Backtesting VaR and ES under the magnifying glass

Christos Argyropoulos
Lancaster University, UK

Ekaterini Panopoulou*
University of Kent, UK

April 8, 2019

Abstract

Backtesting provides the means of determining the accuracy of risk forecasts and the corresponding risk model. Given that the actual return generating process is unknown, the evaluation methods rely on various assumptions in order to quantify the models inefficiencies and proceed with the model evaluation. These method specific assumptions, in conjunction with the regulatory policies can introduce distortions in the evaluation process, which affect the reliability of the evaluation results. To investigate such effects from a practitioner's perspective, this paper reviews the major Value at Risk and Expected Shortfall forecast evaluation methods and evaluates their performance under a common simulation and financial application framework. Our findings suggest that focusing on specific individual hypothesis tests provides a more reliable alternative than the corresponding conditional coverage ones. In addition, selecting a two year out-of-sample period provides a significantly better power to relevance ratio than the more relevant but powerless regulatory one-year specification.

Keywords: Value-at-Risk, Expected Shortfall, Model Accuracy, Backtesting, Forecast Evaluation

* *Corresponding author.* Ekaterini Panopoulou, Kent Business School, University of Kent, Canterbury CT2 7FS, United Kingdom, T824469, Email: A.Panopoulou@kent.ac.uk.

1 Introduction

Driven either by the regulations or its own utility, an institution needs to distinguish the accurate risk forecasting methods from a large pool of proposed specifications. To this end, evaluation or backtesting of risk models provides the means for determining the accuracy of the candidate models. Although backtesting is a crucial component of the internal model approach, there are no specific regulatory recommendations for the type of tests that should be used. On the contrary, the evaluation methodology is freely chosen by the implementing institution. With respect to the evaluation of the candidate risk models, there is a large body of literature proposing two major approaches. The density evaluation approach was established by the results of Diebold et al. (1998) and Berkowitz (2001). It evaluates the fit of the model's implied density, or specific regions of it, to the historical data. On the other hand, the forecast evaluation approach tests a model's accuracy by assessing the properties embedded in the forecasts. This approach is the industry benchmark as it provides intuitive motivation, ease of implementation and small demand for sensitive information.

Despite its advantages and appeal, the forecast evaluation approach suffers from small sample inefficiencies. The scarcity of extreme events reduces the amount of available/testable information. Furthermore, model risk emerges also as another major source of unreliability since it can distort the results in two possible ways. First, Escanciano and Olmo (2010) and Escanciano and Olmo (2011) suggest that the forecast inherited model risk affects the asymptotic variance of the test statistics. Second, the test statistics specifications and assumptions evaluate a predefined structure that accounts only for a small portion of the actual return dynamics. Finally, with respect to risk measure selection there is a large debate regarding the risk measure selection and whether it can be evaluated or not. Some academics suggest that ES forecasts can not be evaluated given the measure's lack of elicitability (see, for example Ziegel (2014)). On the other hand, there is new evidence that elicitability is not necessary for forecast backtesting (see, for

example Emmer et al. (2015)).

In a recent paper, Nieto and Ruiz (2016) survey the VaR forecasting and backtesting literature, however without evaluating the performance of the backtesting methods. Specifically, the authors evaluate the performance of various VaR forecasting methods at the 1% coverage level. Their empirical exercise includes different setups designed to look into the effects of various in-sample and out-of-sample periods. Their findings suggest that there is significant variation in the accuracy of the forecasting methods. In addition, the authors conclude that simpler methods with asymmetric volatility dynamics and error distributions are the most competitive. Our work also relates to Campbell (2007) as it focuses on the performance of the backtesting methods. Specifically, through a simulation study we evaluate various VaR and ES coverage level series and out-of-sample lengths in order to assess the sample properties of the forecast evaluation approach methods. Finally, we use the S&P 500 returns to evaluate the forecasts on real financial data.

Our findings suggest that selecting an intermediate out-of-sample length increases significantly the reliability of the methods since the high coverage levels in conjunction with small out-of-sample periods (one year) leads to distorted size and low power. Specifically, the tests are oversized for almost all the cases under investigation and the corresponding power is low. Simple tests that focus directly on the quantitative perspective of the risk forecast evaluation fail to perform adequately in a small information set environment, while more elaborate specifications suffer in a rich information set environment. Under our simulation and financial application exercise, we find that three individual hypotheses testing specifications are more robust and almost equally powerful to the respective conditional coverage counterparts.

The rest of the paper is structured as follows. In Section 2 we introduce the VaR and ES definitions/notions and main backtesting approaches. Section 3 describes the simulation results and the small sample properties of the methods. In Section 4 we conduct an empirical implementation on the S&P returns and Section 5 concludes.

2 Backtesting

Regardless of the forecasting methodology, the forecaster needs to prove the model's ability to approximate the actual, but unknown, distribution of returns. This can be done by either evaluating each model's implied density fit (density evaluation) or the accuracy of the produced forecasts (forecast backtesting). In the following sections we describe the methods and underlying ideas for the forecast evaluation approach as the density evaluation is beyond the scope of this paper.

2.1 VaR Backtesting

Forecast evaluation methods, or Event Probability Approach methods (Campbell (2007)), examine if the properties implied by the correctly specified model are showed by the forecasts. To set up ideas, let $VaR_t(q)$ be the time t q-conditional quantile of a long position with a continuous return distribution. Given the information set G_{t-1} , VaR is mathematically defined as:

$$P(r_t \leq -VaR_t(q) \,|G_{t-1}) = q. \tag{1}$$

Emanating from equation (1), the probability of losses "violating" an accurate VaR measure is $q\%$ almost surely. In addition, this result should remain independent of the information set. The rationale of such properties boils down to the model capacity to produce an acceptable number of exceedances/violations of VaR, in absence of dependency on the past information set. Specifically, let the violation sequence be a series of random variables defined below:

$$I_t(q) = \begin{cases} 1, \text{ if } \ r_t < -VaR_t(q) \\ \\ 0, \text{ if } \ r_t > -VaR_t(q). \end{cases} \tag{2}$$

Given an accurate risk model the following holds:

$$P(r_t \leq -VaR_t(q) \,|G_{t-1}) = q \Longrightarrow E\left[I_t(q)|G_{t-1}\right] = q. \tag{3}$$

Equation (3) dictates that the number of losses exceeding the VaR are not in excess/less than the one dictated by the confidence level. Equally important, the violations of VaR should be independent of the past information set. In other words, the exceedances should be random events, not derived from the model inadequacy to interpret or adapt to the evolving information set.

2.1.1 Violation Tests

Kupiec (1995) proposes the Percentage of Failure (POF) test, a straightforward method of evaluating the number of exceedances in relation to the expected ones. The null hypothesis of unconditional coverage is defined as $LR_{POF,0} : E\left[I_t(q)\right] = q$ and tested via a simple test statistic. Following Kupiec (1995), Christoffersen (1998) proposes a complete methodology of evaluating the number of exceedances and their independence. The author states that in order to examine the validity of a VaR model, an implication of equation (3) should be put to the test. This implication consists of examining whether the violation sequence is iid Bernoulli(q) which is formally stated as:

$$E[I_t(q)|I_{t-1}(q), I_{t-2}(q), I_{t-3}(q), ...] = q.$$

The latter can be partitioned to the Unconditional Coverage (UC) hypothesis where the null hypothesis $H_{0,uc}$:$E[I_t(q)] = q$ is tested against the alternative $H_{1,uc}$:$E[I_t(q)] \neq q$, and the iid property which is tested through a first order Markov structure. The independence test rationale dictates that, if the violations are dependent then the transition probabilities would not be equal. Finally, Christoffersen (1998) proposes a joint test that combines both hypotheses (Conditional Coverage CC hypothesis). In order to test for the aforementioned hypothesis, the author proposes the following Likelihood Ratios (LR):

$$LR_{uc} = -2ln(\frac{(1-q)^{T_0}q^{T_1}}{(1-\frac{T_1}{T})^{T_0}(\frac{T_1}{T})^{T1}}) \sim \chi^2_1, \tag{4}$$

$$LR_{ind} = -2ln((1-\frac{T_1}{T})^{T_0}(\frac{T_1}{T})^{T1}) + 2ln((1-\pi_{01})^{T_{00}}\pi_{01}^{T_{01}}(1-\pi_{11})^{T_{10}}\pi_{11}^{T_{11}} \sim \chi^2_1, \tag{5}$$

$$LR_{cc} = LR_{uc} + LR_{ind} \sim \chi^2_2, \tag{6}$$

where T is the number of out-of-sample observations, T_0 the number of non violations, T_1 the number of violations and T_{ij} with $i, j = 0$ $(no\ violation)$, $1(violation)$ is the number of observed events with the j event following the i event. The estimates of the probabilities of T_{ij} are marked as π_{01} and π_{11}. Berkowitz et al. (2011) unify and extend the aforementioned tests by redefining equation (3) on the basis of a martingale difference sequence:$E\left[(I_t(q) - q) \otimes Z_{t-1}\right] = 0$, where Z_{t-1} is the variable describing the information test available at the formulation of the VaR forecast. For the cases of $Z_{t-1} = I_{t-k}, k \geq 1$, the authors propose a Portmanteau test in order to evaluate whether the autocorrelations of the violations sequence are zero.

Engle and Manganelli (2004) propose the Dynamic Quantile (DQ) approach, focusing directly on the correlation of VaR forecasts with the available information set. The corresponding evaluation method is based on a quantile regression model, which associates the observed violations with the past violations and any past information according to the following structure:

$$Hit_t(q) = \delta + \sum_{j=1}^{K} \beta_j Hit_{t-j}(q) + \sum_{j=1}^{K} \gamma_j \zeta_{t-j} + \varepsilon_t, \tag{7}$$

where $Hit_t(q) = I_t(q) - q$ denotes the modified violations sequence, δ is a constant term and ζ_{t-j} corresponds to any information derived from the existing information set. The null hypothesis of independence, DQ_{ind}, dictates that $\beta_j = \gamma_j = 0$, $\forall j = 1...K$, while the null hypothesis, DQ_{uc}, for the number of violations dictates that $\delta = 0$. The DQ test expands the information set for the independence evaluation by including explanatory variables and higher orders lags.

With respect to the DQ approach, Dumitrescu et al. (2012) juxtapose the inconsistency of implementing a linear specification model on binary dependent variables, arguing that it will distort the respective hypothesis testing. This is due to the discrete nature of the distribution of linear model errors and their consequent heteroskedasticity. To alleviate these shortcomings, the authors propose a non-linear-Dynamic Binary (DB)

regression model, aiming to improve the finite sample properties through the usage of a more appropriate link function. The proposed model is given below:

$$E\left[I_t(q)|G_{t-1}\right] = P\left[I_t(q) = 1|G_{t-1}\right] = F\left(\pi_t\right), \tag{8}$$

$$\pi_t = c + \sum_{j=1}^{K} \beta_j \pi_{t-j} + \sum_{j=1}^{K} \delta I_{t-j}(q) + \sum_{j=1}^{K} \psi_j l_{t-j}(\zeta_{t-j}) + \sum_{j=1}^{K} \gamma_j l_{t-j}(\zeta_{t-j}) I_{t-j}(q), \tag{9}$$

where $F\left(\right)$ is an arbitrary CDF and π_t is an index that relates the information set with the violation sequence. Dumitrescu et al. (2012) propose seven specifications for π_t ranging from the simple autoregressive case to the one that introduces asymmetric effects to the violation history.[1] The estimation of the above coefficients is conducted using maximum likelihood methods. Under the null hypotheses, equation (3) holds and leads to the following result:

$$H_{0,DB_{CC}} : \beta_j = \delta_j = \psi_j = \gamma_j = 0 \; and \; c = F^{-1}\left(q\right), \forall j \in \left\{1, 2, 3, ..., K\right\}, \tag{10}$$

$$H_{0,DB_{IND}} : \beta_j = \delta_j = \psi_j = \gamma_j = 0 \; , \forall j \in \left\{1, 2, 3, ..., K\right\}. \tag{11}$$

Finally, within the quantile regression framework, Gaglianone et al. (2011) propose a random coefficient test linking the conditional quantile of the return distribution with VaR forecasts.

[1] The link function specifications are the following:

1. $\pi_t = c + \beta_1 \pi_{t-1}$
2. $\pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1}$
3. $\pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1} + \delta_2 I_{t-2}$
4. $\pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1} + \delta_2 I_{t-2} + \delta_3 I_{t-3}$
5. $\pi_t = c + \beta_1 \pi_{t-1} + \psi_1 VaR_{t-1}$
6. $\pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1} + \psi_1 VaR_{t-1}$
7. $\pi_t = c + \beta_1 \pi_{t-1} + \delta_1 I_{t-1} + \psi_1 VaR_{t-1} + \gamma_1 VaR_{t-1} I_{t-1}$

2.1.2 Duration Tests

While the violations tests focus directly on the violations of the VaR threshold, the duration approach takes into account the time interval between two violations. It evaluates the independence and conditional coverage hypotheses by testing the distribution properties of the sequence of time intervals between violations. The duration approach is based on the idea of dependence causing violations to cluster. In more detail, let d_v be the time interval between the $v-1$ and the v violation. In the conditional coverage case, the evaluated forecast series should produce exactly q violations equally spread across the out-of-sample period. Therefore, the violation sequence will be characterized by a distribution with no memory.[2] This entails that the d_v sequence will follow the geometric distribution, i.e.

$$f(d_v, q) = q(1-q)^{d_v-1}, d_v \in N. \tag{12}$$

Thus the probability of a violation at time t does not depend on the elapsed days since the previous violation. The only continuous distribution characterized by the lack of memory is the exponential distribution:

$$f(d_v, q) = qe^{-qd_v}.$$

Christoffersen and Pelletier (2004) formulate a duration test by considering the Weibull distribution with parameters a, b for the alternative hypothesis;[3, 4]

$$w(d_v, a, b) = a^b b d_v^{b-1} e^{-ad_v^b}. \tag{13}$$

Consequently the Independence null hypothesis (Dur_{Ind}) is not rejected if $b = 1$ while the conditional coverage (Dur_{cc}) hypothesis is not rejected if $b = 1$ and $a = q$.

