1, 223–236.

Cotter, J., Dowd, K., 2010. Estimating financial risk measures for futures positions: A nonparametric approach. Journal of Futures Markets 30, 689–703.

Couillard, M., Davison, M., 2005. A comment on measuring the Hurst exponent of financial time series. Physica A: Statistical Mechanics and its Applications 348, 404–418.

Cramér, H., 1946. Mathematical methods of statistics. Princeton University Press, Princeton.

Daniel, K., Hirshleifer, D., Subrahmanyam, A., 1998. Investor psychology and security market under- and overreactions. Journal of Finance 53, 1839–1885.

Daniel, K., Moskowitz, T.J., 2016. Momentum crashes. Journal of Financial Economics 122, 221–247.

Daníelsson, J., 2011. Financial Risk Forecasting. John Wiley and Sons, Chichester.

Daníelsson, J., James, K., Valenzuela, M., Zer, I., 2016. Model risk of risk models. Journal of Financial Stability 23, 79–91.

Daníelsson, J., Jorgensen, B., Samorodnitsky, G., Sarma, M., de Vries, C., 2013. Fat tails, VaR and subadditivity. Journal of Econometrics 172, 283–291.

Daskalaki, C., Kostakis, A., Skiadopoulos, G., 2014. Are there common factors in individual commodity futures returns? Journal of Banking and Finance 40, 346–363.

Daskalaki, C., Skiadopoulos, G., Topaloglou, N., 2017. Diversification benefits of commodities: A stochastic dominance efficiency approach. Journal of Empirical Finance 44, 250–269.

De Souza, C., Gokcan, S., 2004. Hedge fund investing: A quantitative approach to hedge fund manager selection and de-selection. Journal of Wealth Management 6, 52–73.

DeBrota, D., Dittus, R., Swain, J., Roberts, S., Wilson, J., 1989. Modeling input processes with Johnson distributions, in: WSC '89: Proceedings of the 21st conference on winter simulation, pp. 308–318.

Degen, M., Embrechts, P., Lambrigger, D., 2007. The quantitative modeling of operational risk: Between g-and-h and EVT. Astin Bulletin 37, 265–291.

Degiannakis, S., Potamia, A., 2017. Multiple-days-ahead value-at-risk and expected shortfall forecasting for stock indices, commodities and exchange rates: Inter-day versus intra-day data. International Review of Financial Analysis 49, 176–190.

Del Brio, E., Mora-Valencia, A., Perote, J., 2017. Risk quantification for commodity ETFs: Backtesting value-at-risk and expected shortfall. International Review of Financial Analysis , forthcoming.

DeMiguel, V., Nogales, F., Uppal, R., 2014. Stock return serial dependence and out-of-sample portfolio performance. Review of Financial Studies 27, 1031–1073.

Di Matteo, T., Aste, T., Dacorogna, M.M., 2005. Long-term memories of developed and emerging markets: Using the scaling analysis to characterize their stage of development. Journal of Banking and Finance 29, 827–851.

Díaz, A., García-Donato, G., Mora-Valencia, A., 2017. Risk quantification in turmoil markets. Risk Management 19, 202–224.

Dolan, E.D., Moré, J.J., 2002. Benchmarking optimization software with performance profiles. Mathematical Programming 91, 201–213.

Du, Z., Escanciano, J.C., 2017. Backtesting expected shortfall: Accounting for tail risk. Management Science 63, 940–958.

Dumas, M., 1992. Productivity in industry and government, 1990. Monthly Labor Review 115, 48–57.

Ellis, C., 2006. The mis-specification of the expected rescaled adjusted range. Physica A: Statistical Mechanics and its Applications 363, 469–476.

Elton, E., Gruber, M., S.J., B., Goetzmann, W., 2007. Modern portfolio theory and investment analysis. John Wiley and Sons, Hoboken.

Engle, R., 2001. GARCH 101: The use of ARCH/GARCH models in applied econometrics. Journal of Economic Perspectives 15, 157–168.

Engle, R., 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business and Economic Statistics 20, 339–350.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987–1007.

Erb, C., Harvey, C., 2006. The strategic and tactical value of commodity futures. Financial Analysts Journal 62, 69–97.

Ergen, I., 2015. Two-step methods in VaR prediction and the importance of fat tails. Quantitative Finance 15, 1013–1030.

Ergün, A., Jun, J., 2010. Time-varying higher-order conditional moments and forecasting intraday VaR and Expected Shortfall. Quarterly Review of Economics and Finance 50, 264–272.

Erten, B., Ocampo, J., 2013. Super cycles of commodity prices since the mid-nineteenth century. World Development 44, 14–30.

Escanciano, J., Mayoral, S., 2009. Semiparametric estimation of dynamic conditional expected shortfall models. https://ssrn.com/abstract=1342418.

Escanciano, J.C., Lobato, I.N., 2009. An automatic Portmanteau test for serial correlation. Journal of Econometrics 151, 140–149.

Etienne, X., Irwin, S., Garcia, P., 2018. Speculation and corn prices. Applied Economics 50, 4724–4744.

Fama, E.F., 1965. The behavior of stock-market prices. Journal of Business 38, 34–105.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Fama, E.F., French, K.R., 2012. Size, value, and momentum in international stock returns. Journal of Financial Economics 105, 457–472.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fan, J., Qi, L., Xiu, D., 2014. Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. Journal of Business and Economic Statistics 32, 178–191.

Ferguson, M.F., Mann, S.C., 2001. Execution costs and their intraday variation in futures markets. Journal of Business 74, 125–160.

Fernandez-Perez, A., Frijns, B., Fuertes, A., Miffre, J., 2018a. The skewness of commodity futures returns. Journal of Banking and Finance 86, 143–158.

Fernandez-Perez, A., Frijns, B., Fuertes, A.M., Miffre, J., 2018b. The skewness of commodity futures returns. Journal of Banking and Finance 86, 143–158.

Figelman, I., 2007. Stock return momentum and reversal. Journal of Portfolio Management 34, 51–67.

Fintzen, D., Stekler, H., 1999. Why did forecasters fail to predict the 1990 recession? International Journal of Forecasting 15, 309–323.

Fissler, T., Ziegel, J., 2016. Higher order elicitability and Osband's principle. Annals of Statistics 44, 1680–1707.

Francq, C., Zakoïan, J., 2004. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. Bernoulli 10, 605–637.

Francq, C., Zakoïan, J., 2012. QML estimation of a class of multivariate asymmetric GARCH models. Econometric Theory 28, 179–206.

Frazzini, A., Pedersen, L., 2014. Betting against beta. Journal of Financial Economics 111, 1–15.

Frey, R., McNeil, A., 2002. VaR and expected shortfall in portfolios of dependent credit risks: Conceptual and practical insights. Journal of Banking and Finance 26, 1317–1334.

Fuertes, A.M., Miffre, J., Fernandez-Perez, A., 2015. Commodity strategies based on momentum, term structure, and idiosyncratic volatility. Journal of Futures Markets 35, 274–297.

Fuertes, A.M., Miffre, J., Rallis, G., 2010. Tactical allocation in commodity futures markets: Combining momentum and term structure signals. Journal of Banking and Finance 34, 2530–2548.

Furió, D., Climent, F., 2013. Extreme value theory versus traditional GARCH approaches applied to financial data: A comparative evaluation. Quantitative Finance 13, 45–63.

Furlong, F.T., Keeley, M.C., 1989. Capital regulation and bank risk-taking: A note. Journal of Banking and Finance 13, 883–891.

Füss, R., Adams, Z., Kaiser, D., 2010. The predictive power of value-at-risk models in commodity futures markets. Journal of Asset Management 11, 261–285.

Gao, F., Song, F., 2008. Estimation risk in GARCH VaR and ES estimates. Econometric Theory 24, 1404–1424.

Gao, X., Nardari, F., 2018. Do commodities add economic value in asset allocation? New evidence from time-varying moments. Journal of Financial and Quantitative Analysis 53, 365–393.

Gaunt, C., Gray, P., 2003. Short-term autocorrelation in Australian equities. Australian Journal of Management 28, 97–117.

Gebhardt, W., Hvidkjaer, S., Swaminathan, B., 2005. Stock and bond market interaction: Does momentum spill over? Journal of Financial Economics 75, 651–690.

Gençay, R., Selçuk, F., 2004. Extreme value theory and Value-at-Risk: Relative performance in emerging markets. International Journal of Forecasting 20, 287–303.

Genre, V., Kenny, G., Meyler, A., Timmermann, A., 2013. Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting 29, 108–121.

George, F., Ramachandran, K.M., 2011. Estimation of parameters of Johnson's system of distributions. Journal of Modern Applied Statistical Methods 10, 494–504.

Georgopoulou, A., Wang, J., 2017. The trend is your friend: Time-series momentum strategies across equity and commodity markets. Review of Finance 21, 1557–1592.

Geweke, J., Porter-Hudak, S., 1983. The estimation and application of long memory time series models. Journal of Time Series Analysis 4, 221–238.

Gilli, M., Këllezi, E., 2006. An application of extreme value theory for measuring financial risk. Computational Economics 27, 207–228.

Gneiting, T., 2011. Making and evaluating point forecasts. Journal of the American Statistical Association 106, 746–762.

Goetzmann, W.N., Huang, S., 2018. Momentum in imperial Russia. Journal of Financial Economics 130, 579–591.

Goodwin, T.H., 1998. The information ratio. Financial Analysts Journal 54, 34–43.

Gorton, G., Rouwenhorst, K., 2006. Facts and fantasies about commodity futures. Financial Analysts Journal 62, 47–68.

Granger, C.W.J., Ding, Z., 1996. Varieties of long memory models. Journal of Econometrics 73, 61–77.

Greene, W., 2003. Econometric Analysis. 5th ed., Prentice Hall, Upper Saddle River.

Guidolin, M., Pedio, M., 2020. Forecasting commodity futures returns with stepwise regressions: Do commodity-specific factors help? Annals of Operations Research, forthcoming .

Haase, M., Seiler Zimmermann, Y., Zimmermann, H., 2016. The impact of speculation on commodity futures markets - A review of the findings of 100 empirical studies. Journal of Commodity Markets 3, 1–15.

Hammoudeh, S., Nguyen, D., Reboredo, J., Wen, X., 2014. Dependence of stock and commodity futures markets in China: Implications for portfolio investment. Emerging Markets Review 21, 183–200.

Han, Y., Hu, T., Yang, J., 2016. Are there exploitable trends in commodity futures prices? Journal of Banking and Finance 70, 214–234.

Hansen, B.E., 1994. Autoregressive conditional density estimation. International Economic Review 35, 705–730.

Hansen, B.E., Lunde, A., 2005. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? Journal of Applied Econometrics 20, 873–889.

Hansen, P.R., 2005. A test for superior predictive ability. Journal of Business and Economic Statistics 23, 365–380.

Harmantzis, F., Miao, L., Chien, Y., 2006. Empirical study of value-at-risk and expected shortfall models with heavy tails. Journal of Risk Finance 7, 117–135.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning. Springer, New York.

Headrick, T.C., Kowalchuk, R.K., Sheng, Y., 2008. Parametric probability densities and distribution functions for Tukey g-and-h transformations and their use for fitting data. Applied Mathematical Sciences 2, 449–462.

Henderson, B.J., Pearson, N.D., Wang, L., 2015. New evidence on the financialization of commodity markets. The Review of Financial Studies 28, 1285–1311.

Herrera, R., 2013. Energy risk management through self-exciting marked point process. Energy Economics 38, 64–76.

Higuchi, T., 1988. Approach to an irregular time series on the basis of the fractal theory. Physica D: Nonlinear Phenomena 31, 277–283.

Hillebrand, E., 2005. Neglecting parameter changes in GARCH models. Journal of Econometrics 129, 121–138.

Hirshleifer, D., 2015. Residual risk, trading costs, and commodity futures risk premia. Review of Financial Studies 1, 173–193.

Ho, L., Yu, M., Chen, P., 2008. Futures margin requirement: A comparison of value-at-risk with expected shortfall measures. Advances in Financial Planning and Forecasting 3, 211–234.

Hoga, Y., 2018. Confidence intervals for conditional tail risk measures in ARMA-GARCH models. Journal of Business and Economic Statistics 37, 613–624.

Hoga, Y., 2019. Extending the limits of backtesting via the 'vanishing p'-approach. Journal of Time Series Analysis 40, 858–866.

Hong, H., Stein, J.C., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. Journal of Finance 54, 2143–2184.

Hong, K., Satchell, S., 2015. Time series momentum trading strategy and autocorrelation amplification. Quantitative Finance 15, 1471–1487.

Hugonnier, J., Morellec, E., 2017. Bank capital, liquid reserves, and insolvency risk. Journal of Financial Economics 125, 266–285.

Hull, M., McGroarty, F., 2014. Do emerging markets become more efficient as they develop? Long memory persistence in equity indices. Emerging Markets Review 18, 45–61.

Hurst, H.E., 1951. Long-term storage capacity of reservoirs. Transactions of the American Society of Civil Engineers 116, 770–799.

Hwang, S., Rubesam, A., 2015. The disappearance of momentum. European Journal of Finance 21, 584–607.

Hwang, S., Valls Pereira, P., 2006. Small sample properties of garch estimates and persistence. European Journal of Finance 12, 473–494.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688.

Ilmanen, A., Israel, R., Lee, R., Moskowitz, T.J., Thapar, A., 2021. How do factor premia vary over time? A century of evidence. Journal of Investment Management, forthcoming .

Inui, K., Kijima, M., 2005. On the significance of expected shortfall as a coherent risk measure. Journal of Banking and Finance 29, 853–864.

Jacobs, H., 2015. What explains the dynamics of 100 anomalies? Journal of Banking and Finance 57, 65–85.

Jacobsen, B., Marshall, B., Visaltanachoti, N., 2019. Stock market predictability and industrial metal returns. Management Science 65, 2947–3448.

Jadhav, D., Ramanathan, T.V., Naik-Nimbalkar, U.V., 2009. Modified estimators of the expected shortfall. Journal of Emerging Market Finance 8, 87–107.

Jagannathan, R., 1985. An investigation of commodity futures prices using the consumption-based intertemporal capital asset pricing model. Journal of Finance 40, 175–191.

Jalal, A., Rockinger, M., 2008. Predicting tail-related risk measures: The consequences of using GARCH filters for non-GARCH data. Journal of Empirical Finance 15, 868–877.

Jegadeesh, N., Noh, J., Pukthuanthong, K., Roll, R., Wang, J., 2019. Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation. Journal of Financial Economics 133, 273–298.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance 48, 65–91.

Jegadeesh, N., Titman, S., 2001. Profitability of momentum strategies: An evaluation of alternative explanations. Journal of Finance 56, 699–720.

Jensen, G.R., Johnson, R.R., Mercer, J.M., 2000. Efficient use of commodity futures in diversified portfolios. Journal of Futures Markets: Futures, Options, and Other Derivative Products 20, 489–506.

Jiménez Moscoso, J., Arunachalam, V., 2011. Using Tukey's g and h family of distributions to calculate value-at-risk and conditional value-at-risk. Journal of Risk 13, 95–116.

Jobson, J.D., Korkie, B.M., 1981. Performance hypothesis testing with the Sharpe and Treynor measures. Journal of Finance 36, 889–908.

Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 36, 149–176.

Jondeau, E., Rockinger, M., 2003. Conditional volatility, skewness, and kurtosis: Existence, persistence, and comovements. Journal of Economic Dynamics and Control 27, 1699–1737.

Jorion, P., 2007. Value at risk: The new benchmark for managing financial risk. 3 ed., McGraw Hill, New York.

Joy, M., 2011. Gold and the US dollar: Hedge or haven? Finance Research Letters 8, 120–131.

Kamara, A., 1984. The behavior of futures prices: A review of theory and evidence. Financial Analysts Journal 40, 68–75.

Kang, S., Mclver, R., Yoon, S., 2017. Dynamic spillover effects among crude oil, precious metal, and agricultural commodity futures markets. Energy Economics 62, 19–32.

Kantelhardt, J.W., Zschiegner, S.A., Koscielny-Bunde, E., Havlin, S., Bunde, A., Stanley, H.E., 2002. Multifractal detrended fluctuation analysis of nonstationary time series. Physica A: Statistical Mechanics and its Applications 316, 87–114.

Kato, K., 2012. Weighted Nadaraya-Watson estimation of conditional expected shortfall. Journal of Financial Econometrics 10, 265–291.

Kellner, R., Rösch, D., 2016. Quantifying market risk with value-at-risk or expected shortfall? – Consequences for capital requirements and model risk. Journal of Economic Dynamics and Control 68, 45–63.

Kerkhof, J., Melenberg, B., 2004. Backtesting for risk-based regulatory capital. Journal of Banking and Finance 28, 1845–1865.

Kinateder, H., 2016. Basel II versus III: A comparative assessment of minimum capital requirements for internal model approaches. Journal of Risk 18, 25–45.

Klar, B., Lindner, F., Meintanis, S., 2012. Specification tests for the error distribution in GARCH models. Computational Statistics and Data Analysis 56, 3587–3598.

Kon, S.J., 1984. Models of stock returns - a comparison. Journal of Finance 39, 147–165.

Kratz, M., Lok, Y., McNeil, A., 2018. Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. Journal of Banking and Finance 88, 393–407.

Krehbiel, T., Adkins, L., 2005. Price risk in the NYMEX energy complex: An extreme value approach. Journal of Futures Markets 25, 309–337.

