SFB
823

Using the extremal index for value-at-risk backtesting

Discussion Paper

Axel Bücher, Peter N. Posch,
Philipp Schmidtke

Nr. 25/2018

Using the Extremal Index for Value-At-Risk Backtesting

Axel Bücher[*], Peter N Posch[†], Philipp Schmidtke[‡]

First version: Sept 2017. This version: October 15, 2018

Abstract

We introduce a set of new Value-at-Risk independence backtests by establishing a connection between the independence property of Value-at-Risk forecasts and the extremal index, a general measure of extremal clustering of stationary sequences. We introduce a sequence of relative excess returns whose extremal index has to be estimated. We compare our backtest to both popular and recent competitors using Monte-Carlo simulations and find considerable power in many scenarios. In an applied section we perform realistic out-of-sample forecasts with common forecasting models and discuss advantages and pitfalls of our approach.

Key words: VaR Backtesting, Extremal Index, Independence, Risk measures.

JEL Classification: C52, C53, C58.

[*] Heinrich Heine University Düsseldorf, Mathematical Institute, Universitätsstr. 1, 40225 Düsseldorf, Germany

[†] TU Dortmund University, Chair of Finance, Otto-Hahn-Str. 6, 44227 Dortmund, Germany

[‡] TU Dortmund University, Chair of Finance, Otto-Hahn-Str. 6, 44227 Dortmund, Germany, E-Mail: philipp.schmidtke@udo.edu. Corresponding author.

1 Introduction

In spite of its usage as a risk measure for more than 20 years, researchers are still engaged in exploring new forecasting and backtesting procedures for the Value-at-Risk (VaR). The latter procedures are typically based on a statistical test which tries to assess whether a certain desirable property is met for the observed sequence of VaR-violations: first, the concept of *correct unconditional coverage* aims at checking whether the number of overall violations is justifiable. From an academic perspective, we typically seek for a forecasting procedure which yields neither too many nor too few violations. On the other hand, regulators are usually interested in situations where the risk is not underestimated, resulting in a focus on not too many violations. Second, the *correct independence* aspect focuses on possible serial dependence of violations, and aims at checking whether the sequence of violations behaves like an independent sequence. This concept becomes most important if unconditional coverage is statistically satisfied, i.e., an unconditional test cannot be rejected. In that case, a test using information about the way how violations occur has still potential to reject the forecasts. Available independence backtests may have power only with respect to a lack of independence, or with respect to both the lack of independence and of correct unconditional coverage. The latter ones are called *conditional coverage tests.*

In general, tackling the independence property is challenging. This is mainly due to the fact that risk forecasts deal with low probability events and an often short testing sample. As a consequence, observing many violations is unlikely, which naturally results in small effective sample sizes and, therefore, bad power properties. In addition, some of the classical tests explicitly assume an alternative model incorporating a special kind of dependence, which may also result in a loss of power if in fact a more general form of dependence is present.

Despite these natural difficulties, the independence hypothesis itself is relevant.

As a matter of fact, most financial time series exhibit large degrees of heteroscedasticity and therefore require for time-changing risk forecasts. Renouncement would lead to a probably threatening violation clustering, something a sound risk management should always aim to prevent.

We contribute to the backtesting literature by introducing a new test for the independence hypothesis which is particularly sensitive to deviations from independence among the most extreme observations. Unlike standard methods, the new test does not use solely the 0-1-violation sequence. Instead, we assess whether a series of VaR-adjusted returns, coined *relative excess returns*, exhibits a significant tendency for that its most extreme observations occur in clusters. As a measure for that tendency, we employ the *extremal index*, a natural measure of clustering of extreme observations stemming from extreme value theory. We implement the approach with two different extremal index estimators, the first one (Süveges and Davison, 2010) leading to a more classic 0-1-test, while the second one (Northrop, 2015; Berghaus and Bücher, 2017) enables the processing of more detailed information. We find considerable power improvements in many cases in comparison to common competing tests, with the second test often showing the most convincing results.

As is well known, VaR lacks some important features of risk measures. The most common alternative measure is provided by the Expected Shortfall (ES), which will soon replace the VaR as the standard regulatory measure of risk for banks (BCBS, 2016). However, since VaR and ES are closely related, it does not come as a surprise that VaR and its backtests also play a prominent role in some ES backtests. For example, Kratz et al. (2018) propose a joint backtest for several VaR levels as an intuitive way to implicitly backtest ES. A second example is BCBS (2016) itself, where out-of-sample backtesting is based on VaR as well. However, both in general and in the aforementioned examples, the issue of a possible lack of independence is rarely addressed. Since the implementation of our idea is relatively independent of the specific VaR level, we see this as a promising approach in this respect.

The remainder of this paper is structured as follows. Section 2 provides preliminaries about the notation, a more detailed description of the backtesting problem alongside with a short overview of existing tests, and mathematical details on the extremal index. Section 3 introduces our new approach of independence backtesting based on the extremal index. In Section 4, we perform a detailed analysis of the small-sample properties, while Section 5 focuses on some empirical implications. Finally, Section 6 concludes, while less important aspects are deferred to a sequence of appendices.

2 Preliminaries on Backtesting and the Extremal Index

In this section we review the essentials of VaR backtesting and introduce our notation. Then we turn to the extremal index and its estimators.

2.1 Backtesting the Value-at-Risk

Consider a random return r_t of a financial asset in a period t, usually a day. Suppose this return is continuously distributed with c.d.f. F_t, conditional on the information set $\mathcal{F}_{t-1}$ which embodies all information up to period $t-1$. We define the Value-at-Risk at level p as $\mathrm{VaR}_p^{(t)} := -F_t^{-1}(p)$, where F_t^{-1} denotes the inverse of F_t. Throughout the paper, we will refer to p as VaR level, usual values are 5 % and 1 %, whereas $q = 1 - p$ will be called the VaR confidence level. Note that, with this definition, we report large losses and hence VaRs as positive numbers.

A violation at time t occurs if $r_t < -\widehat{\mathrm{VaR}}_p^{(t)}$, where $\widehat{\mathrm{VaR}}_p^{(t)}$ denotes a forecast of the true VaR at period t, calculated based on information from $\mathcal{F}_{t-1}$. Using a series of VaR forecasts corresponding to observed returns $r_1, \ldots, r_n$, we define the violation sequence $(I_t)_{t=1}^n$, by

$$I_t = \begin{cases} 1 \text{ (violation)}, & \text{if } r_t < -\widehat{\mathrm{VaR}}_p^{(t)} \\ 0 \text{ (compliance)}, & \text{if } r_t \geq -\widehat{\mathrm{VaR}}_p^{(t)} \end{cases}. \tag{2.1}$$

The time points t where violations occur, that is $I_t = 1$, are called violation times or violations indices. Suppose there are M_1 violations, that is, $M_1 = \#\{I_t = 1\}$, and order the violation times increasingly $t_1 < \cdots < t_{M_1}$. We define the inter-violation durations D_i as $D_i := t_{i+1} - t_i$, where $i = 1, \ldots, M_1 - 1$. If the VaR forecasts happen to be completely correct, that is $\widehat{\mathrm{VaR}}_p^{(t)} = \mathrm{VaR}_p^{(t)}$ for all t, then the violation sequence forms an i.i.d. Bernoulli sequence with success probability p, that is $I_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. This implies $M_1 = \sum_{t=1}^n I_t \sim \text{Binom}(n, p)$ and $D_i \overset{\text{i.i.d.}}{\sim} \text{Geom}(p)$.

The goal of backtesting is to asses whether a sequence of n ex-ante VaR forecasts are appropriate in relation to the realized returns. This is usually done by stressing one of the above mentioned properties of the violation sequence or the durations.

Since Christoffersen (1998) backtests are classified according to their focus, see also the discussion in the introduction. The property of forecasts being completely unsuspicious is called correct *conditional coverage* (cc) and may be written as

$$\text{cc: } I_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), \quad t = 1, \ldots, n. \tag{2.2}$$

Before this term was introduced, assessing VaR forecasts was solely concerned with the aspect of *unconditional coverage* (uc) which is defined by

$$\text{uc: } E(I_t) = p, \quad t = 1, \ldots, n. \tag{2.3}$$

In other words, uc is concerned about whether the frequency of violations is reasonable in the sense that, for all time points t, the probability of observing an violation equals p, which is the probability had the true VaR been used for the calculation of I_t. A simple way to get a first impression about the latter property is to calculate the actual number of violations M_1 and compare the result to its expectation np under the assumption $\widehat{\mathrm{VaR}}_p^{(t)} = \mathrm{VaR}_p^{(t)}$. See, e.g., Kupiec (1995) for an early test or BCBS (1996b) for the Traffic Light Approach used by the Basel Committee.

Unconditional coverage is complemented by the *independence property* (ind), given by

$$\text{ind: } I_1, \ldots, I_n \text{ are stochastically independent.} \tag{2.4}$$

Rather than on how many violations occur, the focus is on how the violations occur over time. A simple graphical way to check this property is to look on a plot of VaR violations and assess visually whether there are any patterns. However, detecting a failure of the independence property can be fairly hard due to the natural scarcity of violations if the VaR level is sufficiently small or the backtesting sample is not large enough. Still, possible dependence among violations can be extremely important for risk managers, as subsequent violations can sum up and result in an overall loss of threatening magnitude.

Note that, from a more technical perspective, plain independence of violations does not necessarily imply absence of violation clustering. This can be seen by an example in Ziggel et al. (2014) where I_t and I_{t-k} are in fact independent but clustering can still happen.[1].

[1] See equation 28 of Ziggel et al. (2014) The authors argue that the independence property should be replaced by an i.i.d. property. Although the prevention and detection of violation clustering is also our aim, we continue to speak of independence when we mean the absence of violation clustering in the remainder of the paper.

2.2 Existing Backtests for the Independence Hypothesis

In this section, we provide a brief description of the backtests that we use as competitors to our new proposal. The section may be skipped at first reading.

Test Based on a Markov Chain Model. This early test by Christoffersen (1998) employs a first-order Markov chain model to allow for possibly dependent violations. The model permits differing probabilities of a violation at time t, depending on whether a violation has occurred at $t-1$. More precisely, for $i, j \in \{0,1\}$, let $\pi_{ij} = \Pr(I_t = j \mid I_{t-1} = i)$ denote the transition probabilities in the chain (I_t). Denote the number of compliances (zeros) in an observed sequence of length n by $M_0 = n - M_1$ and denote by M_{ij} the number of events in $I_1, \dots, I_n$ where an observation of i is followed by an observation of j. The likelihood of the Markov model is

$$L^{\mathrm{Mar}}(I_1, \dots, I_n; \pi_{01}, \pi_{11}) = (1-\pi_{01})^{M_0 - M_{01}} \pi_{01}^{M_{01}} (1-\pi_{11})^{M_1 - M_{11}} \pi_{11}^{M_{11}}.$$

The null hypothesis of previous-state-independent violations is equivalent to equality of the transition probabilities π_{01} and π_{11}, i.e, the null of independence can be written as

$$H_0 : \pi_{01} = \pi_{11}. \tag{2.5}$$

Under H_0, the likelihood simplifies to $L^{H_0}(I_1, \dots, I_n; \pi_1) = \pi_1^{M_1}(1-\pi_1)^{n-M_1}$. The test statistic $LR_{\mathrm{ind}}^{\mathrm{Mar}}$ is now defined as the ratio of both likelihoods multiplied by 2, with π_{ij} replaced by their maximum likelihood estimators $\hat{\pi}_{ij}$, and can be shown to be asymptotically $\chi^2(1)$ distributed:

$$LR_{\mathrm{ind}}^{\mathrm{Mar}} = 2 \log \left(\frac{L^{\mathrm{Mar}}(I_1, \dots, I_n; \hat{\pi}_{01}, \hat{\pi}_{11})}{L^{H_0}(I_1, \dots, I_n; \hat{\pi}_1)} \right) \overset{\mathrm{asy}}{\sim} \chi^2(1), \tag{2.6}$$

Note that $\hat{\pi}_{01} = M_{01}/M_0$, $\hat{\pi}_{11} = M_{11}/M_1$ and $\hat{\pi}_1 = M_1/n$.

In addition to $LR_{\mathrm{ind}}^{\mathrm{Mar}}$, one often encounters a statistic $LR_{\mathrm{uc}}^{\mathrm{Mar}}$ for the uc property in (2.3). The sum of both statistics is known as the Markov chain conditional coverage test $LR_{\text{cc}}^{\text{Mar}}$ for (2.2) and is widely used in the literature and appraised as a standard method, see Alexander et al. (2013).

The test statistic $LR_{\text{ind}}^{\text{Mar}}$ solely relies on information of directly subsequent observations and how they relate to each other. Hence, only a limited kind of deviation form independence can be detected, which is often viewed as a striking drawback. For example, if there were only two violations in (I_t) which occurred directly subsequently, then $\hat{\pi}_{11} > 0$ which corresponds to evidence against H_0. Instead, if the second violation would have happened only one day later, then immediately $\hat{\pi}_{11} = 0$ which - according to the Markov chain model - no longer provides evidence of clustering. In order to overcome this limitation, duration-based tests as described in the next paragraph have been proposed.

Durations-Based Test Using Explicit Alternative Models. The first usage of inter-violation durations can be found in Christoffersen and Pelletier (2004). If the cc property in (2.2) holds, then the durations D_i are distributed as an independent sequence of geometric random variables, $D_i \overset{\text{i.i.d.}}{\sim} \text{Geom}(p)$. Christoffersen and Pelletier (2004) propose to exploit this in a time-continuous setting. Since the exponential distribution is the natural continuous counterpart of the geometric distribution, alternative models which embed the exponential distribution are eligible, as, e.g., the Weibull distribution or the Gamma distribution. However, both models do not use the order of the durations and are hence incapable of detecting deviations from independence of the durations. This limitation has been tackled by Christoffersen and Pelletier (2004), who proposed to use the exponential autoregressive conditional duration (EACD) model to model dependent durations. A further, more recent, alternative ('geometric test') has been proposed by Berkowitz et al. (2011), where a discrete approach was used to model duration dependence.