Drawing on the duration approach literature, Candelon et al. (2011) propose a test which utilizes the GMM approach to test for the Geometric distribution directly. Specif-

[2]Haas (2005) argues that the duration approach provides a clear cut interpretation of parameters while on the other hand it requires specific distributional assumptions for the alternative case of dependence.

[3]The authors also considered the Gamma distribution as it encapsulates the exponential. However the results reported by Haas (2005) lean towards the Weibull distribution as the alternative.

[4]The Exponential distribution can be derived from the Weibull distribution for $b = 1$.

ically their test employs the orthonormal polynomials associated with the geometric distribution. This enables the separate evaluation of the unconditional coverage and independence hypothesis. Furthermore, there is no need to specify a distribution as an alternative. Specifically the orthonormal polynomials related to the geometric distribution are defined below:

$$M_{j+1}\left(d_v,\beta\right)=\frac{\left(1-\beta\right)\left(2j+1\right)+\beta\left(j-d+1\right)}{\left(j+1\right)\sqrt[2]{1-\beta}}M_j\left(d_v,\beta\right)-\left(\frac{j}{j+1}\right)M_{j-1}\left(d_v,\beta\right), \quad (14)$$

with $M_{-1}\left(d_v,\beta\right)=0$ and $M_0\left(d_v,\beta\right)=1$. Evaluating the unconditional coverage hypothesis is straightforward. Under the null hypothesis, the expected value of the duration variable should be equal to $1/q$. Thus the null hypothesis can be stated as $H_{0,Juc}$:$E[M_1\left(d_i,q\right)]=0$. In order to evaluate the null hypothesis the authors propose the following test:

$$J_{uc}\left(p\right)=\left(\frac{1}{\sqrt[2]{N}}\sum_{i=1}^{N}M_1\left(d_i,q\right)\right)^2\rightarrow\chi_1^2. \quad (15)$$

Independence testing consists of testing the duration sequence for a geometric distribution with a parameter q' not necessarily equal to q. This enables the testing of the independence property separately. The null hypothesis can be stated as $H_{0,Jind}$:$E[M_j\left(d_i,q'\right)]=0$, $j\in\{1,2,3,...,p\}$, with the corresponding test defined as:

$$J_{ind}\left(p\right)=\left(\frac{1}{\sqrt[2]{N}}\sum_{i=1}^{N}M\left(d_i,q'\right)\right)^T\left(\frac{1}{\sqrt[2]{N}}\sum_{i=1}^{N}M\left(d_i,q'\right)\right)\rightarrow\chi_p^2, \quad (16)$$

where $M\left(d_i,q\right)$ is the vector of the orthonormal polynomials $M_j\left(d_i,q\right)$ and $j\in\{1,2,3,...,p\}$. For the conditional coverage property, the corresponding null hypothesis with respect to the aforementioned polynomials is defined as $H_{0,cc}$:$E[M_j\left(d_v,q\right)]=0$, $j\in\{1,2,3,...,k\}$. This implies that the duration sequence $\{d_1.d_2,d_3,...,d_N\}$ follows a geometric distribution with $q\%$ success rate. In order to evaluate this hypothesis the authors propose the following test:

$$J_{cc}\left(p\right)=\left(\frac{1}{\sqrt[2]{N}}\sum_{i=1}^{N}M\left(d_i,q\right)\right)^T\left(\frac{1}{\sqrt[2]{N}}\sum_{i=1}^{N}M\left(d_i,q\right)\right)\rightarrow\chi_p^2, \quad (17)$$

where $M(d_i, q)$ is the vector of the orthonormal polynomials $M_j(d_i, q)$ and $j \in \{1, 2, 3, ..., p\}$.

Within the duration approach literature, Berkowitz et al. (2011) propose a LR test based on a the hazard function of the duration sequence. The authors state that under the null hypothesis of conditional coverage, the duration hazard function should be flat and equal to q. Pelletier and Wei (2016) expand the aforementioned method by including the vector of VaR forecasts. The underlying reasoning suggests that if VaR forecasts are misspecified they might be the cause of the upcoming violation. In other words, the authors expand the information set by including also VaR forecasts. Finally, Santo and Alves (2012) propose an independence testing procedure on the concept of exact distribution. In more detail, the authors propose the $\frac{\max - duration}{median - duration}$ ratio as a means of testing the independency property. Specifically, if the model is accurate the duration sequence would be equally spread and consequently the ratio would be equal to a specific value. If the empirical value of the ratio deviates (is larger) from the theoretical one, then there is strong evidence against the independence hypothesis.

Commenting on the aforementioned methods, Ziggel et al. (2014) argues that the unconditional and independence properties, as stated and tested, suffer from severe restrictions. In order for the unconditional coverage null hypothesis to hold there must be $P[I_t(q) = 1] = q \; \forall t$. Implicitly this result imposes a stationary condition for the violation sequence, which is counterintuitive. For instance, in high volatility periods the probability of a violation is higher even if the total empirical number is equal to the expected one. Furthermore, the authors argue that evaluating the autocorrelations of the violation sequence may not be sufficient. On the contrary, there is a possibility of violation clustering despite the violation sequence iid property. In order to mitigate these distortions, the authors redefine the UC and IND properties as follows:

$$H_{0,MCSuc} : E[\frac{1}{n}\sum_{t=1}^{n} I_t(q)] = q$$

$$\{I_t(q)\} \, iid \; Bernoulli(\widetilde{q}) \; \forall t,$$

where $\widetilde{q}$ is an arbitrary probability. In order to test the aforementioned hypothesis the authors propose the following test statistics:

$$MCS_{uc} = \sum_{t=1}^{n} I_t(q) + \epsilon, \epsilon \sim 0.001N(0,1), \tag{18}$$

$$MCS_{ind} = t_1^2 + (n - t_m)^2 + \sum_{t=2}^{m} (t_i - t_{i-1})^2 + \epsilon, \epsilon \sim 0.001N(0,1), \tag{19}$$

where $\{t_1, t_2, t_3, ..., t_m\}$ are the exact times of violations.[5] The introduction of the random variable ϵ enables the tests to keep their size through infinite Monte Carlo simulations, which is essential for the calculation of the critical values. With respect to the independency test, the main goal is to quantify the distances between the violations. If the violations cluster, the sum part of MCS_{ind} statistic will generate larger values in comparison to the case of non clustering violations. On the other hand, if the violations are equally spread out in the sample, the sum part would acquire its minimum value. Finally, for the conditional coverage case they propose a weighted function of the MCS_{uc} and MCS_{ind} tests:

$$MCS_{cc} = af(MCS_{uc}) + (1-a)g(MCS_{ind}), 0 \leq a \leq 1,$$

$$f(MCS_{uc}) = \left| \frac{MCS_{uc}/(n-q)}{q} \right| = \left| \frac{\left(\sum_{t=1}^{n} I_t(q) + \epsilon\right)/(n-q)}{q} \right|,$$

$$g(MCS_{ind}) = \frac{MCS_{ind} - \widehat{r}}{\widehat{r}},$$

where $\widehat{r}$ is an estimator of the expected value of MCS_{ind} under the null hypothesis. Contrary to the LR_{cc} test described in equation (6), the components of the MCS_{cc} test are both positive which prohibits any offsetting effects. Finally, the authors waive a formal asymptotic distribution for the test and instead calculate the critical values through Monte Carlo simulations.

[5] The formulation of the $H_{0,MCSuc}$ enables the testing of under or over estimating the risk in the form of the alternatives $H_{0,MCSuc} : E[\frac{1}{n}\sum_{t=1}^{n} I_t(q)] > q$ or $H_{0,MCSuc} : E[\frac{1}{n}\sum_{t=1}^{n} I_t(q)] < q$ respectively. In addition the authors state that the power of the one sided test is significantly higher.

2.1.3 Multilevel Tests

Given that an accurate model should describe correctly the whole tail of the distribution, evaluating the performance on a single coverage level may be misleading. Hurlin and Tokpavi (2007) propose a multilevel approach by considering multiple coverage levels and their cross-correlations. The authors suggest that extending the Portmanteau test to the multilevel case would increase the information set and thus provide a more powerful test. More in detail, the martingale difference sequence property of the $Hit_t(q) = I_t(q) - q$ series dictates that $E\left[Hit_t(q)|G_{t-1}\right] = 0$, thus $E\left[Hit_t(q)Hit_{t-k}(q)\right] = 0$ for every $k \in N$ and $E\left[Hit_t(q)Hit_{t-k}(q')\right] = 0$ for every $q \neq q'$ and $k \in N$. Based on this result, Hurlin and Tokpavi (2007) propose a multivariate extension of the Portmanteau tests in order to evaluate the null hypothesis $H_{0,Qcc}$:$E[Hit_t(q_i)Hit_{t-k}(q_j)] = q,\ k = 1, 2, 3, .., K$ and $q_i \neq q_j$:[6]

$$Q_m\left(K\right) = T\sum_{k=1}^{K}\left(vec\widehat{R}_k\right)^T\left(\widehat{R}_0^{-1}\otimes\widehat{R}_0^{-1}\right)\left(vec\widehat{R}_k\right) \rightarrow \chi^2_{km^2}, \tag{20}$$

where $\widehat{R}_k = D\,\widehat{C}_k D$, $\widehat{C}_k$ is the empirical covariance matrix of the Hit_t vector and D is the diagonal matrix containing the standard deviations associated to the $Hit_t(q)$. In terms of VaR confidence levels, the authors consider the 1%, 5% and 10% level of coverage and up to the fifth violation lag. This is done in order to ensure that the matrix of hit sequences would not be singular. Drawing on the multilevel VaR testing literature, Leccadito et al. (2014) propose two methods of conditional coverage testing in order to deal with the cases of singular VaR forecasts matrices. The first consists of the expansion of the Christoffersen (1998) approach to the multilevel case by incorporating multiple VaR coverage levels in the transition matrix. The second method detects whether each coverage level produces the expected violations, while at the same time it evaluates the dependency structure through a Pearson type test.

Focusing on the unconditional coverage case, Perignon and Smith (2008) propose a

[6]This test can be considered as a multivariate extension of the Berkowitz et al. (2011) proposed methodology.

multilevel generalization of the Kupiec (1995) test. The authors define a series of violations sequences that schematically determine the magnitude of the violations. In the same vein, Colletaz et al. (2013) propose a test which takes into account the severity of each violation. Specifically, a second violation sequence is defined as follows:

$$J_t = \begin{cases} 1, \text{ if } \ r_t < -VaR_t(q') \\ 0, \text{ if } \ r_t > -VaR_t(q'), q' < q, \end{cases} \tag{21}$$

where q' is a stricter coverage level. The second violations or super exemptions sequence aims at measuring the number of initial violations that exceed the second threshold $\text{VaR}_t(q')$. Thus, if the risk model produces an acceptable number of violations in conjunction with an increased number of super exemptions (losses of extreme severity) the null hypothesis will be rejected. To perform the test, three indicator functions are introduced:

$$g_{0,t} = 1 - g_{1,t} - g_{2,t} = 1 - I_t,$$

$$g_{1,t} = I_t - J_t = \begin{cases} 1, \text{ if } \ -VaR_t(q') < r_t < -VaR_t(q) \\ 0, \text{ if } \ r_t < -VaR_t(q'), \end{cases}$$

$$g_{2,t} = J_t = \begin{cases} 1, \text{ if } \ r_t < -VaR_t(q') \\ 0, \text{ if } \ r_t > -VaR_t(q'). \end{cases}$$

The above random variables follow the Bernoulli distribution with $1-q$, $q-q'$, q' parameters respectively. The joint null hypothesis of the test is defined as $H_{0,muc}$: $E[I_t(q)] = q$ and $E[J_t(q')] = q'$ and the test is performed via the following likelihood ratio:

$$LR_{muc} = -2ln((1-q)^{N_0}(q-q')^{N_1}\,(q')^{N_2}) + 2ln((1-\frac{N_0}{T})^{N_0}(\frac{N_1}{T})^{N_1}(\frac{N_2}{T})^{N_2}) \sim \chi^2_2, \tag{22}$$

where $N_{i,t} = \sum_{t=1}^{T} g_{i,t}, i = 0, 1, 2.$

2.2 ES Backtesting

Contrary to VaR, backtesting ES can be characterized as a more elaborate process. According to Embrechts et al. (2014), VaR as a frequency oriented measure can be evaluated directly by a hit and miss process. On the other hand, ES as a severity measure requires the specification of the underlying DGP process or at least an assumption about it. To make matters worse, Ziegel (2014) argues that only elicitable risk measures can be meaningfully compared, while Gneiting (2011) proves that although VaR is generally elicitable, ES in not.[7] However, the results of Emmer et al. (2015) and Fissler and Ziegel (2016) suggest that the pair of VaR and ES is jointly elicitable, paving the way for a meaningful comparison of competing ES forecasts.