Kristoufek, L., 2012. How are rescaled range analyses affected by different memory and distributional properties? A Monte Carlo study. Physica A: Statistical Mechanics and its Applications 391, 4252–4260.

Kristoufek, L., Vosvrda, M., 2014. Commodity futures and market efficiency. Energy Economics 42, 50–57.

Kucher, O., McCoskey, S., 2017. The long-run relationship between precious metal prices and the business cycle. Quarterly Review of Economics and Finance 65, 263–275.

Kuester, K., Mittnik, S., Paolella, M., 2006. Value-at-risk prediction: A comparison of alternative strategies. Journal of Financial Econometrics 4, 53–89.

Kupiec, P., 1995. Techniques for verifying the accuracy of risk measurement models. The Journal of Derivates 3, 73–84.

Lahiri, S.N., 1999. Theoretical comparisons of block bootstrap methods. Annals of Statistics 27, 386–404.

Lang, K., Auer, B., 2020. The economic and financial properties of crude oil: A review. North American Journal of Economics and Finance , forthcoming.

Laporta, A., Merlo, L., Petrella, L., 2018. Selection of value at risk models for energy commodities. Energy Economics 74, 628–643.

Le, T., 2020. Forecasting value at risk and expected shortfall with mixed data sampling. International Journal of Forecasting , forthcoming.

Ledoit, O., Wolf, M., 2008. Robust performance hypothesis testing with the Sharpe ratio. Journal of Empirical Finance 15, 850–859.

Lewellen, J., 2002. Momentum and autocorrelation in stock returns. Review of Financial Studies 15, 533–563.

Liu, C.A., 2015. Distribution theory of the least squares averaging estimator. Journal of Econometrics 186, 142–159.

Ljung, G.M., Box, G.E., 1978. On a measure of lack of fit in time series models. Biometrika 65, 297–303.

Lo, A.W., 1991. Long-term memory in stock market prices. Econometrica 59, 1279–1313.

Lo, A.W., MacKinlay, A.C., 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. Review of Financial Studies 1, 41–66.

Lo, A.W., MacKinlay, A.C., 1989. The size and power of the variance ratio test in finite samples: A Monte Carlo investigation. Journal of Econometrics 40, 203–238.

Lo, A.W., MacKinlay, A.C., 1990. When are contrarian profits due to stock market overreaction? Review of Financial Studies 3, 175–205.

Locke, P.R., Venkatesh, P., 1997. Futures market transaction costs. Journal of Futures Markets 17, 229–245.

López-García, M.N., Trinidad-Segovia, J.E., Sánchez-Granero, M.A., Pouchkarev, I., 2021. Extending the Fama and French model with a long term memory factor. European Journal of Operational Research 291, 421–426.

Lösner, R., Wied, D., Ziggel, D., 2019. New backtests for unconditional coverage of expected shortfall. Journal of Risk 21, 39–59.

Lucey, B., Li, S., 2015. Is gold a safe haven? International evidence. Applied Economics Letters 22, 35–45.

Mac Gillivray, H., 1992. Shape properties of the g-and-h and Johnson families. Communications in Statistics - Theory and Methods 21, 1233–1250.

Majewska, E., Olbrys, J., 2017. Formal identification of crises on the Euro area stock markets, 2004-2015, in: N., T., A., V. (Eds.), Advances in applied economic research. Springer, Cham, pp. 167–180.

Mancini, L., Trojani, F., 2011. Robust value at risk prediction. Journal of Financial Econometrics 9, 281–313.

Mandelbrot, B.B., 1971. When can price be arbitraged efficiently? A limit to the validity of the random walk and martingale models. Review of Economics and Statistics 53, 225–236.

Mandelbrot, B.B., Van Ness, J.W., 1968. Fractional Brownian motions, fractional noises and applications. SIAM Review 10, 422–437.

Mandelbrot, B.B., Wallis, J.R., 1968. Noah, Joseph, and operational hydrology. Water Resources Research 4, 909–918.

Mandelbrot, B.B., Wallis, J.R., 1969. Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. Water Resources Research 5, 967–988.

Manera, M., Nicolini, M., Vignati, I., 2013. Financial speculation in energy and agriculture futures markets: A multivariate GARCH approach. Energy Journal 34, 55–81.

Manganelli, S., Engle, R., 2001. Value at risk models in finance. European Central Bank Working Paper No. 75.

Mao, M.Q., Wei, K.C.J., 2014. Price and earnings momentum: An explanation using return decomposition. Journal of Empirical Finance 28, 332–351.

Marcjasz, G., Serafin, T., Weron, R., 2018. Selection of calibration windows for day-ahead electricity price forecasting. Energies 11, 2364.

Marimoutou, V., Raggad, B., Trabelsi, A., 2009. Extreme value theory and value at risk: Application to oil market. Energy Economics 31, 519–530.

Marshall, B.R., Cahan, R.H., Cahan, J.M., 2008. Can commodity futures be profitably traded with quantitative market timing strategies? Journal of Banking and Finance 32, 1810–1819.

Marshall, B.R., Nguyen, N.H., Visaltanachoti, N., 2012. Commodity liquidity measurement and transaction costs. Review of Financial Studies 25, 599–638.

Martinez, J., Iglewicz, B., 1984. Some properties of the Tukey g and h family of distributions. Communications in Statistics - Theory and Methods 13, 353–369.

Martins-Filho, C., Yao, F., 2006. Estimation of value-at-risk and expected shortfall based on nonlinear models of return dynamics and extreme value theory. Studies in Nonlinear Dynamics and Econometrics 10, Article 4, 1–41.

Martins-Filho, C., Yao, F., Torero, M., 2018. Nonparametric estimation of conditional value at risk and expected shortfall based on extreme value theory. Econometric Theory 34, 23–67.

McNeil, A.J., 1997. Estimating the tails of loss severity distributions using extreme value theory. Astin Bulletin 27, 117–137.

McNeil, A.J., Frey, R., 2000. Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. Journal of Empirical Finance 7, 271–300.

McNeil, A.J., Frey, R., Embrechts, P., 2005. Quantitative risk management: Concepts, techniques and Tools. Princeton University Press, Princeton.

Meade, N., 2010. Oil prices - Brownian motion or mean reversion? A study using a one year ahead density forecast criterion. Energy Economics 32, 1485–1498.

Mehlitz, J., Auer, B., 2020. A Monte Carlo evaluation of non-parametric estimators of expected shortfall. Journal of Risk Finance 21, 355–397.

Mehlitz, J.S., Auer, B.R., 2021. Time-varying dynamics of expected shortfall in commodity futures markets. Journal of Futures Markets 41, 895–925.

Memmel, C., 2003. Performance hypothesis testing with the Sharpe ratio. Finance Letters 1, 21–23.

Miffre, J., 2016. Long-short commodity investing: A review of the literature. Journal of Commodity Markets 1, 3–13.

Miffre, J., Rallis, G., 2007. Momentum strategies in commodity futures markets. Journal of Banking and Finance 31, 1863–1886.

Mills, T.C., 1995. Modelling skewness and kurtosis in the London Stock Exchange FT-SE index return distributions. Journal of the Royal Statistical Society: Series D 44, 323–332.

Mögel, B., Auer, B., 2018. How accurate are modern value-at-risk estimators derived from extreme value theory? Review of Quantitative Finance and Accounting 50, 979–1030.

Montanari, A., Taqqu, M.S., Teverovsky, V., 1999. Estimating long-range dependence in the presence of periodicity: An empirical study. Mathematical and Computer Modelling 29, 217–228.

Moraux, F., 2011. How valuable is your VaR? Large sample confidence intervals for normal VaR. Journal of Risk Management in Financial Institutions 4, 189–200.

Moskowitz, J., T., Ooi, Y.H., Pedersen, L.H., 2012. Time series momentum. Journal of Financial Economics 104, 228–250.

Mostaghimi, M., 2004. Monetary policy, composite leading economic indicators and predicting the 2001 recession. Journal of Forecasting 23, 463–477.

Nadarajah, S., Zhang, B., Chan, S., 2014. Estimation methods for expected shortfall. Quantitative Finance 14, 271–291.

Nadaraya, E.A., 1964. Some new estimates for distribution functions. Theory of Probability and Its Applications 9, 497–500.

Nadaraya, E.A., 1965. On non-parametric estimates of density functions and regression curves. Theory of Probability and Its Applications 10, 186–190.

Narayan, P.K., Narayan, S., Sharma, S.S., 2013. An analysis of commodity markets: what gain for investors? Journal of Banking & Finance 37, 3878–3889.

National Bureau of Economic Research, 2010. US business cycle expansions and contractions. http://www.nber.org/cycles.html.

Newey, W., Steigerwald, D., 1997. Asymptotic bias for quasi-maximum likelihood estimators in conditional heteroscedasticity models. Econometrica 65, 587–599.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Nguyen, D., Sousa, R., Uddin, G., 2015. Testing for asymmetric causality between U.S. equity returns and commodity futures returns. Finance Research Letters 12, 38–47.

Nolde, N., Ziegel, J., 2017. Elicitability and backtesting: Perspectives for banking regulation. Annals of Applied Statistics 11, 1833–1873.

Novales, A., Garcia-Jorcano, L., 2017. Volatility specifications versus probability distributions in VaR forecasting. https://ssrn.com/abstract=3023885.

Novales, A., Garcia-Jorcano, L., 2019. Backtesting extreme value theory models of expected shortfall. Quantitative Finance 19, 799–825.

Pagan, A.R., Sossounov, K.A., 2003. A simple framework for analysing bull and bear markets. Journal of Applied Econometrics 18, 23–46.

Pan, M.S., Liano, K., Huang, G.C., 2004. Industry momentum strategies and autocorrelations in stock returns. Journal of Empirical Finance 11, 185–202.

Paraschiv, F., Mudry, P., Andries, A., 2015. Stress-testing for portfolios of commodity futures. Economic Modelling 50, 9–18.

Parzen, E., 1979. Nonparametric statistical data modeling. Journal of the American Statistical Association 74, 105–121.

Paschke, R., Prokopczuk, M., Simen, C., 2020. Curve momentum. Journal of Banking and Finance 113, 105718.

Pastor, L., Stambaugh, R.F., 2003. Liquidity risk and expected returns. Journal of Political Economy 111, 642–685.

Peiró, A., 1999. Skewness in financial returns. Journal of Banking and Finance 23, 847–862.

Peng, C.K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., Goldberger, A.L., 1994. Mosaic organization of DNA nucleotides. Physical Review E 49, 1685–1689.

Peracchi, F., Tanase, A.V., 2008. On estimating the conditional expected shortfall. Applied Stochastic Models in Business and Industry 24, 471–493.

Pérignon, C., Smith, D., 2010. The level and quality of value-at-risk disclosure by commercial banks. Journal of Banking and Finance 34, 362–377.

Peters, E., 1992. R/S analysis using logarithmic returns. Financial Analysts Journal 48, 81–82.

Peters, E.E., 1994. Fractal market analysis: Applying chaos theory to investment and economics. John Wiley and Sons, Hoboken.

Pickands, J.I., 1975. Statistical inference using extreme order statistics. Annals of Statistics 3, 119–131.

Pindyck, R., Rubinfeld, D., 1998. Econometric models and economic forecasts. 4 ed., McGraw-Hill, Singapore.

Pitera, M., Schmidt, T., 2018. Unbiased estimation of risk. Journal of Banking and Finance 91, 133–145.

Plante, M., 2019. OPEC in the news. Energy Economics 80, 163–172.

Politis, D.N., Romano, J.P., 1992. A circular block resampling procedure for stationary data, in: Lepage, R., Billard, L. (Eds.), Exploring the limits of bootstrap. Wiley, New York, pp. 263–270.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. Journal of the American Statistical Association 89, 1303–1313.

Pritsker, M., 2006. The hidden dangers of historical simulation. Journal of Banking and Finance 30, 561–582.

Rachev, S., Jašić, T., Stoyanov, S., Fabozzi, F., 2007. Momentum strategies based on reward-risk stock selection criteria. Journal of Empirical Finance 31, 2325–2346.

Ramos-Requena, J., Trinidad-Segovia, J., Sánchez-Granero, M., 2017. Introducing Hurst exponent in pair trading. Physica A: Statistical Mechanics and its Applications 488, 39–45.

Rapalias, F., Liu, J., Thomaskos, D.D., 2021. Return signal momentum. Journal of Banking and Finance 124, 106063.

Raykar, V.C., Duraiswami, R., 2006. Fast optimal bandwidth selection for kernel density estimation. Proceedings of the 2006 SIAM International Conference on Data Mining , 524–528.

Rea, W., Oxley, L., Reale, M., Brown, J., 2013. Not all estimators are born equal: The empirical properties of some estimators of long memory. Mathematics and Computers in Simulation 93, 29–42.

Reboredo, J.C., 2013. Is gold a safe haven or a hedge for the US dollar? Implications for risk management. Journal of Banking and Finance 37, 2665–2676.

Righi, M., Ceretta, P.S., 2013. Individual and flexible expected shortfall backtesting. Journal of Risk Model Validation 7, 3–20.

Romano, J.P., Wolf, M., 2006. Improved nonparametric confidence intervals in time series regressions. Nonparametric Statistics 18, 199–214.

Rossi, P., 2014. Bayesian non- and semi-parametric methods and applications. Princeton University Press, Princeton.

Rouwenhorst, K.G., Tang, K., 2012. Commodity investing. Annual Review of Financial Economics 4, 447–467.

Scaillet, O., 2004. Nonparametric estimation and sensitivity analysis of expected shortfall. Mathematical Finance 14, 115–129.

Scarrott, C., MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT - Statistical Journal 10, 33–60.

Schaumburg, J., 2012. Predicting extreme value at risk: Nonparametric quantile regression with refinements from extreme value theory. Computational Statistics and Data Analysis 56, 4081–4096.

Schuhmacher, F., Auer, B., 2014. Sufficient conditions under which SSD- and MR-efficient sets are identical. European Journal of Operational Research 239, 756–763.

Schuhmacher, F., Eling, M., 2011. Sufficient conditions for expected utility to imply drawdown-based performance rankings. Journal of Banking and Finance 35, 2311–2318.

Sensoy, A., Hacihasanoglu, E., 2014. Time-varying long range dependence in energy futures markets. Energy Economics 46, 318–327.

Serban, A.F., 2010. Combining mean reversion and momentum trading strategies in foreign exchange markets. Journal of Banking and Finance 34, 2720–2727.

Sharpe, W., 1994. The Sharpe ratio. Journal of Portfolio Management 2, 49–58.

Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density. Journal of the Royal Statistical Society 53, 683–690.

Sheather, S.J., Marron, J.S., 1990. Kernel quantile estimators. Journal of the American Statistical Association 85, 410–416.

Shen, Q., Szakmary, A.C., Sharma, S.C., 2007. An examination of momentum strategies in commodity futures markets. Journal of Futures Markets 27, 227–256.

Silvennoinen, A., Thorp, S., 2013. Financialization, crisis and commodity correlation dynamics. Journal of International Financial Markets, Institutions and Money 24, 42–65.

Simonsen, I., Hansen, A., Nes, O.M., 1998. Determination of the Hurst exponent by use of wavelet transforms. Physical Review E 58, 2779.

Sockin, M., Xiong, W., 2015. Informational frictions and commodity markets. Journal of Finance 70, 2063–2098.

Stavroyiannis, S., 2018. Value-at-risk and related measures for the bitcoin. Journal of Risk Finance 19, 127–136.

Studenmund, A.H., 2017. Using econometrics: A practical guide. 7 ed., Pearson, Boston.

Szakmary, A., Shen, Q., Sharma, S., 2010. Trend-following trading strategies in commodity futures: A re-examination. Journal of Banking and Finance 34, 409–426.

Tang, K., Xiong, W., 2012. Index investment and the financialization of commodities. Financial Analysts Journal 68, 54–74.

Taqqu, M., Teverovsky, V., Willinger, W., 1995. Estimators for long-range dependence: An empirical study. Fractals 3, 785–798.

Taqqu, M.S., Teverovsky, V., 1998. On estimating the intensity of long-range dependence in finite and infinite variance time series, in: Adler, R., Feldman, R., Taqqu, M.S. (Eds.), A practical guide to heavy tails: Statistical techniques and applications. Birkhäuser, Basel, pp. 177–218.

Tasche, D., 2002. Expected shortfall and beyond. Journal of Banking and Finance 26, 1519–1533.

Taylor, J., 2008a. Estimating value at risk and expected shortfall using expectiles. Journal of Financial Econometrics 6, 231–252.

Taylor, J., 2008b. Using exponentially weighted quantile regression to estimate value at risk and expected shortfall. Journal of Financial Econometrics 6, 382–406.

Taylor, N., 2014. The rise and fall of technical trading rule success. Journal of Banking and Finance 40, 286–302.

Teverovsky, V., Taqqu, M.S., Willinger, W., 1999. A critical look at Lo's modified R/S statistic. Journal of Statistical Planning and Inference 80, 211–227.

Thavaneswaran, A., Paseka, A., Frank, J., 2019. Generalized value at risk forecasting. Communications in Statistics - Theory and Methods, forthcoming .

Timmermann, A., 2006. Forecast combinations, in: Elliot, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of economic forecasting. Elsevier, Oxford. volume 1, pp. 135–196.

Trapin, L., 2018. Can volatility models explain extreme events? Journal of Financial Econometrics 16, 297–315.