Of all the aforementioned tests, the approach based on Weibull distribution seems to be most popular one, whence we also choose it as a competitor throughout the simulation study. Recall that the density of the Weibull distribution is given by

$$f_W(d; p, b) = p^b \, b \, d^{b-1} \exp(-(p \, d)^b). \tag{2.7}$$

Plugging in $b = 1$ yields the exponential distribution $f_W(D; p, 1) = f_{\exp}(D, p) = p \exp(-p \, D)$ with expectation $1/p$, where p should be interpreted as the VaR level. Therefore, the null hypothesis ind in (2.4) can be identified with

$$H_0 : b = 1. \tag{2.8}$$

In practice, the sample of observed durations $D_1, \ldots, D_{M_1-1}$ is now fitted to the exponential distribution and to the Weibull distribution, and, similar as in $LR_{\text{ind}}^{\text{Mar}}$, a likelihood ratio test is performed to test for $H_0 : b = 1$. More precisely, the test statistic is defined as

$$LR_{\text{ind}}^{\text{Wei}} = 2 \log \left(\frac{L^{\text{Wei}}(D_1, \ldots, D_{M_1-1}; \hat{p}_{\text{Wei}}, \hat{b}_{\text{Wei}})}{L^{\text{Exp}}(D_1, \ldots, D_{M_1-1}; \hat{p}_{\text{Exp}})} \right) \overset{\text{asy}}{\sim} \chi^2(1), \tag{2.9}$$

where $\hat{p}_{\text{Wei}}$, $\hat{b}_{\text{Wei}}$ and $\hat{p}_{\text{Exp}}$ denote the respective maximum likelihood estimators. Note that, in the Weibull case, these values must be computed by numerical optimization. Moreover, Berkowitz et al. (2011) propose to introduce additional artificial durations at the start and the end of the violation sequence, which are then treated as censored durations. Details are omitted for the sake of brevity.

Durations-Based Test Using a Generalized Method of Moments. Candelon et al. (2011) introduce a duration-based test which needs no specification of any alternative model. Instead, it is directly tested whether the observed durations follow the geometric law, by applying a goodness-of-fit procedure based on the generalized method of moment (GMM, see also Bontemps and Meddahi, 2012).

More precisely, the geometric distribution is associated with some recursively defined sequence of orthonormal polynomials, denoted by $P_j(d;p)$ with $j \in \mathbb{N}$. The connection between these polynomials and the geometric distribution guarantees that

$$\mathrm{E}\left[P_j(D;p)\right] = 0 \quad \text{for all } j \in \mathbb{N}, \tag{2.10}$$

where D denotes a geometrically distributed variable with success probability p. After choosing a maximal order k, the property ind in (2.4) is then tested by checking whether the sample

$$P_j(D_1;\hat{p}), \ldots, P_j(D_{M_1-1},\hat{p})$$

is approximately centred, for all $j = 1, \ldots, k$. The precise test statistic, which is asymptotically $\chi^2(k-1)$ distributed, can be found in Candelon et al. (2011). The test will subsequently be denoted by $GMM_{\text{ind}}^{(k)}$.

Durations-Based Test Using the Sum of Squared Durations. The last test on our list is proposed by Ziggel et al. (2014) and exploits the fact that violation times $t_1, \ldots, t_{M_1}$ should be equally spread across the sample $\{1, \ldots, n\}$ if violation clustering is not existent. This can be measured by calculating the sum of squared durations as

$$MCS_{\text{ind}} = t_1^2 + (n - t_{M_1})^2 + \sum_{i=1}^{M_1-1} D_i^2. \tag{2.11}$$

If the violation sequence exhibits no clustering, then MCS_{ind} is typically small. Note that Equation 2.11 includes both the censored duration t_1, with implicit violation time 0, and $(n - t_{M_1})$, with implicit violation time n. The original work does not provide the asymptotic distribution. Therefore, this test relies necessarily on Monte-Carlo simulations which justifies the abbreviation MCS. However, this should not regarded as a drawback, since it is common to use such simulations also in situations where a test's asymptotic distribution is known. In fact, this is also the approach we apply for all new tests proposed in this paper, see Section 3.2 below for details. Nevertheless, the procedure for MCS differs from the other tests in one aspect. The statistic of a particular application of the test is compared to a distribution depending on its outcome, namely the number of violations M_1, see Appendix A.2. of Ziggel et al. (2014). This is in contrast to the other tests where such a constraint is not used.

2.3 The Extremal Index

Loosely spoken, the extremal index θ, a parameter in the interval $[0, 1]$, measures the tendency of a (strictly) stationary time series to form temporal clusters of extreme values. The formal definition is as follows, see, e.g., Embrechts et al. (1997), p. 416.

Definition 2.1. Let (e_t) be a strictly stationary sequence with stationary c.d.f. $F(x) = \Pr(e_1 \leq x)$ and let θ be a non-negative number. Assume that, for every $\tau > 0$, there exists a sequence $(u_n) = (u_n(\tau))$ such that

$$\lim_{n\to\infty} n \Pr(e_1 > u_n) = \tau,$$

and

$$\lim_{n\to\infty} \Pr(M_n \leq u_n) = \exp(-\theta\tau),$$

where $M_n = \max\{e_1, \ldots, e_n\}$. Then θ is called the **extremal index** of the sequence (e_t), and it can be shown to lie necessarily in $[0, 1]$.

The definition is fairly abstract and certainly needs some explanation. Consider an i.i.d. sequence first, and assume that the c.d.f. F of e_1 is continuous and, for simplicity, invertible. For given $\tau > 0$, we may then choose $u_n = F^{-1}(1 - \tau/n)$

to guarantee that $n\Pr(e_1 > u_n) = n\{1 - F(F^{-1}(1 - \tau/n))\} = \tau$, i.e., the first limiting relationship in the above definition is satisfied. Since $n\Pr(e_1 > u_n) = \mathrm{E}[\sum_{i=1}^n 1(e_t > u_n)]$ by linearity of expectation, we obtain that we can expect, on average, to observe τ exceedances of the threshold u_n in a sequence of length n. At the same time, it may well happen that we do not observe a single exceedance of the threshold, and this event is exactly $\{M_n \le u_n\}$. For the i.i.d. case, we obtain

$$\Pr(M_n \le u_n) = \Pr(e_1 \le u_n) \cdots \Pr(e_n \le u_n) = F(u_n)^n = \Big(1 - \frac{n\{1 - F(u_n)\}}{n}\Big)^n,$$

which converges to $\exp(-\tau)$. As a consequence, we obtain that the extremal index of an i.i.d. sequence is $\theta = 1$.

For more general time series, a similar calculation is typically much more difficult. It has however been shown that, under weak conditions on the serial dependence and if $\Pr(M_n \le u_n)$ does converge, then the limit is always of the form $\exp(-\theta\tau)$ with θ being independent of the level τ, as requested in the above definition (Leadbetter, 1983). The extremal index has been shown to exist for many common time series models, including e.g. GARCH-models (Mikosch and Starica, 2000), and is often smaller than 1 as in the i.i.d. case.

A common interpretation of the extremal index is as follows: the reciprocal of the extremal index, i.e., $1/\theta$, represents, in a suitable elaborated asymptotic framework, the expected size of a temporal cluster of extreme observations, see p. 421 in Embrechts et al. (1997). As a consequence, $\theta = 1$ means that extreme observations typically occur by oneself, while values below 1 mean that extreme observations tend to occur in temporal clusters, that is, close by in time; with the expected number of 'close-by-extreme-observations' being equal to $1/\theta$. It is exactly this interpretation which leads us to consider backtests based on the extremal index in Section 3.

2.4 Estimators for the Extremal Index

Perhaps not surprisingly, a huge variety of estimators for the extremal index has been described in the literature. Early estimators include the blocks and the runs method, see Smith and Weissman (1994) or Beirlant et al. (2004) for an overview. In this section, we describe the classical blocks estimator and two more recent methods which will be applied in the subsequent parts of this paper. In what follows, let $e_1, \dots, e_n$ be an observed stretch from a strictly stationary time series whose extremal index exists and is larger than 0.

2.4.1 The Classical Blocks Estimator

One of the most intuitive estimators is the classical blocks estimator, see Smith and Weissman (1994). This estimator is closely related to the definition of clusters and its relationship to the extremal index and relies on a block size $b = b_n$ and a threshold $u = u_n$ to be chosen by the statistician.

Divide the sample $e_1, \dots, e_n$ into n/b disjoint blocks of size b.[2] Let $M_i^{\mathrm{dj}} = \max\left\{e_{(i-1)b+1}, \dots, e_{ib}\right\}$ denote the maximum of the observations in the ith disjoint block. The set of exceedances within a block containing at least one exceedance (i.e., $M_i^{\mathrm{dj}} > u$) is regarded as a cluster. Since $1/\theta$ is the expected cluster size, this suggests to set

$$\begin{aligned}\hat{\theta}_n^{\mathrm{CB}} &= \left(\frac{1}{\sum_{i=1}^{n/b} \mathbb{1}(M_i^{\mathrm{dj}} > u)} \left(\sum_{i=1}^{n/b} \sum_{j=1}^{b} \mathbb{1}(e_{(i-1)b+j} > u) \right) \right)^{-1} \\ &= \frac{\sum_{i=1}^{n/b} \mathbb{1}(M_i^{\mathrm{dj}} > u)}{\sum_{i=1}^{n} \mathbb{1}(e_i > u)},\end{aligned}$$

which equals the number of clusters over the number of exceedances and yields the classical blocks estimator.

[2] We assume that the number of blocks n/b is an integer. If this is not the case, a possible remainder block of smaller size than b must be discarded (typically at the beginning or the end of the observation period).

2.4.2 The K-Gap Estimator

The K-Gap estimator by Süveges and Davison (2010) is based on inter-exceedance times between the extreme observations (the latter bear some similarities with the duration times introduced in a backtesting context in Section 2.1). The foundations of the estimator are laid in Ferro and Segers (2003) where it is shown that the inter-exceedance times, appropriately standardized, weakly converge to a limiting mixture model. This remains true after truncation by the so-called gap parameter $K \in \mathbb{N}$, as shown for $K = 1$ in Süveges (2007) and in the general K-gap case in Süveges and Davison (2010).

The K-gap estimator does depend on a high threshold $u = u_n$ to be chosen by the statistician and is constructed as follows. Let $M_1 = \sum_{t=1}^{n} \mathbb{1}(e_t > u)$ denote the number of exceedances of the threshold u. Let $1 \leq j_1 < \cdots < j_{M_1} \leq n$ denote the time points at which an exceedance has occured, and let $T_i = j_{i+1} - j_i$ denote the inter-exceedance time, for $i = 1, \ldots, M_1 - 1$. The K-gaps are introduced by truncating with $K > 0$, that is

$$S_i^{(K)} = \max\{T_i - K, 0\}.$$

The mentioned limiting mixture model means that a transformed inter-exceedance time (K-gaps also) follows either an exponential distribution with mean θ (with probability θ, inter-exceedance time positive) or equals zero (with probability $1-\theta$). Under the assumption of independence of the inter-exceedance times, this result can be used to derive the following log likelihood for θ:

$$\log L_K(\theta; S_i^{(K)}) = (M_1 - 1 - M_C)\log(1-\theta) + 2M_C \log\theta - \theta \sum_{i=1}^{M_1-1} \bar{F}(u_n) S_i^{(K)},$$

where $M_C = \sum_{i=1}^{M_1-1} \mathbb{1}(S_i^{(K)} \neq 0)$. The maximization of this log-likelihood yields a closed-form estimator of the extremal index given by

$$\hat{\theta}_n^{\mathrm{G}} = \hat{\theta}_n^{\mathrm{G}}(u, K) = \frac{\Sigma_2 - (\Sigma_2^2 - 8M_C\Sigma_1)^{1/2}}{2\Sigma_1} \tag{2.12}$$

where $\Sigma_1 = \sum_{i=1}^{M_1-1} \bar{F}(u_n) S_i^{(K)}$ and $\Sigma_2 = \Sigma_1 + M_1 - 1 + M_C$. In practice, one must typically replace the unknown function F by the empirical c.d.f. $\hat{F}_n$ and u_n by $\hat{F}_n^{-1}(q)$, for some value $q = q_n$ near 1. The asymptotic behavior of the estimator does seem to be known, unless one imposes additional strong assumptions such as knowledge of the c.d.f. F and independence of the inter-exceedance-times (which is not the case in general).

2.4.3 A Block Based Maximum Likelihood Estimator

A sliding block version of a maximum likelihood estimator for the extremal index has been proposed and theoretically analyzed in Northrop (2015) and Berghaus and Bücher (2017), respectively. Unlike other blocks estimators for the extremal index, it is only depending on one parameter to be chosen by the statistician, namely a block length parameter $b = b_n$. The estimator has a simple closed form expression, and is defined as follows: first, given a block length b, let $M_t^{\mathrm{sl}} = \max\{e_t, \ldots, e_{t+b-1}\}$ and $Z_t^{\mathrm{sl}} = b\{1 - F(M_t^{\mathrm{sl}})\}$, where F denotes the c.d.f. of e_1 and where $t = 1, \ldots, n-b+1$. It can be shown that the transformed block maxima Z_t^{sl} are asymptotically independent and exponentially distributed with mean θ^{-1}. Hence, after replacing F by its empirical counterpart $\hat{F}_n$, the reciprocal of the sample mean of $\hat{Z}_t^{\mathrm{sl}} = b\{1 - \hat{F}_n(M_t^{\mathrm{sl}})\}$ can be used to estimate the extremal index[3]:

$$\hat{\theta}_n^{\mathrm{B}} = \hat{\theta}_n^{\mathrm{B}}(b) = \left(\frac{1}{n-b+1} \sum_{t=1}^{n-b+1} \hat{Z}_t^{\mathrm{sl}} \right)^{-1}. \tag{2.13}$$

[3] Berghaus and Bücher (2017) propose an additional bias correction, which we do not describe here in detail, but which we employ throughout the simulation studies and the empirical applications.

Under regularity conditions on the time series and if $b = b_n \to \infty$ with $b = o(n)$, it follows from Theorem 3.1 in Berghaus and Bücher (2017) that

$$\sqrt{n/b}\,(\hat\theta_n^{\mathrm{B}} - \theta) \to \mathrm{N}(0, \theta^4\,\sigma_{\mathrm{sl}}^2),$$

where θ denotes the true extremal index and where $\sigma_{\mathrm{sl}}^2 > 0$ denotes the asymptotic variance of the sliding blocks estimator. It is worthwhile to mention that $\sigma_{\mathrm{sl}}^2 = 0.2726$ in case the extremal index is equal to 1, that is, the limiting distribution is pivotal; see Example 3.3 in Berghaus and Bücher (2017).

2.4.4 An Example

In Figure 1 we illustrate the classical blocks estimator from Section 2.4.1, the K-Gap estimator from Section 2.4.2, and the disjoint variant of the block based Maximum Likelihood estimator from Section 2.4.3 (obtained by using $\hat Z_t^{\mathrm{dj}} = b\,\{1 - \hat F_n(M_t^{\mathrm{dj}})\}$ with $M_t^{\mathrm{dj}} = \max\{e_{bt-1+1}, \ldots, e_{tb}\}$ for $t = 1, \ldots, n/b$).

The data consist of about three years of negative daily log returns on the S&P 500 index. The solid red horizontal line corresponds to the ex-post 97.5 % empirical quantile of the negative return data, i.e., $u = 0.0225$. Hence, we have exactly 20 values above this threshold. All exceedances of the threshold are labeled with a vertical dotted red line. The gray dashed lines mark the edges of the disjoint blocks. We choose a block length of $b = 40$ returns, resulting in $n/b = 20$ disjoint blocks.

