Regression Based Expected Shortfall Backtesting

Sebastian Bayer and Timo Dimitriadis[†]

University of Konstanz, Department of Economics, 78457 Konstanz, Germany

This Version: January 12, 2018

Abstract

In this article, we introduce a regression based backtest for the risk measure Expected Shortfall (ES) which is based on a joint regression framework for the quantile and the ES. We also introduce a second variant of this ES backtest which allows for testing one-sided hypotheses by only testing an intercept parameter. These two backtests are the first backtests in the literature which solely backtest the risk measure ES as they only require ES forecasts as input parameters. In contrast, the existing ES backtesting techniques require forecasts for further quantities such as the Value at Risk, the volatility or even the entire (tail) distribution. As the regulatory authorities only receive forecasts for the ES, backtests including further input parameters are not applicable in practice. We compare the empirical performance of our new backtests to existing approaches in terms of their empirical size and power through several different simulation studies. We find that our backtests clearly outperform the existing backtesting procedures in the literature in terms of their size and (size-adjusted) power properties throughout all considered simulation experiments. We provide an R package for these ES backtests which is easily applicable for practitioners.

Keywords: Expected Shortfall, Backtesting, Mincer-Zarnowitz Regression, Forecast Evaluation

1. Introduction

Through the transition from Value at Risk (VaR) to Expected Shortfall (ES) as the primary market risk measure in the Basel Accord (Basel Committee, 2016), there is a great demand for reliable methods for estimating, forecasting and backtesting the ES. Formally, the ES at level $\tau \in (0, 1)$ is defined as the mean of the returns smaller than the respective τ-quantile (the VaR), where τ is usually chosen to be 2.5% as stipulated by the Basel Accord. The ES is introduced into the banking regulation because it overcomes several shortcomings of the VaR, such as not being coherent and its inability to capture tail risks beyond the τ-quantile (Artzner et al., 1999; Danielsson et al., 2001; Basel Committee, 2013). In contrast to estimation and forecasting of ES where most of the existing models for the VaR can easily be adapted and generalized to the ES, such a generalization is unfortunately not as straight-forward for backtesting of ES forecasts (Emmer et al., 2015). In general, backtesting of a risk measure is the process of testing whether given forecasts for this risk measure are correctly specified, which is carried out by comparing the history of the issued risk forecasts with the corresponding realized returns. The primary reason for the difficulty to directly backtest ES is its non-elicitability and non-identifiability (Weber, 2006; Gneiting, 2011; Fissler and Ziegel, 2016; Fissler et al., 2016) as consequently, there is no analog to the hit sequence which is the natural identification function of quantiles and which lies at the heart of most VaR backtests.[1]

[†]Corresponding author

Email addresses: sebastian.bayer@uni-konstanz.de, timo.dimitriadis@uni-konstanz.de,

[1]See Yamai and Yoshiba (2002); Kerkhof and Melenberg (2004); Carver (2013); Acerbi and Szekely (2014); Emmer et al. (2015); Ziegel (2016); Fissler et al. (2016); Nolde and Ziegel (2017) for the ongoing discussion on backtestability of the ES.

As a consequence, most of the proposed procedures in the growing literature on backtesting ES use indirect approaches, e.g. based on forecasts for the entire tail distribution or by linear approximations of the ES with VaR forecasts at several probability levels. However, these approaches are in fact either joint backtests for a vector of risk measures such as the triple containing the VaR, the ES, and the volatility or even for the whole tail distribution (Nolde and Ziegel, 2017). As the proposed backtests require further input parameters such as forecasts for the volatility, the tail distribution beyond some quantile, or even the entire distribution, they are not applicable for the regulatory authorities because this additional information is not reported by the financial institutions (Aramonte et al., 2011; Basel Committee, 2016, 2017).

In this paper, we propose a novel backtest for ES forecasts which is based on a regression framework which models the conditional ES as a linear function, where we use financial returns as the response variable and ES forecasts as the explanatory variable including an intercept term. For correct ES forecasts, the intercept and slope parameters should be equal to 0 and 1, respectively. We use a Wald statistic to test for these parameter values, where we apply both, an asymptotic test using the covariance estimator introduced in Dimitriadis and Bayer (2017) and a bootstrap hypothesis test. We call this novel test the *bivariate ESR backtest*. This procedure is the first that backtests the risk measure ES stand-alone, i.e. the first that only uses ES forecasts as input parameters.[2] Through this feature, our new test is the first backtest for the ES which is practically applicable for the regulatory authorities who only have ES forecasts at hand.

Such regression-based forecast evaluation approaches are already used for testing mean forecasts (Mincer and Zarnowitz, 1969), quantile forecasts (Gaglianone et al., 2011; Guler et al., 2017), and expectile forecasts (Guler et al., 2017). In contrast to these functionals where regression techniques are easily available (see e.g. Koenker and Bassett, 1978, Efron, 1991), estimating regression parameters for an ES specific regression equation is more difficult as the ES is not elicitable (Gneiting, 2011). We overcome this difficulty by estimating the parameters of a joint regression procedure for the quantile and the ES, recently proposed by Dimitriadis and Bayer (2017), Patton et al. (2017) and Barendse (2017).

We also introduce a second regression-based ES backtest by fixing the slope parameter in the regression to one, and by only estimating and testing the intercept term, where we call this test the *intercept ESR backtest*. This second backtest allows for both, one-sided and two-sided hypotheses which contrasts with the first backtest that only allows for a two-sided hypothesis as it is generally unclear how underestimated and overestimated ES forecasts respectively influence the slope and intercept parameters. Because the capital requirements that the financial institutions must keep as a reserve depend on the reported risk forecasts, the market participants have an incentive to overestimate[3] the risk forecasts to minimize these expensive capital requirements. In contrast, underestimation of the forecasts results in too conservative risk forecasts and larger capital reserves, which does not have to be punished by the regulatory authorities. Thus, the regulators only have to prevent and penalize the overestimation of risk forecasts, which demonstrates the necessity of one-sided testing procedures. For example, the currently applied traffic light system (Basel Committee, 1996) is in fact a one-sided VaR backtest. As the bivariate ESR backtest, this intercept ESR test also has the desired characteristic to only require ES forecasts as input parameters and consequently is the first procedure in the literature that solely backtests the ES against a one-sided alternative.

