New Backtests for Unconditional Coverage of the Expected Shortfall*

Robert Löser †

Technische Universität Dortmund

Dominik Wied ‡

University of Cologne

Daniel Ziggel §

FOM Hochschule für Oekonomie & Management

February 16, 2018

ABSTRACT

While the Value-at-Risk (VaR) has been the standard risk measure for a long time, the Expected Shortfall (ES) has become more and more popular in recent times, as it provides important information about the tail risk. We present a new backtest for the unconditional coverage property of the ES. The test is based on the so called cumulative violation process and its main advantage is that the distribution is known for finite out-of-sample size. This leads to better size and power properties compared to existing tests. Moreover, we extend the test principle to a multivariate test and analyze its behavior by simulations and an application to bank returns.

Keywords: Expected Shortfall, Model Risk, Multivariate Backtesting.

JEL Classification: C52, C53, C58.

---

*Financial support by the Collaborative Research Center "Statistical Modeling of Nonlinear Dynamic Processes" (SFB 823, project A1) of the German Research Foundation (DFG) is gratefully acknowledged. Moreover, we thank the anonymous referee for the helpful comments on an earlier version of this paper.

†CDI-Gebäude, 44221 Dortmund, Germany, telephone: +49 231 755 5419, e-mail: *loeser@statistik.tu-dortmund.de.*

‡Meister-Ekkehart-Str. 9, 50923 Cologne, Germany, telephone: +49 221 470 4514, e-mail: *dwied@uni-koeln.de.*

§Hauptstraße 5b, 46569 Hünxe, Germany, telephone: +49 2064 456 4394, e-mail: *daniel.ziggel@fom.de.*

1 Introduction

Since the 1996 Market Risk Amendment to the First Basel Accord, the Value at Risk (VaR) was (and still is) the standard measure for risks of financial investments. Besides, it has become the industry standard not only for banks but also, e.g., for insurance companies (due to Solvency II) and asset managers. However, despite its prevalence, conceptual simplicity and easy interpretation, the VaR has several drawbacks based on fundamental deficiencies. On the one hand, it lacks the desirable property of a coherent risk measure (see Artzner et al., 1999) for non-Gaussian Profit & Loss (P/L) distributions. On the other hand, the VaR does not account for tail risks.

As a consequence, alternative risk measures are of increasing importance and interest with a particular focus on the Expected Shortfall (ES). Main reasons for this are that the ES is a coherent risk measure and accounts for tail risks. Moreover, a consultation paper from the Basel Committee (Basel Committee on Banking Supervision, 2012) opted to replace VaR with ES. However, while the estimation of the ES is quite similar compared to the VaR, backtesting ES models remains a major challenge (see Yamai and Yoshiba, 2002, 2005; Kerkhof and Melenberg, 2004). While there are several formal VaR-backtests (see, e.g., Candelon et al., 2011; Berkowitz et al., 2011; Ziggel et al., 2014, for some recently proposed tests), there are only a few studies dealing with ES-backtests (Berkowitz, 2001; Wong, 2008, 2010; Acerbi and Szekely, 2014). Most recently, Du and Escanciano (2016) proposed some backtests for ES forecasts which are easy to implement. We build upon these recently proposed backtests and present a new backtest for the unconditional coverage (UC) property of the ES. This property is of particular interest for regulators whose focus is mainly on UC backtests.

Our test is based on the so called cumulative violation process. Its main advantage is that the distribution of the test statistic is available for finite out-of-sample size which leads to better size and power properties compared to existing tests. Moreover, it can be easily extended to a multivariate setting. To the best of our knowledge, there is currently no multivariate backtesting framework for the ES, although this is a highly important task (see Wied et al. (2016) for a discussion concerning multivariate VaR tests).

2 Methodology

In this section, we introduce the notation used throughout the paper, define the desirable properties of VaR and ES models and present our new uni- and multivariate backtests.

2.1 Notation and ES-Violation Properties

Let $\{y_t\}_{t=1}^n$ be the observable part of a time series $\{y_t\}_{t\in\mathbb{Z}}$, where y_t represents the return of a bank or an asset at day t. Moreover, let $\{VaR_{t|t-1}(p)\}_{t=1}^n$ be VaR-forecasts at level $p \in (0,1)$, implicitly defined by $P(y_t < -\text{VaR}_{t|t-1}(p)|\mathbb{F}_{t-1}) = p$, where $\mathbb{F}_{t-1}$ denotes the information set up to time $t-1$. The ex-post indicator variable $h_t(p)$ for a given VaR-forecast $VaR_{t|t-1}(p)$ is defined as

$$h_t(p) = \mathbb{1}(y_t < -\text{VaR}_{t|t-1}(p)),$$

where $\mathbb{1}$ denotes the indicator function. In this notation, $h_t(p) = 1$ denotes a VaR-violation.

In this paper we focus on backtesting the ES. Following Du and Escanciano (2016) we define the conditional distribution of y_t given the information set $\mathbb{F}_{t-1}$ as $G_{t|t-1}(\cdot) := G_t(\cdot|\mathbb{F}_{t-1})$. The ES is defined as

$$ES_t := -E(y_t|y_t < -\text{VaR}_{t|t-1}(p), \mathbb{F}_{t-1}) = \frac{1}{p}\int_0^p \text{VaR}_{t|t-1}(u)du.$$

Du and Escanciano (2016) focus on the so called cumulative violations process to test the correct specification of the ES. The cumulative violation process

$$\begin{aligned}H_t(p) :&= \frac{1}{p}\int_0^p h_t(u)du = \frac{1}{p}\int_0^p \mathbb{1}(y_t < -\mathrm{VaR}_{t|t-1}(u))du \\ &= \frac{1}{p}\int_0^p \mathbb{1}(G_{t|t-1}(y_t) < u)du = \frac{1}{p}\left(p - G_{t|t-1}(y_t)\right)\mathbb{1}(y_t < -\mathrm{VaR}_{t|t-1}(p))\end{aligned}$$

takes values ranging from zero to one if a VaR violation occurs ($h_t(p) = 1$), otherwise it is equal to zero. If $G_{t|t-1}$ is continuous for all t, then $G_{t|t-1}(y_t)$ is uniformly distributed on $[0, 1]$. In consequence the expected value at each time t is given by

$$E(H_t(p)) = \frac{p}{2}, \quad \forall t. \tag{1}$$

This is called the UC-property for the ES. As stated in Du and Escanciano (2016), $\{H_t\}$ is unobservable because the true model is unknown and has to be estimated. Therefore some assumptions are necessary:

Assumption 1.

1. *There is a parametric model $G_{t|t-1}(y|\theta_0)$ which specifies the distribution $G_{t|t-1}(y)$ $\forall t$ and $y \leq G^{-1}_{t|t-1}(p)$. The parameter θ_0 lies in the interior of a finite-dimensional interval Θ.*

2. *$G_{t|t-1}(x|\theta)$ is continuously differentiable in θ and $x \in \mathbb{R}$ and strictly increasing in $x \in \mathbb{R}$ almost surely.*

3. *The in-sample of size T is used to estimate the parameter $\theta_0 \in \mathbb{R}^p$ with the consistent estimator $\hat{\theta}_T$.*

These assumptions define our framework and should be fulfilled in most situations.
In the next section we present the UC-backtest from Du and Escanciano (2016) and propose our new UC-backtests. The advantage of our tests is that the distribution for finite out-of-sample size is known.

2.2 New UC-Backtest

For testing the hypothesis

$$
\begin{aligned}
&H_0 : E(H_t(p)) = \frac{p}{2},\ \forall\ t = 1, \ldots, n, \\
&vs. \\
&H_1 : \neg H_0
\end{aligned}
\tag{2}
$$

Du and Escanciano (2016) suggest to use a t-test with normal approximation. Since $E(H_t(p)^2) = \frac{p}{3}$ and $\text{Var}(H_t(p)) = \frac{p}{3} - \frac{p^2}{4}$, $\forall t$, the t-test statistic is given by

$$U_{ES} = \sqrt{n}\frac{\overline{H}(p) - p/2}{\sqrt{p(1/3 - p/4)}}$$

with $\overline{H}(p) := \frac{1}{n}\sum_{t=1}^{n} \hat{H}_t(p)$ and $\hat{H}_t(p) = \frac{1}{p}(p - G_{t|t-1}(y_t|\hat{\theta}_T))\mathbb{1}(y_t < G^{-1}_{t|t-1}(p|\hat{\theta}_T))$. Note that $\hat{H}_t(p)$ is used instead of $H_t(p)$ because θ_0 is unknown and has to be estimated by $\hat{\theta}_T$. If $\hat{\theta}_T$ is $\sqrt{T}$-consistent and T increases faster than n (and both tend to infinity) such that $n/T \to 0$, U_{ES} has a standard normal limit distribution.