Tsay, R.S., 2015. Analysis of financial time series. 2 ed., John Wiley & Sons, Inc., Hoboken, New Jersey.

Tukey, J.W., 1977. Exploratory data analysis. Addison-Wesley, Reading.

van Dijk, M., 2011. Is size dead? A review of the size effect in equity returns. Journal of Banking and Finance 35, 3263–3274.

Wand, M.P., Jones, M.C., 1995. Kernel smoothing. Chapman and Hall/CRC, Boca Raton.

Wang, C., 2001. Investor sentiment and return predictability in agricultural futures markets. Journal of Futures Markets 21, 929–952.

Wang, H., Zhang, X., Zou, G., 2009. Frequentist model averaging estimation: A review. Journal of Systems Science and Complexity 22, article 792.

Wang, Y., Ma, F., Wei, Y., Wu, C., 2016. Forecasting realized volatility in a changing world: A dynamic model averaging approach. Journal of Banking and Finance 64, 136–149.

Watugala, S., 2019. Economic uncertainty, trading activity, and commodity futures volatility. Journal of Futures Markets 39, 921–945.

Weron, R., 2002. Estimating long-range dependence: Finite sample properties and confidence intervals. Physica A: Statistical Mechanics and its Applications 312, 285–299.

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

Wheeler, R.E., 1980. Quantile estimators of Johnson curve parameters. Biometrika 67, 725–728.

White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.

Wied, D., Weiß, G., Ziggel, D., 2016. Evaluating Value-at-Risk forecasts: A new set of multivariate backtests. Journal of Banking and Finance 72, 121–132.

Wilcox, R., 2012. Introduction to Robust Estimation and Hypothesis Testing. volume 3. Elsevier, Amsterdam.

Willinger, W., Taqqu, M.S., Teverovsky, V., 1999. Stock market prices and long-range dependence. Finance and Stochastics 3, 1–13.

Wong, W.K., 2008. Backtesting trading risk of commercial banks using expected shortfall. Journal of Banking and Finance 32, 1404–1415.

Woodard, J., Egelkraut, T., Garcia, P., Pennings, J., 2011. Effects of full collateralization in commodity futures investments. Journal of Derivatives and Hedge Funds 16, 253–266.

Xu, Y., Iglewicz, B., Chervoneva, I., 2014. Robust estimation of the parameters of g-and-h distributions, with applications to outlier detection. Computational Statistics and Data Analysis 75, 66–80.

Yamai, Y., Yoshiba, T., 2005. Value-at-risk versus expected shortfall: A practical perspective. Journal of Banking and Finance 29, 997–1015.

Yao, H., Li, Z., Lai, Y., 2013. Mean-CVaR portfolio selection: A nonparametric estimation framework. Computers and Operations Research 40, 1014–1022.

Ye, M., Zyren, J., Shore, J., 2005. A monthly crude oil spot price forecasting model using relative inventories. International Journal of Forecasting 21, 491–501.

Youssef, M., Belkacem, L., Mokni, K., 2015. Value-at-risk estimation of energy commodities: A long-memory GARCH-EVT approach. Energy Economics 51, 99–110.

Yu, K., Ally, A., Yang, S., Hand, D., 2010. Kernel quantile-based estimation of expected shortfall. Journal of Risk 12, 15–32.

Zaremba, A., Mikutowski, M., Szczygielski, J., Karathanasopoulos, A., 2021. The alpha momentum effect in commodity markets. Energy Economics 93, 104421.

Ziggel, D., Berens, T., Weiß, G., Wied, D., 2014. A new set of improved value-at-risk backtests. Journal of Banking and Finance 48, 29–41.

Appendix

A. Supplementary results for Chapter 1

A.1. Additional tables

Table A.1.: MPE of ES estimates for $\gamma = 95\,\%$ and varied n

	ND	POT	H	H1	H2	H3	J1	J2	K1	K2	MV
$n = 21$											
(a)	-21.69	-20.25	-24.79	-22.61	-25.28	-25.28	-24.55	-33.25	-11.56	78.75	-13.05
(b)	-20.73	-23.21	-27.30	-25.35	-27.83	-27.83	-27.28	-37.38	-12.19	66.75	-16.24
(c)	28.15	-12.43	-15.70	-12.62	-16.02	-16.02	-15.74	-21.29	-8.28	88.44	-0.15
(d)	-1.08	-20.16	-23.82	-21.52	-24.29	-24.29	-23.86	-32.63	-10.82	70.69	-11.18
(e)	-1.20	-15.37	-19.35	-16.67	-19.76	-19.76	-19.50	-26.23	-9.46	84.93	-6.24
(f)	-3.31	-18.28	-22.19	-19.75	-22.64	-22.64	-22.19	-30.16	-10.46	77.91	-9.37
$n = 126$											
(a)	-20.15	8.06	-4.79	-1.61	-4.96	-5.86	-4.88	-12.84	-2.00	1.25	-4.78
(b)	-18.03	-4.79	-5.43	-2.56	-5.65	-6.72	-5.47	-16.69	-2.92	-2.07	-7.03
(c)	30.49	-1.75	-3.03	0.76	-3.14	-3.69	-2.98	-7.67	-1.70	10.44	1.77
(d)	2.03	0.29	-4.84	-1.72	-5.02	-5.95	-4.69	-14.24	-2.55	-1.10	-3.78
(e)	-0.20	-1.00	-3.74	-0.15	-3.87	-4.53	-3.72	-8.95	-1.79	6.17	-2.18
(f)	-1.17	0.16	-4.37	-1.06	-4.53	-5.35	-4.35	-12.08	-2.19	2.94	-3.20
$n = 252$											
(a)	-20.07	-0.32	-1.95	1.37	-2.05	-3.17	-7.12	-11.48	-1.27	-1.26	-4.73
(b)	-17.76	0.43	-2.17	0.83	-2.30	-3.66	-10.27	-15.86	-2.58	-3.85	-5.72
(c)	30.70	-6.24	-1.23	2.66	-1.29	-1.96	-4.19	-6.79	-1.26	4.88	1.53
(d)	2.44	-0.42	-1.83	1.41	-1.94	-3.12	-8.76	-13.56	-2.13	-2.27	-3.02
(e)	-0.09	-0.33	-1.46	2.25	-1.53	-2.34	-4.66	-7.69	-1.04	2.70	-1.42
(f)	-0.96	-1.38	-1.73	1.70	-1.82	-2.85	-7.00	-11.08	-1.66	0.04	-2.67
$n = 504$											
(a)	-19.97	-0.12	-1.38	-0.82	-1.43	-1.59	-10.61	-12.06	-0.62	-1.27	-4.99
(b)	-17.47	0.25	-1.58	-1.08	-1.64	-1.84	-15.05	-16.82	-2.57	-3.66	-6.15
(c)	30.86	-0.02	-0.87	-0.22	-0.90	-1.00	-6.17	-7.03	-1.20	2.86	1.63
(d)	2.85	0.19	-1.32	-0.78	-1.38	-1.54	-12.87	-14.40	-2.39	-2.24	-3.39
(e)	-0.04	-0.07	-1.04	-0.41	-1.07	-1.19	-6.86	-7.89	-0.51	1.31	-1.78
(f)	-0.75	0.05	-1.24	-0.66	-1.28	-1.43	-10.31	-11.64	-1.46	-0.60	-2.93
$n = 1008$											
(a)	-19.94	-0.23	-0.58	-0.01	-0.60	-0.79	-10.93	-11.76	-0.49	-1.01	-4.63
(b)	-17.39	-0.21	-0.69	-0.18	-0.73	-0.96	-15.66	-16.66	-3.50	-3.67	-5.97
(c)	30.91	-0.06	-0.37	0.29	-0.38	-0.50	-6.34	-6.83	-2.07	1.69	1.63
(d)	2.94	-0.13	-0.55	0.00	-0.58	-0.78	-13.37	-14.23	-3.61	-1.91	-3.22
(e)	-0.01	-0.16	-0.44	0.19	-0.46	-0.60	-7.00	-7.59	-0.39	0.80	-1.57
(f)	-0.70	-0.16	-0.53	0.06	-0.55	-0.73	-10.66	-11.41	-2.01	-0.82	-2.75

For a confidence level of $\gamma = 95\,\%$ and varied sample sizes n, this table presents the mean percentage error (MPE) of the expected shortfall (ES) estimates produced by our parametric and non-parametric techniques. Simulation settings and methods are specified and abbreviated as in Table 1.1.

Table A.2.: MPE of ES estimates for $\gamma = 97.5\,\%$ and varied n

	ND	POT	H	H1	H2	H3	J1	J2	K1	K2	MV
$n = 21$											
(a)	-25.81	-31.89	-24.97	25.92	-25.28	-31.31	-24.92	-31.61	-24.92	280.75	8.60
(b)	-28.63	-38.20	-28.90	15.36	-29.24	-35.95	-28.52	-36.00	-29.86	227.22	-1.27
(c)	30.20	-20.54	-16.35	50.92	-16.55	-20.65	-16.38	-20.89	-16.37	307.07	26.05
(d)	-8.35	-33.99	-25.77	24.39	-26.07	-32.09	-25.54	-32.23	-26.18	275.33	8.95
(e)	-0.76	-23.95	-18.96	42.73	-19.20	-24.04	-19.25	-24.59	-18.84	335.76	22.89
(f)	-6.67	-29.71	-22.99	31.86	-23.27	-28.81	-22.92	-29.06	-23.23	285.74	13.06
$n = 126$											
(a)	-24.38	-3.42	-8.30	-5.07	-8.42	-9.02	-8.37	-15.96	-2.64	30.22	-5.54
(b)	-26.18	-8.70	-10.12	-7.26	-10.27	-11.01	-10.09	-20.90	-4.25	28.22	-8.06
(c)	32.56	-0.06	-5.33	-1.54	-5.41	-5.78	-5.29	-9.97	-1.83	22.38	1.97
(d)	-5.35	5.29	-9.11	-6.03	-9.24	-9.89	-9.01	-18.44	-3.70	18.00	-4.75
(e)	-0.21	-2.06	-6.14	-2.49	-6.23	-6.66	-6.12	-11.09	-2.02	22.92	-2.01
(f)	-4.71	-1.79	-7.80	-4.48	-7.91	-8.47	-7.78	-15.27	-2.89	24.35	-3.68
$n = 252$											
(a)	-24.27	-1.12	-4.10	-0.67	-4.18	-4.99	-10.54	-14.53	-1.70	3.62	-6.25
(b)	-25.89	-0.13	-4.98	-1.92	-5.07	-6.13	-15.02	-20.19	-3.73	0.99	-8.21
(c)	32.77	-0.99	-2.57	1.36	-2.61	-3.13	-6.53	-9.01	-1.54	8.97	1.67
(d)	-5.01	-1.06	-4.52	-1.25	-4.60	-5.53	-13.21	-17.77	-3.61	-0.49	-5.71
(e)	-0.09	-1.38	-2.93	0.90	-2.99	-3.56	-6.90	-9.66	-1.22	7.34	-2.05
(f)	-4.50	-0.94	-3.82	-0.32	-3.89	-4.67	-10.44	-14.23	-2.36	4.09	-4.11
$n = 504$											
(a)	-24.18	-0.69	-1.59	1.97	-1.63	-2.61	-9.08	-12.00	-1.07	-1.08	-5.20
(b)	-25.64	-0.01	-1.79	1.39	-1.85	-3.14	-13.55	-17.34	-4.32	-4.66	-7.09
(c)	32.91	-0.33	-1.00	3.01	-1.03	-1.63	-5.53	-7.35	-2.29	4.76	2.15
(d)	-4.61	-0.17	-1.62	1.75	-1.67	-2.80	-11.88	-15.21	-4.89	-2.99	-4.41
(e)	-0.04	-0.56	-1.15	2.78	-1.18	-1.85	-5.65	-7.67	-0.82	3.52	-1.26
(f)	-4.31	-0.35	-1.43	2.18	-1.47	-2.41	-9.14	-11.91	-2.68	-0.09	-3.16
$n = 1008$											
(a)	-24.17	-0.43	-1.19	-0.59	-1.21	-1.36	-10.23	-11.29	-0.74	-1.20	-5.24
(b)	-25.57	-0.30	-1.42	-0.89	-1.45	-1.66	-15.45	-16.81	-5.84	-4.93	-7.43
(c)	32.94	-0.12	-0.75	-0.08	-0.77	-0.86	-6.24	-6.90	-4.71	2.22	1.47
(d)	-4.53	-0.33	-1.30	-0.74	-1.33	-1.51	-13.50	-14.70	-7.67	-2.91	-4.85
(e)	-0.02	-0.30	-0.83	-0.17	-0.84	-0.95	-6.36	-7.09	-0.86	1.73	-1.57
(f)	-4.27	-0.30	-1.10	-0.49	-1.12	-1.27	-10.36	-11.36	-3.96	-1.02	-3.52

For a confidence level of $\gamma = 97.5\,\%$ and varied sample sizes n, this table presents the mean percentage error (MPE) of the expected shortfall (ES) estimates produced by our parametric and non-parametric techniques. Simulation settings and methods are specified and abbreviated as in Table 1.1.

Table A.3.: MPE of ES estimates for $\gamma = 99\,\%$ and varied n

	ND	POT	H	H1	H2	H3	J1	J2	K1	K2	MV
$n = 21$											
(a)	-30.58	-51.35	-38.44	3.27	-38.54	-40.53	-38.35	-40.54	-32.19	668.84	36.16
(b)	-38.12	-62.97	-45.67	-12.04	-45.77	-47.84	-45.64	-47.91	-41.07	517.32	13.03
(c)	30.71	-34.22	-26.26	33.08	-26.33	-27.77	-26.28	-27.87	-19.84	770.93	64.62
(d)	-18.15	-57.09	-41.77	-2.54	-41.86	-43.76	-41.56	-43.66	-36.44	621.55	29.47
(e)	-1.22	-38.32	-28.99	24.84	-29.08	-30.80	-29.20	-31.08	-22.15	842.28	65.63
(f)	-11.47	-48.79	-36.23	9.32	-36.32	-38.14	-36.21	-38.21	-30.34	685.80	41.81
$n = 126$											
(a)	-29.33	16.11	-13.41	0.39	-13.48	-15.12	-13.36	-21.02	-6.05	312.93	21.77
(b)	-36.13	-13.64	-17.34	-5.54	-17.42	-19.55	-17.31	-27.92	-9.97	301.37	13.66
(c)	33.19	0.89	-8.91	7.19	-8.95	-10.04	-8.87	-13.92	-2.84	121.15	10.89
(d)	-15.58	18.44	-15.87	-3.23	-15.95	-17.85	-15.81	-25.28	-8.61	281.70	18.20
(e)	-0.21	-2.18	-9.40	6.42	-9.45	-10.61	-9.40	-14.55	-3.53	193.10	14.02
(f)	-9.61	3.92	-12.99	1.05	-13.05	-14.63	-12.95	-20.54	-6.20	242.00	15.83
$n = 252$											
(a)	-29.18	-2.48	-6.39	8.84	-6.44	-8.96	-6.47	-15.62	-3.90	157.24	8.66
(b)	-35.88	1.07	-8.36	4.94	-8.43	-11.89	-8.37	-21.90	-8.79	182.20	8.46
(c)	33.41	-10.37	-4.24	12.84	-4.27	-5.93	-4.26	-10.20	-3.53	34.03	3.75
(d)	-15.19	0.03	-7.62	6.39	-7.68	-10.80	-7.48	-19.71	-8.43	142.40	7.19
(e)	-0.09	-3.03	-4.42	12.56	-4.45	-6.15	-4.43	-10.24	-2.67	57.32	3.44
(f)	-9.39	-2.96	-6.21	9.11	-6.25	-8.75	-6.20	-15.53	-5.46	114.60	6.43
$n = 504$											
(a)	-29.09	-1.75	-4.91	-4.31	-4.94	-5.02	-10.92	-13.52	0.78	25.20	-4.85
(b)	-35.65	-0.84	-6.61	-6.09	-6.65	-6.77	-16.49	-20.13	-3.79	38.93	-6.41
(c)	33.59	-1.33	-3.24	-2.58	-3.26	-3.32	-7.07	-8.76	-4.23	11.93	1.17
(d)	-14.95	-1.20	-6.03	-5.48	-6.06	-6.17	-14.82	-18.10	-4.37	15.91	-6.13
(e)	-0.04	-1.58	-3.33	-2.67	-3.35	-3.41	-6.84	-8.53	-0.15	11.98	-1.79
(f)	-9.23	-1.34	-4.82	-4.23	-4.85	-4.94	-11.23	-13.81	-2.35	20.79	-3.49
$n = 1008$											
(a)	-29.10	-0.99	-2.57	-1.96	-2.59	-2.71	-9.50	-11.20	-0.80	0.68	-6.07
(b)	-35.59	-0.83	-3.42	-2.87	-3.44	-3.60	-15.10	-17.50	-7.06	-1.04	-9.05
(c)	33.63	-0.63	-1.68	-1.01	-1.69	-1.77	-6.10	-7.21	-12.12	4.40	0.58
(d)	-14.85	-0.99	-3.10	-2.53	-3.12	-3.27	-13.51	-15.68	-7.08	-2.84	-6.70
(e)	-0.02	-0.77	-1.71	-1.03	-1.72	-1.80	-5.73	-6.83	-2.87	4.34	-1.81
(f)	-9.19	-0.84	-2.50	-1.88	-2.51	-2.63	-9.99	-11.68	-5.99	1.11	-4.47

For a confidence level of $\gamma = 99\,\%$ and varied sample sizes n, this table presents the mean percentage error (MPE) of the expected shortfall (ES) estimates produced by our parametric and non-parametric techniques. Simulation settings and methods are specified and abbreviated as in Table 1.1.