We introduce several simulation setups to evaluate the empirical properties of our novel ES backtests and compare them to the existing joint VaR and ES backtests of McNeil and Frey (2000) and Nolde and Ziegel (2017). In the first setup, we implement the classical size and power analysis for backtesting risk measures, where we simulate data stemming from a realistic data generating process and evaluate the empirical rejection frequencies of the backtests for forecasts stemming from the true and from some misspecified forecasting models. In the second setup, we introduce a novel technique for evaluating

---

[2]The backtests which come closest to our procedure in this regard are the exceedance residual backtests of McNeil and Frey (2000) and the conditional coverage backtests of Nolde and Ziegel (2017) which are in fact joint backtests for the VaR and the ES.

[3]Throughout the paper, we use the sign convention that losses are denoted by negative numbers and *overestimation* of risk measures is meant in the mathematical sense, i.e. as reporting too large real numbers, which implies that the associated market risk is *underestimated*.

the power of backtests for financial risk measures, where we continuously misspecify certain model parameters of the data generating process to obtain a continuum of alternative models with a gradually increasing degree of misspecification. Misspecifying the different model parameters separately allows us to misspecify certain model characteristics (such as the reaction to shocks) in isolation, which permits a closer examination of the proposed backtesting procedures. To the best of our knowledge, this evaluation technique is new to the literature.

From these simulations, we find that the bivariate and the intercept ESR backtests that we propose in this paper are reasonable sized, especially when the tests are applied using the bootstrap. Moreover, they are more powerful than the existing backtests of McNeil and Frey (2000) and Nolde and Ziegel (2017) in almost all of the considered simulation designs for both, testing against one-sided and two-sided alternatives. Notably, throughout all simulation designs, the two ESR backtests are able to detect the various different misspecifications of the forecasts. In contrast, the existing backtests sometimes completely fail to detect certain misspecifications, for instance when the forecaster reports risk forecasts for a misspecified probability level.

The rest of this paper is organized as follows. Section 2 introduces the theory of our new backtests, and Section 3 reviews the existing ES backtesting techniques. Section 4 contains several simulation studies, and Section 5 applies the backtests to the risk forecasts of the S&P500 index. Section 6 concludes.

2. Theory

2.1. Setup and Notation

Let us consider a stochastic process

$$Z = \left\{ Z_t : \Omega \to \mathbb{R}^{k+1}, k \in \mathbb{N}, t = 1, \ldots, T \right\}, \tag{2.1}$$

defined on some complete probability space $\left(\Omega, \mathcal{F}, \mathbb{P} \right)$, with the filtration $\mathcal{F} = \left\{ \mathcal{F}_t, t = 1, \ldots T \right\}$ and $\mathcal{F}_t = \sigma \{ Z_s, s \leq t \}$ for all $t = 1, \ldots, T$. We partition the stochastic process $Z_t = (Y_t, X_t)$, where Y_t is an absolutely continuous random variable of interest and X_t is a k-dimensional vector of explanatory variables. We denote the conditional cumulative distribution function of Y_t given the past information $\mathcal{F}_{t-1}$ by $F_t(y) = \mathbb{P}(Y_t \leq y \mid \mathcal{F}_{t-1})$ and the corresponding probability density function by f_t. Whenever they exist, the mean and the variance of F_t are denoted by $\mathbb{E}_t[\cdot]$ and $\mathrm{Var}_t(\cdot)$.

For financial applications, the variable Y_t denotes the daily log returns of a financial asset (for instance, a stock or a portfolio), i.e. $Y_t = \log P_t - \log P_{t-1}$, where P_t denotes the price of the asset at day $t = 1, \ldots, T$. This means that throughout this paper, we use the sign convention that positive returns denote profits, and negative returns denote losses. The vector X_t contains further variables that are used to produce forecasts for certain functionals (usually risk measures) of the random variable Y_t.

We are interested in testing whether forecasts for a certain d-dimensional, $d \in \mathbb{N}$ functional (risk measure) $\rho = \rho(F_t)$ of the conditional distribution F_t are correctly specified. For that, we define the most frequently used functionals for financial risk management in the following. The conditional quantile of Y_t given the information set $\mathcal{F}_{t-1}$ at level $\tau \in (0, 1)$ is defined as

$$Q_\tau \left(Y_t \mid \mathcal{F}_{t-1} \right) = F_t^{-1}(\tau) = \inf \left\{ y \in \mathbb{R} : F_t(y) \geq \tau \right\}, \tag{2.2}$$

which is called the VaR at level τ in financial applications. Furthermore, we define the functional ES at level τ of Y_t given $\mathcal{F}_{t-1}$ as

$$\mathrm{ES}_\tau \left(Y_t \mid \mathcal{F}_{t-1} \right) = \frac{1}{\tau} \int_0^\tau F_t^{-1}(s) \, \mathrm{d}s. \tag{2.3}$$

If the distribution function F_t is continuous at its τ-quantile, this definition can be simplified to the truncated tail mean of Y_t,

$$\mathrm{ES}_\tau \left(Y_t \mid \mathcal{F}_{t-1} \right) = \mathbb{E}_t \left[Y_t \mid Y_t \leq Q_\tau \left(Y_t \mid \mathcal{F}_{t-1} \right) \right]. \tag{2.4}$$

We denote an $\mathcal{F}_{t-1}$-measurable one-step-ahead forecast for day t for the risk measure ρ of the distribution F_t, stemming from some external forecaster or from some given forecasting model[4] by $\hat{\rho}_t = \hat{\rho}_t(\mathcal{F}_{t-1})$. Following this notation, we denote forecasts for the τ-VaR by $\hat{v}_t$ and for the τ-ES by $\hat{e}_t$ for some fixed level $\tau \in (0, 1)$. For simplicity of the notation, we drop the dependence on τ as it is a fixed quantity.

As both, the incentive of the forecaster and the underlying method used to generate the forecasts are in general unknown, these forecasts are not necessarily correctly specified. The focus of this paper is to develop statistical tests for correctness of a given series of forecasts $\{\hat{\rho}_t, t = 1, \ldots, T\}$ for the risk measure ρ relative to the realized return series $\{Y_t, t = 1, \ldots, T\}$. This is in the literature usually referred to as *backtesting* of the risk measure ρ without strictly defining this terminology. We provide such a definition in the following. The core message of this definition is that besides the realized return series, a backtest for some risk measure is only allowed to require forecasts for this risk measure as input parameters.