Now we consider the case that T tends to infinity (no estimation error) but n is fixed and relatively small (e.g. $n = 250$ or $n = 500$). To be more precise, we assume $G_{t|t-1}(\cdot|\hat{\theta}_T) = G_{t|t-1}(\cdot|\theta_0)$. In this situation, $\mathbb{1}(y_t < G^{-1}_{t|t-1}(p|\hat{\theta}_T))$ follows a Bernoulli distribution with parameter p. Moreover, $\frac{1}{p}(p - G_{t|t-1}(y_t|\hat{\theta}_T))$ is uniformly distributed on $(0, 1)$. So, we simulated series of (H_t) by first simulating n Bernoulli variables $b_1, \ldots, b_n$ with parameter p. Then, we set H_t to 0 for all t where $b_t = 0$ and to the realization of a uniform distribution on $(0, 1)$ for all t where $b_t = 1$. In this way we simulated 500 000 series of (H_t) for $n = 250, 500$ and $p = 0.025$ and calculated U_{ES} for each. With a kernel density estimation (Gaussian kernel, bandwith by Silverman's 'rule of thumb', (Silverman, 1986, p. 48)) we compare the estimated density function with the standard normal density function. The result is displayed in Figure 1. We see that the simulated distribution is right skewed. Consequently, quantiles for high probabilities obtained from the normal approximation are too small.

– Figure 1 here –

Our main contribution is to tackle this problem. For this, we have to reformulate the test hypothesis. We start with the following observation:
Whenever a hit occurs at time t, $G_{t|t-1}(y_t)$ should be uniformly distributed on $(0, p)$ and therefore $H_t(p)$ should be uniformly distributed on $(0, 1)$. On the other hand, a hit should occur with probability p if the ES-model is reasonable. Consequently, if the risk model is appropriate, the observable series $\{H_t(p)\}_{t=1}^n$ can be modelled as a series of products of Bernoulli distributed with uniformly distributed random variables. This leads to our test hypotheses:

$H_0 : H_t(p) = h_t(p) \cdot u_t,\ h_t(p) \sim \mathcal{B}(p), u_t \sim \mathcal{U}(0, 1)$ if $h_t(p) = 1$ and bounded otherwise ,

$\forall t = 1, \ldots, n$

vs.

$H_1 :\ \neg H_0$

If H_0 holds, the UC property (1) is obviously fulfilled. Our reformulated hypothesis seems to be stronger than the previous one (2). However, as the calculated expected value and variance in the t-test statistic are based on the reasonable assumption that $E(h_t(u)) = u$, $\forall u \in (0, p)$, it is equivalent to our test hypothesis.
Moreover, no information concerning past values of $H_t(p)$ should be helpful in forecasting hits and their characteristics if the ES-model is correctly specified. Hence, we add the following reasonable assumption, which is similar to what Du and Escanciano (2016) assume in order to derive that the variance of $\sqrt{n}\overline{H}(p)$ is given by $\frac{p}{3} - \frac{p^2}{4}$.

Assumption 2.

The random vectors $(h_1(p), u_1), \ldots, (h_n(p), u_n)$ are independent over time.

Under this assumption, with the law of total probability, the cumulative distribution function (cdf in the following) of

$$H_{\cdot n} := \sum_{t=1}^{n} H_t(p) = \frac{1}{p} \sum_{t=1}^{n} (p - G_{t|t-1}(y_t|\theta_0)) \mathbb{1}(y_t < G_{t|t-1}^{-1}(p|\theta_0)) \tag{3}$$

is given by

$$F_{H_{\cdot n}}(x) := P\left(H_{\cdot n} \leq x\right) = \sum_{k=0}^{n} P\left(\sum_{t=1}^{n} h_t(p) = k, \sum_{j=1}^{k} u_j \leq x\right).$$

Here, $\sum_{t=1}^{n} h_t(p)$ is binomial distributed with parameter n and p and the distribution of $\sum_{j=1}^{k} u_j$ is the so called Irwin-Hall distribution (Irwin (1927), Hall (1927)) with parameter k and cumulative distribution function

$$\Upsilon_k(x) := \frac{1}{k!} \sum_{j=0}^{\lfloor x \rfloor} (-1)^j \binom{n}{j} (x-j)^k.$$

Thus, the cdf is given by

$$F_{H_{\cdot n}}(x) = \begin{cases} 0 & \text{, if } x < 0 \\ (1-p)^n & \text{, if } x = 0 \\ (1-p)^n + \sum\limits_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \Upsilon_k(x) & \text{, if } x \in (0, n] \\ 1 & \text{, if } x > n. \end{cases}$$

With increasing k, Υ_k is numerically unstable, because $\binom{n}{k}$ takes huge values and/or $p^k(1-p)^{n-k}$ is close to zero. For implementation, it can be useful to use the normal approximation beginning from an upper bound (e.g. $k \geq 20$). The Irwin-Hall distribution with parameter k presents the distribution of the sum of k independent uniformly distributed random variables each with expected value $\frac{1}{2}$ and variance $\frac{1}{12}$. One can use the central limit theorem to approximate this distribution by a normal distribution with mean $k \cdot \frac{1}{2}$ and variance $k \cdot \frac{1}{12}$ and therefore $\Upsilon_k(x) \approx \Phi\left(\frac{x-k/2}{\sqrt{k/12}}\right)$, where Φ denotes the distribution function of the standard normal distribution.

To demonstrate the usefulness of our test, we simulate $N = 10,000$ series of $\{H_t(p)\}_{t=1}^n$ with length $n = 250$ and calculate high quantiles of the simulated $\{H_{\cdot n,j}\}_{j=1}^N$. The simulated quantiles, the theoretical quantiles from $F_{H_{\cdot n}}$ and the approximation used from Du and Escanciano (2016) are displayed in Table 1. The latter is a normal distribution with expected value $np/2$ and variance $n(p/3 - p^2/4)$. One notices that the theoretical quantiles from $F_{H_{\cdot n}}$ are closer to the simulated ones than the ones from Du and Escanciano (2016).

– Table 1 here –

In the following, we assume that there is at least one hit or rather $\hat{H}_{\cdot n}(p) > 0$, with $\hat{H}_{\cdot n}$ defined as in (3) with the estimated parameter $\hat{\theta}_T$ instead of θ_0. We get the continuous conditional cdf

$$F_{H_{\cdot n}|H_{\cdot n}>0}(x) = (1 - (1-p)^n)^{-1} \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \Upsilon_k(x).$$

The conditional cdf is used to define our test statistic

$$S_{UC} := F_{H_{\cdot n}|H_{\cdot n}>0}(\hat{H}_{\cdot n}) = (1 - (1-p)^n)^{-1} \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \Upsilon_k(\hat{H}_{\cdot n})$$

whose limit distribution is given in the following theorem.

Theorem 3.
Under Assumption 1 and 2, if n is fixed and $T \to \infty$ then

$$S_{UC} \Big| \left\{\hat{H}_{\cdot n} > 0\right\} \xrightarrow{d} \mathcal{U}(0,1).$$

The proof of this theorem can be found in Appendix A.

Thus one would reject H_0 at level α if $S_{UC} > 1 - \alpha$. We expect that our test has better size properties, because the t-test suffers from the approximation error.

2.3 Multivariate Test

Next, we extend the test to the multivariate framework. Instead of the following approach, one could also simply use the univariate test in combination with a Bonferroni-Holm correction. So, we could use the t-test from Du and Escanciano (2016) or the improved test for each business line $i = 1, \ldots, m$ with the following procedure: The p-values for each line are sorted, leading to the values $P_{(i)}$. If $P_{(i)} < \frac{\alpha}{m+1-i}$ and $P_{(i+1)} > \frac{\alpha}{m-i}$, the first i single hypothesis are rejected. However, we reject the global hypothesis if there is at least one hypothesis that is rejected. To be precise we reject if

$$\tilde{P} := \min_{k=1,\ldots,m} (P_{(k)} \cdot (m + 1 - k)) < \alpha. \tag{4}$$

In contrast to that, we propose a generic multivariate procedure which aims at systematic model errors instead of focusing on single business lines. We expect such a procedure to have more power than the former one. The price for this is that we need stronger assumptions, in particular, n must tend to infinity.

With m business lines we define

$$\mathbf{H}_t := (H_{t,1}(p_1), \ldots, H_{t,m}(p_m))' \tag{5}$$

with $H_{t,j}$ the cumulative violation from business line j at day t, $j = 1, \ldots, m$ and $t = 1, \ldots, n$. For simplicity we assume that $p_1 = \ldots = p_m =: p$, but all results can be easily extended for different coverage levels. The test hypothesis regarding the distribution of the expression in equation (5) is formulated as

$$
\begin{aligned}
H_0^m : \mathbf{H}_t &= (h_{t,1}(p) \cdot u_{t,1}, \dots, h_{t,m}(p) \cdot u_{t,m})', \\
& h_{t,i}(p) \sim \mathcal{B}(p) \text{ and } u_{t,i} \sim \mathcal{U}(0,1), \text{ if } h_t(p) = 1 \text{ and bounded otherwise}, \\
& \forall t = 1, \dots, n, i = 1, \dots, m \\
& \text{vs.} \\
H_1^m : & \neg H_0
\end{aligned}
$$

Similarly to Assumption 2, we assume independence over time and the same cross-sectional dependence structure at each time point:

Assumption 4.

1. *$\boldsymbol{H}_1, \dots, \boldsymbol{H}_n$ are independent.*

2. *$Cov(\boldsymbol{H}_1) = \dots = Cov(\boldsymbol{H}_n) =: \Sigma$.*

Clearly, we need that n tends to infinity to estimate Σ consistently. Therefore, it is required that T tends to infinity relatively faster than n. Moreover, we need the following

Assumption 5.