Contrary to the ranking of ES forecasts, Acerbi and Szekely (2014) and Emmer et al. (2015) question the elicitability as a necessary condition for the statistically adequacy of a risk measure. According to Acerbi and Szekely (2014) backtesting ES should not be confused with ranking and comparing a series of competing ES forecasts. Similarly, Kerkhof and Melenberg (2004) and Du and Escanciano (2017) argue that backtesting ES is feasible and not more difficult than backtesting VaR.[8]

Drawing on the early VaR backtesting methods, Christoffersen (2011) proposes (for the continuous case of return distributions) an adaptation of the Christoffersen (1998) method. Specifically, he proposes a regression based test where the deviations from the $ES_t(q)$ during the violations of $VaR_t(q)$ are linked with the vector of variables X_t which correspond to the information set G_{t-1}. The idea is to evaluate whether the risk model utilizes all the available information efficiently in order to forecast ES. The test is based

[7] Consider the loss function $S : (T(\widehat{F_j}), r_t) \rightarrow \mathbb{R}$ where $T(\widehat{F_j}) \rightarrow \mathbb{R}$ is a functional of a competing distribution $\widehat{F_i}$. If g is the true distribution and we assume a non-negative representation of the loss functions, S is a consistent loss function for a specific functional T if $E(S(T(g), R)) \leq E(S(T(\widehat{F_i}), R))$. It is strictly consistent if $E(S(T(g), R)) = E(S(T(\widehat{F_i}), R)) \rightarrow T(g) = T(\widehat{F_i})$. In other words, a consistent scoring function would ensure that the most accurate forecast is selected. Furthermore, a functional T is called elicitable if and only if there is a loss function that is strictly consistent for it.

[8] Acerbi and Szekely (2017) debate the theoretical notion of backtestability and concludes that in a strict sense, ES can not be backtested. However, the authors suggest that ES forecasts can be evaluated statistically in conjunction with an auxiliary statistic since there is a model independent mechanism that guarantees small sensitivity on the auxiliary statistic predictions.

on the predictive ability of the vector series to explain the deviation of the tail losses from the expected ones. If the forecasts are accurate, then there should be no predictive ability from the vector of variables.

Focusing on the properties of the exceedances, McNeil and Frey (2000) define the respective residuals as:

$$res_t = \frac{r_t - \widehat{ES_t}(q)}{\widehat{\sigma}_t}, \tag{23}$$

where $\widehat{\sigma}_t$ is the conditional standard deviation of the utilized model. Under the null hypothesis of correct fit of the model, the residual series $\{res_t\}_{t=1}^{n}$ should have a zero mean distribution. The testing of the null hypothesis is conducted through a bootstrap technique in order to avoid assumptions about the $\{res_t\}_{t=1}^{n}$ series distribution. In the same vein, Righi and Ceretta (2014) propose an adaptation where the dispersion of the exemptions are used in order to standardize the test statistic. Colletaz et al. (2013) utilize the fact that both risk measures are produced by the same model/underlying distribution. Therefore, the authors consider a higher threshold q_0 such that $VaR_t(q_0) = ES_t(q)$. The underlying idea is to evaluate the performance of the risk model on describing the tail of the distribution via the magnitude of the losses.

Although the aforementioned tests are intuitive, they do not tackle the ES forecast accuracy directly. Emmer et al. (2015) focus on the tail area under consideration and suggest the evaluation of the VaR forecasts that represent the quartiles of that specific area. If these four VaR forecasts are accurate, the corresponding ES would be accurate. Kratz et al. (2018) extend this methodology and develop a multinomial VaR threshold approach where multiple coverage level forecasts are jointly evaluated. Acerbi and Szekely (2014) propose a more straightforward ES evaluation approach. The authors consider only the unconditional coverage case since they assume that the independence of tail events is tested separately. Specifically, the null is defined as $P^{[q]} = F^{[q]}$ where $P^{[q]}$ and $F^{[q]}$ are the distribution tails of the model and actual returns respectively. They propose three non parametric specifications in order to evaluate the validity of the ES forecasts. The first test is similar to McNeil and Frey (2000) and it averages the losses at the violations of VaR.

In addition, it requires the testing of the underlying VaR threshold. The second test is more straightforward as it relies on the unconditional definition of ES, while the third test consists of a modification of the density evaluation approach proposed by Diebold et al. (1998) and Berkowitz (2001). More in detail, the authors utilize the ranked probabilities to estimate the ES and compare them against the theoretically correct ones, according to the null hypothesis of Berkowitz (2001). In principle, the critical values of the tests are computed through simulations.

With respect to direct ES forecast evaluation, Du and Escanciano (2017) propose the first conditional coverage testing methodology based on the notion of cumulative violations. To fix up ideas, let

$$H_t(q) = \frac{1}{q}\int_0^q I_t(u)du \tag{24}$$

be the cumulative violation process which accumulates the violations across the distribution's tail. From equations (1) and (2) the following holds:

$$I_t(u) = 1(r_t < VaR_t(q)) = 1(u_t < u), \tag{25}$$

where $u_t = F_{t-1}^{-1}(r_t), F_{t-1}$ is the model imposed conditional CDF and 1() is the indicator function. Thus, equation (24) can be rewritten as:

$$H_t(q) = \frac{1}{q}(q - u_t)1(u_t < u). \tag{26}$$

Equation (26) provides a better insight on the notion of cumulative hits which measures the distance of the returns from the corresponding q quantile during the violations. The authors prove that if $\{1(r_t < \text{VaR}_t(q)) - q\}_{t=1}^{\infty}$ is a martingale difference sequence (mds) then $\{H_t(q) - \frac{q}{2}\}_{t=1}^{\infty}$ is also an mds. This enables the constructions of tests that would evaluate the accuracy of the ES forecasts by evaluating the mds property. Thus the null unconditional coverage hypothesis is defined as $H_{0,U_{ES}} : E[H_t(q)] = \frac{q}{2}$. The conditional coverage hypothesis is defined as $H_{0,C_{ES}} : E[H_t(q)|G_{t-1}] = \frac{q}{2}$. In order to evaluate the aforementioned hypotheses the authors propose the following sample test

statistics:

$$U_{ES} = \frac{\sqrt[2]{n}(\overline{H_t(q)} - \frac{q}{2})}{\sqrt[2]{q(1/3 - q/4)}} \to N(0,1), \tag{27}$$

$$C_{ES} = n \sum_{j=1}^{m} \widehat{\rho}_{nj} \to \chi^2_m, \tag{28}$$

where $\overline{H_t(q)} = \sum_{t=1}^{n} \widehat{H}_t(q)$ is the sample mean of the empirical cumulative violation sequence and $\widehat{\rho}_{nj}$ is the j-th lag of the of the empirical cumulative violation sequence sample autocorrelation. For the CC case the autocovariances of the cumulative violation hits are evaluated.[9] Following Du and Escanciano (2017), we also calculate the similar Box Pierce VaR Conditional test C_{VaR}.

3 Small Sample Properties

In this section we use Monte Carlo simulations to evaluate and compare the small sample properties of the tests described in Section 2. We assume the following $GARCH(1,1) - t_7$ model for the daily returns' Data Generating Process (DGP):

$$\begin{gathered} r_t = \mu + \varepsilon_t,\ \varepsilon_t = \sigma_t z_t(\frac{5}{7}),\ z_t \sim t_7, \\ \sigma_t = \omega + 0.1\varepsilon^2_{t-1} + 0.85\sigma^2_{t-1}. \end{gathered} \tag{29}$$

The model specification is similar to the simulation setup of Du and Escanciano (2017) while in our case μ and ω are estimated on the daily log returns of S&P500 covering the

[9] Loser et al. (2018) improve the Unconditional Coverage test for the cases of finite out of sample periods. Specifically, for finite out-of-sample periods and no estimation error, they generalize the null hypothesis of the ES unconditional coverage based on the concept of the cumulative violation as a product of a Bernoulli and a Uniform random variables. Furthermore, they derive the actual distribution of the respective U_{ES} test which leads to improved small sample properties. Their simulation results suggest that under estimation error the size properties.

period of *2/1/1985-15/10/2008*.[10] The estimates of μ and ω are 2.8×10^{-4} and 7.528×10^{-6}. The daily variance persistence is equal to 0.95 and the annualized unconditional standard deviation is equal to 0.195.

For each simulation sample we calculate the $1\%, 2.5\%, 5\%, 10\%$ $GARCH-Normal$, $GARCH(1,1)-t_7$ and *Historical Simulation* (HS) VaR and ES forecast series of length $R \in \{1000, 750, 500, 250\}$, which we compare with the respective out-of-sample returns of the simulated sample. This enables the computation of the test statistics described in Section 2. With respect to the estimation process, for the $GARCH-Normal$ and $GARCH(1,1)-t_7$ risk forecasts we use a rolling estimation window of length $T = 10000$.[11] Contrary to the in-sample specifications of the parametric models, we calculate the HS risk forecasts with a rolling window of 250 daily observations. In this way, we ensure that the HS risk forecasts are relatively responsive to small changes in the returns volatility levels. The aforementioned process is repeated 10000 times in order to calculate the rejection rates of the forecast series for each coverage level and out-of-sample period. For the size properties, we use the $GARCH(1,1)-t_7$ forecasts rejection cases in order to calculate the false rejection rate. For the power properties, we use the HS forecasts results to calculate the correctly rejected cases and produce the main rsults of our analysis. Furthermore, we complement our analysis with the respective rejections cases against the $GARCH-Normal$ forecasts, which we report in the on-line appendix.

By definition, the non-parametric and unconditional estimation process of HS, creates forecasts that react to the underlying volatility fluctuations rather than anticipate them. In other words, an increasing pattern of the underlying volatility could lead to a successive underestimation of risk. Similarly, a switch from a high to low volatility period, or a numnber of extreme losses embeded in a low volatility period, will not be identified immediately by the HS risk forecasts, which could lead to an extended period of risk over-

[10] Du and Escanciano (2017) utilize an $AR(1)-GARCH(1,1)-t_5$ with the following parameter vector $\theta = (a_0, \omega, a, \beta) = (0.05, 0.05, 0.1, 0.85)$.

[11] The $T = 10000$ in-sample specification targets a $R/T < 10\%$ ratio and therefore makes any estimation risk effects on the asymptotic properties of the implemented tests negligible (see, for example Du and Escanciano (2017), Escanciano and Olmo (2011)).

estimation. Hence the HS forecasts are efficiently violating the conditional coverage and independence properties (Candelon et al. (2011), p.326) without imposing estimation error on the forecasts. Contrary to the HS forecasts, the $GARCH-Normal$ forecasts share the same conditional variance specification with the underlying returns' DGP. Hence, we expect the risk forecasts to adapt faster to the changes of the underlying volatility levels and consequently alleviate, to some extent, the impact of the method's misspecification to the risk forecasts adequacy. On the other hand, the $GARCH-Normal$ forecasts will suffer from the distributional mismatch between the DGP and the forecasting method. The lack of density at the tails of the Normal distribution will lead to an underestimation of the risk, especially for the more extreme coverage levels. On the other hand, the increased density at regions closer to the middle part of the distribution will produce an increased standardised quantile and therefore a possible overestimation of risk.

Table 1 summarizes the evaluated specifications alongside their abbreviations as these are reported in the following sections and figures.

[Table 1 around here]

3.1 Violations Distribution

Implementing an evaluation process may include risks and limitations that influence the final results. For example, the LR tests can not be calculated for cases of no violations. Similarly, the Dur tests needs more than one violation in order to create the duration sequence. Figures 1 and 2 describe the empirical distribution of the number of violations per simulated sample and forecasting method under the $\{10000, R\}$, $R \in \{1000,750,500,250\}$ in-sample/out-of-sample specifications. As expected, the $GARCH-t$ number of violations distribution (Figure 1) have an almost symmetrical distribution which, for each case of coverage level and out-of-sample length, is located approximately around the expected number of violations. However, for the more extreme coverage levels and smaller out-of-sample periods the zero and one violation bins seem to increase in size. This effect

is more pronounced for the one-year out-of-sample period and 1% coverage level, where the zero and one violation bins account for over 25% of the simulated samples.

[Figure 1 around here]

The HS VaR forecasts (Figure 2) results suggest that for 10% and 5% coverage levels, the zero and one violation bins remain unpopulated for almost all the out-of-sample specifications. The 2.5% coverage level results suggest that for the four- and three-year out-of-sample periods the zero and one violations bins are unpopulated while for the two- and one-year cases the number of samples yielding zero or one violation are increasing. For the 1% coverage level the probability of a sample with a zero or one violation is significantly increased for all the out-of-sample specifications, with the one-year period rendering more than 40% of the simulated samples with zero or one violation.

[Figure 2 around here]

These results have an immediate impact on the implementation of the evaluation methods. The shape of the distribution of the number of violations suggest that the high coverage levels combined with short out-of-sample periods may hinder the seamless implementation of the evaluation methods that focus directly on the number of violations. This outcome is sensitive to the underlying leptokurtic DGP and the fact that small out-of-sample periods leave no margin to smooth out the length of the confidence interval of the expected number of violations. Consequently, there is a direct impact to small sample properties of such methods since most of the high coverage level rejections will be attributed mostly to the right tail of the violation number empirical distribution (underestimation of risk).[12] Therefore, high coverage levels and small out-of-sample periods will impact directly such methods since their implementation is not feasible. On

[12]This finding is supported by the distribution of violations of the $GARCH - Normal$ forecasts for the 1% coverage level. Specifically, the underestimation of risk by the $GARCH - Normal$ forecasts creates a more symmetrical violation distribution, located above a larger than the expected number of violations. Since there is not physical restriction by the violations zero bin, there are rejection cases due to a small number of violations. Please refer to Figure 1 of the on-line appendix and the accompanying text for a more detailed analysis.

the other hand, methods that focus more on the properties of the violations sequence may not be directly impacted. However, the lack of testable information may harm the validity of their results since there will be no adequate amount of violations to test for any structure and properties.

3.2 Size Properties

Figure 3 reports the empirical size of the tests for the 5% notional size. The size properties are calculated as the rejection rates of the $GARCH(1,1) - t_7$ risk forecasts series over the total number of the simulated samples. Ideally, we would expect the "perfect" test to deliver false rejection rates equal to the notional size for each of the out-of-sample periods and coverage levels. This is the case for the lower coverage levels and most of the tests considered. However, for the higher coverage levels the size of the tests depends mainly on the length of the evaluation period.