Definition 2.1. A *backtest* for the series of forecasts $\{\hat{\rho}_t, t = 1, \ldots, T\}$ for the d-dimensional risk measure (functional) ρ relative to the realized return series $\{Y_t, t = 1, \ldots, T\}$ is a function

$$f : \mathbb{R}^T \times \mathbb{R}^{T \times d} \to (0, 1), \tag{2.5}$$

which maps the return and forecast series onto the respective p-value of the test.

This strict differentiation becomes relevant in the context of backtesting ES as, in contrast to the existing VaR backtests, the recently proposed ES backtests require further input parameters such as forecasts for the VaR, the volatility, or the entire tail distribution. The demand for these further quantities induces the following practical problems. First, the regulatory authorities who rely on such backtesting methods do not necessarily receive forecasts from the financial institutions for the additional information required by these tests, which makes such backtests inapplicable for the regulatory authorities. Second, a rejection of the tests does not necessarily imply that the ES is misspecified, but that the forecasts for any of the input components are misspecified. Consequently, these tests are in fact not backtests for the ES, but rather backtests for vectors of risk measures, or the entire tail distribution.

The two novel regression based ES backtests we propose in the next section are the first backtests for the ES which follow Definition 2.1 as they only require forecasts for the ES. This makes these tests the first ES backtests in this sense.

2.2. The bivariate ESR Backtest

We propose a new backtest for the risk measure ES that tests whether a series of ES forecasts $\{\hat{e}_t, t = 1, \ldots T\}$, stemming from some external forecaster or forecasting model is correctly specified relative to a series of in due course realized returns $\{Y_t, t = 1, \ldots T\}$. For that, we regress the returns Y_t on the forecasts $\hat{e}_t$ including an intercept term by using a regression equation designed specifically for the functional ES,

$$Y_t = \alpha + \beta \hat{e}_t + u_t^e, \tag{2.6}$$

where $\mathrm{ES}_\tau(u_t^e \mid \mathcal{F}_{t-1}) = 0$. Given the structure in (2.6) and since $\hat{e}_t$ is generated by using the information set $\mathcal{F}_{t-1}$, this condition on the error term is equivalent to

$$\mathrm{ES}_\tau(Y_t \mid \mathcal{F}_{t-1}) = \alpha + \beta \hat{e}_t. \tag{2.7}$$

We then test the hypothesis

$$\mathbb{H}_0 : (\alpha, \beta) = (0, 1) \qquad \text{against} \qquad \mathbb{H}_1 : (\alpha, \beta) \neq (0, 1). \tag{2.8}$$

---

[4]For recent overviews on VaR and ES forecasting approaches, see Komunjer (2004) and Nadarajah et al. (2014).

Under $\mathbb{H}_0$, the ES forecasts are correctly specified as it holds that $\hat{e}_t = \mathrm{ES}_\tau\,(Y_t \mid \mathcal{F}_{t-1})$.[5] Since this ES backtest is based on a regression procedure and simultaneously tests the parameters α and β, we call this test the *bivariate ESR backtest*.

As outlined in Dimitriadis and Bayer (2017), estimating the parameters (α, β) in (2.6) by M- or Z/GMM-estimation stand-alone using a semiparametric method without specifying the full conditional distribution of the error term u_t^e is not possible since the functional ES is not elicitable (Gneiting, 2011). However, these parameters can be estimated through a regression technique which jointly models a regression equation for the quantile and the ES proposed by Dimitriadis and Bayer (2017), Patton et al. (2017) and Barendse (2017), which we briefly review in Appendix A. We use this joint regression framework for the semiparametric estimation of (2.6) by estimating the joint system,

$$Y_t = \gamma + \delta \hat{e}_t + u_t^q, \tag{2.9}$$

$$Y_t = \alpha + \beta \hat{e}_t + u_t^e, \tag{2.10}$$

where $Q_\tau(u_t^q \mid \mathcal{F}_{t-1}) = 0.$ and $\mathrm{ES}_\tau(u_t^e \mid \mathcal{F}_{t-1}) = 0.$ This means we choose Y_t as the response variable and $(1, \hat{e}_t)$ as explanatory variables for this regression procedure. Because our null hypothesis is based on only testing the parameters (α, β) in the ES regression equation given in (2.10), we use a Wald statistic which only incorporates these parameters,

$$T_{\mathrm{ESR}} = \left((\hat{\alpha}, \hat{\beta})' - (0, 1)' \right)' \widehat{\Sigma}_{\mathrm{ES}}^{-1} \left((\hat{\alpha}, \hat{\beta})' - (0, 1)' \right)', \tag{2.11}$$

where $\widehat{\Sigma}_{\mathrm{ES}}$ is an estimator for the (asymptotic) covariance matrix of the M-estimator of the parameters (α, β). Patton et al. (2017) show consistency and asymptotic normality for the M-estimator of the regression parameters for α-mixing time series. Using this, and given that $\hat{\Sigma}_{\mathrm{ES}} \xrightarrow{\mathbb{P}} \Sigma_{\mathrm{ES}}$, the test statistic asymptotically follows a χ^2 distribution with two degrees of freedom,

$$T_{\mathrm{ESR}} \xrightarrow{d} \chi_2^2. \tag{2.12}$$

We implement both, backtests based on estimates for the asymptotic covariance matrix and based on the bootstrap (Efron, 1979). For the asymptotic version, we employ the *scl-sp* covariance estimation method discussed in Dimitriadis and Bayer (2017). We further implement the bootstrap hypothesis testing procedure[6] where in each bootstrap sample, we estimate the model parameters and the asymptotic covariance matrix to compute a total of $B = 1000$ bootstrap Wald statistics as in (2.11), where the bootstrap estimates are centered around the estimate for the original sample. Finally, the bootstrap p-value is the share of the B bootstrap test statistics that are larger than or equal to the test statistic for the original sample. As neither the underlying loss function of the M-estimator, given in (A.5), nor the asymptotic covariance, given in (A.8) - (A.12), depend on the temporal ordering of the pairs $(Y_t, \hat{e}_t)$, we apply the iid bootstrap resampling technique of Efron (1979).