1. *$\sqrt{T}(\hat{\theta}_T - \theta_0) = O_p(1)$.*

2. *The first moments of the random variables $\sup_{\theta \in \Theta} \frac{\partial E(H_t(p)|\mathbb{F}_{t-1})}{\partial \theta}$ are uniformly bounded over $t \in \mathbb{N}$.*

Assumption 5.1 means that $\hat{\theta}_T$ is $\sqrt{T}$ consistent and the limit distribution of $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is bounded in probability. For example, a maximum likelihood estimator with a fixed, rolling or recursive forecasting scheme fulfills this condition as it is shown by Escanciano and Olmo (2010). The more technical Assumption 5.2 is similar to Assumption A.3 in Escanciano and Olmo (2010). Moreover, in Appendix B, we give one example of lower-level assumptions, under which this assumption holds.

The test statistic is based on the standardized sum of the univariate test statistics. More precisely, the test statistic is given by:

$$S_{UC}^m := \frac{1}{\hat{\sigma}} \sum_{i=1}^{m} \Phi^{-1}(S_{UC,i}) = \frac{1}{\hat{\sigma}} \sum_{i=1}^{m} \Phi^{-1}(F_{H_{\cdot n}|H_{\cdot n}>0}(\hat{H}_{\cdot n,i})) \tag{6}$$

with $\hat{\sigma}^2$ a consistent estimator for $\text{Var}(\sum_{i=1}^m \Phi^{-1}(S_{UC,i}))$.

Theorem 6.

Under Assumption 1,2,4 and 5, if $T \to \infty$ and $n \to \infty$, $n/T \to 0$, it holds

$$S_{UC}^m | \left\{ \hat{H}_{\cdot n,j} > 0,\ \forall j \right\} \xrightarrow{d} \mathcal{N}(0,1),$$

with the variance estimator

$$\hat{\sigma}^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\sum_{t=1}^n (\hat{H}_{t,i} - \overline{H}_i)(\hat{H}_{t,j} - \overline{H}_j)}{\sqrt{\sum_{t=1}^n (\hat{H}_{t,i} - \overline{H}_i)} \sqrt{\sum_{t=1}^n (\hat{H}_{t,j} - \overline{H}_j)}}.$$

Here, $\overline{H}_i := \frac{1}{n} \sum_{t=1}^n \hat{H}_{t,i}$ and $\overline{H}_j := \frac{1}{n} \sum_{t=1}^n \hat{H}_{t,j}$.

The proof of this theorem can be found in Appendix A.

With this theorem, we obtain our multivariate UC-test for backtesting the ES. Given that there is at least one hit in each business line, we reject H_0^m if $S_{UC}^m > u_{1-\alpha}$, with $u_{1-\alpha}$ the $1-\alpha$ quantile of the normal distribution.

3 Simulation Study

In our simulation study we examine the power of our proposed backtest in a controllable but realistic scenario. Also we compare the empirical size and power with the t-test proposed by Du and Escanciano (2016). All simulations are computed for significance levels of 5% and 1%, for $n = 250, 500$ and $p = 0.025$. We perform 2,000 repetitions for each setting.

In the univariate setting our test delivers the exact distribution for finite out-of-sample-size, but uses basically the same information like the t-test. Thus we enhanced the size properties, but the size corrected power is the same as the t-test. So we report both the size corrected and uncorrected power in the univariate setting. However, in the multivariate settings we report only the size corrected power.
We consider two scenarios. In the first scenario there is a structural break between time point T and $T+1$ and in the second scenario the risk model is misspecified.

In the first scenario, in order to extend the univariate AR(1)-GARCH(1,1) model from Du and Escanciano (2016) we use a AR(1)-CCC-GARCH(1,1) model with normal- and t-distributed innovations to generate data. The values of the parameters are the same as in Du and Escanciano (2016), thus these are typical values in empirical applications (Du and Escanciano, 2016, p.15). The m-dimensional series $\{Y_t\}_{t=1,\dots,T+n}$ is generated by:

$$
\begin{aligned}
&Y_t = \rho Y_{t-1} + v_t,\ v_t = \sigma_t \epsilon_t, \epsilon_t \sim t(\nu, \Sigma) \\
&\sigma_t^2 = \omega \mathbf{1}_m + \alpha \mathbf{I}_m v_{t-1}^2 + \beta \mathbf{I}_m \sigma_{t-1}^2.
\end{aligned}
$$

Here, $|\rho| < 1, \omega \geq c > 0, \alpha, \beta \geq 0$ for some constant c. We use $\nu = \infty$ (which leads to normally distributed innovations) to generate the in-sample data $\{Y_t\}_{t=1,\dots,T}$ and fit the model. If the out of sample is also driven by normal innovations the VaR and ES can be consistently estimated and H_0^m holds.
To examine the power of our test, we simulate a structural break at time T, after that point the innovations are t-distributed with $\nu \in \{30, 15, 10, 7\}$ degrees of freedom. The chosen covariance matrix Σ will be fixed with $\Sigma_{ij} = 0.4$ if $i \neq j$ and $\Sigma_{ii} = 1$.
In each of the simulations we estimate $\theta_0 = (\rho, \omega, \alpha, \beta)' = (0.05, 0.05, 0.1, 0.85)'$ separately for each business line by $\hat{\theta}_{T,j}$ with the well known conditional maximum likelihood

estimation with in-sample size $T \in \{500, 2500\}$. Thus we get for each business line j and each day t the estimated cdf

$$G_{t|t-1,j}(Y_{t,j}|\hat{\theta}_{T,j}) = \Phi(\hat{\epsilon}_{t,j}) \tag{7}$$

with

$$\hat{\epsilon}_{t,j} = \frac{Y_{t,j} - \hat{\rho}_j Y_{t-1,j}}{\hat{\sigma}_{t,j}},$$

and

$$\hat{\sigma}_{t,j} = \hat{\omega}_j + \hat{\alpha}_j(\hat{\sigma}_{t-1,j}\hat{\epsilon}_{t-1,j})^2 + \hat{\beta}\sigma^2_{t-1,j}. \tag{8}$$

With this cdf we can calculate the estimated cumulative violation for each day t and each business line j:

$$\hat{H}_{t,j}(p) = \frac{1}{p}\left(p - G_{t|t-1,j}(Y_{t,j}|\hat{\theta}_{T,j})\right)\mathbb{1}(G_{t|t-1,j}(Y_{t,j}|\hat{\theta}_{T,j}) < p). \tag{9}$$

The power of our test is compared to the Bonferroni-Holm adjusted t-test-procedure (called Bonferroni-Holm test in the following), i.e., we report the number of simulation runs in which at least one single hypothesis is rejected based on the decision rule in (4). For this, we use the t-test from Du and Escanciano (2016). For power comparison we report the size corrected power. Therefore, we modify the decision rule (4) in the way that we reject the null hypothesis at significance level α if

$$\min_{k=1,\dots,n}(P_{(k)} \cdot (m + 1 - k)) < \tilde{P}_\alpha,$$

where $\tilde{P}_\alpha$ is the α-quantile of the 2,000 values of $\tilde{P}$ (as defined in (4)) simulated under the null hypothesis.

Tables 2 and 3 show the simulation results in the univariate case for $T = 2500$ and $T = 500$, respectively. As mentioned before, the corrected power is the same for both tests. Therefore the uncorrected power for both tests is reported as well as the common corrected power. In this setting our test has better size properties but less power. The size distortions are due to the parameter estimation error.

– Table 2 & 3 here –

The simulation results for the multivariate case are presented in Tables 4-7 below.
As shown, our test clearly outperforms the standard t-test which has extremely bad size properties. Moreover, the size adjusted power of our test is significantly better in all cases. For an in-sample size of length $T = 500$, our test suffers from slight size distortions which vanish for $T = 2500$.

– Tables 4-7 here –

In the second scenario the data generation process is a multivariate Garch in mean model with constant correlations. The m-dimensional series $\{Y_t\}_{t=1\dots,T+n}$ is generated by:

$$Y_t = -\gamma\sigma_t^2 + v_t,\ v_t = \sigma_t\epsilon_t, \epsilon_t \sim N(0, \Sigma)$$
$$\sigma_t^2 = \omega\mathbf{1}_m + \alpha\mathbf{I}_m v_{t-1}^2 + \beta\mathbf{I}_m\sigma_{t-1}^2.$$

As before, the covariance matrix Σ will be fixed with $\Sigma_{ij} = 0.4$ if $i \neq j$ and $\Sigma_{ii} = 1$. The GARCH paramaters are chosen as proposed by Du and Escanciano (2016) and are set to $\omega = 0.01$, $\alpha = 0.29$ and $\beta = 0.7$. With $\gamma = 0$ the AR-GARCH risk model holds ($\rho = 0$) while for $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ we examine the power of the tests. The estimation of the cdf's, residuals and cumulative violations is analogue to scenario 1 and shown in (7), (8) and (9), respectively.
The simulation results of the univariate case are shown in Tables 8 and 9. Similar to the first setting, our test has slightly better size properties whereas the (uncorrected) power is slightly lower. Again, the corrected power is the same for both tests.