The first line at the top of the Figure shows the empirical cluster sizes used for the disjoint classical blocks estimator from Section 2.4.1. The inverse of the average cluster size provides the classical blocks estimator for the extremal index, with a value of $\hat\theta^{\mathrm{CB}} = 0.45$ for the particular example.

The second line corresponds to the K-Gaps estimator from Section 2.4.2 with $K = 6$, which is depending on the threshold u and the gap parameter K, but not

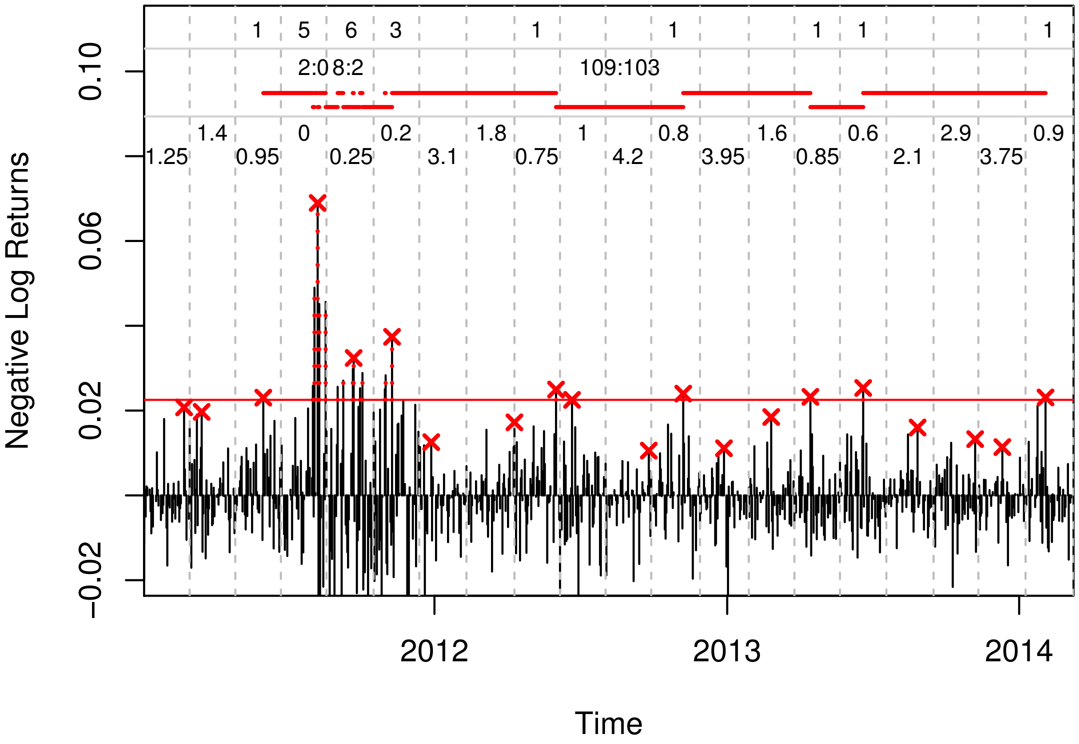

Figure 1: This plot illustrates the three extremal index estimators described in Section 2.4. The data set consists of about 800 daily returns on the S&P 500 index from 2011-01-03 till 2014-03-10. The first line at the top documents cluster sizes, the second line shows three examples for durations and 6-Gaps, and the last line reports transformed block maxima.

the block size b. The partly displaced horizontal red lines serve as an illustration for the durations betweens exceedances. Three numerical examples are provided above those lines. For instance, 109:103 means that the inter-exceedance time was 109 days. This value, truncated with $K = 6$, leads to a K-Gap of 103. Since this inter-exceedance time is quite high, the truncating alters little. However, in the first example the duration is 2 results in a K-Gap of 0. The final estimated value is $\hat{\theta}_n^{\mathrm{G}} = 0.51$.

The blocks estimator $\hat{\theta}_n^{\mathrm{B}}$ from Section 2.4.3 is only depending on the block size

b and is based on computing maxima within each block. In particular, such a block maximum (red crosses) can also be below the threshold, as for example for blocks 1–2 in the picture. The block maxima are then transformed to the pseudo-observations $\hat{Z}_t^{\mathrm{dj}}$, which are reported in the third line of the plot. Here, the inverse of the mean yields an estimate of $\hat{\theta}_n^{\mathrm{B}} = 0.62$.

In this example, all estimates are quite similar and show that the negative S&P 500 returns exhibit a large degree of extremal dependence in their right tail.[4] However, it is well know that such estimates can deviate largely depending on the estimator and parameter choice[5].

3 Backtesting Based on the Extremal Index

If the use of correct VaR forecasts leads to an i.i.d. Bernoulli violation sequence, it seems natural to expect that there exists a VaR-adjusted return series which does not exhibit any (or only low) serial dependence, provided the true $\mathrm{VaR}_p^{(t)}$-values are used. In fact, given some arbitrary forecasts $\widehat{\mathrm{VaR}}_p^{(t)}$, we propose to consider the following negative return-VaR ratio, defined as

$$e_t := e_t(\widehat{\mathrm{VaR}}_p^{(t)}) := -\frac{r_t}{\widehat{\mathrm{VaR}}_p^{(t)}}. \tag{3.1}$$

Note, the negative sign in front of the ratio, which implies that by looking at the right tail of (e_t), we essentially look at the right tail of (r_t). There is an obvious relationship with the violation sequence (I_t) defined in (2.1): we have $\{e_t > 1 \Leftrightarrow I_t = 1\}$ and $\{e_t \leq 1 \Leftrightarrow I_t = 0\}$, but (e_t) obviously contains much more information.

[4] This implies extremal dependence in the left tail of the S&P 500 returns.

[5] See for example Tables 8.1.8 and 8.1.9 in Embrechts et al. (1997).

3.1 Relative Excess Returns of Mean-Scale Models

It is instructive to first consider the relative excess return series (e_t), with $\widehat{\mathrm{VaR}}_p^{(t)} = \mathrm{VaR}_p^{(t)}$, in a general mean-scale model defined by

$$r_t = \mu_t + \sigma_t \, z_t, \qquad \text{where } z_t \overset{\text{i.i.d.}}{\sim} F_z \tag{3.2}$$

and where $\mathrm{E}(z_t) = 0$, $\mathrm{Var}(z_t) = 1$ and μ_t, σ_t are $\mathcal{F}_{t-1}$ -measurable. As a consequence, the conditional VaR using information up to $t-1$ can be written as

$$\mathrm{VaR}_p^{(t)} = -\mu_t - \sigma_t F_z^{-1}(p), \tag{3.3}$$

which implies that

$$e_t = \frac{\mu_t + \sigma_t \, z_t}{\mu_t + \sigma_t \, F_z^{-1}(p)}. \tag{3.4}$$

We are next going to argue that the sequence (e_t) is either an i.i.d. sequence (zero mean case) or at least approximately serially independent (non-zero mean case), in particular when looking only at the left tail. This suggests to backtest the VaR-forecasts by checking for serial independence or the absence of extremal clustering of the relative excesses (e_t), as will be done in later sections.

The Zero Mean Case. If $\mu_t \equiv 0$, then Formula 3.4 simplifies to $e_t = z_t / F_z^{-1}(p)$. As a consequence, (e_t) is an i.i.d. sequence due to the i.i.d. property of the innovations (z_t).

In practice, the possibility of a non-zero mean cannot be ruled out. However, it is often argued that financial returns show only insignificant means, see, e.g., Hansen (2005). In that paper, a large number of mean-scale models is examined with respect to their volatility forecasting performance, relative to simple specifications such as the classical GARCH(1,1)-model. Three different specifications for the conditional mean are used and it is concluded that the performance is almost identical across the three versions. In other words, for financial returns, the mean is often negligible, especially in the short-term. This stylized fact is also assured by the popularity of methods and models which explicitly use the assumption of zero conditional means. A prime example is provided by the famous square-root-of-time rule for time-scaling of the volatility and VaR. The rule is well appreciated among academics and practitioners, and is even implemented in regulatory standards for VaR scaling from daily returns to 10-day-returns (see BCBS, 1996a).

The General Case. Next, consider the general case with $\mu_t \neq 0$ being allowed. The event $\{e_t > y\}$ can then be rewritten as

$$S_t(y) = \left\{ z_t < y\left(\frac{\mu_t}{\sigma_t} + F_z^{-1}(p)\right) - \frac{\mu_t}{\sigma_t} \right\},$$

and this representation suggests that the relative excess returns are in general not serially independent: the events $\{e_t > y\}$ and $\{e_{t+1} > x\}$ are connected through the conditional mean and volatility. However, we argue that the serial dependence is actually either vanishing or low in certain typical cases.

The first case is $x = y = 1$, in which case $S_t(1) = \{z_t < F_z^{-1}(p)\} = \{I_t = 1\}$, which is obviously independent over time. In fact, we are left with the classical violation sequence (I_t).

Next, in case of either $\mu_{t+1} \approx 0$ or $\sigma_{t+1} \to \infty$, we get at least approximate equality of $S_t(y)$ and $\{z_t < F_z^{-1}(p)\}$ and hence approximate serial independence of (e_t). Note that large volatilities σ_t are typically present in periods of financial turmoil, which are in turn associated with our phenomenon of interest, that is, violation clustering.

More generally, the serial dependence vanishes for $x, y \geq 1$ whenever $-F_z^{-1}(p) \gg \mu_t/\sigma_t$ for all t with high probability, which is reasonable for large values of q. In that case, $S_t(y)$ implies $z_t \ll F_z^{-1}(p)$, so that only very small values of z_t may trigger the event $S_t(y)$. Since z_t is an i.i.d. sequence, the events $S_t(y)$ are approximately serially independent too, with high probability.

3.2 The Backtesting Procedure for VaR

The discussion in the preceding section motivates to backtest VaR-forecasts by checking whether the relative excess sequence (e_t) exhibits no or only low serial dependence, especially in the right tail. As a proxy, we propose to check for extremal clustering by using the extremal index, whence the resulting test can in fact be expected to be particularly sensitive to deviations from independence in the far right tail of e_t, i.e., in the most important part for risk management needs. More precisely, recalling that an independent sequence has extremal index 1, we aim at checking for what we coin *no cluster property* (noc):

$$\text{noc: } \theta_e := \theta((e_t)_t) = 1. \tag{3.5}$$

The backesting procedure we propose then is as follows: first, given a sequence of VaR forecasts $\widehat{\mathrm{VaR}}_p^{(t)}$ and observed returns r_t, for $t = 1, \ldots, n$, calculate (e_t) as defined in (3.1). Second, calculate $\hat{\theta}_n = \hat{\theta}_n(e_1, \ldots, e_n)$ with $\hat{\theta}_n$ denoting any of the extremal index estimators from Section 2.4. Finally, reject the VaR-forcasts if $\hat{\theta}_n$ is significantly smaller than 1. Regarding the extremal index estimators, we only consider the sliding blocks estimator from Section 2.4.3 and the K-Gap estimator from Section 2.4.2; the resulting tests will subsequently be denoted by $\Theta^{\mathrm{B}}_{\mathrm{noc}} = \Theta^{\mathrm{B}}_{\mathrm{noc}}(b)$ and $\Theta^{\mathrm{G}}_{\mathrm{noc}} = \Theta^{\mathrm{G}}_{\mathrm{noc}}(u, K)$, respectively.

Regarding test $\Theta^{\mathrm{B}}_{\mathrm{noc}}$, we first need to choose a block length parameter b. A preliminary Monte Carlo simulation study to compare several values of b; details can be found in Table 8 in Appendix A.1; guides us to choose $b = 40$ across all further analyses. Although more suitable choices may be possible depending on the data generating process (DGP), we set a general data-dependent strategy for the choice of b aside, thus possibly resigning power in some cases.

Critical values for test $\Theta^{\mathrm{B}}_{\mathrm{noc}}$ could in principal be calculated based on the normal approximation described in Section 2.4.3: if the extremal index is 1, then

$\hat{\theta}_b - 1$ is approximately centred normal with variance $0.2726 \cdot b/n$, no matter the stationary distribution F_e or the serial dependence of the time series outside the upper tail. However, the fact that the limiting distribution is pivotal also allows to approximate it by simulating from an arbitrary model for which the extremal index is 1. We hence opt for calculating critical values by first simulating $\tilde{e}_1, \ldots, \tilde{e}_n$ from the model

$$\tilde{e}_t = \frac{\tilde{r}_t}{-\operatorname{VaR}_p^{(t)}}, \quad \tilde{r}_t \overset{\text{i.i.d.}}{\sim} N(0,1), \quad \widehat{\operatorname{VaR}}_p^{(t)} = -\Phi^{-1}(p), \tag{3.6}$$

where Φ denotes the c.d.f. of the standard normal distribution, and by then calculating $\hat{\theta}_n^{\mathrm{B}} = \hat{\theta}_n^{\mathrm{B}}(\tilde{e}_1, \ldots, \tilde{e}_n)$ with the same block length parameter as chosen above, i.e., $b = 40$. Note that such a simulation based approach is also common for other classical backtesting procedures where asymptotic distributions are known, e.g., for the test $LR_{\mathrm{ind}}^{\mathrm{Mar}}$ defined in Section 2.2.

Let us next describe details on test $\Theta_{\mathrm{noc}}^{\mathrm{G}}$, which depends on the choice of both K and $u = u_n$. Regarding the latter choice of u, we simply set $u = 1$, which essentially means that we leave the extreme value context and are back to the 0-1-violation sequence from Section 2.1 (note that $e_t > 1$ if and only if $I_t = 1$). The K-gap approach should hence rather be regarded as a classical duration-based test for the hypothesis of independence of the innovation sequence given in (2.4) (though with a different initial motivation), see also Section 2.2 for other duration-based tests.[6] In particular, this viewpoint suggests to obtain critical values of the test simply by generating i.i.d. Bernoulli(p)-sequences $\tilde{I}_t$, and to calculate $\hat{\theta}_n^{\mathrm{G}} = \hat{\theta}_n^{\mathrm{G}}(K)$ by considering each time points where $\tilde{I}_t = 1$ as an exceedance (note that $\hat{\theta}_n^{\mathrm{G}}$ only depends on those time points). Regarding the choice of the K-gap parameter, a further preliminary simulation study (details are presented in Table 7 in Appendix A.1) prompts us choose $K = 6$ for all further analyses.

[6] A more appropriate notation would hence be $\Theta_{\mathrm{ind}}^{\mathrm{G}}$ instead of $\Theta_{\mathrm{noc}}^{\mathrm{G}}$.

Throughout this paper, the two above described simulation-based approaches, as well as all other similar approaches, are based on $N = 10,000$ replications, and corresponding p-values are computed as described in Appendix A.2, see also Dufour (2006).