Table A.4.: MAPE and RSD of ES estimates for $\gamma = 95\,\%$ and varied n

	ND	POT	H	H1	H2	H3	J1	J2	K1	K2	MV
MAPE, $n = 21$											
(a)	26.21	32.39	31.18	30.05	31.37	31.37	31.08	35.84	30.42	93.52	37.34
(b)	28.58	39.22	37.07	36.02	37.22	37.22	37.12	40.97	37.98	86.70	41.81
(c)	29.83	19.73	19.47	17.78	19.61	19.61	19.42	22.76	19.29	90.55	27.81
(d)	20.43	33.97	32.13	30.84	32.27	32.27	32.20	35.73	33.18	83.57	36.66
(e)	15.14	23.52	23.41	22.01	23.60	23.60	23.44	27.90	22.09	92.03	29.67
(f)	24.04	29.77	28.65	27.34	28.81	28.81	28.65	32.64	28.59	89.27	34.66
MAPE, $n = 126$											
(a)	20.36	30.43	12.66	12.22	12.68	12.80	12.63	15.42	12.66	19.37	16.12
(b)	19.90	46.33	16.99	16.44	16.99	17.05	17.09	19.74	16.76	21.74	20.90
(c)	30.49	11.85	7.55	7.16	7.56	7.65	7.53	9.25	7.68	21.71	11.84
(d)	9.90	40.57	14.54	13.94	14.55	14.61	14.58	16.90	14.73	20.20	17.45
(e)	6.12	11.74	8.72	8.34	8.74	8.87	8.67	10.85	8.65	19.13	9.98
(f)	17.35	28.18	12.09	11.62	12.10	12.20	12.10	14.43	12.10	20.43	15.26
MAPE, $n = 252$											
(a)	20.09	9.22	8.97	8.95	8.97	9.06	10.24	12.76	8.99	12.54	10.98
(b)	18.64	15.72	12.31	12.11	12.30	12.33	13.66	17.03	11.91	14.87	14.09
(c)	30.70	11.28	5.30	5.64	5.30	5.37	6.05	7.55	5.48	14.49	9.72
(d)	7.58	13.03	10.58	10.40	10.57	10.6	11.70	14.59	10.67	14.56	11.43
(e)	4.32	6.02	6.05	6.31	6.06	6.15	6.93	8.68	6.08	12.91	6.95
(f)	16.27	11.05	8.64	8.68	8.64	8.70	9.72	12.12	8.63	13.87	10.63
MAPE, $n = 504$											
(a)	19.97	6.29	6.32	6.26	6.32	6.33	11.04	12.31	6.31	8.87	9.00
(b)	17.90	9.12	8.83	8.77	8.84	8.84	15.37	16.99	8.51	10.81	11.40
(c)	30.86	3.71	3.74	3.68	3.74	3.75	6.46	7.21	4.00	10.41	7.76
(d)	5.89	7.75	7.58	7.51	7.58	7.58	13.16	14.56	7.73	10.59	8.99
(e)	3.04	4.22	4.28	4.23	4.29	4.30	7.28	8.15	4.28	9.08	5.32
(f)	15.53	6.22	6.15	6.09	6.15	6.16	10.66	11.84	6.17	9.95	8.49
MAPE, $n = 1008$											
(a)	19.94	4.46	4.48	4.47	4.48	4.49	10.99	11.79	4.47	6.38	7.59
(b)	17.58	6.33	6.29	6.26	6.29	6.30	15.68	16.67	6.41	8.07	9.59
(c)	30.91	2.63	2.65	2.64	2.65	2.65	6.37	6.85	3.37	7.21	6.79
(d)	4.62	5.40	5.38	5.35	5.38	5.38	13.39	14.24	6.09	7.93	7.32
(e)	2.15	2.98	3.02	3.01	3.02	3.03	7.07	7.63	3.03	6.55	4.15
(f)	15.04	4.36	4.36	4.35	4.36	4.37	10.7	11.44	4.67	7.23	7.09
RSD, $n = 21$											
(a)	28.12	42.04	35.52	35.26	35.42	35.42	35.59	34.79	40.72	64.79	26.42
(b)	34.84	55.17	46.95	46.34	46.73	46.73	46.69	41.28	56.08	69.91	32.92
(c)	18.12	23.65	20.08	19.86	20.00	20.00	19.99	19.25	24.37	58.13	16.67
(d)	28.69	45.82	38.49	37.91	38.28	38.28	38.74	33.78	48.06	69.11	27.57
(e)	19.15	28.87	25.02	24.88	24.97	24.97	24.97	25.31	28.06	61.20	24.11
(f)	25.78	39.11	33.21	32.85	33.08	33.08	33.20	30.88	39.46	64.63	25.54
RSD, $n = 126$											
(a)	11.69	1668.37	15.66	15.50	15.63	15.52	15.61	14.93	16.06	32.36	189.73
(b)	16.25	1056.80	21.92	21.58	21.87	21.64	22.21	18.80	21.55	33.58	109.25
(c)	7.60	268.86	9.11	8.99	9.09	9.01	9.12	8.60	9.61	44.69	27.07
(d)	14.22	865.01	18.65	18.32	18.60	18.38	18.65	15.81	19.11	32.84	91.09
(e)	7.67	195.68	10.55	10.47	10.53	10.48	10.50	10.49	10.83	35.85	21.68
(f)	11.49	810.94	15.18	14.97	15.15	15.01	15.22	13.73	15.43	35.86	87.76
RSD, $n = 252$											
(a)	8.25	26.11	11.23	11.11	11.22	11.12	10.95	10.78	11.31	21.16	8.55
(b)	12.23	153.64	15.86	15.61	15.84	15.62	13.97	13.24	15.04	21.85	19.67
(c)	5.39	1164.05	6.58	6.49	6.57	6.50	6.34	6.17	6.83	36.50	107.61
(d)	10.62	192.95	13.53	13.30	13.52	13.31	11.84	11.18	13.46	25.31	21.88
(e)	5.41	7.58	7.51	7.45	7.51	7.46	7.50	7.49	7.59	28.73	6.48
(f)	8.38	308.87	10.94	10.79	10.93	10.8	10.12	9.77	10.85	26.71	32.84
RSD, $n = 504$											
(a)	5.81	7.93	7.89	7.88	7.89	7.88	7.59	7.57	7.93	16.08	5.78
(b)	9.42	13.22	11.27	11.24	11.27	11.24	9.37	9.25	10.51	16.39	8.00
(c)	3.83	4.68	4.64	4.63	4.64	4.63	4.40	4.37	4.93	31.66	4.44
(d)	8.11	10.22	9.63	9.60	9.62	9.60	7.94	7.82	9.52	19.97	6.79
(e)	3.81	5.29	5.30	5.30	5.30	5.30	5.36	5.36	5.36	23.62	4.77
(f)	6.20	8.27	7.75	7.73	7.74	7.73	6.93	6.87	7.65	21.54	5.95
RSD, $n = 1008$											
(a)	4.12	5.60	5.62	5.61	5.61	5.61	5.40	5.39	5.61	12.93	4.16
(b)	6.79	8.03	7.98	7.96	7.98	7.96	6.60	6.55	7.34	12.86	5.62
(c)	2.71	3.30	3.30	3.30	3.30	3.30	3.10	3.09	3.71	26.22	3.40
(d)	5.94	6.83	6.81	6.79	6.81	6.79	5.56	5.52	6.79	16.71	4.86
(e)	2.70	3.74	3.77	3.76	3.77	3.76	3.80	3.80	3.79	19.57	3.55
(f)	4.45	5.50	5.50	5.48	5.50	5.48	4.89	4.87	5.45	17.66	4.32

For the confidence level $\gamma = 95\,\%$, this table extends the results of Table 1.4 to other sample sizes n.

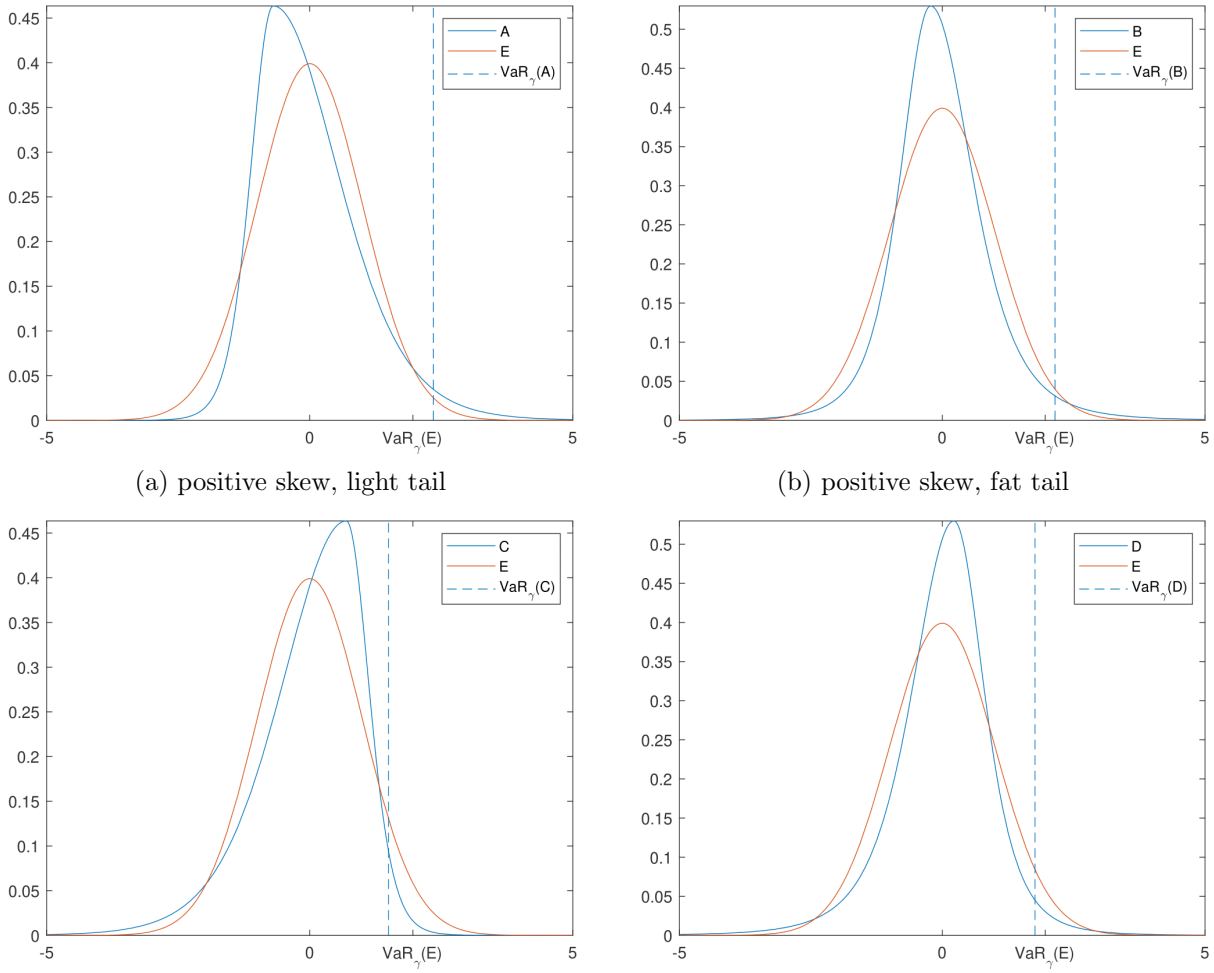
Table A.5.: MAPE and RSD of ES estimates for $\gamma = 99\,\%$ and varied n

	ND	POT	H	H1	H2	H3	J1	J2	K1	K2	MV
MAPE, $n = 21$											
(a)	31.96	53.73	41.57	28.39	41.62	42.78	41.54	42.80	37.70	673.48	103.56
(b)	40.03	66.13	50.70	32.90	50.75	51.73	50.73	51.80	47.08	527.35	96.92
(c)	31.95	35.79	28.23	34.98	28.27	29.17	28.20	29.20	24.47	771.07	104.13
(d)	24.81	60.06	46.33	27.53	46.37	47.29	46.22	47.23	42.82	624.92	101.36
(e)	14.17	39.84	30.87	31.69	30.92	32.04	31.03	32.27	26.82	842.62	111.23
(f)	28.58	51.11	39.54	31.10	39.59	40.60	39.54	40.66	35.78	689.44	96.13
MAPE, $n = 126$											
(a)	29.34	74.35	19.64	16.06	19.65	20.08	19.53	22.94	19.19	324.81	56.56
(b)	36.42	137.72	28.11	23.61	28.12	28.41	28.07	31.05	28.14	319.16	68.88
(c)	33.20	35.53	12.82	11.50	12.83	13.14	12.83	15.16	12.75	125.95	28.57
(d)	17.32	118.10	25.23	20.54	25.24	25.53	25.29	28.06	25.97	297.59	60.89
(e)	5.71	26.54	12.90	11.86	12.92	13.33	12.86	15.72	12.14	200.64	32.46
(f)	24.40	78.45	19.74	16.71	19.75	20.10	19.72	22.59	19.64	253.55	48.13
MAPE, $n = 252$											
(a)	29.19	15.94	14.01	14.70	14.02	14.36	14.06	17.38	14.05	176.30	32.40
(b)	35.99	30.31	21.29	19.58	21.29	21.40	21.21	24.46	21.39	207.67	42.46
(c)	33.41	17.52	9.08	13.72	9.08	9.34	9.10	11.36	10.00	46.13	16.87
(d)	16.14	25.80	19.21	17.69	19.21	19.34	19.17	22.04	20.88	166.85	34.63
(e)	4.03	9.09	8.82	13.69	8.82	9.19	8.78	11.40	8.71	70.83	15.34
(f)	23.75	19.73	14.48	15.88	14.48	14.73	14.46	17.33	15.01	133.55	27.66
MAPE, $n = 504$											
(a)	29.09	10.35	10.16	9.97	10.16	10.18	12.59	14.40	10.95	43.01	16.09
(b)	35.73	17.05	15.81	15.63	15.81	15.82	18.69	21.17	15.84	64.76	23.63
(c)	33.59	6.59	6.60	6.38	6.60	6.61	8.17	9.36	9.70	25.58	11.92
(d)	15.41	15.02	14.15	13.96	14.15	14.16	16.82	19.06	16.38	39.60	17.87
(e)	2.84	6.15	6.34	6.13	6.34	6.35	8.00	9.19	6.82	24.78	8.30
(f)	23.33	11.03	10.61	10.41	10.61	10.62	12.85	14.64	11.94	39.55	15.20
MAPE, $n = 1008$											
(a)	29.10	7.20	7.21	7.07	7.22	7.23	10.33	11.67	8.01	14.73	10.98
(b)	35.62	11.70	11.54	11.41	11.55	11.55	16.01	17.96	12.15	21.10	16.06
(c)	33.63	4.58	4.64	4.49	4.64	4.65	6.66	7.53	13.80	16.69	10.13
(d)	15.06	10.36	10.34	10.20	10.35	10.35	14.34	16.10	12.82	16.66	12.66
(e)	2.01	4.27	4.41	4.27	4.41	4.42	6.35	7.20	6.43	15.35	5.91
(f)	23.08	7.62	7.63	7.49	7.63	7.64	10.74	12.09	10.64	16.91	10.89
RSD, $n = 21$											
(a)	26.45	108.09	40.93	35.69	40.87	39.66	41.02	39.67	41.59	62.29	37.35
(b)	33.23	192.55	58.39	46.63	58.28	56.00	58.88	56.36	53.70	68.17	40.85
(c)	18.21	53.28	23.63	20.01	23.58	22.72	23.51	22.58	25.45	83.45	44.84
(d)	29.24	151.04	50.27	39.12	50.16	47.97	49.75	47.37	48.70	71.05	41.95
(e)	17.91	65.34	27.57	25.01	27.53	26.82	27.62	26.85	28.81	65.54	43.36
(f)	25.01	114.06	40.16	33.29	40.09	38.63	40.16	38.56	39.65	70.10	41.67
RSD, $n = 126$											
(a)	10.89	3710.57	21.94	20.61	21.91	21.26	21.83	19.33	24.83	100.78	355.64
(b)	16.03	5120.62	35.60	32.86	35.54	34.23	35.72	28.13	38.31	92.29	391.19
(c)	7.70	1909.28	13.53	12.59	13.51	13.05	13.61	11.77	16.64	146.16	176.23
(d)	14.10	3028.81	30.91	28.42	30.87	29.66	31.10	24.42	36.04	115.69	306.32
(e)	7.17	765.50	13.61	12.93	13.60	13.25	13.53	12.64	15.15	133.51	75.70
(f)	11.18	2906.96	23.12	21.48	23.09	22.29	23.16	19.26	26.19	117.69	261.02
RSD, $n = 252$											
(a)	7.75	71.35	17.01	15.96	16.99	16.21	17.00	14.76	17.79	128.38	32.59
(b)	11.79	339.34	28.04	25.89	28.01	26.42	27.77	21.33	27.38	114.17	45.88
(c)	5.43	2777.16	10.71	9.97	10.70	10.14	10.76	9.18	13.08	111.57	240.40
(d)	10.74	133.45	25.26	23.18	25.23	23.68	24.95	18.75	27.30	142.82	37.49
(e)	5.05	11.68	10.32	9.81	10.32	9.93	10.29	9.55	10.74	132.62	22.36
(f)	8.15	666.60	18.27	16.96	18.25	17.28	18.15	14.71	19.26	125.91	75.74
RSD, $n = 504$											
(a)	5.47	13.20	12.03	12.00	12.02	12.01	11.12	10.80	13.76	113.33	16.96
(b)	9.20	26.88	20.10	20.03	20.09	20.05	16.23	15.33	20.24	125.19	22.56
(c)	3.88	8.27	7.66	7.64	7.66	7.64	6.97	6.72	12.64	76.09	9.73
(d)	8.16	21.57	17.70	17.64	17.69	17.66	14.39	13.55	21.16	115.71	18.39
(e)	3.56	7.63	7.31	7.29	7.30	7.29	7.12	7.02	8.67	67.30	9.31
(f)	6.06	15.51	12.96	12.92	12.95	12.93	11.17	10.69	15.29	99.52	15.39
RSD, $n = 1008$											
(a)	3.87	9.07	8.77	8.75	8.77	8.75	8.01	7.86	10.06	47.28	7.79
(b)	6.62	15.17	14.68	14.62	14.67	14.63	11.60	11.22	14.36	66.52	11.88
(c)	2.74	5.74	5.59	5.57	5.59	5.57	5.05	4.94	13.97	63.12	7.54
(d)	6.05	13.36	13.20	13.15	13.19	13.15	10.30	9.92	15.65	46.55	9.82
(e)	2.52	5.32	5.29	5.28	5.29	5.28	5.12	5.08	7.96	50.49	6.56
(f)	4.36	9.73	9.50	9.47	9.50	9.48	8.01	7.80	12.40	54.79	8.72

For the confidence level $\gamma = 99\,\%$, this table extends the results of Table 1.4 to other sample sizes n.