Similar tests are already implemented for backtesting of forecasts for the mean (Mincer and Zarnowitz, 1969), for quantiles (Gaglianone et al., 2011) and for expectiles (Guler et al., 2017). As these functionals are elicitable, M-estimation of regression parameters for mean, quantile (Koenker and Bassett, 1978) and expectile (Efron, 1991) regressions is straight-forward. This section shows that introducing the same concept for backtesting ES forecasts is possible, but technically more demanding as we have to estimate the regression parameters through a joint system as given in (2.9) and (2.10).

---

[5] Given that the ES forecasts are correctly specified, i.e. $\hat{e}_t = \mathrm{ES}_\tau\,(Y_t \mid \mathcal{F}_{t-1})$, the correct specification condition (2.7) is equivalent to $\alpha = (1 - \beta)\hat{e}_t$. This results in the remark of Holden and Peel (1990), who claim that the null hypothesis, given in (2.8) is only a sufficient, but not a necessary condition for correctly specified forecasts as $\alpha = (1 - \beta)\hat{e}_t$ is the required necessary condition. However, this more general condition implies that the forecasts $\hat{e}_t$ are constant for all $t = 1, \ldots, T$, which is highly unrealistic given the dynamic nature of financial time series. Consequently, we employ the hypotheses given in (2.8) for our backtesting procedure.

[6]This approach provides an asymptotic refinement, i.e. the error in the rejection probability decreases faster compared to both, the asymptotic distribution and the bootstrapped covariance matrices for the test, see e.g. MacKinnon (2009). In the construction of confidence intervals, this is also known as the percentile-t method.

2.3. The One-sided Intercept ESR Backtest

The bivariate ESR backtest introduced in the previous section only allows for testing two-sided hypotheses as specified in (2.8) because it is generally unclear how too small or too large risk forecasts influence the parameters α and β. Because the capital requirements the financial institutions have to keep as a reserve depend on the reported risk forecasts, the market participants have an incentive to overestimate[7] the risk forecasts in order to keep as little capital requirements as possible. In contrast, underestimation of the risk measures results in too conservative risk forecasts and consequently higher capital requirements, which does not have to be punished by the regulatory authorities.[8] Thus, the regulatory authorities only have to prevent and consequently penalize the overestimation of risk measures, which can be done by using one-sided backtesting procedures. For example, the traffic light system (Basel Committee, 1996), currently implemented in the Basel Accords, is in fact a one-sided backtest for the hit ratios of VaR forecasts.

   Consequently, we also introduce a regression-based backtesting procedure for the ES that allows for both, specifying one-sided and two-sided hypotheses. This backtest is based on regressing the forecast errors, $Y_t - \hat{e}_t$, on an intercept term only,

$$Y_t - \hat{e}_t = \alpha + u_t^e, \tag{2.13}$$

where $\mathrm{ES}_\tau(u_t^e \mid \mathcal{F}_{t-1}) = 0$ and testing whether the parameter α is zero. Estimation of the parameter α in (2.13) is carried out by computing the empirical ES of the forecast errors $Y_t - \hat{e}_t$. By using this restricted regression equation, we can define a one-sided and a two-sided hypothesis,

$$
\begin{aligned}
\mathbb{H}_0^{2s} &: \alpha = 0 \qquad \text{against} \qquad \mathbb{H}_1^{2s} : \alpha \neq 0, \quad \text{and} \\
\mathbb{H}_0^{1s} &: \alpha \geq 0 \qquad \text{against} \qquad \mathbb{H}_1^{1s} : \alpha < 0,
\end{aligned}
\tag{2.14}
$$

which we test by using a t-test based on the asymptotic covariance and based on the bootstrap procedure described above. Note that this is equivalent to setting the slope parameter of the bivariate ESR test given in (2.6) to one and only estimating and testing the intercept term. Consequently, we call this backtest the *intercept ESR backtest*. Both, the bivariate and the intercept ESR backtests proposed in this paper are implemented in our R package esback (Bayer and Dimitriadis, 2017).

3. Existing Backtests

Over the past two decades and especially driven by the recent transition from VaR to ES in the Basel regulatory framework (Basel Committee, 2016, 2017), a large literature on backtesting the ES has emerged. These backtests are usually introduced with financial regulators in mind who need to verify the risk forecasts they receive from the financial institutions. To be applicable for the regulatory authorities, a backtest for the risk measure ES thus follows Definition 2.1 and only requires the observed return series and the ES forecasts as input variables. However, many of the proposed backtests for the ES fail to have this property. In particular, several tests require the whole return distribution (Berkowitz, 2001; Kerkhof and Melenberg, 2004; Wong, 2008; Acerbi and Szekely, 2014; Graham and Pál, 2014), the cumulative violation process $\int_0^\tau \mathbb{1}_{\{Y_t \leq \hat{v}_t(p)\}}\, \mathrm{d}p$ (Costanzino and Curran, 2015; Emmer et al., 2015; Du and Escanciano, 2017; Kratz et al., 2016), the volatility (McNeil and Frey, 2000; Nolde and Ziegel, 2017; Righi and Ceretta, 2013, 2015), or the VaR (McNeil and Frey, 2000; Nolde and Ziegel, 2017) in addition to the ES forecasts. However, this information (except the VaR) is not reported by the financial institutions and therefore, most of these tests can not be used by the regulators (Aramonte et al., 2011; Basel Committee, 2017).

   Furthermore, when more information than solely the ES forecasts is used for backtesting, a rejection of the null hypothesis does not necessarily imply that the ES forecasts are wrong. More precisely, a rejection

---

[7] Overestimation of a risk measure is meant in the mathematical sense, which means reporting a too large real number. As the ES forecasts are strictly negative, this implies in fact underestimation of the associated market risk.

[8] One could interpret the higher capital requirements as a punishment for too conservative risk forecasts.

of the null implies that *some* component of the input parameters is incorrect (cf. Nolde and Ziegel, 2017). A related concern is raised by Aramonte et al. (2011), who note that financial institutions could be tempted to submit forecasts of this additional information chosen such that the tests have particularly low power, so that correctness of their internal model (or their issued ES forecasts) is not doubted.

Strictly following Definition 2.1, we would have to distinguish between backtests for the ES and joint backtests for the pair VaR and ES. However, as the ES is strongly intertwined with the VaR (through its definition and through the joint elicitability), sensible forecasts for the ES are based on correctly specified VaR forecasts. Consequently, it is reasonable to backtest both quantities jointly and thus, we compare the performance of our ES backtests to existing joint VaR and ES backtests in the literature. In the following, we describe the exceedance residual test of McNeil and Frey (2000) and the conditional calibration tests of Nolde and Ziegel (2017) in more detail, since both have versions that only require VaR forecasts in addition to the ES.