3.3 Extensions to More General Risk Measures Including Expected Shortfall

Backtesting the Expected Shortfall (ES) recently received increased attention due to its upcoming implementation as a standard risk measure for regulatory purposes in banking (BCBS, 2016). Most available backtests of ES focus on unconditional coverage, see Kerkhof and Melenberg (2004), Wong (2008), Costanzino and Curran (2015), and Kratz et al. (2018). Only recently, Du and Escanciano (2017) propose to additionally use a Box-Pierce test to test for autocorrelation in a certain sequence of cumulative violations. This test can hence be regarded as the first ES backtest for independence (rather: serial uncorrelation). In this section, we extend the basic idea from Section 3.2 to obtain a further backtest for ES that is particularly sensitive to certain deviations from independence in the tails.

Recall that the main idea of the VaR method from Section 3.2 consists of checking whether the relative excess returns in (3.1) do not show any sign of extremal clustering. The sensibility of such an approach was explained in Section 3.1 for mean-scale models, and the arguments can in fact be generalized to any risk measure which is translation invariant and positively homogeneous. Indeed, recall that a risk measure $\rho : M \to \mathbb{R}$, M a set of random variables, satisfies translation invariance if, for all $R \in M$ and every $c \in \mathbb{R}$, we have $\rho(R + c) = \rho(R) - c$ (the change of the sign stems from interpreting R as a return and not a loss). Positive homogeneity is satisfied if $\rho(c\,R) = c\,\rho(R)$ for all $R \in M$ and $c > 0$ (McNeil et al., 2005). By the same arguments as in Section 3.1, it is sensible to backtest a sequence of forecasts $\hat{\rho}_t$ by checking whether the sequence

$$e_t = -\frac{r_t}{\hat{\rho}_t} \tag{3.7}$$

does not so any sign of extremal clustering. Indeed, for location scale models as defined in (3.2) and for $\hat{\rho}_t = \rho_t = \rho(r_t \mid \mathcal{F}_{t-1})$, we obtain that

$$e_t = -\frac{r_t}{\rho_t} = \frac{\mu_t + \sigma_t z_t}{\mu_t - \sigma_t \rho(z_t)},$$

which simplifies to Equation 3.4 if we use $\rho(z_t) = -F_z^{-1}(p)$, that is, VaR. As a consequence, it is sensible to apply the methodology described in Section 3.2 to the sequence $(e_t)_t$ defined in (3.7), for any translation invariant and homogeneous risk measure. We do not pursue this any further in this document.

3.4 An Extension to Distributional Backtests

The general idea from Section 3.2 may also be applied to backtesting forecasts of the entire conditional distribution (or density), see also Berkowitz (2001). More precisely, suppose that $\hat{F}_t$ is a distributional forecast of the conditional c.d.f. of r_t given $\mathcal{F}_{t-1}$, the latter being denoted by F_t. The role of the VaR-adjusted return series (e_t) may then be played by the probability integral transform sequence

$$u_t = 1 - \hat{F}_t(r_t), \quad t = 1, \dots, n.$$

In case $\hat{F}_t = F_t$, the sequence is known to constitute an i.i.d. sequence of uniformly distributed random variables on the interval $[0, 1]$, see Rosenblatt (1952). As in the previous section, a distributional backtest that is particular sensitive to deviations from independence in the upper right tail of u_t is obtained by comparing the estimated extremal index of $u_1, \dots, u_n$ with 1.

4 Size and Power Analysis

In this section, we compare our new approach to several classical backtesting procedures in terms of their empirical size and power properties. The employed alternative backtests are briefly described in Section 2.2. The results of large scale Monte Carlo simulation studies are then presented in Sections 4.1–4.5, under varying circumstances of interest.

4.1 Power Properties when True Unconditional VaRs are Used

The first simulation experiment is guided by Table 4 in Ziggel et al. (2014), the purpose being to compare backtests in situations where clustered violations are likely due to the use of unconditional instead of conditional VaRs. The comparison is only carried out for *independence* backtesting procedures (including the tests for noc, as the latter also have power against most deviations from ind). The data generating process is as follows:

$$r_t = \sigma_t \, z_t, \qquad t = 1, \ldots, n,$$

with $z_t \overset{\text{i.i.d.}}{\sim} N(0,1)$, $\sigma_1 = 1$ and

$$\sigma_t^2 = \lambda \sigma_{t-1}^2 + (1-\lambda) z_{t-1}^2, \qquad t = 2, \ldots, n.$$

As in Ziggel et al. (2014), the parameter λ is chosen from the set $\{0.8706, 0.9829, 0.9914, 1\}$ (case $\lambda = 1$ will correspond to the null hypothesis) and the sample size n is chosen from the set $\{252, 1000, 2500\}$. Note that results for sample sizes as small as 252 should be regarded with a little caution, at least for test $\Theta^{\mathrm{B}}_{\mathrm{noc}}$: taking blocks of length $b = 40$ results in only 6 disjoint block maxima (and slightly more distinct values for the sliding block maxima), to which an exponential distribution is then fit. However, since we do not rely on asymptotics but rather on simulation to calculate critical values, we still think that an application to such small sample sizes is sensible (and in fact, the results confirm our intuition).

Recall from Section 3.1 that the true conditional VaR of the above described model is given by $\mathrm{VaR}_p^{(t)} = -\sigma_t \Phi^{-1}(p)$ and that the use of $\widehat{\mathrm{VaR}}_p^{(t)} = \mathrm{VaR}_p^{(t)}$ in formula (3.1) would result in an i.i.d. sequence of relative excess returns. Serial dependence (and in particular extremal clustering) is now introduced by instead setting $\widehat{\mathrm{VaR}}_p^{(t)} = \widehat{\mathrm{VaR}}_p$ (independent of t) to the empirical VaR computed from a preliminary simulation with 100,000 returns. The simulation study is performed conditional on the restriction of at least two violations per backtesting sample.[7] As described in Section 3, the parameter b is set to $b = 40$ for test $\Theta_{\mathrm{noc}}^{\mathrm{B}}$ and to $K = 6$ for test $\Theta_{\mathrm{noc}}^{\mathrm{K}}$.

The results, namely empirical rejection rates of the competing independence backtests (calculated based on 5000 Monte Carlo repetitions), are reported in Table 1. All tests exhibit a reasonable approximation of the intended level (case $\lambda = 1$). In terms of power, the proposed extremal index tests typically yield the largest power, which on top is often much larger than for the classical competitors. In the few cases the non-extremal index tests yield larger power, the improvement over the extremal index versions is rather small.

The rejection rates of all approaches except $\Theta_{\mathrm{noc}}^{\mathrm{B}}$ are decreasing in the VaR level. Clearly, the reason is that a small number of violations cannot yield the same evidence for serial dependence like a large number of violations is capable of. For $\Theta_{\mathrm{noc}}^{\mathrm{B}}$, the rejection rates change barely for varying level, a likely explanation being that the input data (relative excess returns) are approximately the same up to a scaling factor. Moreover, by construction, $\Theta_{\mathrm{noc}}^{\mathrm{B}}$ is also able to use information

[7] The probability that a generated sample of size n violates this condition is negligible in all cases except for the 1 % VaR level and $n = 252$ case, where approximately 28.2 % of randomly drawn samples exhibit at most one violation.

λ	n	Significance level: 1%						Significance level: 5%						Significance level: 10%					
		$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind}}^{\text{Wei}}$	$GMM_{\text{ind}}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc}}^{\text{G}}(6)$	$\Theta_{\text{noc}}^{\text{B}}(40)$	$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind}}^{\text{Wei}}$	$GMM_{\text{ind}}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc}}^{\text{G}}(6)$	$\Theta_{\text{noc}}^{\text{B}}(40)$	$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind}}^{\text{Wei}}$	$GMM_{\text{ind}}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc}}^{\text{G}}(6)$	$\Theta_{\text{noc}}^{\text{B}}(40)$
Panel A: 5 % VaR																			
0.8706	252	0.056	0.008	0.110	0.093	0.115	**0.138**	0.147	0.040	0.235	0.229	**0.276**	0.269	0.200	0.086	0.327	0.340	**0.388**	0.385
	1000	0.130	0.037	0.204	0.231	0.396	**0.528**	0.210	0.143	0.579	0.505	0.621	**0.738**	0.284	0.251	0.705	0.653	0.728	**0.833**
	2500	0.325	0.158	0.568	0.588	0.772	**0.920**	0.514	0.382	0.903	0.837	0.901	**0.975**	0.626	0.524	0.951	0.919	0.939	**0.991**
0.9828	252	0.011	0.017	0.021	0.023	0.019	**0.028**	0.061	0.051	0.068	0.090	0.080	**0.100**	0.105	0.097	0.125	0.160	0.144	**0.169**
	1000	0.014	0.005	0.050	0.069	0.046	**0.097**	0.045	0.023	0.157	0.200	0.140	**0.241**	0.089	0.054	0.242	0.301	0.239	**0.370**
	2500	0.013	0.002	0.085	0.134	0.086	**0.225**	0.056	0.009	0.264	0.328	0.234	**0.479**	0.110	0.022	0.354	0.472	0.338	**0.610**
0.9914	252	0.009	0.011	0.011	0.013	0.010	**0.017**	0.056	0.052	0.055	**0.067**	0.055	**0.067**	0.098	0.103	0.106	**0.127**	0.114	0.117
	1000	0.010	0.005	0.021	0.028	0.021	**0.036**	0.042	0.032	0.090	0.108	0.081	**0.133**	0.082	0.068	0.154	0.180	0.147	**0.224**
	2500	0.013	0.003	0.048	0.062	0.034	**0.091**	0.050	0.017	0.136	0.175	0.106	**0.231**	0.102	0.035	0.210	0.287	0.185	**0.346**
1 (H_0)	252	0.009	0.011	0.014	0.012	0.008	0.014	0.052	0.044	0.047	0.058	0.046	0.056	0.102	0.096	0.095	0.113	0.099	0.109
	1000	0.014	0.008	0.009	0.012	0.013	0.010	0.053	0.047	0.048	0.054	0.060	0.054	0.097	0.101	0.105	0.103	0.121	0.104
	2500	0.006	0.012	0.011	0.010	0.008	0.012	0.048	0.057	0.060	0.059	0.054	0.054	0.098	0.103	0.114	0.113	0.112	0.103
Panel B: 1 % VaR																			
0.8706	252	0.069	0.036	0.089	0.050	0.072	**0.152**	0.197	0.109	0.211	0.139	0.227	**0.322**	0.263	0.165	0.291	0.220	0.354	**0.445**
	1000	0.099	0.022	0.089	0.046	0.184	**0.527**	0.216	0.120	0.215	0.167	0.372	**0.739**	0.337	0.216	0.272	0.277	0.497	**0.848**
	2500	0.204	0.168	0.138	0.076	0.439	**0.914**	0.397	0.374	0.321	0.255	0.644	**0.974**	0.493	0.491	0.409	0.401	0.748	**0.989**
0.9828	252	0.019	0.018	0.028	0.015	0.019	**0.039**	**0.115**	0.078	0.112	0.082	0.077	0.112	**0.213**	0.169	0.196	0.150	0.180	0.191
	1000	0.016	0.015	0.050	0.035	0.020	**0.096**	0.064	0.050	0.118	0.120	0.091	**0.236**	0.162	0.104	0.186	0.209	0.160	**0.352**
	2500	0.019	0.016	0.065	0.051	0.035	**0.229**	0.091	0.078	0.176	0.180	0.125	**0.452**	0.155	0.142	0.259	0.303	0.214	**0.591**
0.9914	252	0.017	0.016	**0.022**	0.015	0.011	0.016	**0.107**	0.073	0.092	0.057	0.067	0.064	**0.199**	0.144	0.170	0.117	0.153	0.127
	1000	0.016	0.019	0.031	0.027	0.014	**0.045**	0.061	0.065	0.091	0.096	0.069	**0.131**	0.130	0.120	0.142	0.170	0.132	**0.223**
	2500	0.014	0.008	0.043	0.044	0.024	**0.080**	0.070	0.055	0.131	0.149	0.090	**0.229**	0.124	0.115	0.196	0.237	0.170	**0.344**
1 (H_0)	252	0.014	0.012	0.009	0.006	0.015	0.011	0.068	0.076	0.061	0.047	0.065	0.043	0.144	0.148	0.137	0.100	0.134	0.091
	1000	0.011	0.009	0.009	0.008	0.012	0.007	0.049	0.051	0.053	0.049	0.050	0.043	0.100	0.103	0.102	0.103	0.099	0.097
	2500	0.011	0.008	0.008	0.008	0.012	0.012	0.054	0.051	0.044	0.047	0.052	0.049	0.104	0.100	0.100	0.098	0.097	0.103

Table 1: Rejection rates for several DGPs, backtesting samples sizes, VaR levels, significance levels and backtesting procedures. The setting is borrowed from Table 4 in Ziggel et al. (2014). The DGPs produce clustered violations by the usage of a constant VaR forecast obtained as the unconditional empirical VaR of a simulated path of length 100,000. The additional DGP with $\lambda = 1$ allows to check for the sizes of the tests. Furthermore, the DGPs are simulated subject to at least 2 violations. The rejection rates are based on 5,000 Monte-Carlo replications. The tests use Monte-Carlo p-values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

of events where violations did occur almost, see also Figure 1. This questions whether it is meaningful to asses the independence property of 1 % VaR forecasts in small samples based solely on violations sequences.

4.2 How Often can we Reject Historical Simulation?

The second simulation is inspired by Candelon et al. (2011) and Christoffersen and Pelletier (2004). Again, returns are simulated using a mean scale model with $\mu_t \equiv 0$ and innovations $z_t \overset{\text{i.i.d.}}{\sim} \sqrt{\frac{d-2}{d}}\,\varepsilon_t$, where ε_t follows a student t-distribution with d degrees of freedom. The conditional variance involves an asymmetric leverage effect and is given by

$$\sigma_t^2 = \omega + \gamma\sigma_{t-1}^2\left(z_{t-1} - \theta\right)^2 + \beta\sigma_{t-1}^2, \qquad t \in \mathbb{N}_{\geq 2},$$

where $\gamma = 0.1, \theta = 0.5, \beta = 0.85, \omega = 3.9683 \cdot 10^{-6}$ and $d = 8$. We set $\sigma_1^2 = \omega$ and use a burn-in period of length $N_{\text{burn-in}} = 200$ before forecasting is started.

Time-varying forecasts are obtained by applying the popular and realistic VaR forecasting technique of (unconditional) 'Historical Simulation' (HS): given an integer $T_e \in \{250, 500\}$, we estimate the conditional VaR at time t by the respective empirical quantile (multiplied with -1) of the T_e observations prior to time point t. The experiment is hence in contrast to the scenario from Section 4.1, where one fixed VaR forecast was used for all t. Still, since HS is not able to capture the dynamics of the time-varying volatility adequately either, the forecasting method should be rejected. We hence allow for an assessment of the methods' power in a more realistic environment. For completeness, we also use as additional competitors to the independence tests the *conditional coverage* backtests described in Section 2.2, which also exploit information about the number of violations. Note that we have to simulate $N_{\text{burn-in}} + T_e + n$ returns in total per replication to obtain the results of the simulation experiment.