A.2. Additional figures

Figure A.1.: Densities of distributional settings



(a) positive skew, light tail

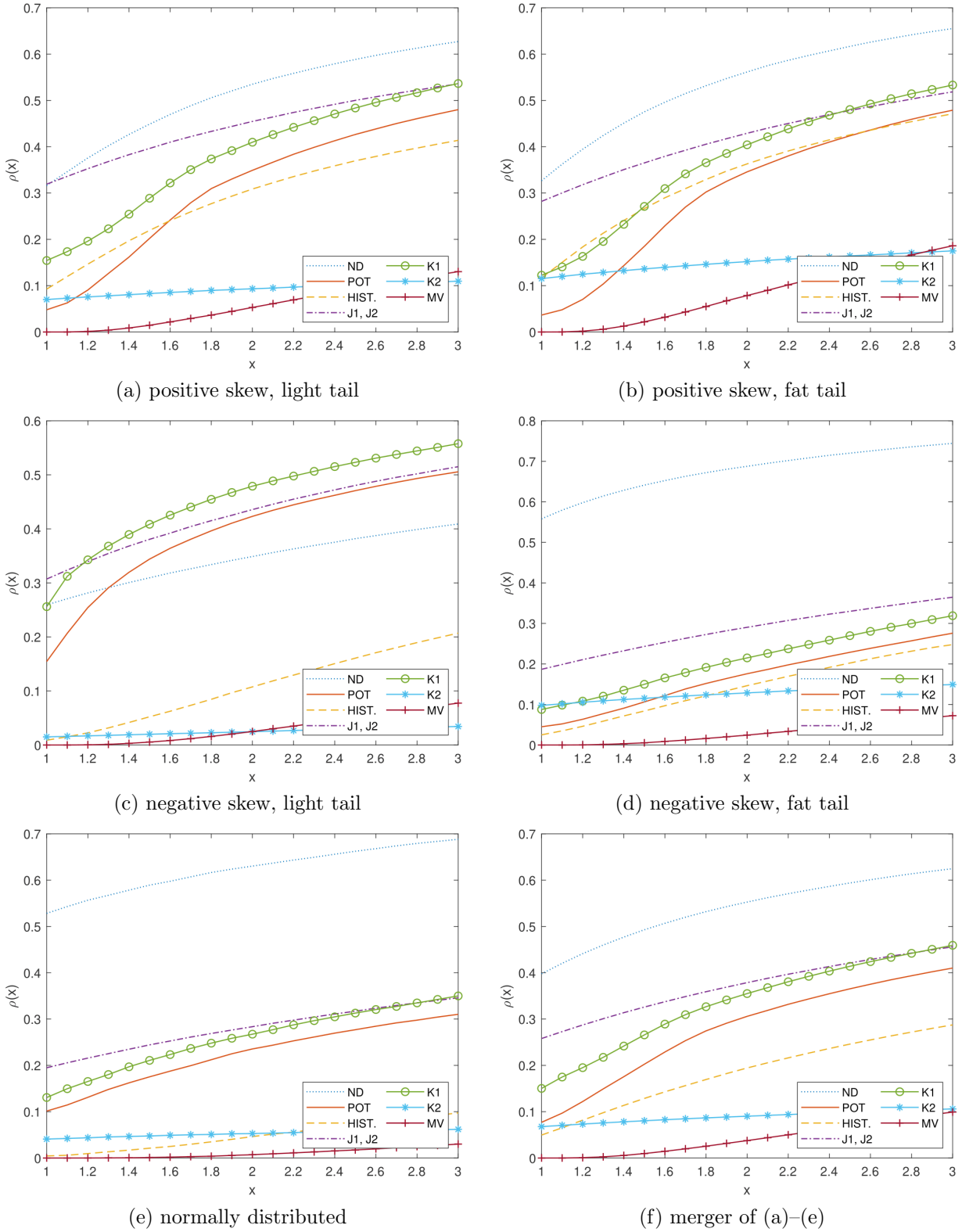
(b) positive skew, fat tail

(c) negative skew, light tail

(d) negative skew, fat tail

This figure illustrates the densities of the four distributional settings (a)–(d) defined in Section 1.3.3 in comparison to the normally distributed setting (e). The dashed line (entry on the x-axis) visualizes the value at risk (VaR) of the former (latter) for an exemplary confidence level of $\gamma = 97.5\,\%$.

Figure A.2.: Performance profiles for $\gamma = 97.5\,\%$ and $n = 21$



(a) positive skew, light tail

(b) positive skew, fat tail

(c) negative skew, light tail

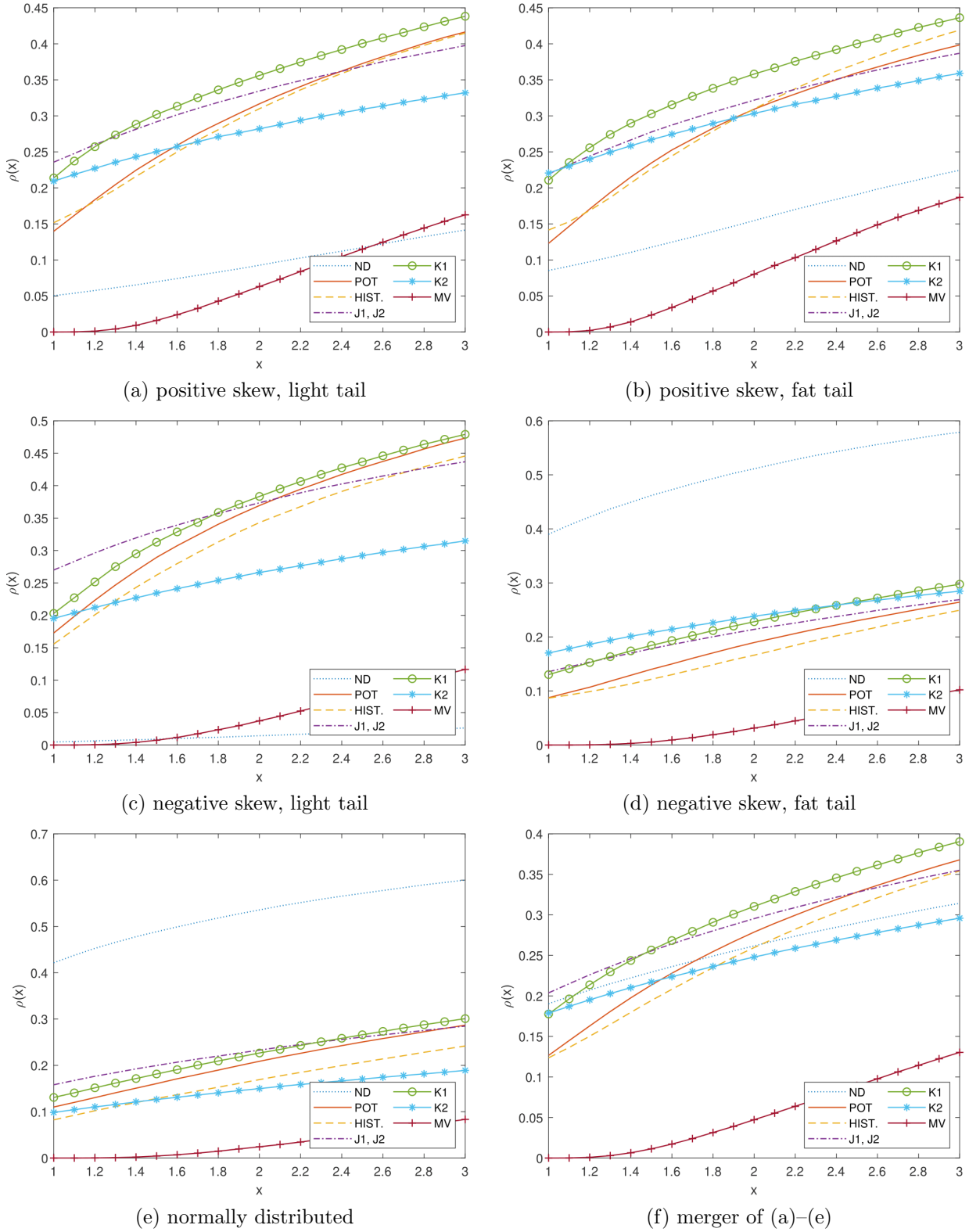
(d) negative skew, fat tail

(e) normally distributed

(f) merger of (a)–(e)

For a confidence level of $\gamma = 97.5\,\%$ and a sample size of $n = 21$, this figure plots the performance profiles (defined in Section 1.3.4.2) of our expected shortfall estimators. Each subfigure concentrates on a specific distributional setting (specified in Section 1.3.3). The estimators are abbreviated as in Table 1.1 and grouped as in Figure 1.1.

Figure A.3.: Performance profiles for $\gamma = 97.5\,\%$ and $n = 126$



(a) positive skew, light tail

(b) positive skew, fat tail

(c) negative skew, light tail

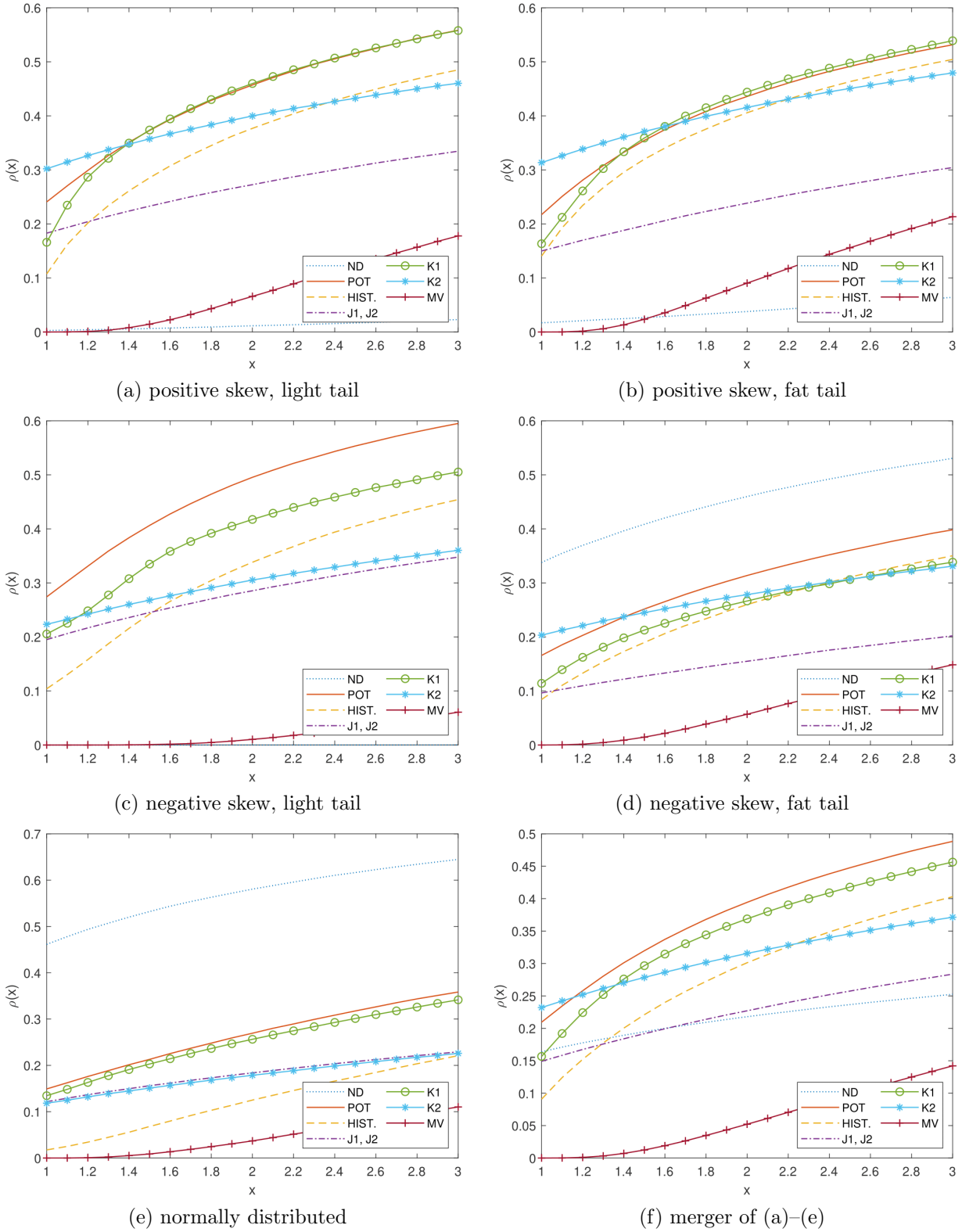
(d) negative skew, fat tail

(e) normally distributed

(f) merger of (a)–(e)

For a confidence level of $\gamma = 97.5\,\%$ and a sample size of $n = 126$, this figure plots the performance profiles (defined in Section 1.3.4.2) of our expected shortfall estimators. Each subfigure concentrates on a specific distributional setting (specified in Section 1.3.3). The estimators are abbreviated as in Table 1.1 and grouped as in Figure 1.1.

Figure A.4.: Performance profiles for $\gamma = 97.5\,\%$ and $n = 504$



(a) positive skew, light tail

(b) positive skew, fat tail

(c) negative skew, light tail

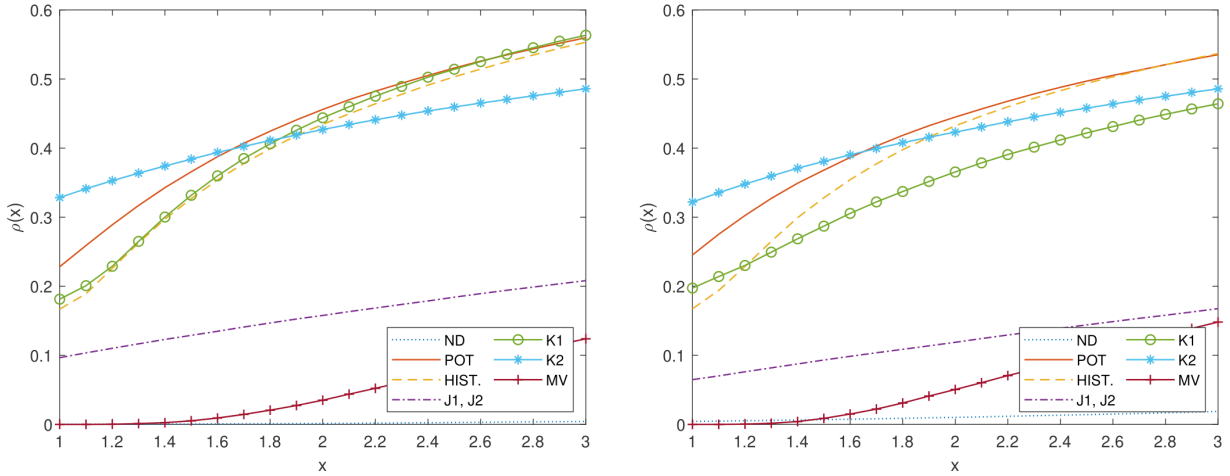
(d) negative skew, fat tail

(e) normally distributed

(f) merger of (a)–(e)

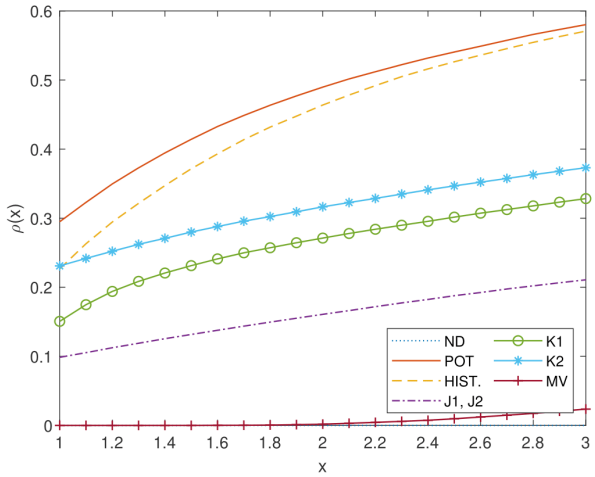
For a confidence level of $\gamma = 97.5\,\%$ and a sample size of $n = 504$, this figure plots the performance profiles (defined in Section 1.3.4.2) of our expected shortfall estimators. Each subfigure concentrates on a specific distributional setting (specified in Section 1.3.3). The estimators are abbreviated as in Table 1.1 and grouped as in Figure 1.1.