– Tables 8 & 9 here –

Tables 10-13 present the multivariate cases. Here, the Bonferroni-Holm test is clearly outperformed by our multivariate backtest in terms of size and power. Nevertheless, even the multivariate backtest shows a size distortion in case of the small in-sample of size $T = 500$. This distortion vanishes when the in-sample size increases.

– Tables 10-13 here –

4 Application

In this section, we apply the two multivariate backtests discussed in the last section on $m = 13$ time series of bank returns in the time span 2006 until 2016. To be more precise, we consider a subset with sufficiently long history of the stocks from the 20 largest banks that have been previously analyzed in Wied et al. (2016) namely: Citigroup Inc., HSBC Holdings, Barclays, BNP Paribas, The Royal Bank of Scotland Group, Bank of America Corporation, JPMorgan Chase & Co, Deutsche Bank AG, Société Générale, Morgan Stanley, Banco Santander, UniCredit and Credit Suisse. As already mentioned in Wied et al. (2016) the motivation of this empirical study is straightforward. On the one hand, a regulator of a set of banks could be interested in testing the overall fragility of the banking sector. On the other hand, regulators focus mainly on uc backtests. Thus, multivariate backtests could be of significant help to regulators to forecast times of contagion in the financial system, thereby supplementing current efforts to stress-test banking sectors (Acharya and Steffen, 2014).

We consider each year separately and fit three different univariate GARCH models (standard GARCH (S-GARCH), E-GARCH and GJR-GARCH) based on the last 10 years, respectively. From this, we calculate the cumulative violation processes $H_{t,i}(p)$ for $p = 2.5\%$, as used in equation (5). To be more precise, for each stock we model

$$Y_{t,i} = \mu_i + \rho_i Y_{t-1,i} + \gamma_i v_{t-1,i} + v_{t,i},\ \ v_{t,i} = \sigma_{t,i}\epsilon_{t,i}$$

individually with skewed t-distributed innovations $\epsilon_{t,i}$ and depending on the model

$$\sigma^2_{t,i} = \omega_i + \alpha_i v^2_{t-1,i} + \beta_i \sigma^2_{t-1,i} \qquad \text{(S-GARCH)}$$

$$\log \sigma^2_{t,i} = \omega_i + \alpha_i(|\epsilon_{t-1,i}| - E(|\epsilon_{t-1,i}|)) + \theta_i \epsilon_{t-1,i} + \beta_i \log(\sigma^2_{t-1,i}) \qquad \text{(E-GARCH)}$$

$$\sigma^2_{t,i} = \omega_i + (\alpha_i + \theta_i \mathbb{1}(v_{t-1,i} < 0))v^2_{t-1,i} + \beta_i \sigma^2_{t-1,i} \qquad \text{(GJR-GARCH).}$$

The innovations $\epsilon_{t,i}$ are assumed to be independent over time. As the risk of each bank is modelled separately, we estimate the models separately for each series. This approach reflects the situation that each bank is analyzed individually by the regulators. Due to its construction, cross-sectional dependence and, in particular, potential contagion effects are taken into account in our test.

Table 14 presents the empirical means and volatilities of the returns for each year. Moreover, it shows the amount of VaR-violations, the average of $H_{t,i}(p)$ for a year and all stocks, $\frac{1}{m \cdot n}\sum_{t=1}^{n}\sum_{i=1}^{m} H_{t,i}(p)$, as well as the maximal average over the stocks, $\max_{i=1,\dots,m} \frac{1}{n}\sum_{t=1}^{n} H_{t,i}(p)$. The second quantity is what basically enters (6), while the third quantity gives information about the worst business line which basically enters the Bonferroni-Holm test.

In addition, Table 14 shows the p-value of the multivariate test. Moreover, for the Bonferroni-Holm test, we present the quantity

$$\tilde{P} = \min_{k=1,\dots,m} (P_{(k)} \cdot (m + 1 - k)),$$

where $P_{(k)}$ is the k-th sorted p-value, see also equation (4).

With significance level $\alpha = 0.05$, as a whole, the test (6) rejects 4 times and the Bonferroni-Holm test 1 time. Given that we would expect 1.65 rejections for 33 tests, the amount of rejections is small.[1] This result is in line with the intuition that the Bonferroni-Holm test rejects in the presence of one extreme business line, while the multivariate test aims for systematic problems with the model fit. As expected, the p-values of the tests are smaller in the time of the financial crisis around 2008. Moreover, there is evidence that the GJR-GARCH model provides the best fit to the data as there is no rejection for $\alpha = 0.05$. This finding is in line with the existing literature (see, e.g., Berens et al., 2018) and due to the fact that not only the size but also the direction of a shock has an impact on the volatility forecast. This is important as numerous studies have shown that asset returns and conditional volatility are negatively correlated (see, e.g., Bekaert and Wu, 2000). Nevertheless, also here, the p-value is lower compared to other times, which shows that the model fit is worse compared to bullish market times.

– Table 14 here –

5 Conclusion

We present a new backtest for the unconditional coverage property of the ES. The distribution of the test statistic is available for finite out-of-sample size which leads to better size and power properties compared to existing tests. Moreover, it can be easily extended to a multivariate test. Our test is easy to implement and should be used whenever the in-sample size is large compared to the out-of sample size. To the best of our knowledge this is the first proposed ES backtest for the multivariate setting.

[1] Note, the small rejection rates in 2008/2009 are explainable by the large in-sample size of ten years as the models are fitted by means of the volatile markets around millennium. If we decrease the in-sample size (e.g. fit 2003-2007 to predict 2008) the prediction results become much worse.

References

Acerbi, C. and B. Szekely (2014): "Backtesting expected shortfall," *Risk Magazine*, 32 (7), 1404–1415.

Acharya, V. V. and S. Steffen (2014): "Falling Short of Expectations? Stress-Testing the European Banking System," *CEPS Policy Brief*, 315, 1–21.

Artzner, P., F. Delbaen, and D. Heath (1999): "Coherent Measures of Risk," *Mathematical Finance*, 9(3), 203–228.

Bekaert, G. and G. Wu (2000): "Asymmetric volatility and risk in equity markets," *Review of Financial Studies*, 13, 1–42.

Berens, T., G. Weiss, and D. Ziggel (2018): "Estimation Window Strategies for Value-at-Risk and Expected Shortfall Forecasting," *The Journal of Risk*, forthcoming.

Berkowitz, J. (2001): "Testing Density Forecasts With Applications to Risk Management," *Journal of Business and Economic Statistics*, 19, 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.

Candelon, B., G. Colletaz, and S. Tokpavi (2011): *Backtesting Value-at-Risk: A GMM Duration-Based Test*, vol. 9, Wiley, New York.

Du, Z. and J. C. Escanciano (2016): "Backtesting Expected Shortfall: Accounting for Tail Risk," *Management Science, DOI: 10.1287/mnsc.2015.2342*.

Escanciano, J. C. and J. Olmo (2010): "Backtesting Parametric Value-at-Risk With Estimation Risk," *Journal of Business & Economic Statistics*, 28(1), 36–51.

Hall, P. (1927): "The Distribution of Means for Samples of Size N Drawn from a Population in which the Variate Takes Values Between 0 and 1," *Biometrika*, 9(3/4), 240–245.

Irwin, J. O. (1927): "On the Frequency Distribution of the Means of Samples from a Population Having any Law of Frequency with Finite Moments, with Special Reference to Pearson's Type II," *Biometrika*, 9(3/4), 225–239.

Kerkhof, J. and B. Melenberg (2004): "Backtesting for Risk-Based Regulatory Capital," *Journal of Banking & Finance*, 28, 1845–1865.

Lindner, A. and A. Szimayer (2005): "A Limit Theorem for Copulas," *Discussion Paper 433, SFB 823.*

Schweizer, B. and E. F. Wolff (1981): "On Nonparametric Measures of Dependence for Random Variables," *The Annals of Statistics*, 9 (4), 879–885.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Basel Committee on Banking Supervision (2012): *Fundamental review of the trading book: A revised market risk framework*, Consultative document, Bank for International Settlements, Basel.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge, New York.

Wied, D., G. Weiss, and D. Ziggel (2016): "Evaluating Value-at-Risk Forecasts: A New Set of Multivariate Backtests," *Journal of Banking & Finance*, 72, 121–132.

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

——— (2010): "Backtesting value-at-risk based on tail losses," *Journal of Banking & Finance*, 17, 526–538.

Yamai, Y. and T. Yoshiba (2002): "On the Validity of Value-at-Risk: Comparative Analyses with Expected Shortfall," *Monetary and Economic Studies*, 20, 57–85.

——— (2005): "Value at Risk versus expected shortfall: a practical perspective," *Journal of Banking and Finance*, 29, 997–1015.

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.

A Proofs

Proof of Theorem 3

First we prove that for each $t = 1, \dots, n$ it holds that $\hat{H}_t(p) \xrightarrow{p} H_t(p)$:

From Assumption 1 we get $\theta_T \xrightarrow{p} \theta_0$. With the continuous mapping theorem and due to the continuity and strict monotonicity of $G_{t|t-1}(\cdot|\hat{\theta}_T)$, $G^{-1}_{t|t-1}(\cdot|\hat{\theta}_T)$ uniformly converges to $G^{-1}_{t|t-1}(\cdot|\theta_T)$.