The 10% coverage level size results (Figure 3 Part A) suggest that most of the VaR unconditional coverage tests ($LR_{uc} - MCS_{uc}$) have size properties close (less than 1.5% distortion) or equal to the notional size regardless of the out-of-sample period. The only exception is the LR_{muc} test as its size distortions are quite large and positively correlated with the length of the out-of-sample period. The 10% coverage level in conjunction with the larger out-of-sample period may lead to an increased amount of DGP-related super exemptions. The likelihood of a loss violating both the 10% and 1% coverage level is larger than the likelihood of a loss violating the 1% and the 0.5% coverage levels. The independence tests ($LR_{ind} - MCS_{ind}$) false rejection rates are similar to the unconditional coverage case as most of the tests' size is close or equal to the notional size. For the LR_{ind} specification the one-year out-of-sample period leads to an increased size. This can be attributed to a large number of consecutive violations of the moderate 10% VaR threshold during periods of increased volatility. For the Dur_{ind} test, the size of the test seems to have extreme distortions for all the out-of-sample periods. The abundance of violations leads to a misestimation of the Weibull parameters and therefore to the false rejection of

the correct method.

Turning to the conditional coverage tests ($LR_{cc} - C_{VaR10}$), the rejection rates suggest that simple specifications seem to benefit from the abundance of violations. Specifically, the LR_{cc} and $DQ_{cc1} - DQ_{cc3}$ tests' size is close to the notional size for all the out-of-sample specifications. The 10% coverage level allows for an adequate amount of violations that, even for the case of risk underestimation will provide a testable information set. On the other hand, the Dur_{cc} results imply that the independence test distortions persist. This is expected since both Dur_{ind} and Dur_{cc} test statistics share the same estimation process and assumptions. Similarly, the elaborate DB specifications fail to perform adequately on the rich violation samples of the 10% coverage level. This is possibly attributed to the method's implied relationship which may be too restrictive for the abundance of data at hand. Finally, the density based C_{VaR1} method seems to produce small size distortions which are eliminated as the information included in the test expands.[13]

Turning to the ES evaluation methods, the unconditional coverage results suggest that the res method is significantly oversized while the U_{ES} has minor distortions. This is expected since the res test statistic is dependent on the magnitude of the difference between the ES and the returns during a violation of VaR. Given that the risk forecasts model specification is identical with the DGP's specification, random outliers will lead the violation sequence which in turn will render the res test statistic significantly different from zero. Contrary to the res method, the U_{ES} statistic is related to the density of the model and therefore it is more robust to possible extreme losses. Finally, as expected, the results of the $C_{ES1} - C_{ES10}$ tests are similar to the $C_{VaR1} - C_{VaR10}$ ones since both tests rely on the density of the evaluated model.

Turning to the 5% coverage level (Figure 3 Part A), the results suggest that the size properties of the tests change slightly in comparison to the 10% coverage case. Specifically, the simple specifications' size is either equal or very close to the notional size. In addition, the distortions of the LR_{muc}, although large, are significantly smaller than the 10%

[13]The Q_2 methods results are not valid for the 10% case since the secondary threshold is also set equal to 10% and remains constant for the remaining simulation exercise.

coverage level with the one-year out-of-sample test being undersized. This is in support of the excessive amount of super exemptions produced by the low coverage VaR. Turning to the independence tests ($LR_{ind} - MCS_{ind}$), the LR_{ind} specification is undersized for the smaller out-of-sample periods while for the larger periods the test is oversized. On the other hand, the Dur_{ind} test seems to perform better than the 10% coverage level case, although still oversized. Interestingly, the DQ_{ind} approach seems to be slightly undersized while for the 10% coverage level the tests are always slightly oversized.

For the conditional coverage tests ($LR_{cc} - C_{VaR10}$) the results suggest that there are increased size distortions for the majority of the tests, while there are less size distortions for the more elaborate DB specifications. In addition, it must be noted that the size distortions are negatively correlated to the out-of-sample length. Finally, the Q_2 method is significantly oversized. This is due to augmented volatility periods that cause violations at the 10% VaR level which, given the simulation set up, will probably cause violations at the higher 5% coverage level. Hence the likelihood of detecting a linear dependency is quite large. Similarly to the VaR backtesting methods results, the ES methods produce similar to the 10% size results with slightly more pronounced distortions for the U_{ES} and C_{ES5} methods. Finally, the res method's size remains extremely distorted.

[Figure 3 Part A around here]

Moving towards the deeper parts of the returns distribution increases the size distortions of the tests. For the 2.5% case (Figure 3 Part B), the LR_{uc} and $DQ_{UC1} - DQ_{UC3}$ tests' size seems to be more distorted in comparison to the previous cases. These distortions are more pronounced for the one-year out-of-sample period than the rest pot-of-sample periods. On the other hand, the LR_{muc} test's size performs better than the previous 10% and 5% coverage case since for the deeper parts of the tail the number of super exemptions is smaller. Interestingly, the J and MCS test maintain their size at the notional size level for all out-of-sample periods. This result is consistent with the 10% and 5% results where their size is again equal or very close to the nominal size. For

the independence tests, the distortions of the LR_{ind} and $DQ_{ind1} - DQ_{ind3}$ test are larger when compared with the 5% and 10% cases. Specifically, the LR_{ind} test is undersized for all out-of-sample periods indicating that the first order Markov property may not be suitable for higher coverage levels and the scarce violations produced by the accurate forecasts. Similarly, the sightly increased distortions of the $DQ_{ind1} - DQ_{ind3}$ tests suggest that the methods linear assumption may not be appropriate when the focus is turned to more extreme and not consecutive losses since the distortions are higher for the augmented DQ specifications. On the other hand, the Dur_{ind} test is still oversized but within the same range of the aforementioned 5% coverage level. Finally, for the J and MCS methods the results suggest that their size is at the notional size level.

The results of the conditional coverage tests have a similar pattern to the 5% results. However, the distortions are increasing for the simpler specifications such as the LR_{cc} and DQ_{cc} test while they decrease for the more elaborate ones such as the DB and Q methods. Specifically, due to the independence tests reduced rejection rates the LR_{cc} test is undersized. On the other hand, the DQ and C_{VaR} tests seem oversized with the distortions increasing when the linear dependency of violations is expanded to the second and third lag. Contrary to the aforementioned methods, the Dur_{cc}, DB and Q_2 methodologies seem to produce less distortions when compared to the 5% and 10% coverage levels. In addition, the results for the $DB4 - DB7$ specifications suggest that increasing the information set may not be beneficial for the test's performance since the size results are more sensitive to the out-of-sample length. Finally, the results of the ES evaluation methods suggest that the decreased amount of information have a direct impact on the U and C methods' size properties as the size distortions increase in comparison to the 5% coverage level. Contrary, the res method size results suggest that although significantly oversized, the smaller sample of violations reduces slightly the false rejection rates.

For the 1% coverage level (Figure 3 Part B) the size distortions are more pronounced and more sensitive to the out-of-sample period. The unconditional coverage tests that

focus directly on the violations (LR_{uc}, $DQ_{UC1} - DQ_{UC3}$, LR_{muc} and MCS) have a more pronounced profile in comparison to the previous coverage levels. Specifically, the LR_{uc} is significantly undersized for the one-year out-of-sample specification while the three- and four-year specifications are closer to the notional size. Similar results are reported for the $DQ_{UC1} - DQ_{UC3}$ tests with the corresponding size being within a 1.5% interval from the notional size. In the same vein, the LR_{muc} and MCS tests are significantly undersized for the one-year out-of-sample case while the J method's distortions are insignificant. The independence tests results suggest that the distortions for the LR_{ind} and DQ_{ind1} specifications are larger than the previous cases. However, the DQ_{ind2} and DQ_{ind3} specifications seem to keep their size within a 1.5% interval over and under the 5% notional size. Finally, the J and MCS_{ind} test keep their size almost equal to the notional size.

The results for the conditional coverage tests suggest that the LR_{cc}, $DQ_{cc1} - DQ_{cc3}$ tests have significant distortions which seem to depend on the specification at hand and the out-of-sample period. On the other hand, the $DB7$ specification seems to produce small distortions within a 1% interval around the notional size while the Q_2 tests remain significantly oversized. With respect to the ES tests, the res method remains significantly oversized while the density related specifications are dependent on the size of the out-of-sample period.

[Figure 3 Part B around here]

To sum up, the size properties of the tests reveal that simple tests such as the LR method perform adequately only when there is an adequate amount of data available to test. On the other hand, methods with elaborate specifications such as the DB, or methods that approximate the required testable property such as the Dur may be too restrictive for the cases with large number of violations. Furthermore, the conditional coverage tests reveal larger distortions than the tests that focus on the single hypothesis which have a more robust profile. Out of the full set of methods evaluated, only the J and MSC methods produce rejection rates equal to the notional size for each of the considered

coverage levels and out-of-sample periods while the DQ method revealed extended size distortions at the 1% coverage level and the conditional coverage test. In the same spirit, the ES tests revealed increasing distortions as the coverage level increased.

3.3 Power Properties

Figure 4 reports the power of the tests for the 5% notional size. Since the practitioners are facing the raw test results and not the size corrected, the power properties are calculated as the raw rejection rates of the misspecified HS risk forecasts. For the 10% coverage level, the power of the evaluation methods varies with the specification of the test and the out-of-sample length. For the unconditional coverage tests, the most consistent performing methods are the LR_{uc}, J_{uc} and MCS_{uc} ones, with the largest power achieved by the J_{uc3} specification for each out-of-sample period (<69%). For the J_{uc2}, J_{uc3} and LR_{muc} methods, the power of the tests is positively correlated with the length of the out-of-sample period while for the rest of the methods the differences between the out-of-sample periods are minimal. Contrary, the results of the DQ specifications suggest that expanding the information set will decrease the power of the test. The latter can be attributed to the nature of HS's non responsive violation sample.[14]

For the independence evaluation methods, the J_{ind} and MCS_{ind} specifications produce the larger rejection rates with the latter being marginal superior to the J_{ind3} method's ones (<67%). Contrary to the unconditional case, the DQ_{ind} specifications' power is an increasing function of the information set tested with, however, an inferior power profile than the top performing methods. With respect to the Dur_{ind} method, the results suggest that it produces the smallest rejection rate than all the alternative testing specifications. For the conditional coverage tests, the LR_{cc} and DQ_{cc} specifications share the same power properties for the one-year out-of-sample length while for the rest out-of-sample lengths the power is increasing with the information set and the length of the out-of-sample

[14]Consecutive violations during persistent augmented volatility and isolated violations caused by non persistent moderate losses could possibly lead to misestimation of the β of the DQ test linear specification.

periods. Interestingly, the Dur_{cc} test's power is higher than the independence test case but still inferior to the aforementioned tests' power. For the DB tests, the $DB2 - DB4$ specifications have a higher power profile than their linear counterparts $DQ2 - DQ4$ while the $DB5 - DB7$ specifications have a significantly lower power especially for the one-year out-of-sample period. The C_{VaR} approach seems to produce the largest power results ($C_{VaR3} < 0.78$).[15] Finally, for the ES case the U_{ES} results are expected since the HS cannot approximate correctly the tail parts of the DGP's density. On the other hand, the res methods power is robust to the out-of-sample period length and for each case larger than 24%.

For the 5% coverage level, the unconditional coverage results are quite similar since minor differences are reported when compared to the previous coverage level. For the unconditional coverage case, the J_{uc} seems to improve marginally its power for each out-of-sample period (J_{uc3}<73%). Similarly, slightly increased power is reported for the remaining specifications especially for the larger out-of-sample periods. This is expected since the 10% coverage level makes it easier for HS, possibly through overestimation of risk, to produce the "expected" number of violations. On the other hand, the larger out-of-sample specifications and stricter coverage levels require a more responsive nature, which HS can not provide.

The rejection rates for the independence tests follow the same pattern with the J_{ind3} and MCS_{ind} specifications providing the most powerful tests for each out-of-sample period(<67%). The J_{ind3} test seems to have a marginal advantage for the larger out-of-sample periods while the MCS_{ind} seems to perform better for the one-year out-of-sample period. Interestingly, the Dur method seems to increase significantly its power especially for the large out-of-sample periods. As discussed earlier, the 5% coverage level requires a more responsive risk estimation method. Hence, the Dur_{ind} can pick-up easier cluster of violations. The results for the conditional coverage paint the same picture with minor differences from the 10% coverage level. The largest power is reported by the Q method

[15]For the HS density estimation we use a normal kernel estimator.

($Q_2(3) < 0.99$) followed by the larger C_{VaR} specifications($C_{VaR10} < 0.74$). Finally, for the unconditional coverage of the ES forecasts, the U method delivers the highest power while the res methods power is significantly lower but again robust to the out-of-sample periods length.

[Figure 4 Part A around here]

The results for the 2.5% and 1% coverage levels (Figure 4 Part B) do not suggest a change of the relationship between the methods' power and the out-of-sample period. However, the more extreme coverage levels reduce the level of the rejection rates especially for the one-year out-of-sample period. For the 2.5% unconditional coverage level, the LR_{uc} rejections rates are almost similar to the previous coverage level. The same outcome holds for the DQ_{uc} approach but only for the larger out-of-sample specifications. For the two-year out-of-sample period the results suggest a small decrease in the power of the test while for the one-year out-of-sample period the power of the test is significantly diminished especially for the larger information sets. This is an indication of the reduced amount of violations impact on the power of the test. For the J method the rejection rates suggest that it is again amongst the most powerful ones ($J_{uc3} < 0.63$). Finally the LR_{muc} reveals a less powerful profile for the larger out-of-specifications while the MSC_{uc} has diminished rejection rates for the one-year out-of-sample specifications.