Figure A.5.: Performance profiles for $\gamma = 97.5\,\%$ and $n = 1008$

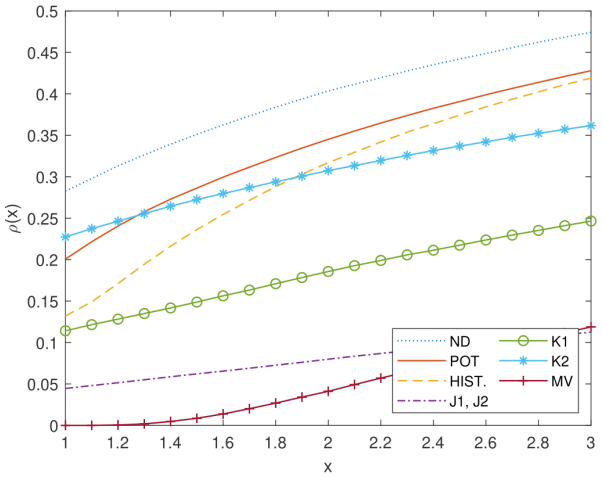


(a) positive skew, light tail

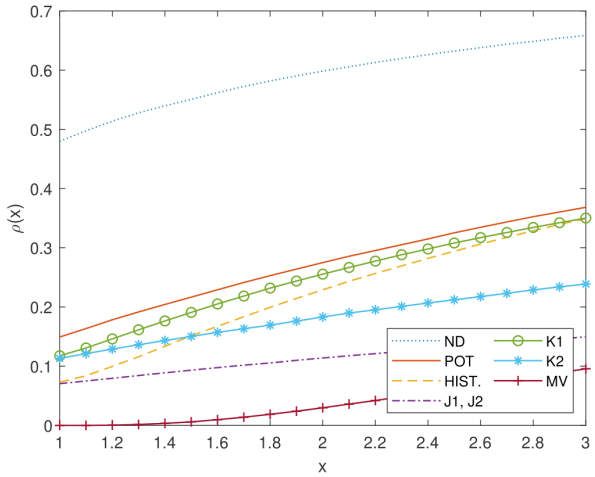
(b) positive skew, fat tail



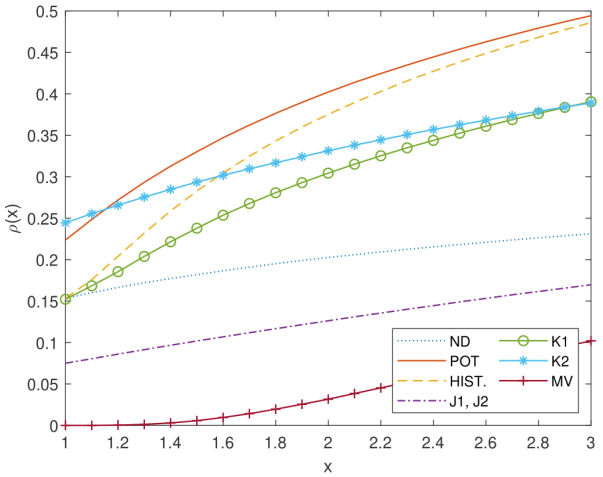
(c) negative skew, light tail



(d) negative skew, fat tail



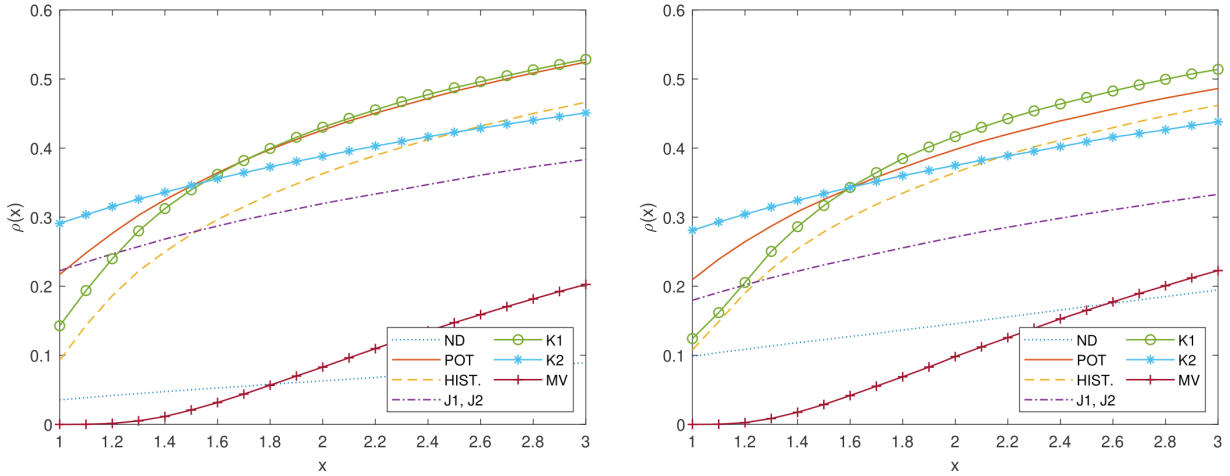
(e) normally distributed



(f) merger of (a)–(e)

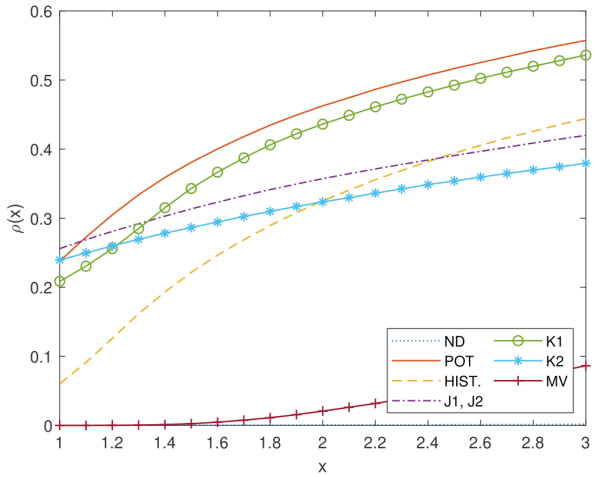
For a confidence level of $\gamma = 97.5\,\%$ and a sample size of $n = 1008$, this figure plots the performance profiles (defined in Section 1.3.4.2) of our expected shortfall estimators. Each subfigure concentrates on a specific distributional setting (specified in Section 1.3.3). The estimators are abbreviated as in Table 1.1 and grouped as in Figure 1.1.

Figure A.6.: Performance profiles for $\gamma = 95\,\%$ and $n = 252$

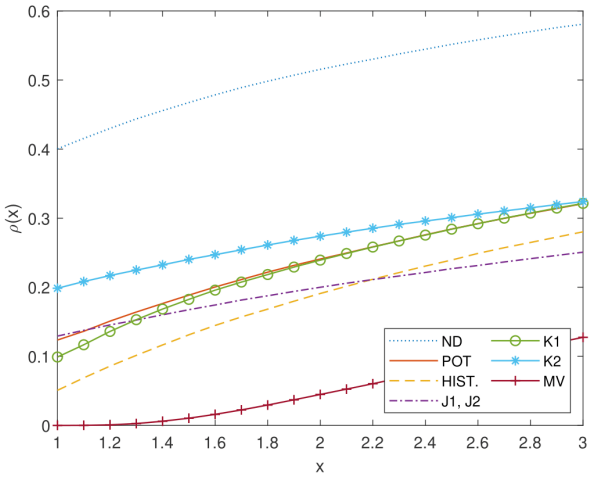


(a) positive skew, light tail

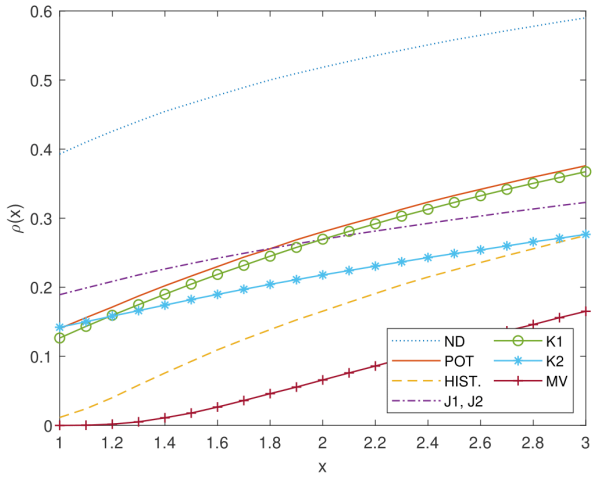
(b) positive skew, fat tail



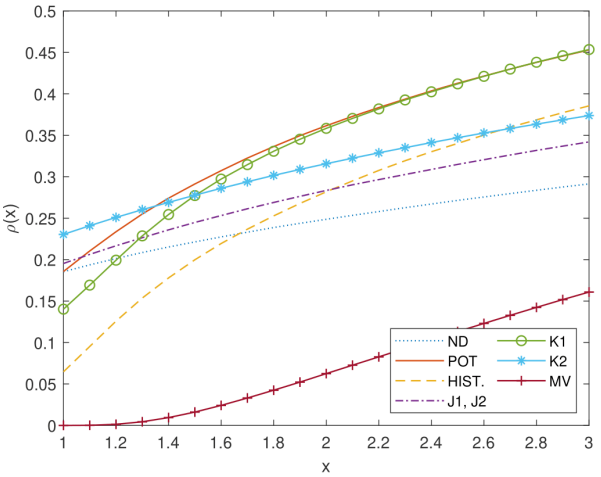
(c) negative skew, light tail



(d) negative skew, fat tail



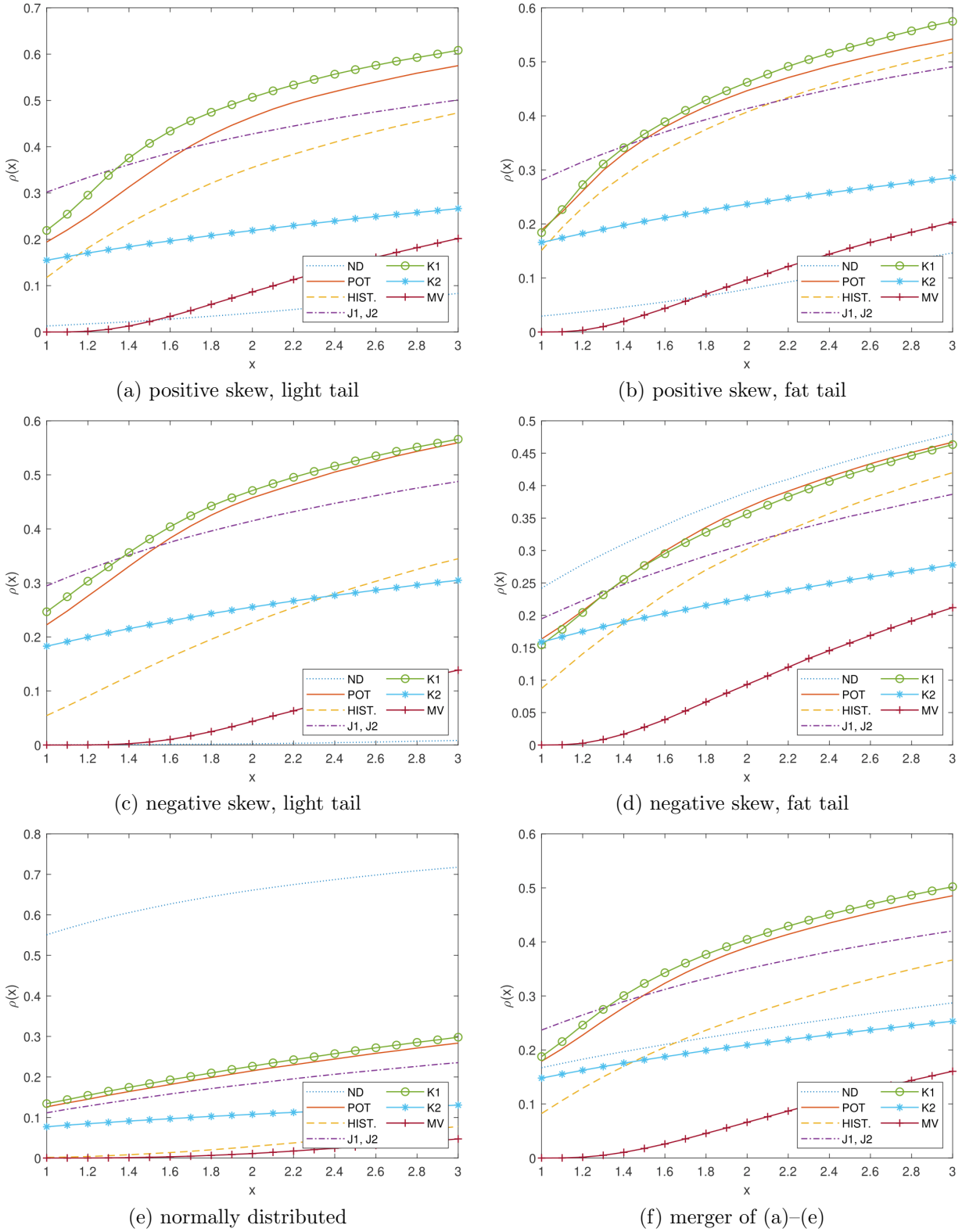
(e) normally distributed



(f) merger of (a)–(e)

For a confidence level of $\gamma = 95\,\%$ and a sample size of $n = 252$, this figure plots the performance profiles (defined in Section 1.3.4.2) of our expected shortfall estimators. Each subfigure concentrates on a specific distributional setting (specified in Section 1.3.3). The estimators are abbreviated as in Table 1.1 and grouped as in Figure 1.1.

Figure A.7.: Performance profiles for $\gamma = 99\,\%$ and $n = 252$



(a) positive skew, light tail

(b) positive skew, fat tail

(c) negative skew, light tail

(d) negative skew, fat tail

(e) normally distributed

(f) merger of (a)–(e)

For a confidence level of $\gamma = 99\,\%$ and a sample size of $n = 252$, this figure plots the performance profiles (defined in Section 1.3.4.2) of our expected shortfall estimators. Each subfigure concentrates on a specific distributional setting (specified in Section 1.3.3). The estimators are abbreviated as in Table 1.1 and grouped as in Figure 1.1.

B. Supplementary results for Chapter 2

B.1. Additional tables

Table B.1.: Estimated ES of standardized losses for $\alpha = 0.01$

	Hansen	POT	g-and-h	Johnson	GM	KDE
S&P GSCI	3.259	3.259	**3.788**	***2.805***	3.207	3.240
Energy	3.292	3.311	**3.837**	***2.816***	3.209	3.275
Precious metals	3.775	3.711	**4.677**	***2.858***	3.973	3.754
Industry metals	3.243	3.327	**3.991**	***2.745***	3.297	3.333
Agriculture	3.074	3.045	**3.562**	***2.710***	2.991	3.074
Livestock	2.968	2.952	**3.086**	***2.804***	2.868	2.927

This table repeats the calculations of Table 2.2 for the coverage level $\alpha = 0.01$ and reports the estimated ES of standardized losses related to the S&P GSCI and its sector sub-indices. Again, the lowest ES estimates for each index are marked in italics and the highest bold.

Table B.2.: AR(1)-GARCH(1,1) results

	Parameters			LB(5)	LB(20)
AR	α_0	α_1		4.168	28.033
	-0.024	-0.001		(0.526)	(0.109)
GARCH	β_0	β_1	β_2	9.489	26.341
	0.005	0.056	0.943	(0.091)	(0.155)

For the S&P GSCI and from January 10, 1983 to December 31, 2018, this table reports the QML estimated parameters of an AR(1)-GARCH(1,1) model based on daily losses (i.e. negative daily percentage log returns). Furthermore, it presents the Ljung-Box test statistics LB(m) for serial correlation up to m lags in the standardized and squared standardized model residuals. p-values are given in parentheses.