3.1. Testing the Exceedance Residuals

One of the first and still most frequently used tests for the ES is the exceedance residual (ER) backtest of McNeil and Frey (2000). This approach is based on the ES residuals that exceed the VaR, $er_t = (Y_t - \hat{e}_t)\mathbb{1}_{\{Y_t \leq \hat{v}_t\}}$, which form a martingale difference sequence given that $\hat{v}_t$ and $\hat{e}_t$ are the true $\mathcal{F}_{t-1}$-measurable quantile and ES respectively. McNeil and Frey (2000) furthermore consider a second version that uses exceedance residuals standardized by the volatility, i.e. $er_t/\hat{\sigma}_t$.

This backtest tests whether the expected value of the (raw or standardized) ER, $\mu = \mathbb{E}[er_t]$, is zero using the estimate $\hat{\mu} = 1/(\sum_{t=1}^{T} \mathbb{1}_{\{Y_t \leq \hat{v}_t\}}) \sum_{t=1}^{T} er_t$ in conjunction with a bootstrap hypothesis test (see Efron and Tibshirani, 1994, p. 224). In the original paper, McNeil and Frey (2000) propose to test μ against the one-sided alternative that μ is negative, i.e. that the ES is overestimated. However, in this paper we discuss both, tests based on one-sided and two-sided hypotheses, so that in addition to the original proposal, we also include a two-sided test,

$$
\begin{aligned}
\mathbb{H}_0^{2s} &: \mu = 0 \qquad \text{against} \qquad \mathbb{H}_1^{2s} : \mu \neq 0, \quad \text{and} \\
\mathbb{H}_0^{1s} &: \mu \geq 0 \qquad \text{against} \qquad \mathbb{H}_1^{1s} : \mu < 0.
\end{aligned}
\tag{3.1}
$$

By Definition 2.1, the test using the standardized ER is in fact a joint backtest for the triple VaR, ES and volatility, whereas the test using the raw ER is a joint backtest for the pair VaR and ES. In light of the discussion above, the test using the raw ER is therefore preferred. Nevertheless, in the simulation studies and the empirical application we apply both approaches and find that they perform alike.

Even though the intercept ESR test introduced in Section 2.3 and the ER backtest appear to be similar, there is a subtle difference between the two test statistics. For the intercept ESR test, we compute the empirical ES of $Y_t - \hat{e}_t$, i.e. the average of $Y_t - \hat{e}_t$ given that $Y_t - \hat{e}_t$ is smaller than its empirical τ-quantile. In contrast, the ER backtest computes the average of $Y_t - \hat{e}_t$, given that Y_t is smaller than the respective forecast for its τ-quantile $\hat{v}_t$. This difference seems marginal, but it has severe consequences for the theoretical and empirical properties of the tests.

As we can write $\hat{\mu} = 1/\tilde{T} \sum_{t=1}^{T} Y_t \mathbb{1}_{\{Y_t \leq \hat{v}_t\}} - 1/\tilde{T} \sum_{t=1}^{T} \hat{e}_t \mathbb{1}_{\{Y_t \leq \hat{v}_t\}}$, where $\tilde{T} = \sum_{t=1}^{T} \mathbb{1}_{\{Y_t \leq \hat{v}_t\}}$, the ER backtest in fact compares the empirical average of Y_t truncated at $\hat{v}_t$ to the average ES forecast $\hat{e}_t$, whenever there is a VaR violation. Thus, this backtest rejects whenever the distance/relation between the VaR and ES-forecasts is incorrect. However, simultaneous misspecifications of both forecasts, such as e.g. generated by misspecification of the volatility process in location scale models cannot be detected. In the same spirit, the ER backtest cannot distinguish between correct forecasts for the VaR and ES at level τ and (correct) forecasts for a misspecified probability level $\tilde{\tau} \neq \tau$, as the given level τ does not influence the ER test statistic at all. In contrast, by computing the empirical τ-quantile of $Y_t - \hat{e}_t$ (instead of using the forecast $\hat{v}_t$), the intercept ESR test does not suffer from these shortcomings as can be observed in the simulation results in Section 4.2.

3.2. Conditional Calibration Backtests

Nolde and Ziegel (2017) introduce the concept of conditional calibration (CC) based on strict identification functions (also known as moment conditions or estimating equations) of the respective functional and show that many classical backtests for risk measures can be unified using this concept. For the pair VaR and ES at level $\tau \in (0, 1)$, they choose the strict identification function

$$V(Y, v, e) = \begin{pmatrix} \tau - \mathbb{1}_{\{Y \leq v\}} \\ e - v + \mathbb{1}_{\{Y \leq v\}}(v - Y)/\tau \end{pmatrix}, \tag{3.2}$$

whose expectation is zero if and only if v and e equal the true VaR and ES of Y respectively. The CC backtest for forecasts for the VaR, $\hat{v}_t$ and for the ES, $\hat{e}_t$ is based on the hypotheses

$$\begin{aligned} \mathbb{H}_0^{2s} &: \mathbb{E}\big[V(Y_t, \hat{v}_t, \hat{e}_t) \mid \mathcal{F}_{t-1}\big] = 0 &\text{against}& \quad \mathbb{E}\big[V(Y_t, \hat{v}_t, \hat{e}_t) \mid \mathcal{F}_{t-1}\big] \neq 0, \quad \text{and} \\ \mathbb{H}_0^{1s} &: \mathbb{E}\big[V(Y_t, \hat{v}_t, \hat{e}_t) \mid \mathcal{F}_{t-1}\big] \geq 0 &\text{against}& \quad \mathbb{E}\big[V(Y_t, \hat{v}_t, \hat{e}_t) \mid \mathcal{F}_{t-1}\big] < 0, \end{aligned} \tag{3.3}$$

component-wise and almost surely for all $t = 1, \ldots, T$. This is equivalent to testing $\mathbb{E}\big[h_t' V(Y_t, \hat{v}_t, \hat{e}_t)\big] = 0$ for all $\mathcal{F}_{t-1}$ measurable $\mathbb{R}^2$-valued functions h_t. As this is infeasible, Nolde and Ziegel (2017) propose to use an $\mathcal{F}_{t-1}$-measurable sequence of $q \times 2$-matrices of test functions $\boldsymbol{h}_t$ for some $q \in \mathbb{N}$ and to use the Wald-type test statistic