T_e	n	Significance level: 1%						Significance level: 5%						Significance level: 10%					
		$LR^{\text{Mar}}_{\text{cc}}$	$LR^{\text{Wei}}_{\text{cc}}$	$GMM^{(5)}_{\text{cc}}$	MCS_{ind}	$\Theta^{\text{G}}_{\text{noc}}(6)$	$\Theta^{\text{B}}_{\text{noc}}(40)$	$LR^{\text{Mar}}_{\text{cc}}$	$LR^{\text{Wei}}_{\text{cc}}$	$GMM^{(5)}_{\text{cc}}$	MCS_{ind}	$\Theta^{\text{G}}_{\text{noc}}(6)$	$\Theta^{B}_{\text{noc}}(40)$	$LR^{\text{Mar}}_{\text{cc}}$	$LR^{\text{Wei}}_{\text{cc}}$	$GMM^{(5)}_{\text{cc}}$	MCS_{ind}	$\Theta^{\text{G}}_{\text{noc}}(6)$	$\Theta^{\text{B}}_{\text{noc}}(40)$
Panel A: 5 % VaR																			
250	250	0.179	0.154	0.146	0.295	0.265	**0.430**	0.322	0.291	0.430	0.484	0.429	**0.610**	0.411	0.363	0.543	0.580	0.515	**0.695**
	500	0.172	0.223	0.275	0.548	0.523	**0.782**	0.320	0.389	0.654	0.746	0.693	**0.878**	0.413	0.486	0.757	0.833	0.764	**0.920**
	750	0.203	0.313	0.389	0.724	0.676	**0.914**	0.328	0.524	0.795	0.881	0.811	**0.962**	0.438	0.630	0.870	0.931	0.875	**0.978**
	1000	0.262	0.446	0.548	0.812	0.775	**0.967**	0.408	0.629	0.879	0.938	0.893	**0.988**	0.470	0.728	0.935	0.969	0.937	**0.993**
	1500	0.352	0.624	0.764	0.931	0.916	**0.997**	0.479	0.789	0.962	0.982	0.966	**1.000**	0.623	0.857	0.983	0.995	0.982	**1.000**
500	250	0.252	0.198	0.182	0.266	0.269	**0.394**	0.409	0.303	0.473	0.445	0.425	**0.578**	0.492	0.383	0.561	0.555	0.513	**0.675**
	500	0.283	0.284	0.344	0.527	0.513	**0.766**	0.441	0.456	0.716	0.736	0.657	**0.876**	0.532	0.554	0.812	0.826	0.734	**0.921**
	750	0.282	0.369	0.436	0.703	0.680	**0.903**	0.436	0.586	0.819	0.860	0.805	**0.958**	0.552	0.682	0.896	0.923	0.858	**0.976**
	1000	0.337	0.510	0.622	0.818	0.775	**0.967**	0.489	0.696	0.901	0.932	0.879	**0.988**	0.552	0.778	0.950	0.962	0.915	**0.995**
	1500	0.428	0.705	0.830	0.945	0.923	**0.995**	0.562	0.841	0.978	0.983	0.964	**0.998**	0.700	0.899	0.991	0.993	0.980	**1.000**
Panel B: 1 % VaR																			
250	250	0.105	0.079	0.191	0.097	0.085	**0.469**	0.210	0.195	0.327	0.209	0.207	**0.635**	0.288	0.251	0.395	0.310	0.306	**0.717**
	500	0.071	0.116	0.210	0.178	0.166	**0.812**	0.173	0.193	0.417	0.374	0.366	**0.892**	0.270	0.228	0.512	0.490	0.486	**0.932**
	750	0.103	0.119	0.104	0.206	0.247	**0.919**	0.163	0.237	0.421	0.409	0.476	**0.970**	0.290	0.364	0.547	0.544	0.590	**0.983**
	1000	0.105	0.207	0.072	0.234	0.336	**0.975**	0.198	0.376	0.491	0.469	0.563	**0.989**	0.276	0.479	0.612	0.614	0.672	**0.994**
	1500	0.137	0.372	0.142	0.287	0.492	**0.998**	0.301	0.549	0.596	0.559	0.701	**1.000**	0.358	0.641	0.725	0.694	0.782	**1.000**
500	250	0.122	0.091	0.194	0.099	0.087	**0.403**	0.224	0.183	0.284	0.197	0.197	**0.579**	0.366	0.228	0.341	0.274	0.282	**0.677**
	500	0.148	0.187	0.257	0.225	0.180	**0.791**	0.279	0.266	0.425	0.396	0.349	**0.880**	0.386	0.319	0.515	0.494	0.447	**0.920**
	750	0.175	0.235	0.197	0.321	0.326	**0.924**	0.253	0.373	0.517	0.507	0.510	**0.973**	0.377	0.479	0.615	0.621	0.612	**0.985**
	1000	0.161	0.354	0.167	0.395	0.423	**0.968**	0.299	0.512	0.611	0.614	0.628	**0.989**	0.385	0.608	0.708	0.718	0.717	**0.993**
	1500	0.213	0.532	0.260	0.507	0.606	**0.997**	0.383	0.683	0.708	0.719	0.768	**1.000**	0.444	0.746	0.807	0.808	0.833	**1.000**

Table 2: Rejection rates of Historical Simulation VaR forecasts with two estimation window sizes $T_e \in \{250, 500\}$ across several backtesting samples sizes, VaR levels and backtesting procedures. The setting is borrowed from Table 3 in Candelon et al. (2011). Compared to this, we dropped the GMM test with $p = 3$ and added instead the MCS independence test. The rejection rates are based on 5,000 Monte-Carlo replications. The tests use Monte-Carlo p-values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

Results of the simulation experiment are reported in Table 2. Focusing only at the classical methods first, we find that typically one of $LR_{\text{cc}}^{\text{Wei}}$, $GMM_{\text{cc}}^{(5)}$, or MCS_{ind} backtests yields the largest power. The 0-1-Extremal Index approach $\Theta_{\text{noc}}^{\text{G}}$ shows overall comparable rejection rates to these - sometimes the power is larger, sometimes smaller and sometimes there is barely any difference. In addition, note that $\Theta_{\text{noc}}^{\text{G}}$ shows slight improvements over its related backtest $LR_{\text{cc}}^{\text{Wei}}$.[8] Finally, the second extremal index test $\Theta_{\text{ind}}^{\text{B}}$ is able to improve the power in every case under consideration, sometimes by a considerable amount.

4.3 Rejection Rates of (Misspecified) Stochastic Volatility Models

In this section, we study backtest rejection rates for forecasts based on estimated, but possibly misspecified stochastic volatility models. The underlying data generating process is fixed as a certain GJR-GARCH(1,1)-model with student t innovations and a non-zero mean parameter μ, with the model parameters being chosen as the estimated values obtained by fitting the model to daily S&P 500 log returns from 1st January 2012 to 1st January 2015 (754 observations), see Appendix A.3 for details. Note that all parameters of the model where found to be highly significant in the latter fit, including the mean parameter $\mu = 6.91 \times 10^{-4}$. Hence, in light of the discussion in Section 3.1, the present setting also serves as a robustness check of the extremal index tests against a non-zero mean.

For each Monte Carlo repetition, a time series of length $n + 1000$, with $n \in \{252, 1000, 2500\}$, is simulated from the above described model. Three forecasting methods are then investigated, based on either a GJR-GARCH(1,1), a GARCH(1,1) or an ARCH(1)-model fit to the first 1,000 observations of the time series, and a subsequent VaR forecast based on the respective estimated model and the realized

[8] Under the null both use an exponential distribution of durations/K-Gaps.

FC	n	Significance level: 1%						Significance level: 5%						Significance level: 10%					
		$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind}}^{\text{Wei}}$	$GMM_{\text{ind}}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc}}^{\text{G}}(6)$	$\Theta_{\text{noc}}^{\text{B}}(40)$	$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind}}^{\text{Wei}}$	$GMM_{\text{ind}}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc}}^{\text{G}}(6)$	$\Theta_{\text{noc}}^{\text{B}}(40)$	$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind}}^{\text{Wei}}$	$GMM_{\text{ind}}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc}}^{\text{G}}(6)$	$\Theta_{\text{noc}}^{\text{B}}(40)$
Panel A: 5 % VaR																			
True VaRs	252	0.015	0.007	0.009	0.007	0.010	0.012	0.058	0.041	0.040	0.044	0.045	0.050	0.114	0.089	0.086	0.094	0.085	0.100
	1000	0.006	0.010	0.014	0.015	0.010	0.007	0.038	0.051	0.057	0.046	0.050	0.048	0.086	0.104	0.110	0.093	0.118	0.106
	2500	0.010	0.007	0.006	0.006	0.012	0.006	0.049	0.039	0.048	0.041	0.052	0.042	0.096	0.089	0.086	0.085	0.098	0.085
GJR-GARCH(1,1)	252	0.013	0.013	0.014	0.011	0.013	**0.023**	0.060	0.058	0.047	0.049	**0.061**	0.060	0.110	0.100	0.105	0.096	**0.113**	0.105
	1000	0.016	0.008	0.013	0.014	0.025	**0.026**	0.052	0.055	0.063	0.063	0.070	**0.075**	0.093	0.102	**0.138**	0.117	0.127	0.132
	2500	0.018	0.024	0.019	0.022	**0.042**	0.031	0.076	0.075	0.089	0.088	**0.108**	0.097	0.136	0.138	0.157	0.144	**0.170**	0.167
GARCH(1,1)	252	0.034	0.006	0.058	0.046	**0.077**	0.060	0.109	0.032	0.138	0.142	**0.200**	0.135	0.166	0.062	0.201	0.238	**0.298**	0.225
	1000	0.080	0.014	0.066	0.072	**0.225**	0.134	0.132	0.038	0.253	0.217	**0.428**	0.294	0.194	0.072	0.368	0.339	**0.574**	0.426
	2500	0.169	0.021	0.157	0.186	**0.487**	0.369	0.315	0.062	0.495	0.419	**0.688**	0.566	0.406	0.109	0.611	0.557	**0.775**	0.680
ARCH(1)	252	0.017	0.016	0.154	0.119	0.200	**0.220**	0.067	0.051	0.312	0.301	**0.405**	0.398	0.114	0.102	0.388	0.441	**0.504**	0.500
	1000	0.050	0.136	0.431	0.464	0.653	**0.759**	0.098	0.312	0.764	0.704	0.785	**0.882**	0.151	0.430	0.850	0.814	0.845	**0.930**
	2500	0.106	0.448	0.836	0.821	0.901	**0.984**	0.205	0.651	0.976	0.944	0.950	**0.997**	0.296	0.745	0.988	0.971	0.966	**1.000**
Panel B: 1 % VaR																			
True VaRs	252	0.010	0.006	0.012	0.006	0.011	0.012	0.045	0.052	0.054	0.048	0.057	0.048	0.100	0.117	0.107	0.110	0.118	0.094
	1000	0.007	0.004	0.013	0.010	0.008	0.014	0.050	0.060	0.052	0.052	0.045	0.044	0.096	0.106	0.110	0.096	0.085	0.108
	2500	0.010	0.014	0.014	0.011	0.010	0.020	0.050	0.052	0.055	0.058	0.050	0.079	0.101	0.104	0.111	0.110	0.102	0.146
GJR-GARCH(1,1)	252	0.015	0.006	**0.016**	0.014	0.014	0.015	0.061	0.059	0.056	0.054	0.054	**0.063**	0.106	0.097	0.108	0.098	0.107	**0.120**
	1000	0.011	0.014	0.014	0.011	0.009	**0.025**	0.046	0.062	0.066	0.057	0.058	**0.078**	0.118	0.107	0.117	0.107	0.104	**0.145**
	2500	0.012	0.011	0.013	0.010	0.021	**0.040**	0.059	0.063	0.051	0.056	0.070	**0.119**	0.110	0.124	0.113	0.109	0.126	**0.188**
GARCH(1,1)	252	0.031	0.015	0.027	0.014	0.030	**0.050**	0.111	0.056	0.088	0.067	0.106	**0.138**	0.174	0.096	0.140	0.118	0.187	**0.235**
	1000	0.048	0.018	0.047	0.027	0.070	**0.178**	0.122	0.063	0.112	0.094	0.190	**0.333**	0.258	0.121	0.171	0.179	0.286	**0.472**
	2500	0.071	0.026	0.043	0.033	0.141	**0.355**	0.219	0.098	0.123	0.119	0.307	**0.570**	0.289	0.164	0.176	0.198	0.423	**0.685**
ARCH(1)	252	0.019	0.052	0.130	0.071	0.080	**0.230**	0.140	0.088	0.228	0.159	0.200	**0.399**	0.208	0.129	0.276	0.232	0.284	**0.514**
	1000	0.029	0.121	0.269	0.148	0.360	**0.780**	0.076	0.296	0.394	0.339	0.547	**0.889**	0.198	0.409	0.474	0.470	0.631	**0.932**
	2500	0.028	0.504	0.452	0.316	0.712	**0.989**	0.125	0.693	0.644	0.550	0.820	**0.997**	0.204	0.771	0.712	0.685	0.869	**0.998**

Table 3: Rejection rates for several forecasting models using stochastic volatility are shown. We use simulated data from an estimated GJR-GARCH(1,1) model with student t innovations, see Appendix A.3. For each iteration the corresponding model is fitted using the first 1,000 simulated returns, remaining returns are used for forecasting and backtesting. The rejection rates are based on 5,000 Monte-Carlo replications. The tests use Monte-Carlo p-values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

returns up to time $t-1$. Note that the three models are included in each other, and that the latter two models are, by construction, misspecified. We hence expect increasing rejection rates in this chronology.

The results are presented in Table 3. For comparison, the first forecasting method (FC) corresponds to the usage of the true VaRs of the simulated data, which is not available in practice but for which the null hypothesis is met (hence, we do not report bold marks here). The remaining methods correspond to the three mentioned forecasting models. The main findings are summarized in the next three paragraphs.

Size. The forecasting method 'True VaRs' serves as a size benchmark. Overall, all methods exhibit a reasonable approximation of the nominal size. Deviations may in most cases be explained by simulation variance of the simulated distributions of the test statistics, as well as the Monte-Carlo simulations itself. However, we also observe a larger deviation for $\Theta_{\mathrm{noc}}^{\mathrm{B}}$ at the 1 % VaR level and a backtesting sample size of $n = 2,500$. For example, the rejection rate is 14.6 % at the 10 % significance level. A possible explanation is the non-zero mean in the DGP, whence we further investigate this issue in Section 4.5 below. This kind of oversizedness does not seem to occur for the other extremal index test $\Theta_{\mathrm{noc}}^{\mathrm{G}}$.