The results for the independence tests suggest a diminished power profile for the J and MCS_{ind} methods and all but the one-year out-of-sample case of the LR_{ind}, Dur_{ind2} and Dur_{ind3} specifications. Interestingly, the results of the DQ_{ind1} and Dur_{ind} tests suggest that there is a slight increase of their power. This is indicative of the restrictive nature of the methods specification since they increase their ability to reject a misspecified method when a more sparsely populated violation sequence is available. Finally, the conditional coverage tests have equal or reduced power for most of the methods. However, the Dur_{cc} results suggest a diminished power for the one-year out-of-sample case and increased power for the rest of the sample lengths. Similar results are reported for the ES testing

process where the power of the test is reduced significantly for every specification under consideration.

For the 1% coverage level, all the methods reveal a diminished power. For instance, the one-year out-of-sample results suggest that the rejection rates fluctuate between the 5% and 20% if we don't account for the Q method. Overall, for the unconditional coverage tests J_{uc3} specification classifies as the most powerful while the LR_{uc} shares the smallest power alongside the LR_{muc}. For the independence case, the results also reveal a diminished power with the DQ, Dur and J methods having the most powerful profile for all the out-of-sample lengths. Similar decreases are observed for the conditional coverage case where again the DQ method provides the most powerful specification (DQ_{cc3}) for all but the largest out-of-sample length. For the four-year out-of-sample length the results suggest that the C_{VaR10} is slightly more powerful. Finally, for the ES case the results suggest a diminished power. Interestingly, the small out-of-sample period provides the largest power for the res method. This is due to the fact that for the small out-of-sample period the chances are for more extreme violations if there are any. Therefore, the corresponding statistic will have increased probability of being non-zero on average.

[Figure 4 Part B around here]

To sum up, the power results suggest that the methods' reliability to detect a mis-specified series of forecasts is positively correlated to the amount of data inserted into the evaluation process. Therefore, for larger out-of-sample periods we expect the power of the test to be larger especially for the higher coverage levels. The significance level affects the amount of data included in the evaluation process given the specification of the risk model. As with the size case, the DQ, J and MCS methods reveal a superior and more robust profile across the different specifications.[16] There are no major differences between the conditional coverage tests and their individual hypotheses testing counter-

[16]The superiority of these methods is robust to the risk forecasts specification. The power results against the $GARCH-Normal$ forecasts reveal a qualitatively similar result as the DQ, J and MCS are superior to the remaining methods. Please refer to the on-line appendix for a more detailed analysis of the power properties against the $GARCH-Normal$ forecasts.

parts, although the conditional coverage tests seem to be slightly more powerful. Finally, the ES tests provide similar results with the exception of the U_{ES} test which, as expected, rejects constantly the non responsive HS forecasts.

4 Application To Financial Data

In this section we examine the performance of the backtesting methods on real financial data. Specifically, we use the the $HS-250$, $GARCH-N$, $GARCH-T$ and $RiskMetrics$ methods and a rolling estimation sample of $R \in \{1000, 750, 500, 250\}$ in order to calculate the series of $5\%, 2.5\%$ and 1% VaR and ES forecasts. We exclude the 10% case since it is of rather small empirical importance. The estimation sample length is 1000 for the parametric models and 250 for the HS. We estimate 1000 out-of sample forecasts corresponding to the $S\&P500$ returns for the 07/01/2011-31/12/2014 period. For the evaluation of the methods, we use the full 1000 observation period and three sub periods of three-, two- and one-year length. All the evaluation periods share a common sample of the last 250 observations.

Figure 5 reports the backtesting results for all methods. For the HS 1000 out-of-sample period (Figure 5 Part A), most of the unconditional coverage tests do not reject the null with the exception of the J method's larger specifications. The non responsive nature of HS leads to more rejections for the independence tests where with the exception of the first lag specifications (LR_{ind},DQ_{ind1},J_{ind1}) the rest are rejecting the null hypothesis of independence. Similar results are reported for the conditional coverage case where the majority of the methods reject the HS forecasts for every coverage level. Regarding the ES tests, the results are mixed since the res test does not reject the ES forecasts series while the remaining ones do.

The results for the three-year out-of-sample period are in line with the full sample results since there is still strong evidence against the suitability of HS's forecasts. In addition to the previous case, for the three-year out-of-sample period the unconditional

tests also reject the null hypothesis since the produced violations are significantly lower than the expected. However, the lower number of violations have some implications for the independence tests where, for the extreme coverage levels, the independence tests do not always reject the null hypothesis (i.e. DQ, J_{ind}). Interestingly, the extreme coverage cases are not rejected by the conditional coverage tests. The results for the two- and one-year out-of-sample periods suggest that even fewer tests reject the misspecified HS.

Figure 5 Part B reports the results for the Riskmetrics forecasts. The tests reveal a robust profile against the unconditional coverage of both VaR and ES across the out-of-sample specifications and the high coverage levels. Interestingly, the independence test and the methods focusing on the density do not seem to reject the method. That is expected since the explosive nature of the IGARCH dynamics can cope with the fat-tailed distribution of the actual returns. Similar results are reported for the GARCH-N forecasts (Figure 5 Part C) although the rejection cases for the smaller out-of-sample specifications are less than the Riskmetrics ones. Interestingly, for the smaller out-of-sample periods and higher coverage levels the conditional coverage methods rarely reject the forecasts while the J_{uc} and DQ_{uc} do.

Figure 5 Part D reports the backtesting results for the GARCH-t forecasts which according to the violation profile seems to fit better the dynamics of the $S\&P500$ returns. For the large out-of-sample specifications only the density focused tests seem to reject consistently the respective forecasts series while the rest of the methods provide small evidence against the forecasts. As with the previous cases, the smaller the out-of-sample period the fewer the rejection cases. Interestingly, for the one-year out-of-sample period the only methods rejecting the respective forecasts are the Dur method the Q methods.

[Figure 5 around here]

To sum up, our empirical findings are indicative of the inefficiencies of the VaR/ES evaluation methods. This is particularly true for the one-year out-of-sample period where almost none of the methods reject the misspecified ones especially for the high coverage

levels. Furthermore, using only the individual hypothesis testing seems more reliable than using directly the conditional coverage test as there are cases where the unconditional or independence hypothesis are rejected while the conditional hypothesis does not reject the misspecified methods.

5 Conclusions

This paper reviews several risk forecast backtesting methods and their performance in detecting misspecified models. The regulatory directives require the validation of the selected risk model for both internal and external reporting reasons. In order to assess the performance of the forecast evaluation methods, we create a simulation exercise where returns are generated according to a specific DGP and the corresponding VaR and ES forecasts are evaluated under various out-of-sample and coverage level specifications.

The simulation results provide three major findings. First, for the higher coverage levels and smaller out-of-sample periods the backtesting methodologies are physically restrained and detect mainly the underestimation of risk. Second, the size findings suggest that the individual hypothesis tests have a more robust profile than the conditional coverage ones. The latter produce significant distortions for almost each case under consideration. Third the power results suggest that the one-year out-of-sample period reduces the power of the tests especially for the higher coverage level. On the other hand, the difference in power between the lower 10% and intermediate 5% coverage levels is quite small.

To complement our simulation results we implement a financial data application where under the same out-of-sample specification a subset of the simulation coverage levels are utilized. The results confirm the simulation findings since the tests fail to reject the misspecified methods for the high coverage levels and small out-of-sample periods. The combination of the simulation and financial data application findings suggest that implementing a couple of individual hypothesis testing specifications for intermediate

coverage levels and two-year out-of-sample periods provides the best trade-off between the low power of the high coverage and small out-of-sample specification and the unimportant low coverage level and large out-of-sample one. Furthermore, implementing the whole set of size accurate evaluation methods and setting the zero rejections as an accuracy criterion can lead to a robust evaluation strategy. Alternatively, all evaluation methods can be utilized with the accuracy threshold set at a specific and small number of total rejections.

References

Acerbi, C. and Szekely, B. (2014). Backtesting expected shortfall. *Risk*, pages 42–47.

Acerbi, C. and Szekely, B. (2017). General properties of backtestable statistics. *Working Paper*.

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

Berkowitz, J., Christoffersen, P., and Pelletier, D. (2011). Evaluating value at risk models with desk level data. *Management Science*, (57):2213–2272.

Campbell, S. D. (2007). A review of backtesting and backtesting procedures. *Journal of Risk*, (9):1–17.

Candelon, B., Colletaz, G., Hurlin, C., and Tokpavi, S. (2011). Backtesting value-at-risk: A GMM duration-based test. *Journal of Financial Econometrics*, (9):314–343.

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review*, (39):841–862.

Christoffersen, P. (2011). *Elements of Financial Risk Management*. Burlington Academic Press.

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

Colletaz, G., Hurlin, C., and Pérignon, C. (2013). The risk map: A new tool for backtesting value-at-risk models. *Journal of Banking and Finance*, (37):3843–3854.

Diebold, F. X., A.Gunther, T., and Tay, A. S. (1998). Evaluating density forecasting. *International Economic Review*, (39):863–883.

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

Dumitrescu, E.-I., Hurlin, C., and Pham, V. (2012). Backtesting value-at-risk: From dynamic quantile to dynamic binary tests. *Finance*, (33):79–111.

Embrechts, P., Puccetti, G., Ruschendorf, L., Wang, R., and Beleraj, A. (2014). An academic response to basel 3.5. *Risks*, (2):25–48.

Emmer, S., Kratz, M., and Tasche, D. (2015). What is the best risk measure in practice? a comparison of standard measures. *Journal of Risk*, (18):31–60.

Engle, R. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics*, (22):367–381.

Escanciano, J. C. and Olmo, J. (2010). Backtesting parametric value-at-risk with estimation risk. *Journal of Business and Economic Statistics*, (28):36–51.

Escanciano, J. C. and Olmo, J. (2011). Robust backtesting for value at risk. *Journal of Financial Econometrics*, (9):132–161.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and osbands principle. *The Annals of Statistics*, (44):1680–1707.

Gaglianone, W. P., Lima, L. R., Linton, O., and Smith, O. R. (2011). Evaluating value at risk modelis via quantile rgression. *Journal of Business and Economic Statistics*, (29):150–160.

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

Haas, M. (2005). Improved duration-based backtesting of value-at-risk. *Journal of Risk*, (2):13–36.

Hurlin, C. and Tokpavi, S. (2007). Backtesting value at risk accuracy: A simple new test. *The Journal of Risk*, (9):19–37.

Kerkhof, J. and Melenberg, B. (2004). Backtesting for risk based regulatory capital. *Journal of Banking and Finance*, (34):267–279.

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

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk management models. *Journal of Derivatives*, (9):73–84.

Leccadito, A., Boffelli, S., and Urga, G. (2014). Evaluating the accuracy of value-at-risk forecasts: New multilevel tests. *International Journal of Forecasting*, (30):206–216.

Loser, R., Wied, D., and Ziggel, D. (2018). New backtests for unconditional coverage of the expected shortfall. *Journal of Risk*, (21):1–21.

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

Nieto, M. R. and Ruiz, E. (2016). Frontiers in VaR forecasting and backtesting. *International Journal of Forecasting*, (32):475–501.

Pelletier, D. and Wei, W. (2016). The geometric-VaR backtesting method. *Journal of Financial Econometrics*, (14):725–745.

Perignon, C. and Smith, D. (2008). A new approach to comparing VaR estimation methods. *The Journal of Derivatives*, (16):54–66.

Righi, M. B. and Ceretta, P. S. (2014). A comparison of expected shortfall estimation models. *Journal of Economics and Business*, (78):14–47.

Santo, P. and Alves, M. (2012). A new class of independence tests for interval forecasts evaluation. *Computational Statistics and Data Analysis*, (56):3366–3380.

Ziegel, J. F. (2014). Coherence and elicitability. *Mathematical Finance*, (26):901–918.

Ziggel, D., Berens, T., Weiss, G. N. F., and Wied., D. (2014). A new set of improved value-at-risk backtests. *Journal of Banking and Finance*, (48):29–41.

Notes to Figures

Figure 1 reports the empirical distribution of the number of violations per simulated sample for the accurate $GARCH(1,1) - t_7$ (DGP) forecasts series. Each row represents an out-of-sample period and each column represents a coverage level.

Figure 2 reports the empirical distribution of the number of violations per simulated sample for the HS forecasts series. Each row represents an out-of-sample period and each column represents a coverage level.

Figure 3 reports the empirical size of the each evaluated method. The empirical size has been calculated as the rejection cases of the accurate $GARCH(1,1) - t_7$ (DGP) forecasts series over the number of simulated samples. Part A of figure 3 reports the rejection rates for the 10% and 5% coverage level and each out-of-sample period. Part B reports the rejection rates for the 2.5% and 1% coverage level and each out-of-sample period. When necessary, the maximum observed rejection rates are reported in the parenthesis.

Figure 4 reports the empirical power of the each evaluated method. The empirical power has been calculated as the rejection cases of the misspecified HS forecasts series over the number of simulated samples. Part A of figure 4 reports the rejection rates for the 10% and 5% coverage level and each out-of-sample period. Part B reports the rejection rates for the 2.5% and 1% coverage level and each out-of-sample period. When necessary, the maximum observed rejection rates are reported in the parenthesis.

Figure 5 reports the evaluation methods p-value for each forecasting method and out-of-sample period the p-values of the 1%, 2.5% and 5% coverage levels.