On the one hand we get $\forall x \in \mathbb{R}$

$$\frac{1}{p}(p - G_{t|t-1}(x|\hat{\theta}_T)) \xrightarrow{p} \frac{1}{p}(p - G_{t|t-1}(x|\theta_0)).$$

Moreover with $p \in (0,1)$ and y continuously distributed and $0 < \epsilon < 1$

$$\begin{aligned}
&P\left(|\mathbb{1}(y < G^{-1}_{t|t-1}(p|\hat{\theta}_T)) - \mathbb{1}(y < G^{-1}_{t|t-1}(p|\theta_0))| > \epsilon\right) \\
=&P(y \in [\min\{G^{-1}_{t|t-1}(p|\hat{\theta}_T), G^{-1}_{t|t-1}(p|\theta_0)\}, \max\{G^{-1}_{t|t-1}(p|\hat{\theta}_T), G^{-1}_{t|t-1}(p|\theta_0)\}]) \\
\xrightarrow{T\to\infty}&P(y = G^{-1}_{t|t-1}(p|\theta_0)) = 0
\end{aligned}$$

so we get $\mathbb{1}(y < G^{-1}_{t|t-1}(p|\hat{\theta}_T)) \xrightarrow{p} \mathbb{1}(y < G^{-1}_{t|t-1}(p|\theta_0))$. Now we use the following lemma that extends the well known Slutzky-Theorem.

Lemma 7. *If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ then $X_n Y_n \xrightarrow{p} XY$.*

Proof There are at least two ways to prove this. The first is a direct calculation:

$$\begin{aligned}
P(|X_nY_n - XY| > \epsilon) &= P(|X_nY_n - X_nY + X_nY - XY| > \epsilon) \\
&\le P(|X_n(Y_n - Y)| + |Y(X_n - X)| > \epsilon) \\
&\le P(|\underbrace{X_n(Y_n - Y)}_{\xrightarrow{p}0}| > \epsilon/2) + P(|\underbrace{Y(X_n - X)}_{\xrightarrow{p}0}| > \epsilon/2) \xrightarrow{n\to\infty} 0.
\end{aligned}$$

Second, one can use that, under the assumptions, $(X_n, Y_n) \xrightarrow{p} (X, Y)$ so that an application of the continuous mapping theorem yields the result. □

Using this lemma we get

$$\hat{H}_t(p) = \frac{1}{p}(p - G_{t|t-1}(x|\hat{\theta}_T)) \cdot \mathbb{1}(y < G^{-1}_{t|t-1}(p|\hat{\theta}_T))$$
$$\xrightarrow{p} \frac{1}{p}(p - G_{t|t-1}(x|\theta_0)) \cdot \mathbb{1}(y < G^{-1}_{t|t-1}(p|\theta_0)) = H_t(p).$$

Since n is fixed we immediately obtain

$$\hat{H}_{\cdot n} \xrightarrow{p} H_{\cdot n},$$

if T tends to infinity. In the second step we show that

$$F_{H_{\cdot n}}\left(\hat{H}_{\cdot n}\middle|\, H_{\cdot n} > 0\right) \Big| \left\{\hat{H}_{\cdot n} > 0\right\} \xrightarrow{d} \mathcal{U}(0,1).$$

Therefore we use another lemma:

Lemma 8.

Let $X_0, X_1, X_2, \ldots$ be continuous random variables on $(\mathbb{R}, \mathbb{F}, P)$ and $X_n \xrightarrow{d} X_0$ and $g : \mathbb{R} \to \mathbb{R}$ absolute continuous and strictly increasing. Then

$$g(X_n)|\{X_n \in B\} \xrightarrow{d} g(X_0)|\{X_0 \in B\}.$$

Proof

Since g is absolutely continuous and strictly increasing, $g^{-1} : \mathbb{R} \to \mathbb{R}$ exists. Because X is absolutely continuous it follows by definition of the convergence in distribution and Portmanteau Lemma (see eg. van der Vaart (1998), p.6) for every $B \in \mathbb{F}$

$$P(X_n \in B) \overset{n\to\infty}{\longrightarrow} P(X_0 \in B).$$

Let $A \subset \mathbb{R}$ and $g^{-1}(A) := \{x \in \mathbb{R} | g(x) \in A\} \in \mathbb{F}$, $B \in \mathbb{F}$ with $P(X_i \in B) > 0 \ \forall i$.

$$\begin{aligned} P(g(X_n) \in A | X_n \in B) = & \frac{P(g(X_n) \in A, X_n \in B)}{P(X_n \in B)} \\ = & \frac{P(X_n \in g^{-1}(A), X_n \in B)}{P(X_n \in B)} \\ = & \frac{P(X_n \in \{g^{-1}(A) \cap B\})}{P(X_n \in B)} \\ \overset{n\to\infty}{\longrightarrow} & \frac{P(X_0 \in \{g^{-1}(A) \cap B\})}{P(X_0 \in B)} \\ = & P(g(X_0) \in A | X_0 \in B). \end{aligned}$$

So we get $g(X_n)|\{X_n \in B\} \xrightarrow{d} g(X_0)|\{X_0 \in B\}$ and Lemma 8 is proved.

With this lemma the proof of Theorem 1 is clear. Per definition it holds

$$F_{H_{\cdot n}}\left(H_{\cdot n}\middle| \, H_{\cdot n} > 0\right) \Big| \left\{H_{\cdot n} > 0\right\} \sim \mathcal{U}(0,1)$$

and with Lemma 8 it follows easily

$$F_{H_{\cdot n}}\left(\hat{H}_{\cdot n}\middle| \, H_{\cdot n} > 0\right) \Big| \left\{\hat{H}_{\cdot n} > 0\right\} \xrightarrow{d} \mathcal{U}(0,1).$$

□

Proof of Theorem 6

To prove this theorem, we use a copula theorem from Lindner and Szimayer (2005):

Theorem 9.

Let $(X_n)_{n\in\mathbb{N}}$ and X be m-dimensional random vectors, where $X_n = (X_{n,1}, \ldots, X_{n,m})'$ and $X = (X_1, \ldots, X_m)'$.

Then X_n converges weakly to X as $n \to \infty$, if and only if the margins $X_{n,j}$ converge weakly to X_j as $n \to \infty$ for $j = 1, \ldots, m$, and if the copulas C_n of X_n converge pointwisely to the copula C of X on $RanF_1 \times \ldots \times RanF_m$ as $n \to \infty$, where F_j denotes the distribution function of X_j.

Thus the proof will be done in two steps: First we show that if $T \to \infty$, $n \to \infty$ and $n/T \to 0$ all margins

$$\Phi^{-1}(F_{H\cdot n|H\cdot n>0}(\hat{H}_{\cdot n,j}))\Big|\Big\{\hat{H}_{\cdot n,j} > 0,\ \forall j\Big\},\ j = 1, \dots, m,$$

have standard normal limit distribution. In the second step we show that the copula of $(\Phi^{-1}(F_{H\cdot n|H\cdot n>0}(\hat{H}_{\cdot n,1})), \dots, \Phi^{-1}(F_{H\cdot n|H\cdot n>0}(\hat{H}_{\cdot n,m})))'$ converge pointwisely to a Gaussian copula.

If n tends to infinity, we observe $\sup_{x\in\mathbb{R}} |F_{H\cdot n|H\cdot n>0}(x) - F_{H\cdot n}(x)| \longrightarrow 0$ and $P(\hat{H}_{\cdot n,j} > 0, \forall j) \longrightarrow 1$. Thus the conditions $\{\hat{H}_{\cdot n,j} > 0,\ \forall j\}$ have no effect on the asymptotic behavior and we continue the proof without them. But note that in finite samples this condition is needed to estimate the dependence structure.

$H_{\cdot n,j}$ is under Assumption 3 a sum of independent random variables, therefore the central limit theorem holds:

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| F_{H\cdot n}(x) - \Phi\left(\frac{x - np/2}{\sqrt{n(p/3 - p^2/4)}}\right)\right| = 0.$$

Moreover for $c \in \mathbb{R}$ fixed and $\epsilon \in (0, 0.5)$ it holds:

$$\begin{aligned}
&\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| F_{H\cdot n}(x + c\cdot n^\epsilon) - \Phi\Big(\frac{x + c\cdot n^\epsilon - np/2}{\sqrt{n(p/3 - p^2/4)}}\Big)\right| &= 0\\
\Leftrightarrow &\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| F_{H\cdot n}(x + c\cdot n^\epsilon) - \Phi\Big(\frac{x - np/2}{\sqrt{n(p/3 - p^2/4)}} + \underbrace{\frac{c}{\sqrt{n^{1-2\epsilon}(p/3 - p^2/4)}}}_{\to 0}\Big)\right| &= 0\\
\Leftrightarrow &\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| F_{H\cdot n}(x + c\cdot n^\epsilon) - \Phi\Big(\frac{x - np/2}{\sqrt{n(p/3 - p^2/4)}}\Big)\right| &= 0.
\end{aligned}$$

So if $c_n = o_P(\sqrt{n})$ we observe

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \Big| F_{H\cdot n}(x) - F_{H\cdot n}(x + c_n)\Big| = 0\ (a.s.).$$