Estimation Risk. The forecast based on estimating the (true) GJR-GARCH(1,1) is slightly more likely to be rejected than the true VaRs. We further check the sensitivity of the backtests to estimation uncertainty in the next Section 4.4.

Rejection Rates of Misspecified Models. As expected, ARCH(1) is most likely to be rejected, followed by the standard GARCH model. Interestingly, $\Theta_{\mathrm{noc}}^{\mathrm{G}}$ performs often better than $\Theta_{\mathrm{noc}}^{\mathrm{B}}$ in the 5% VaR Panel. For the 1% Panel, the decrease in the number of violations leads to a better performance of $\Theta_{\mathrm{noc}}^{\mathrm{B}}$. In most cases, both tests are able to improve the power substantially compared to the classical competitors.

4.4 Estimation Risk

The results in Table 3 reveal that estimating the correct model is not sufficient to get correct forecasts. Hence, an additional aspect of the forecasting task in general is the ability to estimate the potentially correct model sufficiently accurate. To shed light on this issue we report in Table 4 results of a similar task as in the previous section. We estimate the true model using varying lengths of sample sizes n_{Est} ranging from 500 to 5,000 (recall that we used a fixed value of $n_{\mathrm{Est}} = 1,000$ in the previous section). The table reveals that, across all tests, the extremal index approaches are most likely to reject the estimated model. A large amount of data is hence needed for the rejection rates to approach the nominal significance level.

n_{Est}	n	Significance level: 1%						Significance level: 5%						Significance level: 10%					
		$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind}}^{\text{Wei}}$	$GMM_{\text{ind}}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc}}^{\text{G}}(6)$	$\Theta_{\text{noc}}^{\text{B}}(40)$	$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind}}^{\text{Wei}}$	$GMM_{\text{ind}}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc}}^{\text{G}}(6)$	$\Theta_{\text{noc}}^{\text{B}}(40)$	$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind}}^{\text{Wei}}$	$GMM_{\text{ind}}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc}}^{\text{G}}(6)$	$\Theta_{\text{noc}}^{\text{B}}(40)$
Panel A: 5 % VaR																			
∞	252	0.013	0.010	0.008	0.007	0.010	0.013	0.058	0.040	0.042	0.041	0.044	0.050	0.114	0.092	0.085	0.091	0.091	0.100
	1000	0.016	0.007	0.009	0.007	0.009	0.009	0.063	0.052	0.034	0.046	0.047	0.046	0.105	0.112	0.081	0.095	0.091	0.101
	2500	0.013	0.011	0.013	0.006	0.008	0.005	0.053	0.048	0.045	0.041	0.039	0.050	0.094	0.102	0.086	0.092	0.089	0.097
5000	252	0.010	0.008	0.012	0.012	0.011	**0.016**	0.044	0.041	0.049	**0.061**	0.047	0.054	0.096	0.089	0.104	**0.118**	0.115	0.101
	1000	0.010	0.008	0.010	**0.013**	0.013	0.010	0.051	0.053	**0.055**	0.051	0.054	0.048	0.097	**0.116**	0.113	0.098	0.108	0.102
	2500	0.012	0.013	**0.015**	0.013	**0.015**	0.012	0.052	0.057	0.056	0.045	0.063	**0.063**	0.107	0.114	0.104	0.099	**0.117**	0.117
1000	252	0.011	0.009	0.014	0.010	0.011	**0.016**	0.055	0.041	0.059	0.063	0.056	**0.065**	0.115	0.092	0.117	0.116	0.110	**0.120**
	1000	0.018	0.013	0.014	0.018	0.021	**0.024**	0.057	0.050	0.077	0.068	**0.080**	0.078	0.115	0.110	0.136	0.126	**0.148**	0.134
	2500	0.014	0.019	0.012	0.016	**0.035**	0.035	0.070	0.079	0.074	0.066	**0.095**	0.086	0.126	0.131	0.135	0.119	**0.158**	0.153
500	252	0.010	0.013	0.021	0.019	0.021	**0.022**	0.058	0.051	0.071	0.074	0.073	**0.078**	0.115	0.102	0.131	0.125	**0.144**	0.130
	1000	0.027	0.017	0.017	0.023	**0.050**	0.044	0.064	0.061	0.088	0.085	**0.128**	0.107	0.107	0.124	0.160	0.143	**0.203**	0.173
	2500	0.029	0.032	0.031	0.041	**0.096**	0.076	0.095	0.087	0.129	0.113	**0.188**	0.162	0.158	0.138	0.209	0.185	**0.247**	0.239
Panel B: 1 % VaR																			
∞	252	0.010	0.006	0.011	0.007	0.010	0.011	0.044	0.053	0.055	0.046	0.055	0.055	0.096	0.118	0.109	0.101	0.111	0.103
	1000	0.010	0.008	0.011	0.010	0.013	0.013	0.046	0.056	0.057	0.052	0.048	0.051	0.100	0.109	0.106	0.103	0.096	0.101
	2500	0.010	0.011	0.011	0.010	0.006	0.014	0.043	0.046	0.045	0.042	0.042	0.063	0.102	0.097	0.094	0.090	0.093	0.134
5000	252	**0.015**	0.007	0.011	0.007	0.011	0.013	**0.053**	0.047	0.044	0.045	0.050	0.048	**0.113**	0.110	0.091	0.087	0.097	0.095
	1000	0.009	0.013	0.008	0.007	0.009	**0.021**	0.045	0.051	0.047	0.041	0.049	**0.069**	0.106	0.110	0.107	0.099	0.094	**0.124**
	2500	0.010	0.010	0.011	0.009	0.012	**0.027**	0.043	0.053	0.047	0.050	0.041	**0.090**	0.090	0.103	0.105	0.099	0.078	**0.150**
1000	252	0.013	0.010	0.011	0.014	0.008	**0.016**	**0.064**	0.052	0.052	0.052	0.057	0.063	0.119	0.105	0.104	0.100	0.110	**0.124**
	1000	0.008	0.010	0.012	0.011	0.006	**0.028**	0.041	0.042	0.042	0.042	0.051	**0.089**	0.102	0.097	0.090	0.091	0.107	**0.171**
	2500	0.017	0.007	0.006	0.004	0.026	**0.043**	0.064	0.046	0.051	0.046	0.079	**0.122**	0.115	0.109	0.107	0.096	0.125	**0.196**
500	252	0.010	0.011	0.014	0.012	0.009	**0.023**	**0.071**	0.040	0.058	0.046	0.056	0.059	**0.141**	0.086	0.101	0.086	0.117	0.120
	1000	0.011	0.016	0.023	0.013	0.022	**0.056**	0.039	0.058	0.073	0.067	0.079	**0.133**	0.147	0.103	0.129	0.125	0.133	**0.201**
	2500	0.021	0.019	0.017	0.014	0.033	**0.077**	0.092	0.069	0.073	0.059	0.094	**0.175**	0.161	0.125	0.126	0.113	0.164	**0.254**

Table 4: Rejection rates for VaRs forecasts computed with an estimated correct model are shown. We use simulated data from an estimated GJR-GARCH(1,1) model with student t innovations, see Appendix A.3. For each iteration the corresponding model is fitted using the first n_{Est} simulated returns, remaining returns are used for forecasting and backtesting. The rejection rates are based on 5,000 Monte-Carlo replications. The tests use Monte-Carlo p-values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

4.5 Impact of a non-zero Mean

In this section, we investigate the influence of a non-zero mean on the various extremal index backtests. To do so, we choose again the GJR-GARCH(1,1)-model from Section 4.3, with all parameters being the same as in the that section except for the mean parameter, which we choose as $\mu = k \times 6.91 \times 10^{-4}$ with $k \in \{-2, -1, 0, 1, 2\}$ (note that we used $k = 1$ in Section 4.3).

In addition, we present a way to deal with potential size distortions related to the mean. Instead of using the relative excess returns defined in Equation 3.1 we suggest to include an estimate of the sample mean $\hat{\mu}$ readily available from the realized returns $r_1, \ldots, r_n$. This leads to a slightly altered version of the relative excess returns given by

$$e_t^{\mu \neq 0} := -\frac{r_t - \hat{\mu}}{\widehat{\mathrm{VaR}}_p^{(t)} + \hat{\mu}}. \tag{4.1}$$

The corresponding test is denoted $\Theta_{\mathrm{noc}}^{\mathrm{B},\mu \neq 0}$ in order to differentiate it from the standard one $\Theta_{\mathrm{noc}}^{\mathrm{B},\mu=0}$ which was solely used in all previous power sections. Note that we do not perform this distinction for $\Theta_{\mathrm{noc}}^{\mathrm{G}}$ since this test is effectively 0-1-violation/duration based as the existing competitors. Still, we report its results as a benchmark.

The results, presented in Table 5, reveal that the influence of the mean seems rather small for $k = -1, 0, 1$. Instead, for $k \in \{-2, 2\}$ the test $\Theta_{\mathrm{noc}}^{\mathrm{B},\mu=0}$ shows some larger deviations from the nominal significance levels. For a negative mean, we observe a tendency of rejecting too infrequently. Hence, the test becomes conservative. For a positive mean, the contrary is true since the null is rejected too often. The effect is more pronounced for the 1 % VaR.

The most serious case $k = 2$ corresponds to a rather unusual scenario. Most assets do not experience upswings of this magnitude over such a long period.[9]

[9] Note that $n = 2,500$ daily returns are roughly 10 years of data.

k	n	Significance level: 1%			Significance level: 5%			Significance level: 10%		
		$\Theta^{\mathrm{G}}_{\mathrm{noc}}(6)$	$\Theta^{\mathrm{B},\mu=0}_{\mathrm{noc}}(40)$	$\Theta^{\mathrm{B},\mu\neq0}_{\mathrm{noc}}(40)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(6)$	$\Theta^{\mathrm{B},\mu=0}_{\mathrm{noc}}(40)$	$\Theta^{\mathrm{B},\mu\neq0}_{\mathrm{noc}}(40)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(6)$	$\Theta^{\mathrm{B},\mu=0}_{\mathrm{noc}}(40)$	$\Theta^{\mathrm{B},\mu\neq0}_{\mathrm{noc}}(40)$
Panel A: 5 % VaR										
-2	252	0.005	0.014	0.016	0.040	0.049	0.051	0.086	0.100	0.106
	1000	0.012	0.012	0.014	0.058	0.051	0.056	0.106	0.109	0.113
	2500	0.008	0.007	0.007	0.050	0.041	0.044	0.105	0.087	0.094
-1	252	0.007	0.011	0.013	0.040	0.038	0.041	0.100	0.096	0.099
	1000	0.008	0.007	0.008	0.050	0.045	0.046	0.101	0.099	0.104
	2500	0.008	0.010	0.010	0.053	0.055	0.055	0.095	0.107	0.108
0	252	0.012	0.015	0.016	0.044	0.051	0.051	0.106	0.110	0.110
	1000	0.008	0.011	0.012	0.054	0.054	0.054	0.101	0.101	0.102
	2500	0.015	0.008	0.008	0.056	0.047	0.047	0.122	0.108	0.107
1	252	0.007	0.016	0.015	0.034	0.051	0.049	0.079	0.098	0.095
	1000	0.008	0.013	0.010	0.047	0.053	0.052	0.111	0.105	0.101
	2500	0.008	0.009	0.008	0.052	0.050	0.048	0.109	0.103	0.101
2	252	0.008	0.016	0.015	0.034	0.068	0.063	0.093	0.112	0.109
	1000	0.009	0.009	0.007	0.048	0.041	0.037	0.108	0.093	0.088
	2500	0.009	0.012	0.008	0.049	0.056	0.051	0.092	0.112	0.101
Panel B: 1 % VaR										
-2	252	0.008	0.006	0.010	0.051	0.034	0.046	0.113	0.083	0.099
	1000	0.008	0.004	0.008	0.044	0.021	0.036	0.095	0.059	0.084
	2500	0.008	0.003	0.007	0.044	0.018	0.039	0.099	0.044	0.091
-1	252	0.008	0.013	0.016	0.038	0.048	0.055	0.084	0.100	0.109
	1000	0.009	0.004	0.008	0.049	0.028	0.038	0.098	0.072	0.092
	2500	0.007	0.006	0.010	0.051	0.040	0.061	0.094	0.088	0.114
0	252	0.011	0.011	0.012	0.052	0.048	0.047	0.097	0.089	0.093
	1000	0.005	0.008	0.007	0.039	0.042	0.042	0.081	0.103	0.106
	2500	0.008	0.008	0.008	0.044	0.045	0.047	0.092	0.094	0.098
1	252	0.007	0.018	0.014	0.036	0.047	0.043	0.081	0.099	0.096
	1000	0.008	0.015	0.012	0.043	0.070	0.053	0.092	0.131	0.110
	2500	0.012	0.009	0.006	0.054	0.060	0.034	0.103	0.117	0.090
2	252	0.006	0.019	0.014	0.039	0.067	0.053	0.094	0.127	0.102
	1000	0.008	0.024	0.013	0.047	0.086	0.055	0.092	0.149	0.111
	2500	0.007	0.025	0.011	0.043	0.102	0.055	0.089	0.190	0.107

Table 5: Rejection rates for true VaRs for different DGPs with varying mean are shown. We show only results for extremal index backtests. In contrast to all previous settings, we add the backtest $\Theta^{\mathrm{B},\mu\neq0}_{\mathrm{noc}}$ which includes de-meaning. The base case ($k = 1$) corresponds to a fitted GJR-GARCH(1,1) model with student t innovations, see Appendix A.3. The parameter k controls the mean of the DGPs by multiplying the mean of the base case. The rejection rates are based on 5,000 Monte-Carlo replications. The tests use Monte-Carlo p-values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

Furthermore, such large average returns are even more unusual for a diversified portfolio. Hence, we do not think that the observed liberality of the test is a too severe issue. Nevertheless, the results of $\Theta_{\text{noc}}^{\text{B};\mu\neq 0}$ show that a simple correction can help making a liberal test more conservative and a conservative test more liberal.

Summarizing, we find that all tests hold their sizes at least approximately in the discussed, usual finance-settings. However, in case of concerns whether those conditions are indeed met, we recommend to use the correction introduced in Equation 4.1.

5 Empirical Applications

After we have investigated the power of the extremal index approach and competing methodologies in theoretical setups, we now shed light onto practical implications of our approach. Our focus is on three questions, which one might summarize under the title "Historical Simulation, Few Violations, and the Rejection of GARCH Models".