Table 1: Abbreviations Table

VaR Unconditional Coverage Tests

Abbreviation	In-Text Equation	Notes	Refference
LR_{uc}	Equation 4		Christoffersen (1998)
DQ_{uc1}	Equation 7	K=1, ,$\gamma_j = 0$	Engle and Manganelli (2004)
DQ_{uc2}	Equation 7	K=2, $\gamma_j = 0$	Engle and Manganelli (2004)
DQ_{uc3}	Equation 7	K=3, $\gamma_j = 0$	Engle and Manganelli (2004)
J_{uc1}	Equation 15	M_1	Candelon et al. (2011)
J_{uc2}	Equation 15	M_2	Candelon et al. (2011)
J_{uc3}	Equation 15	M_3	Candelon et al. (2011)
LR_{muc}	Equation 22		Colletaz et al. (2013)
MCS_{uc}	Equation 18		Ziggel et al. (2014)

VaR Independence Tests

Abbreviation	In-Text Equation	Notes	Refference
LR_{ind}	Equation 5		Christoffersen (1998)
DQ_{ind1}	Equation 7	K=1, ,$\gamma_j = 0$	Engle and Manganelli (2004)
DQ_{ind2}	Equation 7	K=2, $\gamma_j = 0$	Engle and Manganelli (2004)
DQ_{ind3}	Equation 7	K=3, $\gamma_j = 0$	Engle and Manganelli (2004)
Dur_{ind}	Equation 15		Christoffersen and Pelletier (2004)
J_{ind1}	Equation 16	M_1	Candelon et al. (2011)
J_{ind2}	Equation 16	M_2	Candelon et al. (2011)
J_{ind3}	Equation 16	M_3	Candelon et al. (2011)
MCS_{ind}	Equation 19		Ziggel et al. (2014)

VaR Conditional Coverage Tests

Abbreviation	In-Text Equation	Notes	Refference
LR_{cc}	Equation 6		Christoffersen (1998)
DQ_{cc1}	Equation 7	K=1, ,$\gamma_j = 0$	Engle and Manganelli (2004)
DQ_{cc2}	Equation 7	K=2, $\gamma_j = 0$	Engle and Manganelli (2004)
DQ_{cc3}	Equation 7	K=3, $\gamma_j = 0$	Engle and Manganelli (2004)
Dur_{cc}	Equation 13		Christoffersen and Pelletier (2004)
DB_1	Equation 9	Footnote 1, Specification 1	Dumitrescu et al. (2012)
DB_2	Equation 9	Footnote 1, Specification 2	Dumitrescu et al. (2012)
DB_3	Equation 9	Footnote 1, Specification 3	Dumitrescu et al. (2012)
DB_4	Equation 9	Footnote 1, Specification 4	Dumitrescu et al. (2012)
DB_5	Equation 9	Footnote 1, Specification 5	Dumitrescu et al. (2012)
DB_6	Equation 9	Footnote 1, Specification 6	Dumitrescu et al. (2012)
DB_7	Equation 9	Footnote 1, Specification 7	Dumitrescu et al. (2012)
$Q_2(1)$	Equation 20	K=1, 10% baseline coverage level	Hurlin and Tokpavi (2007)
$Q_2(2)$	Equation 20	K=2, 10% baseline coverage level	Hurlin and Tokpavi (2007)
$Q_2(3)$	Equation 20	K=3, 10% baseline coverage level	Hurlin and Tokpavi (2007)
C_{VaR1}	Equation 28	m=1	Du and Escanciano (2017)
C_{VaR5}	Equation 28	m=5	Du and Escanciano (2017)
C_{VaR10}	Equation 28	m=10	Du and Escanciano (2017)

ES Tests

Abbreviation	In-Text Equation	Notes	Refference
res	Equation 23		McNeil and Frey (2000)
U_{ES}	Equation 27		Du and Escanciano (2017)
C_{ES1}	Equation 28	m=1	Du and Escanciano (2017)
C_{ES5}	Equation 28	m=5	Du and Escanciano (2017)
C_{ES10}	Equation 28	m=10	Du and Escanciano (2017)

Note: Table 1 summarizes the abbreviations and specifications of each evaluated method reported in sections 3 and 4 and in the respective figures.