We now consider $d_n := \hat{H}_{\cdot n} - H_{\cdot n}$. Du and Escanciano (2016) mentioned that

$$\frac{1}{\sqrt{n}} d_n = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} E(\hat{H}_t(p) - H_t(p) | \mathbb{F}_{t-1}) + o_P(1),$$

which follows directly from previous results from Escanciano and Olmo (2010). With similar arguments and the mean value theorem we get:

$$\begin{aligned} d_n &= \sum_{t=1}^{n} (\hat{H}_t(p) - H_t(p)) \\ &= \sqrt{n} \Big[\frac{1}{\sqrt{n}} \sum_{t=1}^{n} E(\hat{H}_t(p) - H_t(p) | \mathbb{F}_{t-1}) + o_P(1) \Big] \\ &= \sum_{t=1}^{n} (\hat{\theta}_T - \theta_0)' \cdot \frac{\partial E(H_t(p)|\mathbb{F}_{t-1})}{\partial \theta} \Big|_{\theta = \tilde{\theta}} + o_P(\sqrt{n}) \\ &= \Big[\underbrace{\sqrt{T}(\hat{\theta}_T - \theta_0)'}_{O_P(1)} \cdot \frac{n}{\sqrt{T}} \cdot \underbrace{\frac{1}{n} \sum_{t=1}^{n} \frac{\partial E(H_t(p)|\mathbb{F}_{t-1})}{\partial \theta} \Big|_{\theta = \tilde{\theta}}}_{O_P(1) \text{ with Assumption 5.2}} + o_P(\sqrt{n}), \Big] \end{aligned}$$

with $\tilde{\theta}$ between $\hat{\theta}_T$ and θ_0.
If $n/\sqrt{T} = o(\sqrt{n}) \Leftrightarrow n/T \to 0$, we get $d_n = o_P(\sqrt{n})$ and

$$F_{H_{\cdot n}}(\hat{H}_{\cdot n_j}) \xrightarrow{d} F_{H_{\cdot n}}(H_{\cdot n_j}) \sim \mathcal{U}(0,1), \ j = 1 \ldots, m,$$

and the first step of the proof is completed with the continuous mapping theorem:

$$\Phi^{-1}(F_{H_{\cdot n}}(\hat{H}_{\cdot n_j})) \to \Phi^{-1}(F_{H_{\cdot n}}(H_{\cdot n_j})) \sim \mathcal{N}(0,1), \ j = 1 \ldots, m.$$

To complete the proof we determine the dependence structure. Under H_0^m and Assumption 3 we get with the central limit theorem for $\overline{\mathbf{H}} := (\overline{H}_1, \ldots, \overline{H}_m)'$:

$$\widetilde{\mathbf{H}}_n := \sqrt{n} \left(\overline{\mathbf{H}} - \mu \right) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

with $\mu = \frac{p}{2} \cdot 1_m$ and a positive definite matrix $\Sigma \in \mathbb{R}^{m\times m}$.

With Theorem 9 the copula C_n of $\widetilde{\mathbf{H}}_n$ converges pointwisely to a Gaussian copula with a correlation matrix R that corresponds to Σ.

We define $g_n : \mathbb{R}^m \to \mathbb{R}^m$, $g_n((x_1, \ldots, x_m)') \to (g_{1,n}(x_1), \ldots, g_{m,n}(x_n))'$, with

$$g_{i,n}(x_i) = \Phi^{-1}(F_{H\cdot n|H\cdot n>0}(\sqrt{n}x_i + np/2)),\ i = 1, \ldots, m.$$

It is easy to see that $g_{i,n}$ is strictly increasing for all $i = 1, \ldots, m$ and $n > 0$ and therefore the distribution of $g_n(\widetilde{\mathbf{H}}_n)$ is also determined by the marginal distributions and the same Copula C_n as before (see eg. Schweizer and Wolff, 1981). Thus we get directly that the copula of $(\Phi^{-1}(F_{H\cdot n|H\cdot n>0}(\hat{H}_{\cdot n,1})), \ldots, \Phi^{-1}(F_{H\cdot n|H\cdot n>0}(\hat{H}_{\cdot n,m})))'$ is also given by C_n and converges to the Gaussian Copula C. Applying Theorem 9 one more time it holds

$$g_n(\widetilde{\mathbf{H}}_n) \xrightarrow{d} \mathcal{N}(0_m, R)$$

and

$$1_m'\ g_n(\widetilde{\mathbf{H}}_n) = \sum_{j=1}^{m} \Phi^{-1}(F_{H\cdot n}(\hat{H}_{\cdot n,j})) \xrightarrow{d} \mathcal{N}(0, 1_m' R 1_m).$$

As mentioned before, using the conditional cdf and the condition $\{\hat{H}_{\cdot n,j} > 0,\ \forall j\}$ does not change the asymptotic behavior. Returning to the conditional case and using the consistent estimator

$$\hat{\sigma}^2 = \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\sum_{t=1}^{n}(\hat{H}_{t,i} - \overline{H}_i)(\hat{H}_{t,j} - \overline{H}_j)}{\sqrt{\sum_{t=1}^{n}(\hat{H}_{t,i} - \overline{H}_i)}\sqrt{\sum_{t=1}^{n}(\hat{H}_{t,j} - \overline{H}_j)}},$$

for $1_m' R 1_m$, we derive the limit distribution of the test statistic:

$$\frac{1}{\hat{\sigma}} \sum_{i=1}^{m} \Phi^{-1}(F_{H\cdot n|H\cdot n>0}(\hat{H}_{\cdot n,i}))|\left\{\hat{H}_{\cdot n,j} > 0,\ \forall j\right\} \xrightarrow{d} \mathcal{N}(0, 1).$$

□

B Remark on Assumption 5.2

In this section, we want to discuss the validity of Assumption 5.2. which states that the first moments of the random variables $\sup_{\theta\in\Theta} \frac{\partial E(H_t(p)|\mathbb{F}_{t-1})}{\partial\theta}$ are uniformly bounded over $t \in \mathbb{N}$.

For example, this assumption holds under the sufficient conditions

i) $(Y_t)_{t\in\mathbb{Z}}$ is a strict stationary process

ii) $(Y_t)_{t\in\mathbb{Z}}$ follows a location scale model $Y_t = \mu_t(\theta_0, \mathbb{F}_{t-1}) + \sigma_t(\theta_0, \mathbb{F}_{t-1}) \cdot \epsilon_t$, $\theta_0 \in \Theta \subset \mathbb{R}^p$

iii) the cdf of ϵ_t is given by F_ϵ and its density by f_ϵ

iv) f_ϵ is bounded: $\sup_{x\in\mathbb{R}} f_\epsilon(x) < \infty \; \forall t$

v) The terms Y_t, $\sup_{\theta\in\Theta} |\left(\frac{\partial\mu_t(\theta,\mathbb{F}_{t-1})}{\partial\theta}\right)|$, $\sup_{\theta\in\Theta} |\left(\frac{\partial\sigma_t(\theta,\mathbb{F}_{t-1})}{\partial\theta}\right)|$, $\mu_t(\theta_0, \mathbb{F}_{t-1})$ and $\sigma_t(\theta_0, \mathbb{F}_{t-1})$ have finite second moments, uniformly over t

vi) $P\left(\sigma_t(\theta, \mathbb{F}_{t-1}) > c\right) = 1$ for some $c > 0$.

Sketch of the proof in five steps:

1. In the following steps we show that under conditions i)-vi) the first moment of the term

$$\sup_{\theta\in\Theta} \left|\frac{\partial}{\partial\theta} H_t(p)\right|$$

is finite and therefore

$$\frac{\partial E(H_t(p)|\mathbb{F}_{t-1})}{\partial\theta} = E\left(\frac{\partial H_t(p)}{\partial\theta}|\mathbb{F}_{t-1}\right).$$

2. Because of $\frac{\partial}{\partial\theta}\mathbb{1}(G_t(Y_t|\theta,\mathbb{F}_{t-1})<p)=0$ (a.s) we obtain by product differentiation

$$\begin{aligned}\frac{\partial H_t(p)}{\partial\theta}\Big|\mathbb{F}_{t-1} =& \frac{\partial}{\partial\theta}\left[\frac{1}{p}\left(p-G_t(Y_t|\theta,\mathbb{F}_{t-1})\right)\cdot\mathbb{1}(G_t(Y_t|\theta,\mathbb{F}_{t-1})<p)\right]\\ \overset{\text{a.s.}}{=}& -\left[\frac{\partial}{\partial\theta}G_t(Y_t|\theta,\mathbb{F}_{t-1})\right]\cdot\mathbb{1}(G_t(Y_t|\theta,\mathbb{F}_{t-1})<p).\end{aligned}$$

3. The gradient is given by

$$\begin{aligned}\frac{\partial}{\partial\theta}G_t(Y_t|\theta_t,\mathbb{F}_{t-1}) &= \frac{\partial}{\partial\theta}F_\epsilon\left(\frac{Y_t-\mu_t(\theta,\mathbb{F}_{t-1})}{\sigma_t(\theta,\mathbb{F}_{t-1})}\right)\\ &= \frac{-\sigma_t(\theta,\mathbb{F}_{t-1})\left(\frac{\partial\mu_t(\theta,\mathbb{F}_{t-1})}{\partial\theta}\right)-(Y_t-\mu_t(\theta,\mathbb{F}_{t-1}))\left(\frac{\partial\sigma_t(\theta,\mathbb{F}_{t-1})}{\partial\theta}\right)}{\sigma_t^2(\theta,\mathbb{F}_{t-1})}\cdot f_\epsilon\left(\frac{Y_t-\mu_t(\theta,\mathbb{F}_{t-1})}{\sigma_t(\theta,\mathbb{F}_{t-1})}\right).\end{aligned}$$

4. Under the given conditions and by applications of the Cauchy-Schwarz-inequality above it holds that the first moment of

$$\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta}G_t(Y_t|\theta,\mathbb{F}_{t-1})\right|$$

is finite and therefore

$$\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta}H_t(p)\right| = O_p(1).$$

5. With $\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta}H_t(p)\right| = O_p(1)$ we can interchange derivation and expectation (as mentioned in step 1.) and Assumptions 5.2 follows by applications of the Cauchy-Schwarz-inequality and the uniform boundedness of the second moments.