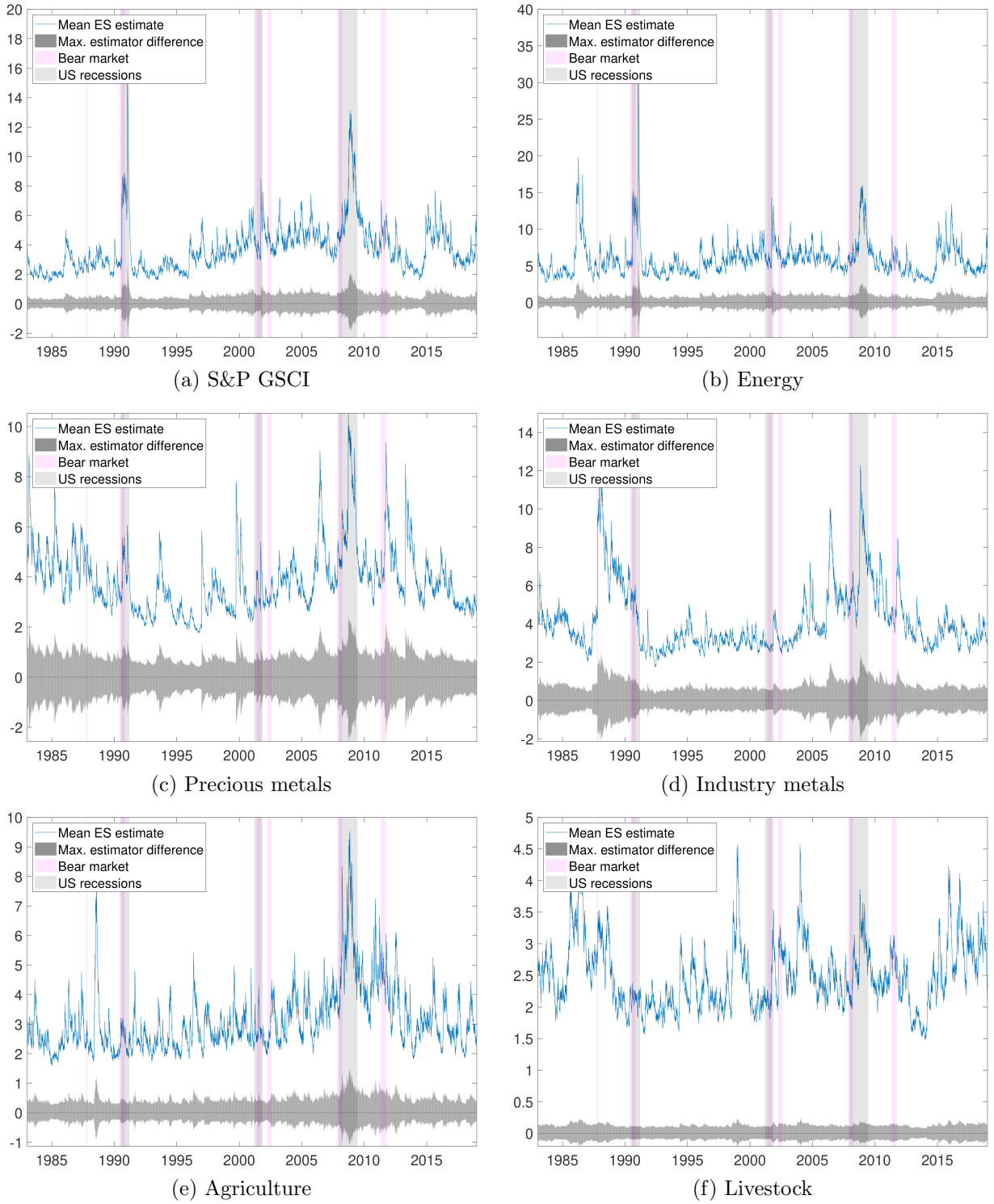
Table B.3.: Mean correlations between commodity futures and stock returns

	General	Falling	Extreme
S&P GSCI	0.175	0.083	0.002
Energy	0.164	0.078	0.002
Precious metals	-0.059	-0.031	-0.002
Industry metals	0.226	0.106	0.002
Agriculture	0.147	0.069	0.001
Livestock	0.147	0.069	0.001

For the period of recession in the global financial crisis (December 2007 to June 2009), this figure presents the mean value of the correlations presented in Figure 2.3 in general, in falling markets (i.e., when S&P 500 returns are negative) and in very extreme market conditions (i.e., when S&P 500 returns are below their 1 % quantile).

B.2. Additional figures

Figure B.1.: Time-varying ES estimates, $\alpha = 0.01$



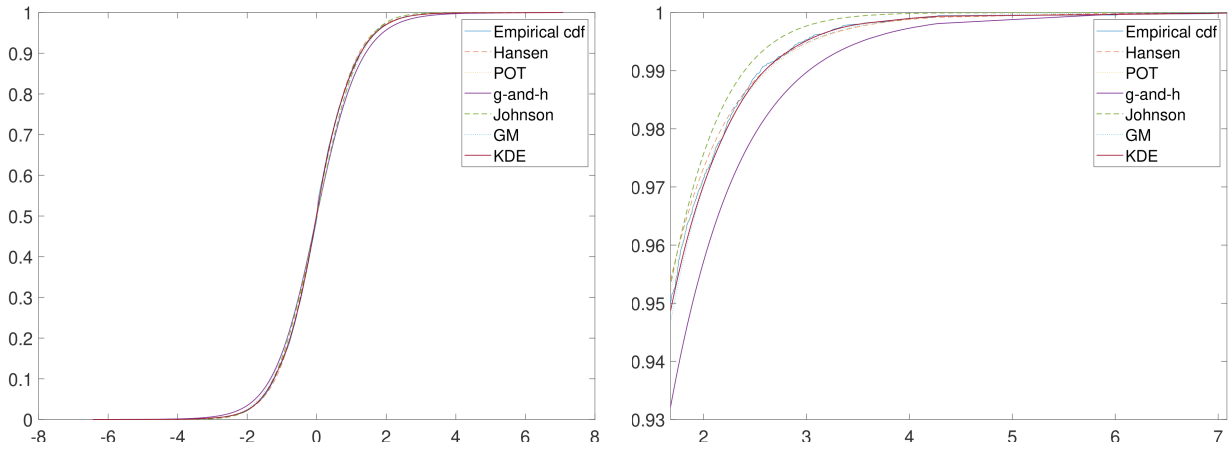
(a) S&P GSCI (b) Energy

(c) Precious metals (d) Industry metals

(e) Agriculture (f) Livestock

This figure extends the results of Figure 2.2 to ES estimation for $\alpha = 0.01$. That is, for our commodity futures index series, it presents the time-varying averages of our six conditional ES estimators, the largest and smallest deviations of individual estimators from the averages as well as US recessions and S&P 500 bear markets.

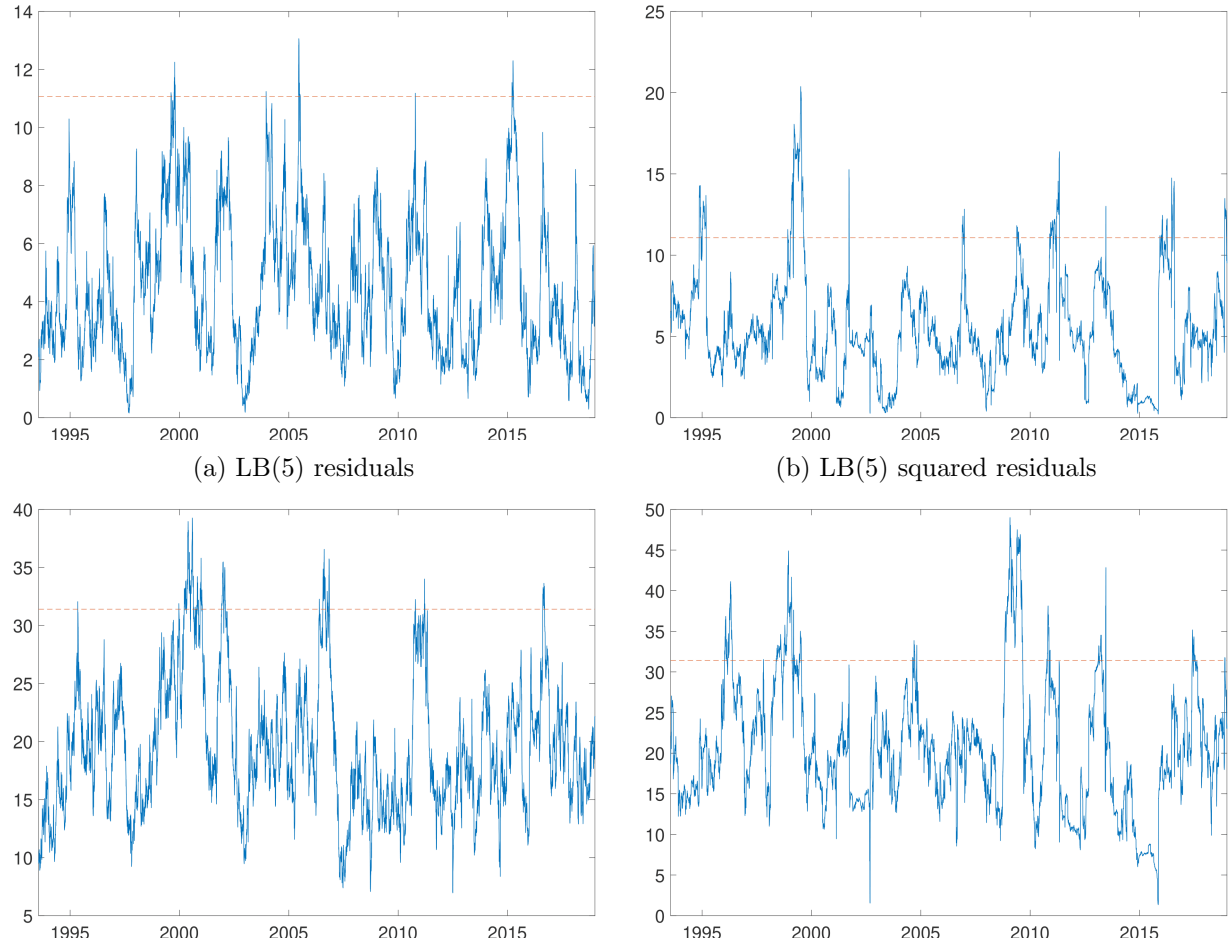
Figure B.2.: Empirical vs. fitted distribution functions



(a) Full cumulative distribution functions (b) Zoom into empirical 5 % tail

This figure presents the empirical cdf of the standardized residuals corresponding to the S&P GSCI AR(1)-GARCH(1,1) model of Table B.2 and the cdfs of our fitted distribution models. While the first subfigure shows the full functions, the second provides a zoomed-in perspective starting from the 95 % quantile of the empirical cdf. Note that, in the POT method, the fitted function concerns only the tail.

Figure B.3.: Ljung-Box test results over time

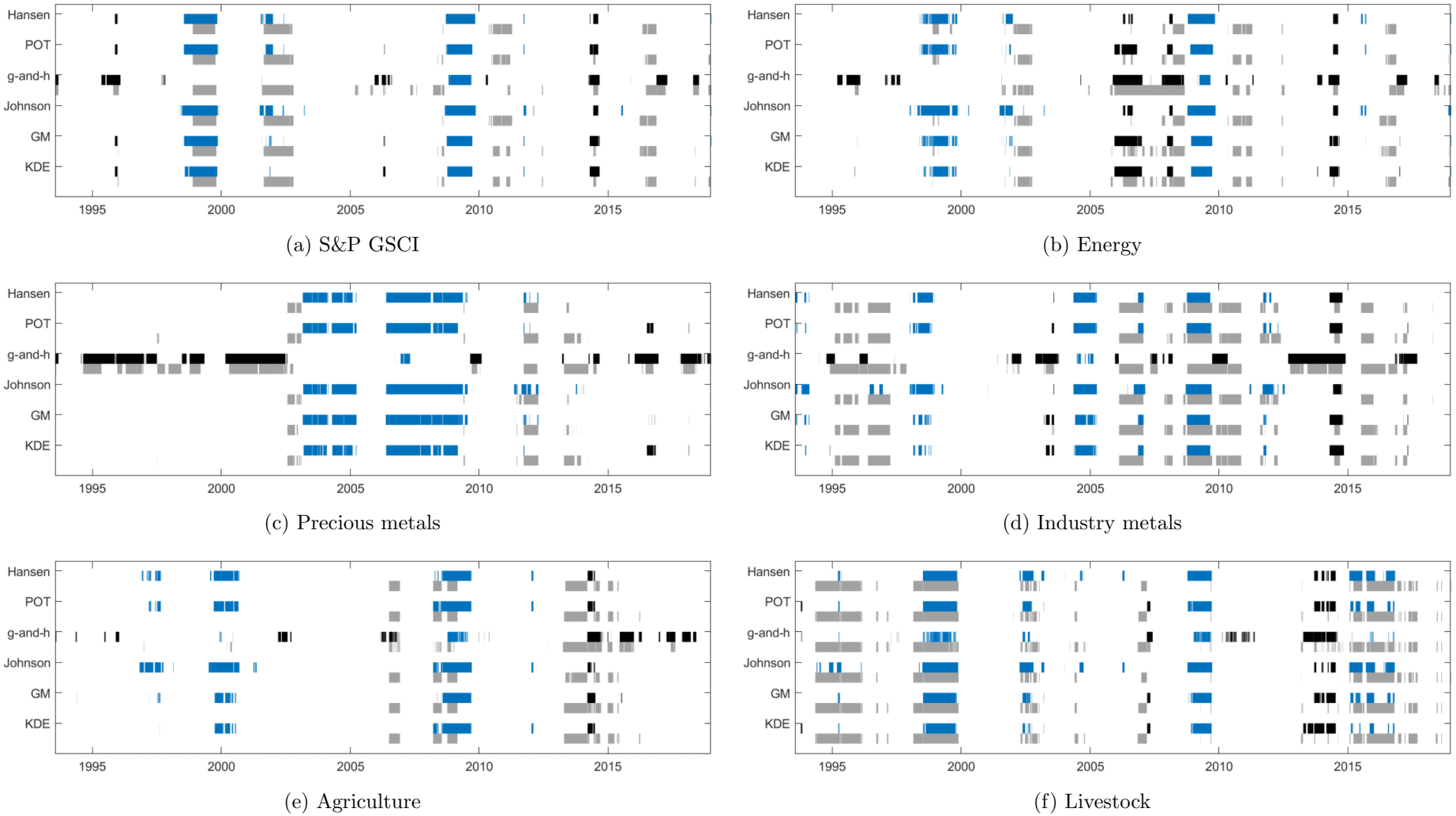


(a) LB(5) residuals (b) LB(5) squared residuals

(c) LB(20) residuals (d) LB(20) squared residuals

For the in-sample periods of our rolling windows and with a focus on the S&P GSCI, this figure presents the Ljung-Box (LB) test statistics (5 and 20 lags) of the AR(1)-GARCH(1,1) standardized residuals and squared standardized residuals. Critical chi-square values for a significance level of $\bar{\alpha} = 0.05$ are given as dashed horizontal lines.

Figure B.4.: Backtest rejection dates



(a) S&P GSCI

(b) Energy

(c) Precious metals

(d) Industry metals

(e) Agriculture

(f) Livestock

This figure presents the rejection dates behind the percentages of Figure 2.5. That is, a blue (black) bar reflects a date where the unconditional backtest statistic lies above (below) the upper (lower) bound of its critical region such that the corresponding method underestimates (overestimates) the ES. Similarly, a gray bar corresponds to a rejection via the conditional backtest.

B.3. Extended discussion of Figure 2.2

While our general discussion of Figure 2.2 looked at ES behavior in recessions and bear markets, this appendix turns to an inspection of the individual commodity subsectors which allows us to highlight their highs and lows and to discuss the impact of purely commodity-specific events. Starting with Figure 2.2(b) for the energy sub-index, in 1986 and 2014/2015, we can detect the effects of Saudi Arabia's decision to temporarily abandon its role as a swing producer in the oil market and to significantly increase its production levels (see Plante, 2019). The mean ES quadrupled in these years and reached local maxima of 13.9 % and 8.8 %, respectively. The highest surge in investment risk can be detected in the oil price crisis during the First Gulf War. Between August and September 1990 (January 02-18, 1991), ES increased from 4 % to more than 10 % (6 % to 23.2 %). However, in June 1991, the ES quickly returned to its pre-crisis level. As mentioned above, other notable risk peaks are connected to the terrorist attacks of September 11, 2001 and the global financial crisis of 2008.

As far as the risk of the precious metals sub-index in Figure 2.2(c) is concerned, its first peak occurs in 1983 (a rise from 2.6 % to 5.7 % between January and March), after the early 1980s recession (not covered by our sample period), when the US announced its Strategic Defensive Initiative during the Cold War and the Israel bank stock crisis started (see Kucher and McCoskey, 2017). In contrast to the other sectors, the ES of precious metals was highly volatile in this year. In the remaining sample, some of the clearest peaks are in March 1985; August 1990 (recession of the early 1990s); August 1993; December 1996; September 1999; September 2001; between June 2005 and 2006 (corresponding to a crucial drop in US housing prices); 2008 (global financial crisis); summer 2011 (US debt-ceiling crisis and bear market with dropping silver and gold prices after large increases); and April 2013 (another US debt-ceiling crisis).[1]

Figure 2.2(d) is dedicated to the risk of industry metals futures. Besides a slight spike at the beginning of 1983, indicating a industry metal price recovery after the recession of the 1980s, and the Black Monday crash in October 1987, where the ES doubled from 3.3 % to 6.6 % between October 20, 1987 and October 23, 1987, another important event influenced investment risk in the 1980s. In February, 1989 the ES reached a local maximum of 8.4 % because labor unrest in South American mines had restricted the availability of copper (see Dumas, 1992). With respect to other maxima, the spring of 2006 highlights some additional spikes just before the peaks of the global financial crisis in autumn 2008. Again, jumps primarily in copper prices caused the ES to rise from 2.4 % in February to 7 % in May. Interestingly, and similar to precious metals, we can document increased risk in the industry metal sector during the US bear market and debt-ceiling crisis in the summer of 2011, which is not evident in the energy, agriculture and livestock sectors.