Figure 1: Empirical Distribution of the Number of Violations-DGP

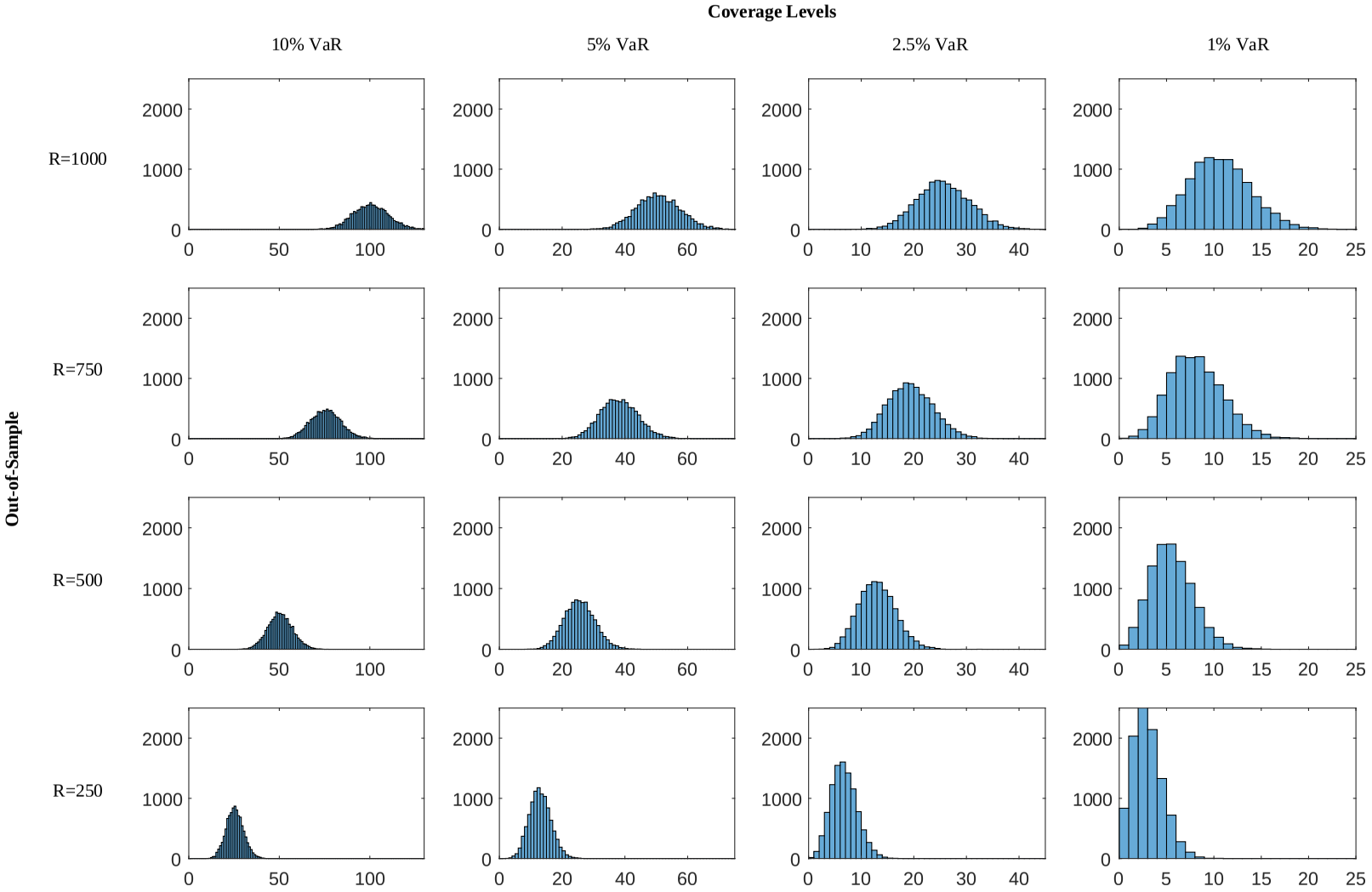


Figure 2: Empirical Distribution of the Number of Violations-HS

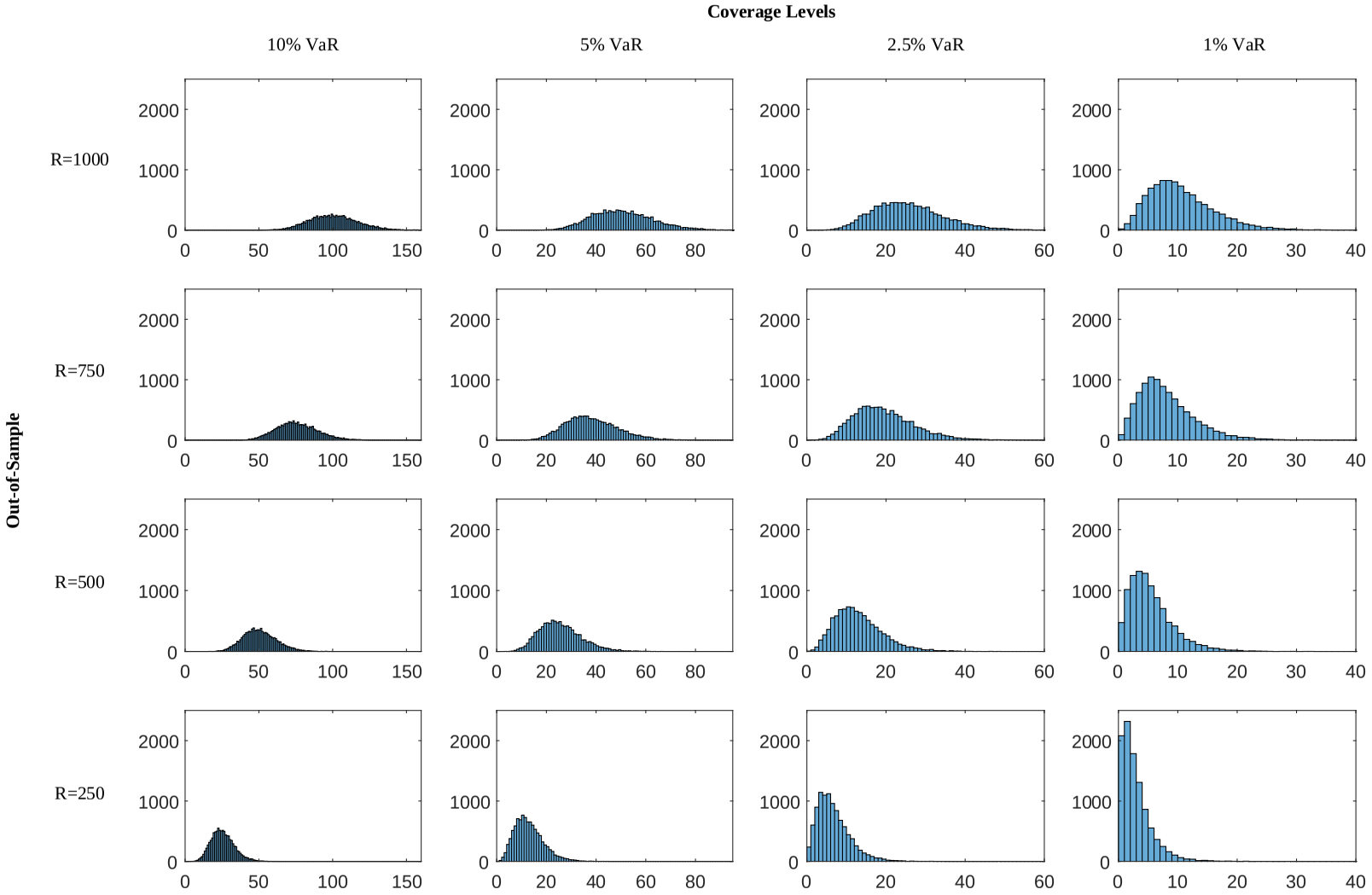


Figure 3: Size Properties - Part A

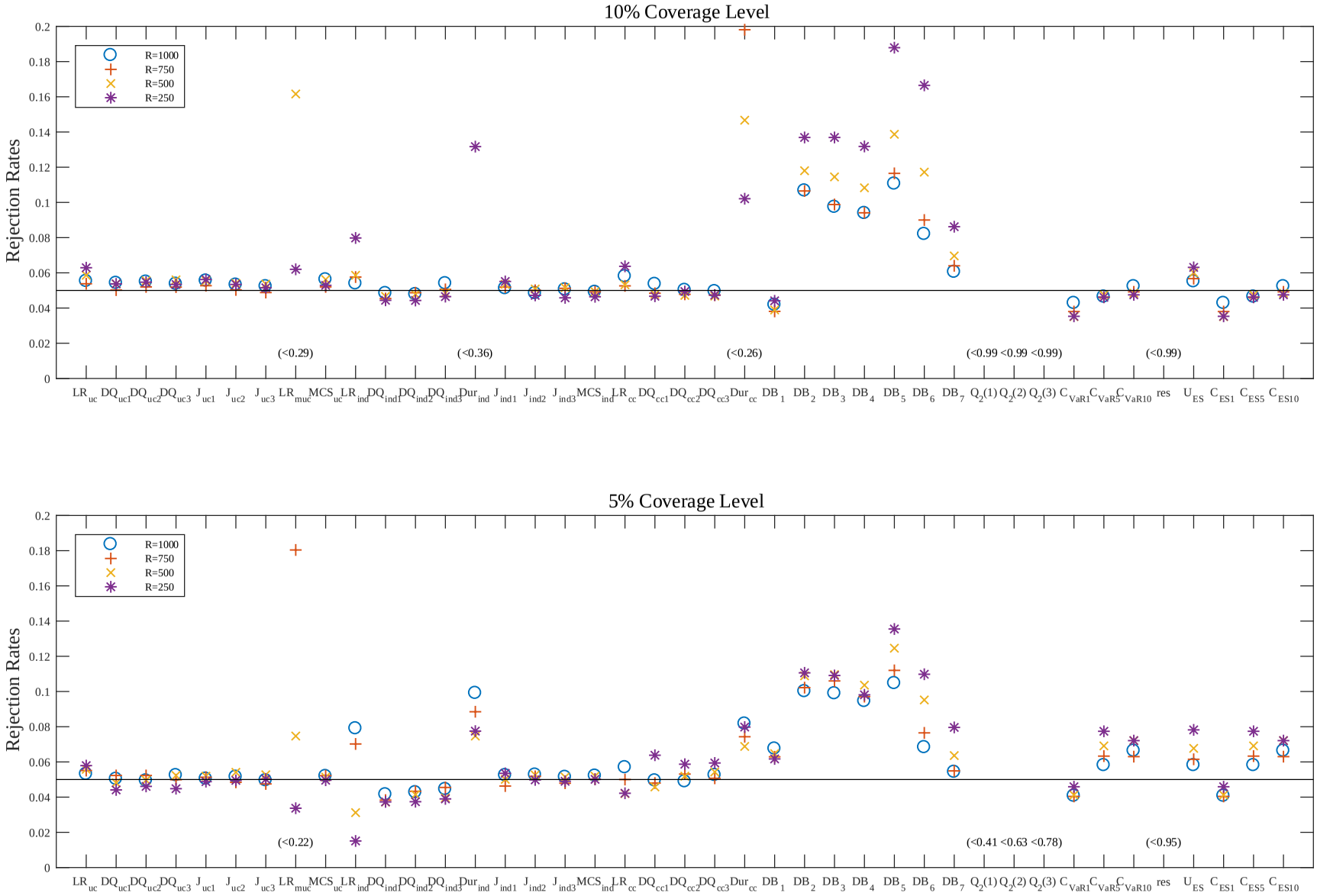
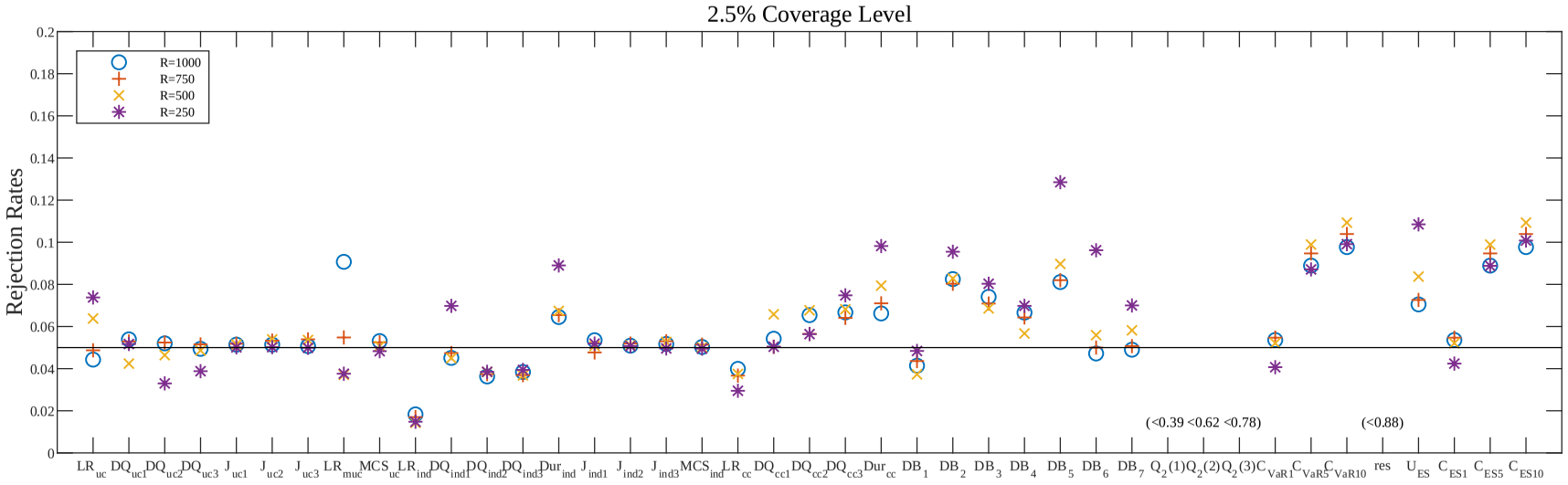


Figure 3 (Continued): Size Properties - Part B



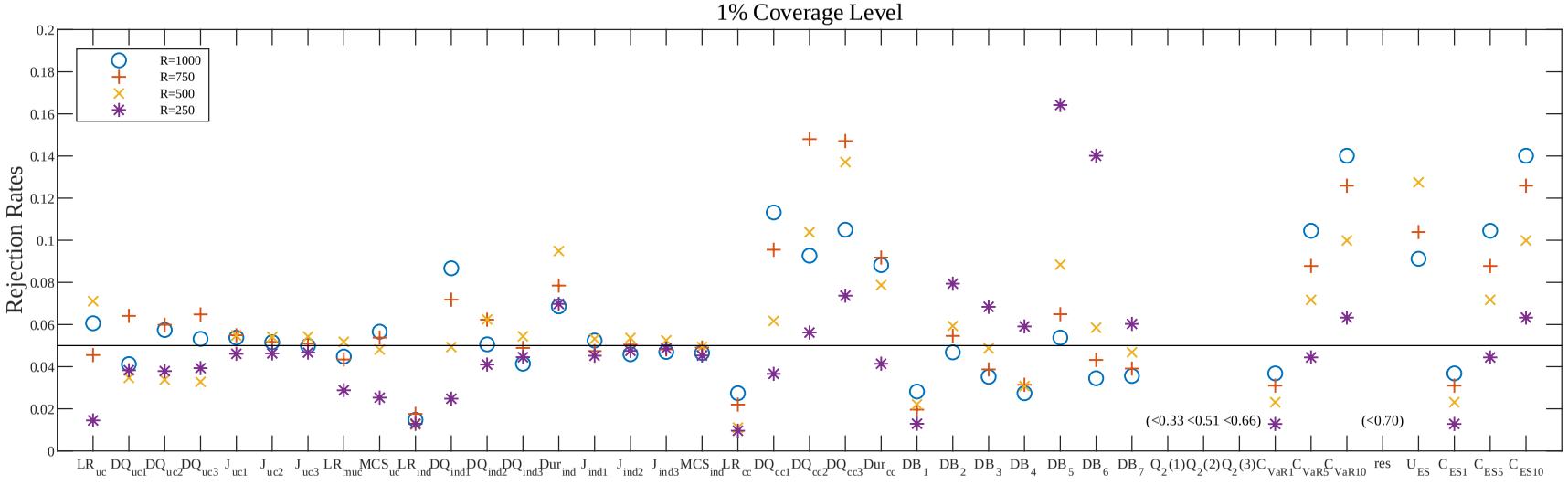


Figure 4: Power Properties - Part A

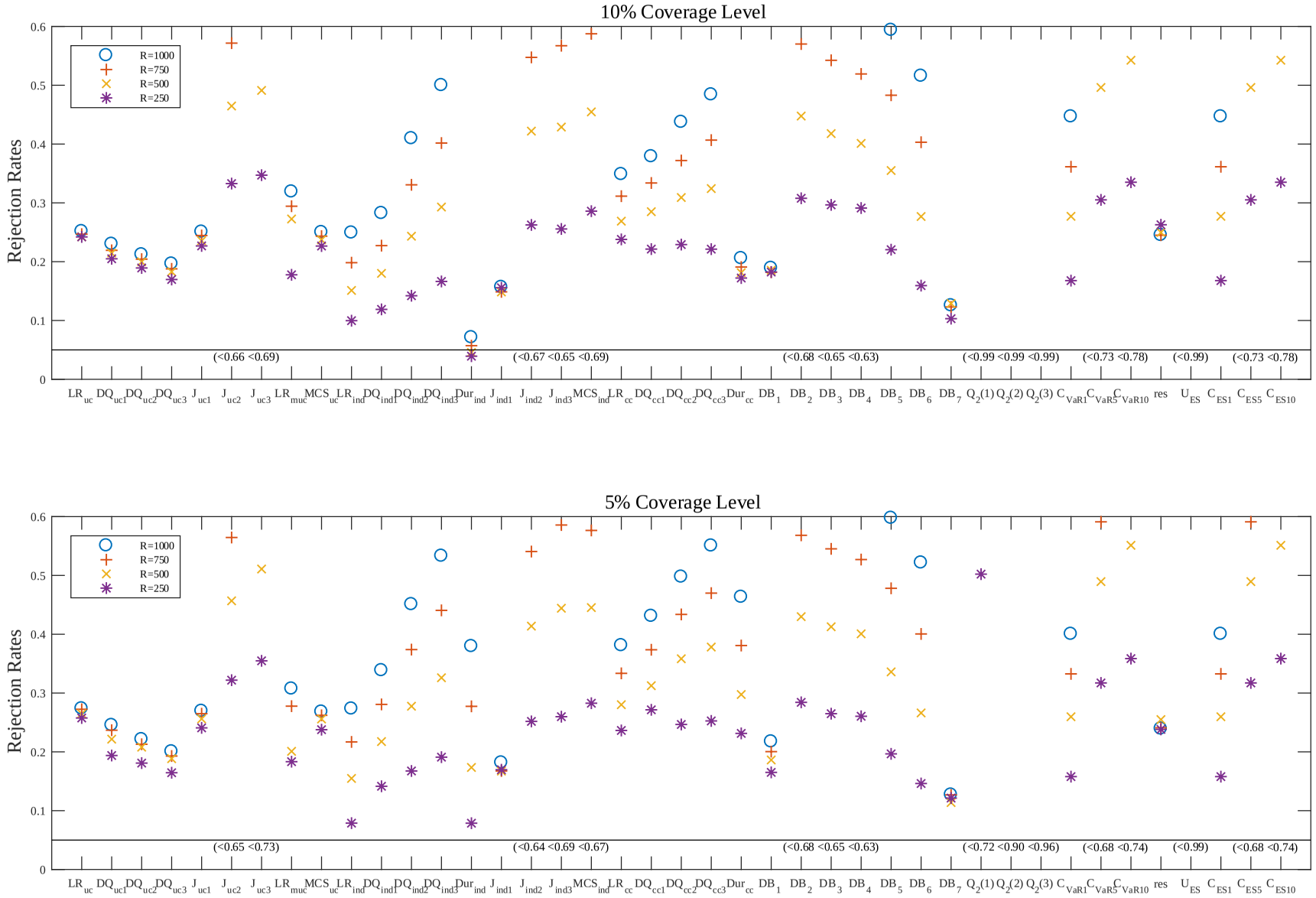
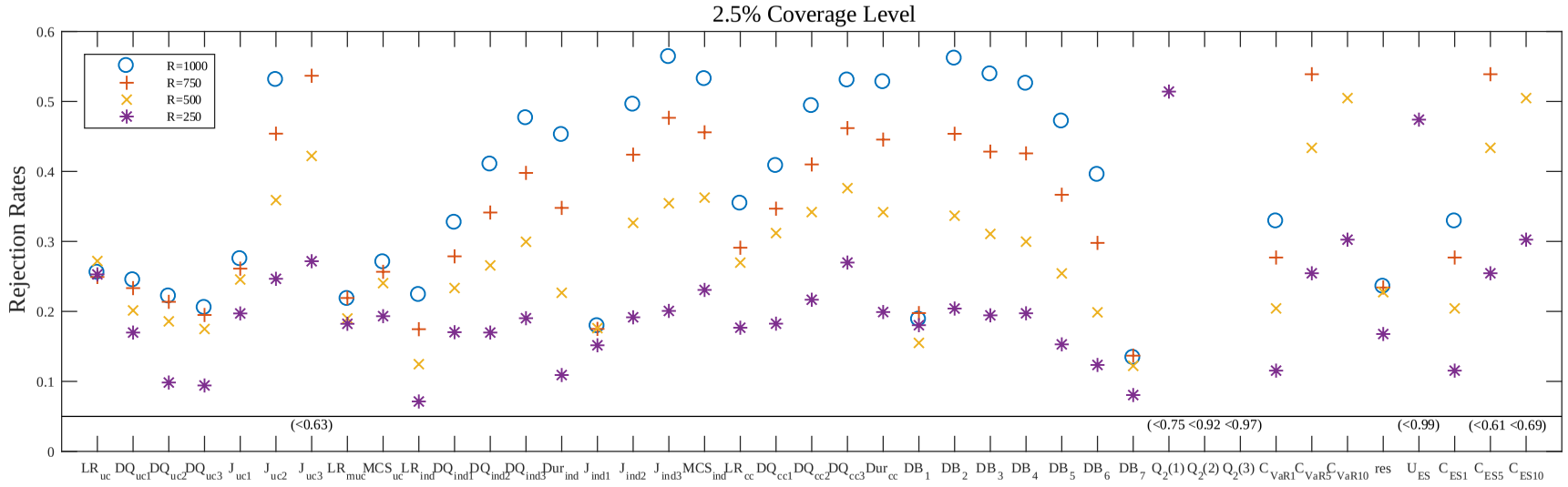