$$T_{\text{CC}} = T \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{h}_t V(Y_t, \hat{v}_t, \hat{e}_t) \right)' \widehat{\Omega}^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{h}_t V(Y_t, \hat{v}_t, \hat{e}_t) \right), \tag{3.4}$$

where $\widehat{\Omega} = \frac{1}{T} \sum_{t=1}^{T} (\boldsymbol{h}_t V(Y_t, \hat{v}_t, \hat{e}_t)) (\boldsymbol{h}_t V(Y_t, \hat{v}_t, \hat{e}_t))'$ is a consistent estimator of the covariance of the q-dimensional vector $\boldsymbol{h}_t V(Y_t, \hat{v}_t, \hat{e}_t)$. Under $\mathbb{H}_0$, the test statistic asymptotically follows a χ_q^2 distribution with q degrees of freedom.

Nolde and Ziegel (2017) propose two versions of this test, where the first uses no information besides the risk forecasts (termed *simple CC test*), and where the second additionally requires volatility forecasts (termed *general CC* test). For the simple CC test, the test function is the identity matrix, $\boldsymbol{h}_t = I_2$, for both, the one- and two-sided hypotheses. For the general CC test, they propose to choose

$$\boldsymbol{h}_t = \hat{\sigma}_t \big((\hat{e}_t - \hat{v}_t)/\tau, 1 \big) \quad \text{and} \quad \boldsymbol{h}_t = \begin{pmatrix} 1 & |\hat{v}_t| & 0 & 0 \\ 0 & 0 & 1 & \hat{\sigma}_t^{-1} \end{pmatrix}', \tag{3.5}$$

for the two-sided and for the one-sided test, respectively, where $\hat{\sigma}_t$ is a forecast for the volatility. As with the standardized ER test, the general CC test is strictly speaking a backtest for the triple VaR, ES, and volatility, but we nevertheless include both versions in our empirical comparisons. We provide implementations of the two ESR backtests proposed in this paper, both ER backtests of McNeil and Frey (2000) and both CC backtests of Nolde and Ziegel (2017) in our R package `esback` (Bayer and Dimitriadis, 2017).

4. Monte-Carlo Simulations

In this section, we evaluate the empirical performance of our proposed ES backtests and compare them to the tests of McNeil and Frey (2000) and Nolde and Ziegel (2017). For that, we first assess the empirical size of the tests, defined as the rejection frequency of the test under the null hypothesis and which should equal the nominal significance level. Then, we analyze the empirical power of the tests which is defined as the rejection frequency of forecasts stemming from some misspecified model and which is optimally as close to one as possible.

This comparison is conducted using two different approaches. The first, presented in Section 4.1, follows the typical strategy in the related literature of first assessing the size of the backtests with

some realistic location-scale data generating process (DGP), followed by an evaluation of the power by backtesting forecasts stemming from overly simplified model, in our case the Historical Simulation model. In the second setup, presented in Section 4.2, we continuously misspecify certain parameters of the true model and thereby obtain alternative models with a continuously increasing degree of misspecification. This approach of evaluating backtests has two main advantages. First, we obtain power curves which can be used to draw conclusions how an increasing model misspecification influences the test decisions. Second, misspecifying the different model parameters separately allows us to misspecify certain model characteristics while leaving the remaining model unchanged. Thus, we can evaluate the capability of the backtests to identify certain model misspecifications, which allows for a closer examination of the backtesting procedures.

4.1. Traditional Size and Power Comparisons

For the first simulation study, we simulate returns from an EGARCH(1,1) model (Nelson, 1991) with t-distributed innovations, where the parameter values are calibrated using daily returns of the S&P 500 index. This model is given by

$$
\begin{aligned}
Y_t &= \sigma_t z_t, \\
\log(\sigma_t^2) &= -0.160 - 0.125 z_{t-1} + 0.130 \left(|z_{t-1}| - E[|z_{t-1}|] \right) + 0.983 \log(\sigma_{t-1}^2), \\
z_t &\sim t_{7.24},
\end{aligned}
\tag{4.1}
$$

where z_t are innovations stemming from the standardized Student-t distribution with 7.24 degrees of freedom. As the EGARCH model is highly flexible and due to its calibrated parameter values, this DGP accurately replicates the distributional properties of daily financial returns. Conditional VaR and ES forecasts at level τ for the DGP in (4.1) are given by

$$
\hat{v}_t = \hat{\sigma}_t q_z(\tau) \quad \text{and} \quad \hat{e}_t = \hat{\sigma}_t \xi_z(\tau),
\tag{4.2}
$$

where $\hat{\sigma}_t$ is a volatility forecast generated through the model given in (4.1) and $q_z(\tau)$ and $\xi_z(\tau)$ are the τ-quantile, respectively the τ-ES of the innovations z_t. For the following size and power analysis of the backtests, we simulate the process (4.1) 10,000 times with sample sizes of 250, 500, 1000, 2500, and 5000 observations and 250 additional pre-sample values required for the power analysis. As stipulated by the Basel Accords, we forecast the two risk measures for the probability level $\tau = 2.5\%$. In this part of the study, we focus on two-sided hypotheses and defer the one-sided case to Section 4.3.

Table 1 presents the empirical sizes of the considered backtests for the different sample sizes and for nominal test sizes of 1%, 5%, and 10%. We find that in large samples, all backtests display rejection rates close to the respective nominal sizes. However, in small samples all backtests are oversized and they differ with respect to their speed of convergence. Looking at the individual tests in greater detail, we find that especially the tests relying on asymptotic quantities (i.e. the ESR and CC tests) are substantially oversized in small samples and converge to the nominal sizes comparably slow. However, by using the bootstrap for the intercept and bivariate ESR tests (indicated by (b) in the table), the empirical sizes are much closer to the nominal sizes in small samples than for the asymptotic versions. Comparing the intercept and the bivariate ESR test, we find that the former has better size properties in small samples, presumably because less parameters need be estimated and the covariance is simpler. Furthermore, also the two ER tests (which also rely on bootstrapping) exhibit good empirical sizes and there are hardly any differences between the raw and the standardized version.