In contrast to the other sectors, the ES swings of the agriculture sub-index in Figure 2.2(e) are less clear-cut. Nonetheless, we can observe several pronounced increases. For example, between May and July 1988, the ES rose from 1.2 % to 5.8 %. Furthermore, in the world food price crisis (often presumed to be triggered by speculation; see Etienne et al., 2018), it had magnitudes up to 6 % (6.8 %) in March (October) 2008 and 5.2 % in November 2010. In the aftermath of this crisis, ES levels returned to values as low as 1.5 % in February 2013.

As can be seen in Figure 2.2(f), the ES in the livestock sector stands out by distinctively lower maximum risk levels and the fact that the ES values produced by our different estimators are almost identical. This is not surprising because Table 2.1 shows that the losses of livestock futures are close to being normally distributed and many of our estimators contain the normal distribution as a special case. Outstanding peaks appeared around May 1986, December 1998, January 2004 (where the mean ES reaches its maximum of 3.4 %), during the food price crisis in October 2008 as well as in November 2015 and October 2016. In January 2014, the ES realized its minimum of 1.1 %.

[1] For a general discussion of US debt-ceiling crises, see Buchanan (2013).

C. Supplementary results for Chapter 3

C.1. Additional tables

Table C.1.: Sharpe ratio significance

	R1-H1	R3-H1	R6-H1	R12-H1	R1-q2	R3-q2	R3-q4	R6-q2	R6-q4	R6-q7	R12-q2	R12-q4	R12-q7	R12-q13	Benchmark
R1-H1		0.58	0.89	0.43	**0.02**	0.34	0.78	0.53	0.36	0.07	0.77	0.62	**0.04**	**0.02**	0.08
R3-H1	0.94		0.71	0.21	**0.05**	0.49	0.97	0.76	0.53	**0.04**	0.55	0.43	**0.03**	**0.01**	**0.04**
R6-H1	0.97	0.97		0.17	**0.03**	0.40	0.87	0.60	0.43	0.06	0.71	0.56	**0.03**	**0.02**	0.05
R12-H1	0.65	0.69	0.52		**0.01**	0.15	0.39	0.24	0.16	0.30	0.76	0.91	0.12	0.08	0.45
R1-q2	**0.01**	**0.01**	**0.00**	**0.00**		0.09	**0.04**	**0.03**	0.15	**0.00**	**0.01**	**0.01**	**0.00**	**0.00**	**0.00**
R3-q2	0.25	0.21	0.20	0.12	**0.01**		0.40	0.59	0.94	**0.01**	0.13	0.14	**0.01**	**0.00**	0.05
R3-q4	0.72	0.63	0.65	0.40	**0.01**	0.30		0.69	0.34	0.06	0.55	0.29	**0.02**	**0.01**	0.18
R6-q2	0.43	0.41	0.33	0.24	**0.03**	0.67	0.55		0.73	**0.03**	0.21	0.22	**0.01**	**0.01**	0.09
R6-q4	0.38	0.34	0.30	0.19	**0.02**	0.66	0.39	0.96		**0.01**	0.24	**0.03**	**0.01**	**0.00**	0.07
R6-q7	0.72	0.75	0.69	0.97	**0.00**	0.10	0.36	0.14	0.09		0.17	0.21	0.24	0.30	0.58
R12-q2	0.57	0.55	0.52	0.32	**0.00**	0.43	0.79	0.73	0.80	0.31		0.80	0.05	**0.03**	0.41
R12-q4	0.99	0.96	0.98	0.68	**0.00**	0.18	0.60	0.30	0.09	0.64	0.49		0.07	**0.05**	0.55
R12-q7	0.73	0.78	0.74	0.94	**0.00**	0.13	0.46	0.23	0.16	0.95	0.30	0.67		0.69	0.29
R12-q13	0.89	0.94	0.91	0.75	**0.00**	0.19	0.61	0.33	0.28	0.74	0.42	0.88	0.69		0.19
Benchmark	0.56	0.60	0.56	0.82	**0.00**	0.13	0.37	0.27	0.22	0.82	0.31	0.59	0.79	0.63	

Extending Tables 3.2 and 3.4, this table presents the p-values of the Ledoit and Wolf (2008) bootstrap test for investigating the null hypothesis of equal Sharpe ratios. The Sharpe ratios of long (long-short) strategies are compared pairwise in the right (left) triangle. Because varying bootstrap block sizes within {1, 3, 6, 9, 12} does not alter the rejection decision, we present the results for a block size of 6. Bootstrapping is performed with 5,000 repetitions. Significant entries at a 5 % level are marked in bold.

Table C.2.: Conditional multi-factor models

	R1-q2	R3-q2	R3-q4	R6-q2	R6-q4	R6-q7	R12-q2	R12-q4	R12-q7	R12-q13
Panel A: Long										
Model 1 (term spread)										
α_1	0.12	0.02	-0.06	-0.04	-0.13	0.01	0.11	-0.18	0.01	0.10
t_{α_1}	0.70	0.14	-0.30	-0.25	-0.71	0.07	0.64	-0.93	0.06	0.46
Model 2 (term spread)										
α_1	0.22	0.07	0.01	0.01	-0.03	-0.00	0.03	-0.12	0.01	0.06
t_{α_1}	1.35	0.45	0.03	0.09	-0.15	-0.02	0.16	-0.70	0.07	0.34
Model 3 (term spread)										
α	**1.49**	**1.25**	**1.06**	**0.93**	**1.08**	-0.01	**0.81**	**0.74**	-0.18	-0.09
t_α	6.37	5.58	4.06	4.65	4.61	-0.05	3.56	2.97	-0.67	-0.33
σ_ε	4.40	4.20	4.90	3.77	4.42	4.92	4.26	4.69	5.00	5.15
IR	0.34	0.30	0.22	0.25	0.24	-0.00	0.19	0.16	-0.04	-0.02
Model 1 (default spread)										
α_1	0.00	0.49	1.05	-0.02	0.24	0.01	-0.61	-0.68	-1.16	-0.83
t_{α_1}	0.00	0.80	1.53	-0.04	0.38	0.02	-1.02	-1.03	-1.67	-1.15
Model 2 (default spread)										
α_1	0.07	0.24	1.03	-0.10	0.31	-0.08	-0.62	-0.26	-1.09	-0.88
t_{α_1}	0.11	0.43	1.59	-0.20	0.52	-0.12	-1.08	-0.42	-1.65	-1.30
Model 3 (default spread)										
α	**1.54**	**1.17**	**0.99**	**0.91**	**1.07**	0.04	**0.77**	**0.75**	-0.16	-0.06
t_α	6.55	5.06	3.76	4.44	4.48	0.14	3.32	2.96	-0.59	-0.22
σ_ε	4.41	4.32	4.93	3.83	4.46	4.91	4.32	4.72	5.00	5.14
IR	0.34	0.30	0.22	0.25	0.24	-0.00	0.19	0.16	-0.04	-0.02
Panel B: Long-short										
Model 1 (term spread)										
α_1	0.28	0.01	-0.37	-0.04	-0.09	-0.15	0.14	-0.48	-0.44	-0.10
t_{α_1}	1.14	0.03	-1.24	-0.15	-0.33	-0.51	0.57	-1.68	-1.52	-0.36
Model 2 (term spread)										
α_1	0.24	-0.11	-0.28	-0.21	-0.06	-0.14	-0.15	**-0.53**	-0.45	-0.14
t_{α_1}	1.04	-0.52	-1.04	-0.90	-0.24	-0.53	-0.65	-2.03	-1.69	-0.55
Model 3 (term spread)										
α	**2.51**	**1.63**	**1.19**	**1.22**	**1.48**	0.48	**1.25**	**1.03**	0.51	0.66
t_α	7.71	5.24	3.07	3.67	3.99	1.23	3.85	2.73	1.35	1.78
σ_ε	6.12	5.87	7.29	6.23	6.98	7.38	6.11	7.08	7.18	6.96
IR	0.41	0.28	0.16	0.20	0.21	0.07	0.20	0.15	0.07	0.09
Model 1 (default spread)										
α_1	-0.41	-0.16	-0.06	-0.60	-0.18	-0.25	-1.14	-1.25	-2.45	-1.17
t_{α_1}	-0.49	-0.20	-0.05	-0.69	-0.19	-0.25	-1.34	-1.27	-2.47	-1.21
Model 2 (default spread)										
α_1	0.75	0.56	0.60	0.22	0.69	0.72	-0.50	-0.32	**-2.12**	-0.82
t_{α_1}	0.93	0.72	0.62	0.26	0.75	0.74	-0.61	-0.34	-2.24	-0.88
Model 3 (default spread)										
α	**2.57**	**1.54**	**1.10**	**1.16**	**1.41**	0.45	**1.19**	**0.99**	0.45	0.59
t_α	8.02	4.95	2.83	3.51	3.81	1.16	3.64	2.63	1.17	1.60
σ_ε	5.99	5.80	7.25	6.19	6.90	7.28	6.12	7.03	7.16	6.96
IR	0.41	0.28	0.16	0.20	0.21	0.07	0.20	0.15	0.07	0.09

Extending Table 3.5 and focusing on long and long-short strategies, this table presents the results of multi-factor regressions where alphas and betas are functions of a conditioning variable Z. That is, we take into account that investment performance and/or systematic connection to the market may be time-varying. Following Christopherson et al. (1998), we specify $\alpha_t = \alpha_0 + \alpha_1 Z_{t-1}$ and $\beta_{i,t} = \beta_{i,0} + \beta_{i,1} Z_{t-1}$ for $i \in \{S, B, C\}$. Model 1 allows for both time-varying alpha and betas. Model 2 (3) assumes time-varying alpha (betas) but constant betas (alpha). The term spread (i.e., the difference between the 30-year US TBond yield and the 3-month US TBill rate) and the default spread (i.e., the yield difference between Moody's seasoned Baa and Aaa US corporate bonds) are used for conditioning. Limited by the availability of the conditioning variables, the regressions are based on data from January 1986 to December 2019. The t-statistics are based on Newey-West standard errors. Significance at a 5 % level is marked in bold.

Table C.3.: Worst losses of traditional momentum

	R1-H1	R1-q2
07/1975	-30.79	0.48
09/2004	-28.20	13.60
08/1981	-27.16	-6.40
08/2000	-23.00	0.50
03/1976	-20.93	0.40
03/1987	-20.40	1.48
06/1999	-19.72	0.40
03/2003	-18.91	0.10
11/1978	-18.64	0.70
07/1988	-17.97	0.51
10/1974	-17.17	0.51
05/2019	-16.94	18.92
02/1983	-16.38	0.62
09/1981	-16.30	-0.54
07/1981	-15.48	0.58
07/1985	-15.47	8.10
11/2004	-15.41	17.92
02/1996	-15.06	0.39
05/2003	-14.91	0.09
02/1995	-14.89	0.40

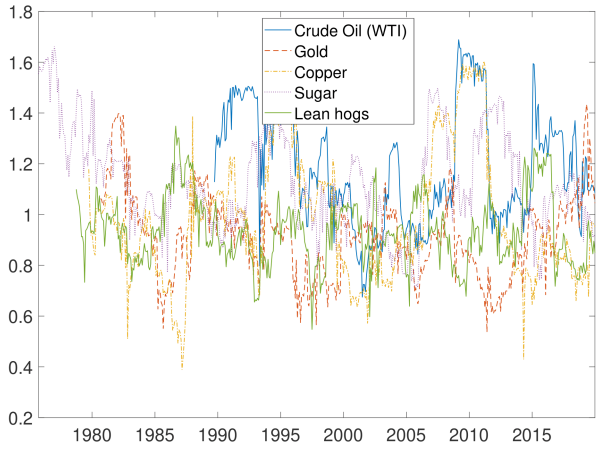
This table reports the 20 worst losses of the traditional momentum strategy R1-H1 and contrasts them with the returns of the memory-enhanced strategy R1-q2.

C.2. Additional figures

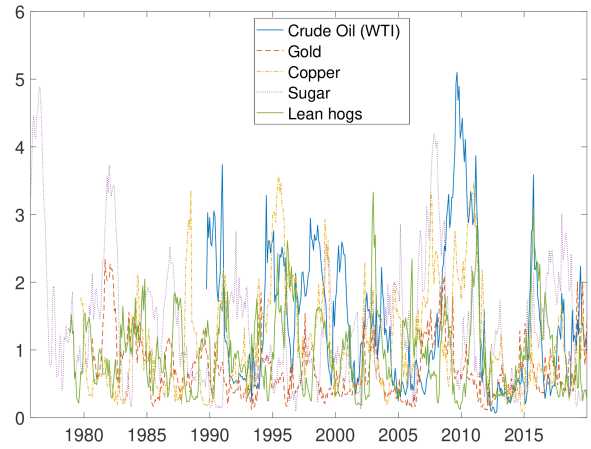
Figure C.1.: Traditional momentum positioning

For each month t in our sample, this figure shows the top (bottom) 20 % commodities, as suggested by the return in month t, and thus the winners (losers) that are chosen for investment in the $t+1$ long (short) leg of a R1-H1 momentum strategy.

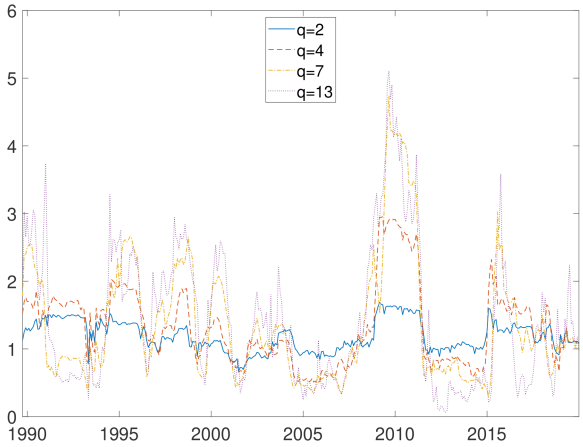
Figure C.2.: Variance ratios over time



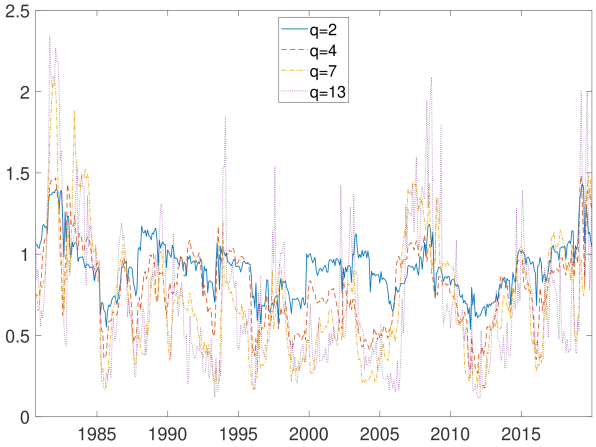
(a) Selected commodities, $q = 2$



(b) Selected commodities, $q = 13$



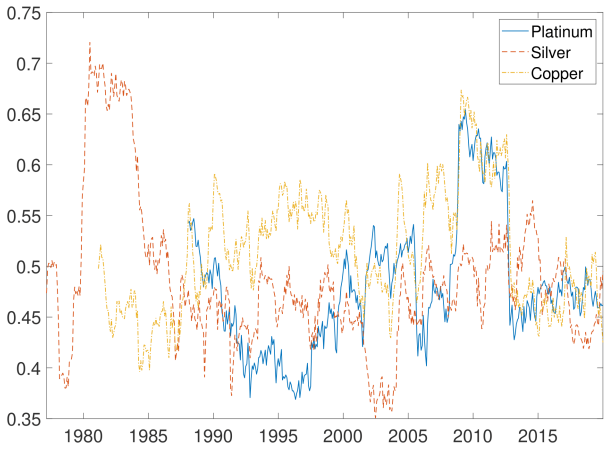
(c) Crude oil (WTI), various q



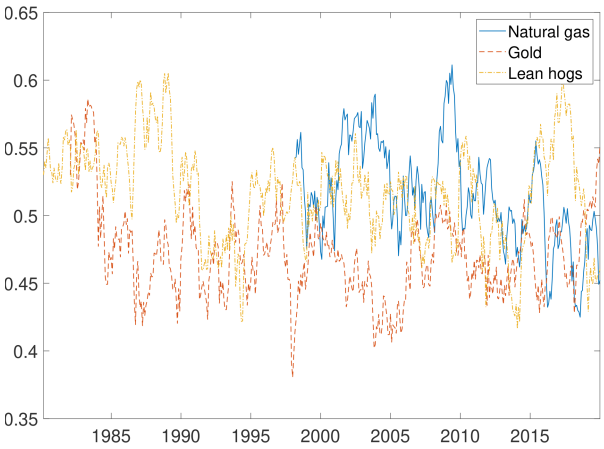
(d) Gold, various q

For several commodities and orders q, this figure presents the monthly evolution of variance ratios (estimated as outlined in Section 3.3.2.1) used in our short memory momentum strategies.

Figure C.3.: Hurst coefficients over time



(a) Actively traded commodities



(b) Untraded commodities

For selected commodities, this figure presents the monthly evolution of Hurst coefficients (estimated via the averaging approach of Section 3.3.2.3) used in our long memory momentum strategies.