Figure 4 (Continued): Power Properties - Part B



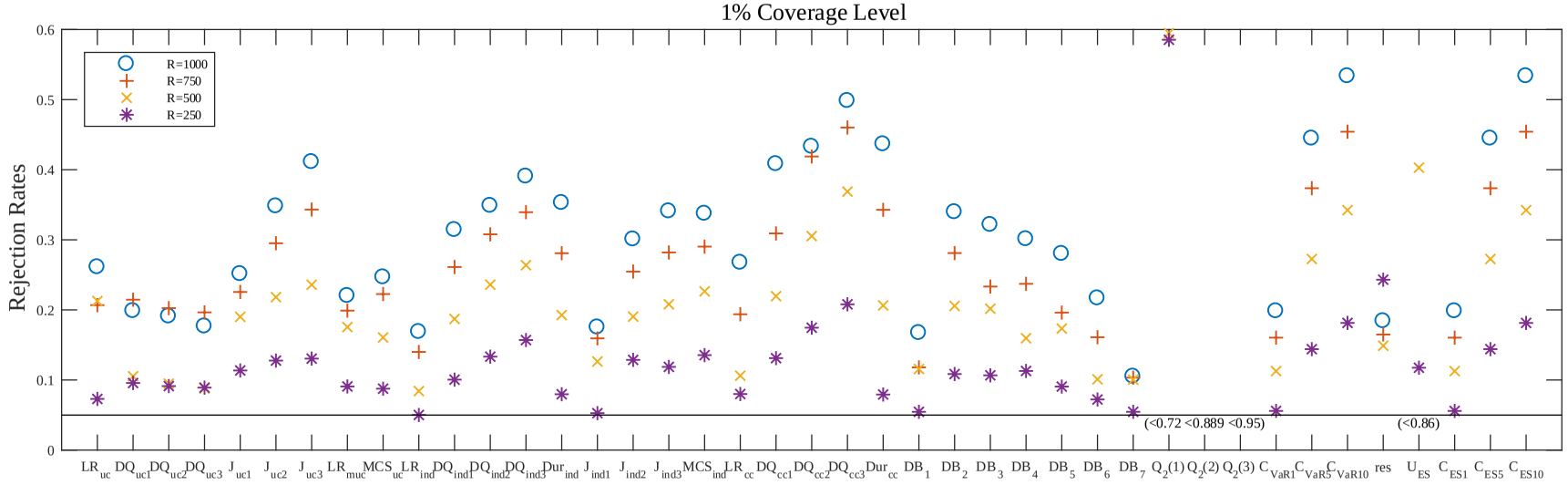


Figure 5: Empirical Implementation Part A - HS

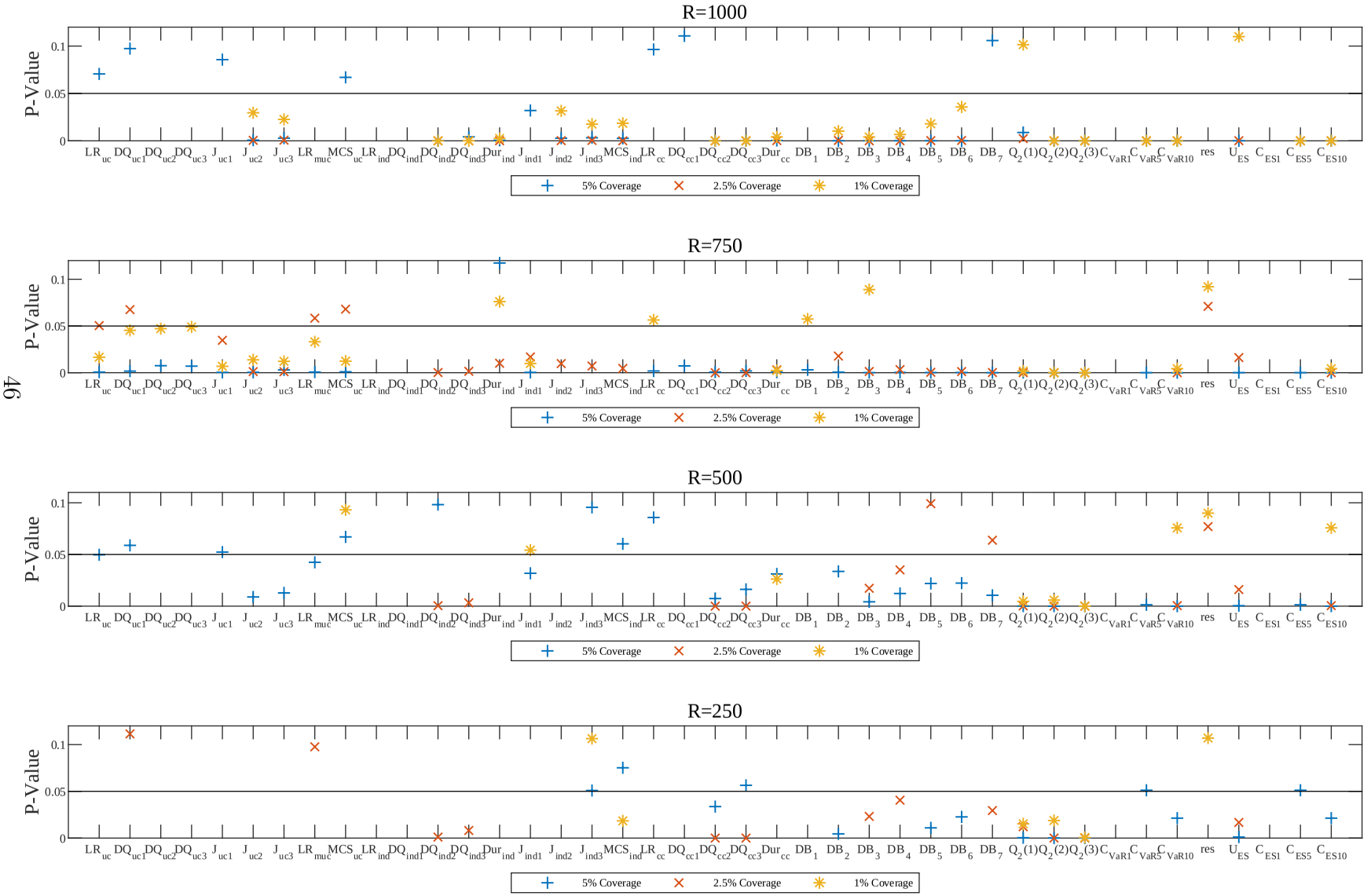


Figure 5 (Continued): Empirical Implementation Part B - Risk Metrics

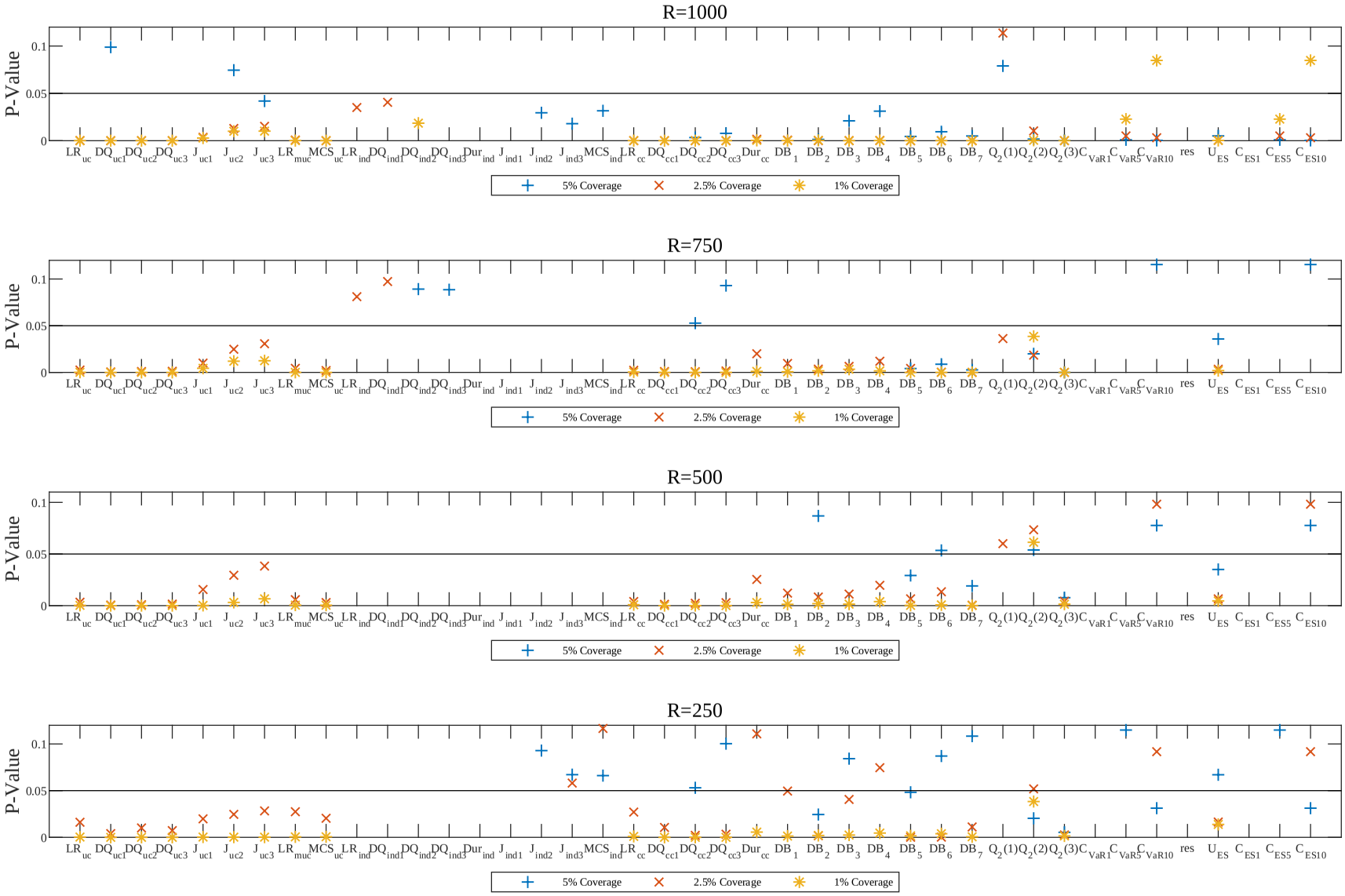


Figure 5 (Continued): Empirical Implementation Part C - GARCH-N

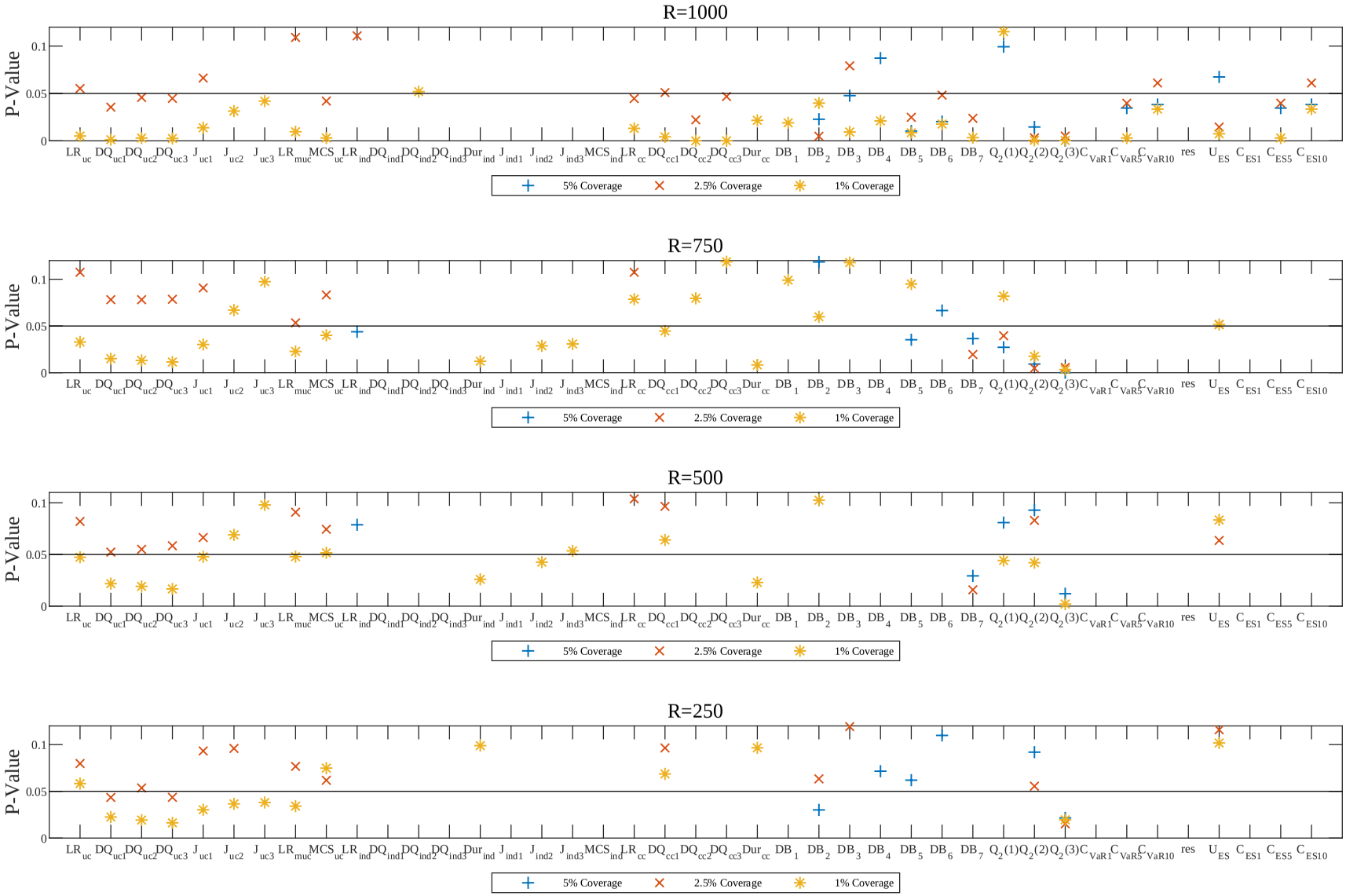


Figure 5 (Continued): Empirical Implementation Part D - GARCH-t