For a comparison of the power of the backtests, we evaluate their ability to reject the null hypothesis for risk models producing incorrect ES forecasts. We utilize the Historical Simulation approach which forecasts the VaR and ES by using their empirical counterparts from previous trading days,

$$
\hat{v}_t = \widehat{Q}_\tau \left(Y_{t-1}, Y_{t-2}, \cdots, Y_{t-w} \right) \quad \text{and} \quad \hat{e}_t = \frac{1}{\sum_{i=1}^{w} \mathbb{1}_{\{Y_{t-i} \le \hat{v}_{t-i}\}}} \sum_{i=1}^{w} Y_{t-i} \cdot \mathbb{1}_{\{Y_{t-i} \le \hat{v}_{t-i}\}},
\tag{4.3}
$$

Table 1: Empirical sizes of the backtests

Nominal Size	Sample Size	bivariate ESR (b)	bivariate ESR	intercept ESR (b)	intercept ESR	General CC	Simple CC	Std. ER	ER
	250	0.03	0.18	0.02	0.10	0.01	0.21	0.04	0.05
	500	0.02	0.09	0.02	0.06	0.03	0.12	0.00	0.01
1%	1000	0.02	0.05	0.01	0.03	0.03	0.07	0.01	0.01
	2500	0.01	0.03	0.01	0.02	0.02	0.04	0.01	0.01
	5000	0.01	0.02	0.01	0.01	0.02	0.03	0.01	0.01
	250	0.09	0.27	0.07	0.16	0.08	0.28	0.07	0.08
	500	0.07	0.17	0.06	0.11	0.10	0.19	0.04	0.06
5%	1000	0.06	0.11	0.05	0.07	0.08	0.13	0.04	0.06
	2500	0.05	0.07	0.04	0.06	0.07	0.09	0.05	0.06
	5000	0.06	0.06	0.05	0.07	0.06	0.07	0.05	0.06
	250	0.15	0.33	0.12	0.20	0.17	0.32	0.12	0.14
	500	0.13	0.22	0.10	0.16	0.16	0.24	0.09	0.11
10%	1000	0.12	0.16	0.09	0.12	0.13	0.19	0.10	0.12
	2500	0.11	0.12	0.09	0.12	0.12	0.14	0.10	0.11
	5000	0.11	0.11	0.11	0.13	0.11	0.12	0.10	0.11

Notes: The table reports the empirical sizes of the backtests for the EGARCH(1,1)-t process given in (4.1). The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

where $\widehat{Q}_\tau$ is the empirical τ-quantile and w is the length of a rolling window, that we set to 250, i.e. one year of data. Since the standardized ER and the general CC backtest both require forecasts of the volatility, we estimate this quantity with the sample standard deviation of the returns over the same rolling window.

For a meaningful and fair comparison of the power of the backtests to reject the null hypothesis, we compare the *size-adjusted power*[9] of the backtests (Lloyd, 2005). For this, the original critical values of the tests are either increased or decreased such that the rejection frequencies of the true model equal the nominal test sizes. The size-adjusted power is then given by the rejection frequencies of the alternative models using these modified critical values.

Figure 1a contains the size-adjusted power of the backtests for all empirical sizes in the unit interval for a sample size of 1000.[10] The black line depicts the case of equal empirical size and power, which can be seen as a lower bound for any reasonable test: whenever the power is below this line, randomly guessing the test decision is more accurate than performing the test. We observe that both, the bivariate and intercept version of the ESR backtest clearly dominate the others at almost all empirical sizes, including the most relevant region of test sizes between 1% and 10%. Furthermore, the ESR tests using asymptotic quantities are slightly more powerful than their bootstrap versions (indicated by (b)), but the loss in power is negligible compared to the improvements in the sizes we find in Table 1.

In order to present results for all considered sample sizes in condensed form for the relevant area of empirical sizes between 1% and 10%, we summarize the size-adjusted power by the partial area under the curve (PAUC), as proposed by Lloyd (2005). For that, we numerically compute the area under each power curve for the empirical sizes between 1% and 10% which is thus the power to reject a false model averaged over the considered test sizes. In Figure 1b, we present the PAUC for all backtests and sample sizes. As expected, the average power increases with the sample size, so that using more information leads

---

[9] A comparison of the *raw power*, i.e. the raw rejection rate of the null hypotheses, could be misleading due to the differences in the empirical sizes of the backtests. In particular, an oversized test would exhibit unrealistically large rejection rates. For completeness, Table C.5 reports the raw power of the tests.

[10] These plots are known as the receiver operating characteristic (ROC) curves and origin from the psychometrics literature (Lloyd, 2005). They are an effective presentation method for general binary classification tasks such as hypothesis testing as they show the size-adjusted power simultaneously for all significance levels.
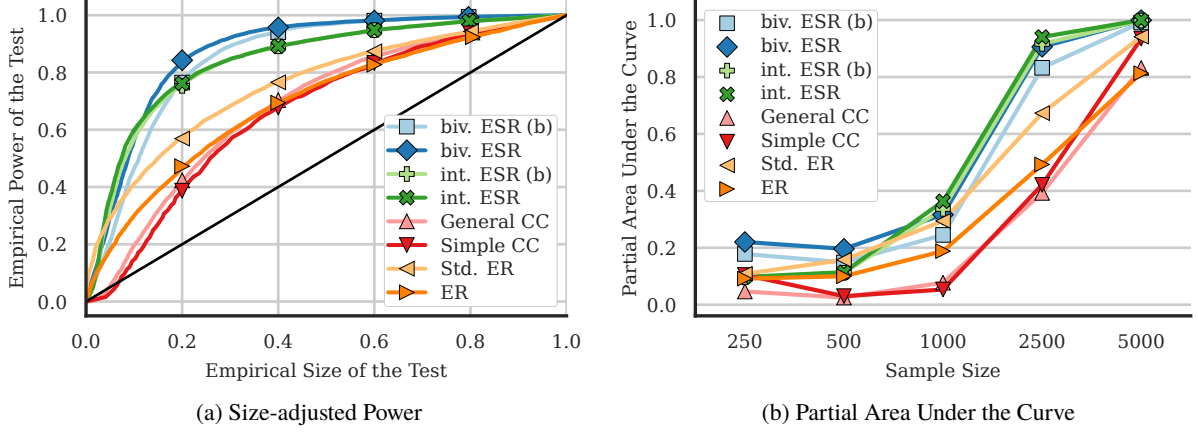
Figure 1: Size-adjusted power and Partial Area Under the Curve plots against Historical Simulation for a sample size of 1000 days. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

to more reliable decisions about the quality of a forecast. We find that for all considered sample sizes, the ESR backtests dominate the other testing approaches. As a robustness check for these findings, we repeat this simulation experiment with the DGP used by Gaglianone et al. (2011) and find that the results, presented in Appendix B, are similar to the findings of this section.