C Figures and Tables

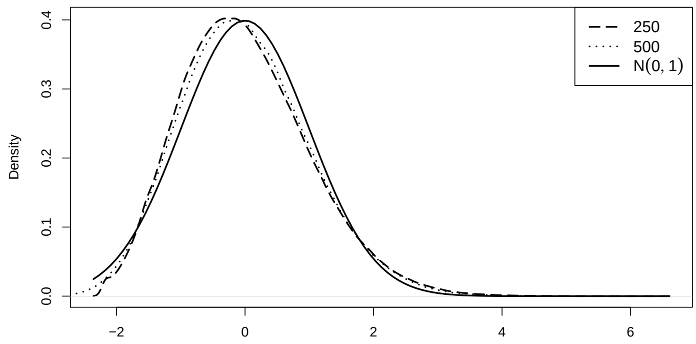

Figure 1: Kernel smoothed density function of U_{ES} with out-of-sample-size $n = 250$ and $n = 500$.

Table 1: Comparison between simulated quantiles and the theoretical quantiles for the univariate test, $n = 250$.

quantile $1 - \alpha$	0.95	0.96	0.97	0.98	0.99
simulated $1 - \alpha$- quantile	5.68	5.87	6.11	6.43	6.96
$F_{H.n}^{-1}(1 - \alpha)$	5.67	5.86	6.10	6.43	6.95
$np/2 + \sqrt{(n(p/3 - p^2/4))}\Phi^{-1}(1 - \alpha)$	5.48	5.63	5.81	6.06	6.45

Table 2: Empirical rejection probabilities for the univariate t-test and the new finite sample backtest with in-sample size $T = 2500$, $\alpha \in \{0.01, 0.05\}$ and scenario 1 (structural break).

	$\alpha = 0.05$						$\alpha = 0.01$					
	$n = 250$			$n = 500$			$n = 250$			$n = 500$		
ν	UC-Test	t-test	corr.	UC-Test	t-test	corr.	UC-Test	t-test	corr.	UC-Test	t-test	corr.
∞	0.06	0.07	0.05	0.05	0.06	0.05	0.01	0.03	0.01	0.01	0.03	0.01
30	0.17	0.19	0.15	0.23	0.25	0.21	0.06	0.09	0.05	0.09	0.12	0.07
15	0.34	0.37	0.32	0.48	0.51	0.46	0.17	0.23	0.14	0.28	0.34	0.24
10	0.54	0.57	0.51	0.75	0.77	0.73	0.32	0.39	0.29	0.56	0.63	0.52
7	0.73	0.76	0.71	0.92	0.93	0.91	0.53	0.62	0.49	0.81	0.85	0.78

Table 3: Empirical rejection probabilities for the univariate t-test and the new finite sample backtest with in-sample size $T = 500$, $\alpha \in \{0.01, 0.05\}$ and scenario 1 (structural break).

	$\alpha = 0.05$						$\alpha = 0.01$					
	$n = 250$			$n = 500$			$n = 250$			$n = 500$		
ν	UC-test	t-test	corr.	UC-test	t-test	corr.	UC-test	t-test	corr.	UC-test	t-test	corr.
∞	0.09	0.11	0.05	0.11	0.13	0.05	0.04	0.06	0.01	0.05	0.06	0.01
30	0.21	0.24	0.11	0.32	0.35	0.17	0.09	0.12	0.03	0.17	0.21	0.06
15	0.40	0.43	0.26	0.53	0.56	0.36	0.22	0.28	0.09	0.36	0.40	0.16
10	0.57	0.60	0.44	0.77	0.78	0.61	0.39	0.46	0.20	0.61	0.66	0.36
7	0.75	0.78	0.63	0.92	0.93	0.83	0.59	0.65	0.38	0.82	0.85	0.60

Table 4: Empirical rejection probabilities for the Bonferroni-Holm test and the new multivariate test for $m = 10$ business lines, in-sample size $T = 2500$, different degrees of freedom ν for the t-distribution, different out-of-sample sizes n, $\alpha \in \{0.01, 0.05\}$ and scenario 1 (structural break).

	$\alpha = 0.05$				$\alpha = 0.01$			
	$n = 250$		$n = 500$		$n = 250$		$n = 500$	
ν	UC_{10}-Test	t-test	UC_{10}-Test	t-test	UC_{10}-Test	t-test	UC_{10}-Test	t-test
∞	0.06	0.12	0.06	0.12	0.01	0.05	0.01	0.05
30	0.32	0.26	0.52	0.30	0.16	0.07	0.25	0.13
15	0.73	0.58	0.93	0.74	0.54	0.28	0.78	0.51
10	0.94	0.85	1.00	0.95	0.85	0.60	0.97	0.86
7	1.00	0.97	1.00	1.00	0.98	0.87	1.00	0.99

Table 5: Empirical rejection probabilities for the Bonferroni-Holm test and the new multivariate test for $m = 50$ business lines, in-sample size $T = 2500$, different degrees of freedom ν for the t-distribution, different out-of-sample sizes n, $\alpha \in \{0.01, 0.05\}$ and scenario 1 (structural break).

	$\alpha = 0.05$				$\alpha = 0.01$			
	$n = 250$		$n = 500$		$n = 250$		$n = 500$	
ν	UC_{50}-test	t-test	UC_{50}-test	t-test	UC_{50}-test	t-test	UC_{50}-test	t-test
∞	0.05	0.23	0.06	0.20	0.01	0.12	0.01	0.10
30	0.43	0.26	0.63	0.37	0.22	0.08	0.39	0.15
15	0.86	0.65	0.99	0.87	0.67	0.32	0.93	0.62
10	0.98	0.89	1.00	0.99	0.94	0.67	1.00	0.94
7	1.00	0.99	1.00	1.00	1.00	0.93	1.00	1.00

Table 6: Empirical rejection probabilities for the Bonferroni-Holm test and the new multivariate test for $m = 10$ business lines, in-sample size $T = 500$, different degrees of freedom ν for the t-distribution, different out-of-sample sizes n, $\alpha \in \{0.01, 0.05\}$ and scenario 1 (structural break).

	$\alpha = 0.05$				$\alpha = 0.01$			
	$n = 250$		$n = 500$		$n = 250$		$n = 500$	
ν	UC_{10}-test	t-test	UC_{10}-test	t-test	UC_{10}-test	t-test	UC_{10}-test	t-test
∞	0.11	0.30	0.15	0.35	0.03	0.17	0.06	0.22
30	0.32	0.21	0.43	0.23	0.14	0.04	0.14	0.05
15	0.72	0.46	0.85	0.57	0.48	0.14	0.59	0.18
10	0.92	0.73	0.98	0.85	0.79	0.36	0.90	0.46
7	0.99	0.93	1.00	0.98	0.97	0.64	1.00	0.83

Table 7: Empirical rejection probabilities for the Bonferroni-Holm test and the new multivariate test for $m = 50$ business lines, in-sample size $T = 500$, different degrees of freedom ν for the t-distribution, different out-of-sample sizes n, $\alpha \in \{0.01, 0.05\}$ and scenario 1 (structural break).

	$\alpha = 0.05$				$\alpha = 0.01$			
	$n = 250$		$n = 500$		$n = 250$		$n = 500$	
ν	UC_{50}-test	t-test	UC_{50}-test	t-test	UC_{50}-test	t-test	UC_{50}-test	t-test
∞	0.14	0.53	0.15	0.60	0.05	0.39	0.06	0.47
30	0.35	0.18	0.56	0.27	0.16	0.05	0.30	0.08
15	0.78	0.44	0.95	0.66	0.58	0.21	0.84	0.32
10	0.97	0.77	1.00	0.94	0.90	0.48	1.00	0.71
7	1.00	0.95	1.00	1.00	0.99	0.79	1.00	0.96

Table 8: Size and power for the univariate t-test and the new test with in-sample size $T = 2500$, $\alpha \in \{0.01, 0.05\}$ and scenario 2 (model misspecification).