The first question aims again at Historical Simulation (HS), which is not only wide-spread in the academic literature as is evident from the frequent use in simulation studies (as in this paper and others) but also one of the most popular forecasting approaches used in practice. See, e.g., Pérignon and Smith (2010) who report that HS was the most used procedure in 2005 with a percentage of 47.4% among their sample. Despite its prevalence, HS in its classical form should be rejected as a correct conditional approach due to its lack of a quick reaction to changing volatility. Therefore, we check backtesting results of HS in two different periods. First, we backtest HS for a 1% VaR on the S&P 500 index in a phase containing the last financial crisis (2008-01-15 till 2011-12-31), and second the subsequent relatively calm phase (2012-01-01 till 2015-12-22). Both backtesting

Turbulent Period from 2008-01-15 till 2011-12-31 (1,000 Obs)

FC	M_1: p-value	$LR^{\text{Mar}}_{\text{ind}}$	$LR^{\text{Wei}}_{\text{ind}}$	$GMM^{(\text{VD})}_{\text{ind}}$	MCS_{ind}	$\Theta^{\text{G}}_{\text{noc}}(6)$	$\Theta^{\text{B}}_{\text{noc}}(40)$
Panel A: 1 % VaR							
HS (250)	22: 0.0010 ***	0.0932	0.0000 ***	0.0000 ***	0.0000 ***	0.0000 ***	0.0000 ***
HS (500)	26: 0.0000 ***	0.0729	0.0000 ***	0.0000 ***	0.0000 ***	0.0000 ***	0.0000 ***
Panel B: 0.05 % VaR							
skew-t (250)	2: 0.1103	0.0402 *	0.0611	0.0190 *	0.2532	0.0631	0.0000 ***
skew-t (500)	0: 0.3175	0.5215	-	-	-	-	0.0000 ***
Panel C: 1 % VaR							
GJR-GARCH(1,1)	18: 0.0221 *	0.1057	0.8109	0.9237	0.4345	0.4318	0.6984
GARCH(1,1)	20: 0.0051 **	0.0954	0.1338	0.0387 *	0.0673	0.0221 *	0.2554
ARCH(1)	31: 0.0000 ***	0.9994	0.0000 ***	0.0000 ***	0.0000 ***	0.0002 ***	0.0000 ***

Calm Period from 2012-01-01 till 2015-12-22 (1,000 Obs)

FC	M_1: p-value	$LR^{\text{Mar}}_{\text{ind}}$	$LR^{\text{Wei}}_{\text{ind}}$	$GMM^{(\text{VD})}_{\text{ind}}$	MCS_{ind}	$\Theta^{\text{G}}_{\text{noc}}(6)$	$\Theta^{\text{B}}_{\text{noc}}(40)$
Panel A: 1 % VaR							
HS (250)	13: 0.3604	0.0039 **	0.2192	0.4600	0.3444	0.0067 **	0.0003 ***
HS (500)	8: 0.5121	0.0001 ***	0.0028 **	0.0032 **	0.0050 **	0.0002 ***	0.0002 ***
Panel B: 0.05 % VaR							
skew-t (250)	1: 0.5336	0.2977	-	-	0.1695	-	0.0000 ***
skew-t (500)	1: 0.5336	0.2073	-	-	0.1701	-	0.0000 ***
Panel C: 1 % VaR							
GJR-GARCH(1,1)	11: 0.7520	0.3778	0.2059	0.1879	0.9014	0.6124	0.9525
GARCH(1,1)	13: 0.3604	0.0645	0.3275	0.1331	0.9349	0.3551	0.7950
ARCH(1)	3: 0.0091 **	0.9932	0.0278 *	0.0249 *	0.0015 **	0.6344	0.0000 ***

Table 6: This table presents backtesting results for several one-step-ahead forecasts adopting a rolling window scheme assessed by different backtests. In Panels A 1 % VaR forecasts are performed with an unconditional non-parametric method. In Panels B 0.05 % VaR forecasts are made using a skew-t distribution. In each case, the numbers in brackets report the size of the rolling window. GARCH model refits are done every 5 days, unconditional methods are refitted on a daily basis. Panels C belong to 1 % VaR forecasts using three different GARCH models with rolling window size 1,000. The out-of-sample periods are a troubled and a calm market period of 1,000 returns each. Hence, in Panels A 10 violations, in Panels B 0.5 violations, and in Panels C again 10 violations have to be expected. Column FC on the left side reports forecasting methods. The second column reports both, the number of violations M_1 and the corresponding asymptotic p-value of the unconditional backtest by Christoffersen (1998). The remaining columns show results of several independence backtests. The numbers are Monte-Carlo p-values. Asterisks mark levels of significance: *** at 0.1 %, ** at 1 %, and * at 5 %.

periods consist of exactly 1,000 returns which leads to an expectation of 10 violations. We re-use these periods for questions 2 and 3. The data was downloaded from Yahoo Finance.

The second question addresses a finding of Pérignon et al. (2008), and Pérignon and Smith (2010). In the first named paper, the authors report that disclosed VaR numbers of Canadian banks were way too conservative in the past (74 violations expected, only 2 violations happened) and suggest two explanations. First, it is hypothesized that markets will severely punish banks who underestimate risk which makes them possibly very conservative. Second, a lack of correct accounting for diversification across departments of a bank could yield too large risk estimates, too. In the second paper, this conservatism is also found in another sample containing US, Canadian, and International Banks. Interestingly, in the subsequent financial crisis, almost all banks suffered substantial losses which is surprising when market risk measurements of banks were considered conservative before. Therefore, we analyze how the extremal index approach enables to asses independence even in the absence of many violations. We achieve this by calculating incorrect conditional VaR forecasts at an unconventional low level of 0.05% by simply fitting a skewed Student t-distribution (Azzalini and Capitanio, 2003) to the $T_e \in \{250, 500\}$ observations prior to time t. Note that Eling (2014) found that this distribution can provide a good fit for asset returns. Due to the low VaR level, only 0.5 violations can be expected throughout the considered time periods.

The third question we consider is about distinguishing different specifications of the volatility process of GARCH-type models at the 1% VaR level. Note that GARCH model-based forecasts possibly provide the most common alternative to HS, aiming at more accurate forecasts due to their particular focus on time-changing aspects. However, there are many different models available and a modeler has to asses which of them captures the dynamic beahavior the best. Hence, we adopt two very popular GARCH specifications (Bollerslev, 1986; Glosten et al., 1993) as well as ARCH(1) and report how backtesting results differ. Of course, it is important to note that backtesting is not the appropriate method for comparing the accuracy of two or more forecasters. Nevertheless, it is interesting to see whether disparities can be made visible by backtesting.

For each forecasting exercise we perform one-day ahead forecasts based on a rolling window scheme of previous returns. Questions 1 and 2 use estimation sample sizes T_e of 250 and 500. In the GARCH case we chose windows of $T_e = 1,000$ returns.

Respective results for all three questions are presented in Table 6. Panels A correspond to question 1, panels B to question 2, and panels C to question 3.

First, we focus on Question 1. Panel A of the turbulent period shows that both HS approaches yield way too many violations. Most independence backtests are able to reject both methods. Here, the only failing backtest is $LR_{\text{ind}}^{\text{Mar}}$. Throughout the calm period, HS forecasts are more appropriate (see the smaller number of violations M_1) but can still be rejected by the use of independence backtests. Especially, both extremal index approaches reject the forecasts, but also $LR_{\text{ind}}^{\text{Mar}}$ which failed in the turbulent phase. This somewhat surprising change can be explained by the fact that $LR_{\text{ind}}^{\text{Mar}}$ can only detect violations that occurred on subsequent days. Moreover, we observe that the rejection of the longer estimation sample using $T_e = 500$ returns appears to be easier, as expected from the simulation results and their interpretation in Section 4.

Next, we turn to Question 2. Apparently, most independence backtests can reject unconditional forecasts. However, if violations are rare, then a correct assessment can become impossible. As expected the skew-t distribution at this low VaR level yields few violations. For zero or one violation, classical violations or durations-based backtests do not lead to a statement, since they are considered to be unfeasible in these cases (at least two violations are necessary to change this). Instead, the extremal index approach decouples the result from the particular VaR level and yields very similar p-values as for the HS scenarios.

Finally, panels C show that a differentiation between several GARCH models can be hard. In almost no case, an independence backtest is able to reject one of these two models. This is to some extent in line with the literature (Hansen, 2005). Only in two cases the p-values are below 5%. The extremal index backtest $\Theta_{\text{noc}}^{\text{B}}$ shows barely a sign of misspecification which is noteworthy due to its often large power. However, especially in the turbulent phase, the GARCH models yield too many violations and can therefore be rejected with an unconditional test. Instead, the ARCH(1) model is quite easily rejected with independence tests.

6 Conclusion

In this paper, a new idea for the assessment of VaR forecasts with respect to violation clustering is worked out in detail. For that purpose, we implement two recently proposed estimators for the extremal index, derive corresponding new backtests, and compare them to existing ones. The results show that especially the sliding blocks estimator from Northrop (2015); Berghaus and Bücher (2017) is suitable for this task. The corresponding backtest exhibits substantial power improvements in many theoretical scenarios and can easily reject unconditional forecasters even in the absence of violations, a feature lacked by many other backtests. The latter feature is possibly interesting to detect bad forecasts even if they are conservative, since conservative forecasts can fail to accurately adapt to changing markets, too. Furthermore, we briefly hint at possible extensions to backtesting Expected Shortfall, which may become more important in the future.

A Appendix

A.1 Parameter Choice for Extremal Index Estimators

For the sake of simplicity, we want to choose the parameters for the extremal index tests constant across all scenarios. To justify our choices we conduct the simulation experiment from Section 4.1 again for several choices of K and b. In contrast to the DGPs underlying Table 1 we perform the simulations here without any restrictions with respect to the number of violations. This allows us to study both the size and power of the tests in the particular setting. Given a reasonably sized test we want a test maximizing the power.

Furthermore, the results of this Section give an idea about the sensitivity of the tests. This can be useful to detect eventual size anomalies. In addition, we can assess the extent of potential power losses and improvements. The latter is useful in order to investigate whether a more sophisticated parameter choice procedure provides space for enhancements.

Table 7 reports results for $\Theta^{\mathrm{G}}_{\mathrm{noc}}(K)$. As the varying bold values for $\lambda < 1$ indicate, the 'optimal' value for K depends on the concrete setting. However, the actual power differences are often quite small, especially for $K > 2$. As a rough tendency, it can be conjectured that K should be increasing in λ, though the sensitivity to power changes seems relatively small. The H_0 DGP reveals no too large deviations from the corresponding significance levels. Overall, the results suggest that a fixed value of $K = 6$ would a reasonable global choice.

Table 8 reports results for $\Theta^{\mathrm{B}}_{\mathrm{noc}}(b)$. The rough tendency suspected in Table 7, that the optimal parameter depends on the DGP, is more visible. Besides, the sensitivity with respect to b is more pronounced, particularly for small sample sizes and low significance levels. This shows the stronger potential for a setting dependent parameter choice procedure for this estimator. Yet, we choose also

λ	n	Significance level: 1%					Significance level: 5%					Significance level: 10%				
		$\Theta^{\mathrm{G}}_{\mathrm{noc}}(2)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(4)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(6)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(8)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(10)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(2)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(4)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(6)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(8)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(10)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(2)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(4)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(6)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(8)$	$\Theta^{\mathrm{G}}_{\mathrm{noc}}(10)$
Panel A: 5 % VaR																
0.8706	252	0.101	0.098	**0.113**	0.098	0.096	0.226	0.250	**0.264**	0.258	0.244	0.328	0.370	**0.395**	0.374	0.354
	1000	0.259	0.371	**0.376**	0.370	0.321	0.480	**0.603**	0.596	0.575	0.525	0.600	**0.700**	0.695	0.677	0.639
	2500	0.602	**0.752**	0.748	0.727	0.629	0.810	0.882	**0.891**	0.868	0.829	0.878	**0.934**	0.930	0.916	0.882
0.9828	252	0.017	0.018	0.017	0.024	**0.026**	0.058	0.066	0.075	0.092	**0.094**	0.123	0.127	0.140	0.157	**0.162**
	1000	0.019	0.024	0.041	0.048	**0.053**	0.101	0.111	0.129	0.141	**0.142**	0.170	0.188	0.218	0.218	**0.224**
	2500	0.040	0.068	0.076	**0.104**	0.090	0.138	0.169	0.214	**0.234**	0.229	0.226	0.281	0.313	**0.325**	0.324
0.9914	252	0.010	0.018	0.018	0.014	**0.019**	0.051	0.067	0.076	**0.077**	**0.077**	0.106	0.134	**0.143**	0.143	0.137
	1000	0.018	0.023	0.026	0.033	**0.036**	0.082	0.106	0.107	0.125	**0.131**	0.159	0.186	0.198	0.209	**0.215**
	2500	0.023	0.044	0.053	**0.071**	0.069	0.096	0.132	0.158	0.182	**0.200**	0.186	0.235	0.256	0.282	**0.300**
1 (H_0)	252	0.014	0.012	0.017	0.014	0.018	0.056	0.046	0.060	0.060	0.053	0.098	0.099	0.102	0.098	0.109
	1000	0.009	0.013	0.009	0.007	0.007	0.044	0.057	0.048	0.043	0.037	0.092	0.104	0.097	0.090	0.084
	2500	0.004	0.009	0.008	0.006	0.005	0.036	0.034	0.040	0.037	0.036	0.082	0.075	0.078	0.078	0.085
Panel B: 1 % VaR																
0.8706	252	0.034	0.029	0.037	0.045	**0.053**	**0.146**	0.142	0.140	0.137	0.127	0.183	**0.217**	0.214	0.210	0.213
	1000	0.127	0.171	**0.181**	0.177	0.173	0.278	0.339	**0.364**	0.363	0.354	0.379	0.442	0.472	**0.479**	0.475
	2500	0.290	0.392	0.419	**0.437**	0.406	0.488	0.594	0.626	**0.633**	0.632	0.594	0.704	0.722	**0.743**	0.733
0.9828	252	0.010	0.009	0.010	0.012	**0.018**	**0.069**	0.052	0.058	0.062	0.063	0.104	**0.123**	0.116	0.110	**0.123**
	1000	0.021	0.025	0.021	0.033	**0.036**	0.086	0.088	0.099	0.113	**0.118**	0.142	0.148	0.174	0.186	**0.190**
	2500	0.025	0.034	0.049	0.055	**0.061**	0.090	0.128	0.144	0.158	**0.160**	0.181	0.214	0.236	0.248	**0.264**
0.9914	252	0.008	0.010	**0.011**	0.009	0.006	**0.057**	0.044	0.046	0.038	0.037	**0.102**	0.095	0.084	0.076	0.089
	1000	0.017	0.018	0.020	0.021	**0.022**	0.060	0.065	0.063	0.068	**0.074**	0.105	0.118	**0.130**	0.118	0.129
	2500	0.016	0.017	0.022	**0.027**	0.019	0.069	0.081	0.079	0.087	**0.090**	0.129	0.145	0.137	0.152	**0.159**
1 (H_0)	252	0.011	0.010	0.010	0.010	0.010	0.047	0.054	0.048	0.050	0.045	0.095	0.090	0.102	0.100	0.094
	1000	0.010	0.017	0.013	0.012	0.008	0.054	0.050	0.051	0.060	0.049	0.102	0.104	0.096	0.098	0.101
	2500	0.013	0.006	0.010	0.011	0.012	0.041	0.049	0.045	0.040	0.046	0.090	0.089	0.094	0.083	0.088

Table 7: Analysis of the influence of the K-gap parameter on the power and size of $\Theta^{\mathrm{G}}_{\mathrm{noc}}$. DGPs are as in Table 1 but without any restriction regarding the number of violations. The rejection rates are based on 5,000 Monte-Carlo replications. The tests use Monte-Carlo p-values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

here a fixed $b = 40$ for all further investigations since this value seems to yield reasonably good power results across all DGPs. Again, looking at the H_0 DGP no obvious problems regarding the size are detected.