	$\alpha = 0.05$						$\alpha = 0.01$					
	$n = 250$			$n = 500$			$n = 250$			$n = 500$		
γ	UC-test	t-test	corr.	UC-test	t-test	corr.	UC-test	t-test	corr.	UC-test	t-test	corr.
0.0	0.06	0.07	0.05	0.06	0.07	0.05	0.01	0.02	0.01	0.01	0.02	0.01
0.1	0.11	0.13	0.10	0.14	0.16	0.12	0.03	0.06	0.03	0.05	0.07	0.04
0.2	0.17	0.19	0.16	0.25	0.27	0.22	0.06	0.10	0.07	0.11	0.15	0.10
0.3	0.25	0.27	0.24	0.33	0.36	0.30	0.11	0.16	0.12	0.17	0.22	0.16
0.4	0.31	0.34	0.30	0.45	0.48	0.42	0.16	0.21	0.17	0.26	0.30	0.25
0.5	0.36	0.40	0.34	0.54	0.56	0.50	0.19	0.24	0.20	0.32	0.39	0.31

Table 9: Size and power for the univariate t-test and the new test with in-sample size $T = 500$, $\alpha \in \{0.01, 0.05\}$ and scenario 2 (model misspecification).

	$\alpha = 0.05$						$\alpha = 0.01$					
	$n = 250$			$n = 500$			$n = 250$			$n = 500$		
γ	UC-test	t-test	corr.	UC-test	t-test	corr.	UC-test	t-test	corr.	UC-test	t-test	corr.
0.0	0.08	0.10	0.05	0.10	0.12	0.05	0.02	0.04	0.01	0.03	0.05	0.01
0.1	0.16	0.19	0.12	0.19	0.22	0.12	0.07	0.09	0.04	0.09	0.12	0.03
0.2	0.21	0.24	0.16	0.31	0.33	0.21	0.10	0.14	0.06	0.18	0.21	0.07
0.3	0.29	0.32	0.24	0.41	0.43	0.31	0.15	0.20	0.10	0.26	0.31	0.13
0.4	0.36	0.38	0.30	0.49	0.52	0.38	0.19	0.25	0.13	0.32	0.37	0.17
0.5	0.40	0.43	0.34	0.55	0.57	0.42	0.23	0.29	0.16	0.38	0.42	0.21

Table 10: Empirical rejection probabilities for the Bonferroni-Holm test and the new multivariate test for $m = 10$ business lines, in-sample size $T = 2500$, out-of-sample sizes $n \in \{250, 500\}$, $\alpha \in \{0.01, 0.05\}$, different values of γ and scenario 2 (model misspecification).

	$\alpha = 0.05$				$\alpha = 0.01$			
	$n = 250$		$n = 500$		$n = 250$		$n = 500$	
γ	UC_{10}-test	t-test	UC_{10}-test	t-test	UC_{10}-test	t-test	UC_{10}-test	t-test
0.0	0.05	0.13	0.05	0.13	0.01	0.05	0.01	0.05
0.1	0.17	0.16	0.26	0.21	0.06	0.06	0.08	0.08
0.2	0.36	0.32	0.57	0.41	0.18	0.16	0.27	0.19
0.3	0.54	0.46	0.81	0.62	0.34	0.27	0.55	0.36
0.4	0.71	0.56	0.92	0.77	0.49	0.36	0.75	0.51
0.5	0.83	0.68	0.97	0.88	0.65	0.45	0.89	0.67

Table 11: Empirical rejection probabilities for the Bonferroni-Holm test and the new multivariate test for $m = 50$ business lines, in-sample size $T = 2500$, out-of-sample sizes $n \in \{250, 500\}$, $\alpha \in \{0.01, 0.05\}$, different values of γ and scenario 2 (model misspecification).

	$\alpha = 0.05$				$\alpha = 0.01$			
	$n = 250$		$n = 500$		$n = 250$		$n = 500$	
γ	UC_{50}-test	t-test	UC_{50}-test	t-test	UC_{50}-test	t-test	UC_{50}-test	t-test
0.0	0.06	0.22	0.05	0.20	0.01	0.12	0.01	0.09
0.1	0.19	0.21	0.32	0.24	0.06	0.08	0.12	0.08
0.2	0.45	0.43	0.71	0.53	0.22	0.23	0.48	0.27
0.3	0.72	0.62	0.93	0.78	0.46	0.39	0.80	0.48
0.4	0.86	0.74	0.99	0.90	0.65	0.50	0.95	0.67
0.5	0.95	0.84	1.00	0.97	0.84	0.63	0.99	0.82

Table 12: Empirical rejection probabilities for the Bonferroni-Holm test and the new multivariate test for $m = 10$ business lines, in-sample size $T = 500$, out-of-sample sizes $n \in \{250, 500\}$, $\alpha \in \{0.01, 0.05\}$, different values of γ and scenario 2 (model misspecification).

	$\alpha = 0.05$				$\alpha = 0.01$			
	$n = 250$		$n = 500$		$n = 250$		$n = 500$	
γ	UC_{10}-test	t-test	UC_{10}-test	t-test	UC_{10}-test	t-test	UC_{10}-test	t-test
0.0	0.10	0.25	0.12	0.30	0.03	0.14	0.04	0.17
0.1	0.18	0.13	0.25	0.17	0.06	0.04	0.09	0.06
0.2	0.34	0.25	0.51	0.32	0.14	0.10	0.27	0.11
0.3	0.51	0.37	0.71	0.48	0.27	0.16	0.48	0.21
0.4	0.66	0.46	0.87	0.65	0.42	0.21	0.70	0.34
0.5	0.77	0.55	0.94	0.74	0.55	0.27	0.82	0.39

Table 13: Empirical rejection probabilities for the Bonferroni-Holm test and the new multivariate test for $m = 50$ business lines, in-sample size $T = 500$, out-of-sample sizes $n \in \{250, 500\}$, $\alpha \in \{0.01, 0.05\}$, different values of γ and scenario 2 (model misspecification).

	$\alpha = 0.05$				$\alpha = 0.01$			
	$n = 250$		$n = 500$		$n = 250$		$n = 500$	
γ	UC_{50}-test	t-test	UC_{50}-test	t-test	UC_{50}-test	t-test	UC_{50}-test	t-test
0.0	0.13	0.46	0.16	0.54	0.03	0.33	0.06	0.39
0.1	0.19	0.19	0.26	0.15	0.08	0.05	0.11	0.04
0.2	0.43	0.32	0.61	0.30	0.22	0.10	0.36	0.10
0.3	0.65	0.46	0.86	0.49	0.43	0.18	0.66	0.19
0.4	0.82	0.58	0.95	0.67	0.64	0.27	0.85	0.28
0.5	0.92	0.70	0.98	0.75	0.78	0.33	0.94	0.38

Table 14: Some yearly summary statistics for the empirical study and the corresponding results for the two multivariate backtests.

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Return per day	0.10 %	-0.07 %	-0.42 %	0.12 %	-0.05 %	-0.22 %	0.11 %	0.10 %	-0.03 %	-0.06 %	-0.02 %
Volatility per day	1.19 %	1.69 %	5.30 %	5.26 %	2.56 %	3.44 %	2.68 %	1.73 %	1.43 %	1.71 %	2.59 %
S-GARCH											
Number -Hits	61	120	117	102	113	133	85	99	99	111	88
mean H_t	0.96 %	1.77 %	1.86 %	1.53 %	1.71 %	1.98 %	1.22 %	1.56 %	1.39 %	1.77 %	1.34 %
max $\bar{H}_n$	1.72 %	2.54 %	2.34 %	2.64 %	1.95 %	2.68 %	1.96 %	2.43 %	1.96 %	2.43 %	1.73 %
p-value m-test	82.24 %	7.24 %	**3.30** %	18.11 %	9.71 %	**1.18** %	52.13 %	15.21 %	27.82 %	8.19 %	38.13 %
$\tilde{P}$-value t-test	95.47 %	13.92 %	33.83 %	8.30 %	53.39 %	8.56 %	94.73 %	6.79 %	82.27 %	23.10 %	80.48 %
E-GARCH											
Number -Hits	61	103	115	114	106	122	77	90	84	91	69
mean H_t	0.96 %	1.49 %	1.86 %	1.78 %	1.50 %	1.54 %	1.13 %	1.49 %	1.18 %	1.41 %	1.08
max $\bar{H}$	1.90 %	2.32 %	2.63 %	3.06 %	2.05 %	2.35 %	1.94 %	2.26 %	1.76 %	2.09 %	1.44
P-value m-test	82.20 %	23.25 %	**3.40** %	**4.73** %	21.77 %	12.68 %	64.41 %	20.38 %	55.96 %	30.24 %	63.04
$\tilde{P}$-value t-test	96.72 %	35.82 %	8.65 %	**0.80** %	88.86 %	32.65 %	94.33 %	46.94 %	83.02 %	71.49 %	93.32 %
GJR-GARCH											
Number -Hits	63	104	112	108	107	120	83	99	88	96	70
mean H_t	0.91 %	1.57 %	1.76 %	1.60 %	1.57 %	1.70 %	1.22 %	1.61 %	1.26 %	1.50 %	1.12
max $\bar{H}$	1.71 %	2.30 %	2.37 %	2.49 %	2.18 %	2.56 %	1.78 %	2.57 %	1.76 %	2.06 %	1.60
Pvalue m-test	85.86 %	17.10 %	5.81 %	10.75 %	16.26 %	5.97 %	51.30 %	12.06 %	43.23 %	22.46 %	59.47
$\tilde{P}$-value t-test	95.65 %	39.10 %	29.56 %	17.38 %	61.95 %	12.85 %	93.31 %	11.75 %	79.96 %	59.71 %	94.47 %