A.2 Computation of p-values

Subsequently, let $T_- = T_-(0)$ denote a generic test statistic for which small values provide evidence against some null hypothesis H_0. Both tests explained in Section 3.2 fall into this category, with $T_- = \hat{\theta}_n^{\mathrm{M}}$ with $\mathrm{M} \in \{\mathrm{B}, \mathrm{G}\}$. Critical values, or equivalently p-values, are obtained by simulating the H_0-distribution of the test statistic. More precisely, let $T_-(1), \ldots, T_-(N)$ denote the simulated values of the test statistic;[10] with N fixed to $N = 10,000$ throughout this paper. To account for possible ties, we follow Dufour (2006) and Ziggel et al. (2014) and define $T_-^*(i) = T_-(i) + \varepsilon_i$, where $(\varepsilon_i) \sim_{\text{i.i.d.}} 0.001 \cdot N(0,1)$. The p-value of the test associated to T_- is then given by

$$p = \frac{NG + 1}{N + 1}, \tag{A.1}$$

where

$$G = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{T_-^*(i) < T_-^*(0)\}.$$

Similarly, let $T_+ = T_+(0)$ denote a generic test statistic for which large values provide evidence against some null hypothesis H_0, which is for instance the case for $LR_{\text{ind}}^{\text{Mar}}, LR_{\text{ind}}^{\text{Wei}}, GMM_{\text{ind}}$ or MCS_{ind}. The H_0 distribution of those tests is simulated

[10] Some tests, especially the duration-based ones, are not always feasible. This happens typically due to a lack of violations in small backtesting samples. If a test is not feasible during our power simulations, we count this always as a non-rejection. Regarding the simulation of the test statistics for p-value computation, we deal with potential non-feasible tests by setting their values artificially to that extreme which corresponds to a non-rejection. Thus, for instance, for the GMM test this would be the minimum of all N obtained values during the simulation.

λ	n	Significance level: 1%					Significance level: 5%					Significance level: 10%				
		$\Theta^{\mathrm{B}}_{\mathrm{noc}}(10)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(20)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(40)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(80)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(120)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(10)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(20)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(40)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(80)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(120)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(10)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(20)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(40)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(80)$	$\Theta^{\mathrm{B}}_{\mathrm{noc}}(120)$
Panel A: 5 % VaR																
0.8706	252	0.142	**0.183**	0.137	0.056	0.025	0.325	**0.383**	0.293	0.155	0.097	0.464	**0.527**	0.404	0.218	0.164
	1000	0.579	**0.724**	0.536	0.231	0.122	0.779	**0.886**	0.752	0.440	0.298	0.876	**0.944**	0.846	0.588	0.423
	2500	0.951	**0.988**	0.917	0.594	0.337	0.992	**0.998**	0.977	0.796	0.599	0.998	**1.000**	0.990	0.874	0.714
0.9828	252	0.018	0.030	**0.040**	0.030	0.012	0.066	0.090	**0.108**	0.088	0.075	0.129	0.152	**0.172**	0.152	0.147
	1000	0.033	0.066	0.089	**0.091**	0.075	0.119	0.186	**0.225**	0.218	0.191	0.206	0.297	**0.338**	0.316	0.295
	2500	0.064	0.153	0.230	**0.235**	0.174	0.196	0.345	**0.440**	0.430	0.376	0.301	0.470	**0.579**	0.551	0.504
0.9914	252	0.014	0.020	**0.021**	0.018	0.013	0.058	0.061	**0.071**	0.065	0.066	0.121	0.124	**0.130**	0.125	0.130
	1000	0.016	0.026	0.036	**0.041**	0.039	0.069	0.097	0.117	**0.131**	0.120	0.141	0.178	0.213	**0.225**	0.208
	2500	0.023	0.059	0.095	**0.110**	0.095	0.102	0.172	0.238	**0.287**	0.253	0.183	0.277	0.357	**0.412**	0.378
1 (H_0)	252	0.014	0.012	0.010	0.009	0.016	0.050	0.052	0.044	0.050	0.060	0.107	0.109	0.103	0.101	0.114
	1000	0.011	0.012	0.012	0.014	0.014	0.056	0.058	0.054	0.048	0.050	0.113	0.117	0.110	0.088	0.102
	2500	0.008	0.008	0.008	0.008	0.012	0.047	0.043	0.042	0.056	0.056	0.102	0.084	0.094	0.110	0.110
Panel B: 1 % VaR																
0.8706	252	0.118	**0.179**	0.122	0.060	0.029	0.295	**0.363**	0.266	0.152	0.115	0.454	**0.519**	0.375	0.231	0.172
	1000	0.614	**0.731**	0.531	0.236	0.131	0.804	**0.888**	0.737	0.436	0.294	0.894	**0.937**	0.842	0.580	0.429
	2500	0.950	**0.989**	0.935	0.606	0.348	0.989	**0.997**	0.978	0.814	0.601	0.996	**0.999**	0.994	0.898	0.718
0.9828	252	0.024	0.030	**0.037**	0.025	0.019	0.076	0.087	0.098	**0.102**	0.082	0.152	0.166	**0.174**	0.162	0.148
	1000	0.032	0.064	**0.110**	0.098	0.084	0.110	0.190	**0.245**	0.238	0.216	0.211	0.300	**0.364**	0.350	0.320
	2500	0.065	0.148	0.224	**0.228**	0.175	0.196	0.348	0.439	**0.452**	0.380	0.306	0.474	**0.585**	0.573	0.524
0.9914	252	0.009	0.018	**0.023**	0.014	0.011	0.046	0.066	0.064	**0.075**	0.064	0.106	0.116	0.128	**0.140**	0.131
	1000	0.015	0.030	0.047	**0.052**	0.042	0.065	0.104	0.128	**0.139**	0.136	0.135	0.184	0.224	**0.225**	0.214
	2500	0.024	0.042	0.070	**0.082**	0.074	0.095	0.154	0.206	**0.228**	0.221	0.182	0.265	0.326	0.332	**0.337**
1 (H_0)	252	0.007	0.007	0.010	0.006	0.012	0.042	0.051	0.057	0.062	0.059	0.096	0.102	0.106	0.102	0.118
	1000	0.012	0.016	0.011	0.008	0.008	0.052	0.060	0.048	0.051	0.049	0.104	0.112	0.112	0.106	0.109
	2500	0.012	0.012	0.009	0.008	0.006	0.043	0.052	0.042	0.042	0.048	0.094	0.100	0.095	0.091	0.096

Table 8: Analysis of the influence of the block size parameter b on the power and size of $\Theta^{\mathrm{B}}_{\mathrm{noc}}$. DGPs are as in Table 1 but without any restriction regarding the number of violations. The rejection rates are based on 5,000 Monte-Carlo replications. The tests use Monte-Carlo p-values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

by drawing $N = 10,000$ samples from

$$\tilde{r}_t \overset{\text{i.i.d.}}{\sim} N(0,1), \quad \widehat{\text{VaR}}_p^{(t)} = -\Phi^{-1}(p),$$

as already described in Equation 3.6 of Section 3.2 which leads to the required violations (see the definition of conditional coverage in Equation 2.2), and durations. Denote the respective values by $T_+(1), \ldots, T_+(N)$ and let $T_+^*(i) = T_+(i) + \varepsilon_i$ with $(\varepsilon_i) \sim_{\text{i.i.d.}} 0.001 \cdot N(0,1)$ as above. We then compute critical values as in (A.1), but with G replaced by

$$G = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{T_+(i) \geq T_+^*(i)\}.$$

A.3 GARCH Model Details and Estimates

In the applied Sections 4 and 5 we employ a variety of GARCH-type models. In this section, we report model specifications and estimation results used for our DGPs where not done before. The GJR-GARCH(1,1) model by Glosten et al. (1993) is defined as follows:

$$\sigma_t^2 = \omega + \alpha z_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma \mathbb{1}(z_{t-1} \leq 0) z_{t-1}^2 \tag{A.2}$$

with returns being generated by $r_t = \mu + \sigma_t z_t$. The innovations are chosen as described in Section 4.2, that is, $z_t \overset{\text{i.i.d.}}{\sim} \sqrt{\frac{d-2}{d}}\, \varepsilon_t$, where ε_t follows a student t-distribution with d degrees of freedom. Note that $\gamma = 0$ yields a simple GARCH(1,1) model and $\beta = \gamma = 0$ an ARCH(1) model. Results of a fit of the model to S&P 500 log-return data from 2012-01-01 till 2015-01-01 are shown in Table 9.

Parameter	Estimate	Robust Std. Error	Robust t stat.	p-value
μ	6.91×10^{-4} ***	1.96×10^{-4}	3.52	4.25×10^{-4}
ω	5.04×10^{-6} ***	3.12×10^{-7}	16.13	0
α	7.83×10^{-8}	1.86×10^{-2}	0	1
β	0.75 ***	2.68×10^{-2}	28.03	0
γ	0.35 ***	7.59×10^{-2}	4.59	4.46×10^{-6}
d	6.87 ***	1.46	4.72	2.41×10^{-6}

Table 9: This table shows our estimates of the GJR-GARCH(1,1) model using S&P 500 data from 2012-01-01 to 2015-01-01 (754 obs).

Except for α, all parameters are highly significant. Therefore, we believe that these parameter values yield a fairly realistic example where the features of the GJR-GARCH(1,1) are able to take shape. In particular, the mean paramater is significant as well, which qualifies this DGP as a model to the check the robustness of our backtest to a non-zero mean.

Acknowledgements This work has been supported in part by the Collaborative Research Center "Statistical Modeling of Nonlinear Dynamic Processes" (SFB 823, Tproject A7) of the German Research Foundation (DFG).

References

Alexander, C., E. Lazar, and S. Stanescu (2013). Forecasting VaR using analytic higher moments for GARCH processes. *International Review of Financial Analysis 30*, 36–45.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol. 65*(2), 367–389.

BCBS (1996a). Overview of the Amendment to the Capital Accord to incorporate Market Risks. *Basel Committee on Banking Supervision.*

BCBS (1996b). Supervisory Framework for the Use of Backtesting in Conjunction with the Internal Models Approach to Market Risk Capital Requirements. *Basel Committee on Banking Supervision.*

BCBS (2016). Minimum capital requirements for market risk. *Basel Committee on Banking Supervision.*

Beirlant, J., Y. Goegebeur, J. Segers, and J. Teugels (2004). *Statistics of extremes: Theory and Applications.* Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd.

Berghaus, B. and A. Bücher (2017+). Weak convergence of a pseudo maximum likelihood estimator for the extremal index.

Berkowitz, J. (2001). Testing Density Forecasts, With Applications to Risk Management. *Journal of Business & Economic Statistics 19*(4), 465–474.

Berkowitz, J., P. Christoffersen, and D. Pelletier (2011). Evaluating Value-at-Risk Models with Desk-Level Data. *Management Science 57*(12), 2213–2227.

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

Bontemps, C. and N. Meddahi (2012). Testing distributional assumptions: A GMM aproach. *Journal of Applied Econometrics 27*(6), 978–1012.

Candelon, B., G. Colletaz, C. Hurlin, and S. Tokpavi (2011). Backtesting Value-at-Risk: A GMM Duration-Based Test. *Journal of Financial Econometrics 9*(2), 314–343.

Christoffersen, P. and D. Pelletier (2004). Backtesting Value-at-Risk: A Duration-Based Approach. *Journal of Financial Econometrics 2*(1), 84–108.

Christoffersen, P. F. (1998). Evaluating Interval Forecasts. *International Economic Review 39*(4), 841.

Costanzino, N. and M. Curran (2015). Backtesting general spectral risk measures with application to expected shortfall. *The Journal of Risk Model Validation 9*(1), 21–31.

Du, Z. and J. C. Escanciano (2017). Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science 63*(4), 940–958.

Dufour, J.-M. (2006). Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *J. Econometrics 133*(2), 443–477.

Eling, M. (2014). Fitting asset returns to skewed distributions: Are the skew-normal and skew-student good models? *Insurance Math. Econom. 59*, 45–56.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997). *Modelling extremal events for insurance and finance*, Volume 33 of *Applications of mathematics.* Berlin u.a.: Springer.

Ferro, C. A. T. and J. Segers (2003). Inference for clusters of extreme values. *J. R. Stat. Soc. Ser. B Stat. Methodol. 65*(2), 545–556.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance 48*(5), 1779.

Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics 23*(4), 365–380.

Kerkhof, J. and B. Melenberg (2004). Backtesting for risk-based regulatory capital. *Journal of Banking & Finance 28*(8), 1845–1865.

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

Kupiec, P. H. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives 3*(2), 73–84.

Leadbetter, M. R. (1983). Extremes and local dependence in stationary sequences. *Z. Wahrsch. Verw. Gebiete 65*(2), 291–306.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative risk management.* Princeton Series in Finance. Princeton, NJ: Princeton University Press.

Mikosch, T. and C. Starica (2000). Limit Theory for the Sample Autocorrelations and Extremes of a GARCH (1, 1) Process. *Ann. Statist. 28*(5), 1427–1451.

Northrop, P. J. (2015). An efficient semiparametric maxima estimator of the extremal index. *Extremes 18*(4), 585–603.

Pérignon, C., Z. Y. Deng, and Z. J. Wang (2008). Do banks overstate their Value-at-Risk? *Journal of Banking & Finance 32*(5), 783–794.

Pérignon, C. and D. R. Smith (2010). The level and quality of Value-at-Risk disclosure by commercial banks. *Journal of Banking & Finance 34*(2), 362–377.

Rosenblatt, M. (1952). Remarks on a Multivariate Transformation. *Ann. Math. Statist. 23*(3), 470–472.

Smith, R. L. and I. Weissman (1994). Estimating the Extremal Index. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(3), 515–528.

Süveges, M. (2007). Likelihood estimation of the extremal index. *Extremes 10*(1-2), 41–55.

Süveges, M. and A. C. Davison (2010). Model misspecification in peaks over threshold analysis. *The Annals of Applied Statistics 4*(1), 203–221.

Wong, W. K. (2008). Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking & Finance 32*(7), 1404–1415.

Ziggel, D., T. Berens, G. N. Weiß, and D. Wied (2014). A new set of improved Value-at-Risk backtests. *Journal of Banking & Finance 48*, 29–41.