4.2. Continuous Model Misspecification

In the second simulation study, we use a GARCH(1,1) model with standardized Student-t distributed innovations,

$$
\begin{aligned}
Y_t &= \sigma_t z_t, \\
\sigma_t^2 &= \gamma_0 + \gamma_1 Y_{t-1}^2 + \gamma_2 \sigma_{t-1}^2, \\
z_t &\sim t_\nu,
\end{aligned}
\tag{4.4}
$$

with the parameter values $\gamma_0 = 0.01$, $\gamma_1 = 0.1$, $\gamma_2 = 0.85$, and $\nu = 5$ for the true model. For the analysis of the backtests, we simulate 10,000 times from this model with a fixed sample size of 2500 observations and consider the probability level $\tau = 2.5\%$ for the VaR and the ES.

Table 2: Empirical sizes for the second simulation study.

Null Hypothesis	bivariate ESR (b)	bivariate ESR	intercept ESR (b)	intercept ESR	General CC	Simple CC	Std. ER	ER
Two-Sided	0.06	0.07	0.05	0.06	0.08	0.09	0.05	0.06
One-Sided	–	–	0.07	0.03	0.02	0.02	0.06	0.06

Notes: This table shows the empirical sizes of the backtests for the GARCH(1,1)-t model given in (4.4), for a nominal test size of 5% and for both, one-sided and two-sided hypotheses. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000). Note that the bivariate ESR test does not permit a one-sided hypothesis and therefore, we only present sizes for the two-sided hypothesis.

Table 2 presents the empirical sizes of the backtests for a nominal size of 5% for both, the two- and one-sided hypotheses. As in the first simulation study, we find that most of the backtests are reasonably sized with rejection frequencies close to the nominal value. However, the two CC tests reject the true model slightly too often in the two-sided, respectively too rarely in the one-sided case.

11

For a detailed analysis of the power of the backtests, we continuously misspecify the true model according to the following five designs:

(a) We misspecify how the conditional variance reacts to the squared returns by varying the ARCH parameter γ_1. We choose $\tilde{\gamma}_1$ between 0.03 and 0.2 and let $\tilde{\gamma}_2 = 0.95 - \tilde{\gamma}_1$, such that the persistence of the GARCH process remains constant. When $\tilde{\gamma}_1 < \gamma_1$, there is too little variation in the ES forecasts due to the reduced response to shocks and the GARCH process approaches a constant volatility model.

(b) We alter the unconditional variance of the GARCH process $\mathbb{E}[\sigma_t^2] = \gamma_0/(1 - \gamma_1 - \gamma_2)$ between 0.5 and 0.01 by varying the parameter γ_0 while holding γ_1 and γ_2 constant. Since the conditional variance is a weighted combination of the unconditional variance, the past squared returns and the past conditional variance, this change implies that the ES is always underestimated when the unconditional variance is larger as its true value, and vice versa.

(c) We vary the persistence of shocks between 0.9 and 0.999 by setting $\tilde{\gamma}_1 = c \cdot \gamma_1$ and $\tilde{\gamma}_2 = c \cdot \gamma_2$ for a constant c that we vary, and $\tilde{\gamma}_0 = \mathbb{E}[\sigma_t^2](1 - \tilde{\gamma}_1 - \tilde{\gamma}_2)$ to keep the unconditional variance constant. A higher persistence causes a stronger and longer reaction to shocks.

(d) We vary the degrees of freedom of the underlying Student-t distribution between 3 and ∞. Since the conditional variance is unaffected, this modification implies a relative horizontal shift of the ES forecasts.

(e) We misspecify the probability level $\tilde{\tau}$ of the ES forecasts between 0.5% and 5%. This represents the scenario that a forecaster submits (accidentally or on purpose) predictions for some level $\tilde{\tau} \neq \tau$. Similar to changing the degrees of freedom, this modification implies a relative horizontal shift of the ES forecasts.

As an illustrative example of these misspecifications, Figures 2a to 2e show 250 realizations of the returns of the true DGP (4.4), together with the corresponding ES forecasts of the true model (black dashed line) and two models following the parameter misspecifications described in the points (a) to (e) above.

We present the size-adjusted rejection rates plotted against the respective misspecified parameters for these five designs in Figures 3a to 3e. The true model is indicated by the gray vertical line and, induced by the results of Figure 2, the x-axis is oriented such that too small (too risky) ES forecasts are on the left side of the true model.[11] Even though there is no backtest that dominates the others throughout all considered designs, several conclusions can be drawn from this figure.

(1) Overall, the bivariate and intercept ESR tests perform similar and in four out of the five considered designs, their performance is superior compared to the general CC and both ER backtesting approaches. (Figures 3a to 3c and 3e). The ESR backtests outperform the competitors especially when we misspecify the volatility dynamics of the underlying GARCH process (Figures 3a to 3c). This shows that, in contrast to the existing approaches, our new ESR backtests can be used to detect misspecifications in the dynamics used to construct the ES forecasts which go beyond level shifts.

(2) The application of the bootstrap for our ESR tests mainly affects the empirical sizes whereas the empirical size-adjusted power of the asymptotic and the bootstrap ESR tests is similar throughout all designs.

(3) The two ER tests (and the general CC test that is constructed to be similar to the ER backtest) cannot discriminate between forecasts for the VaR and ES issued through misspecified volatility processes (Figures 3a to 3c) and through misspecified probability levels $\tilde{\tau} \neq \tau$ (Figure 3e). This confirms the theoretical results discussed in Section 3.1 that these backtests only reject misspecifications which affect the relation (distance) between the VaR and ES forecasts. In contrast, these backtests perform well in the

---

[11]Notice that this inequality of the forecast magnitude only holds on average in the cases of Figures 3a and 3c whereas it holds strictly for Figures 3b, 3d and 3e.

(a) Changing the reaction to the squared returns

(b) Changing the unconditional variance

(c) Changing the persistence

(d) Changing the degrees of freedom

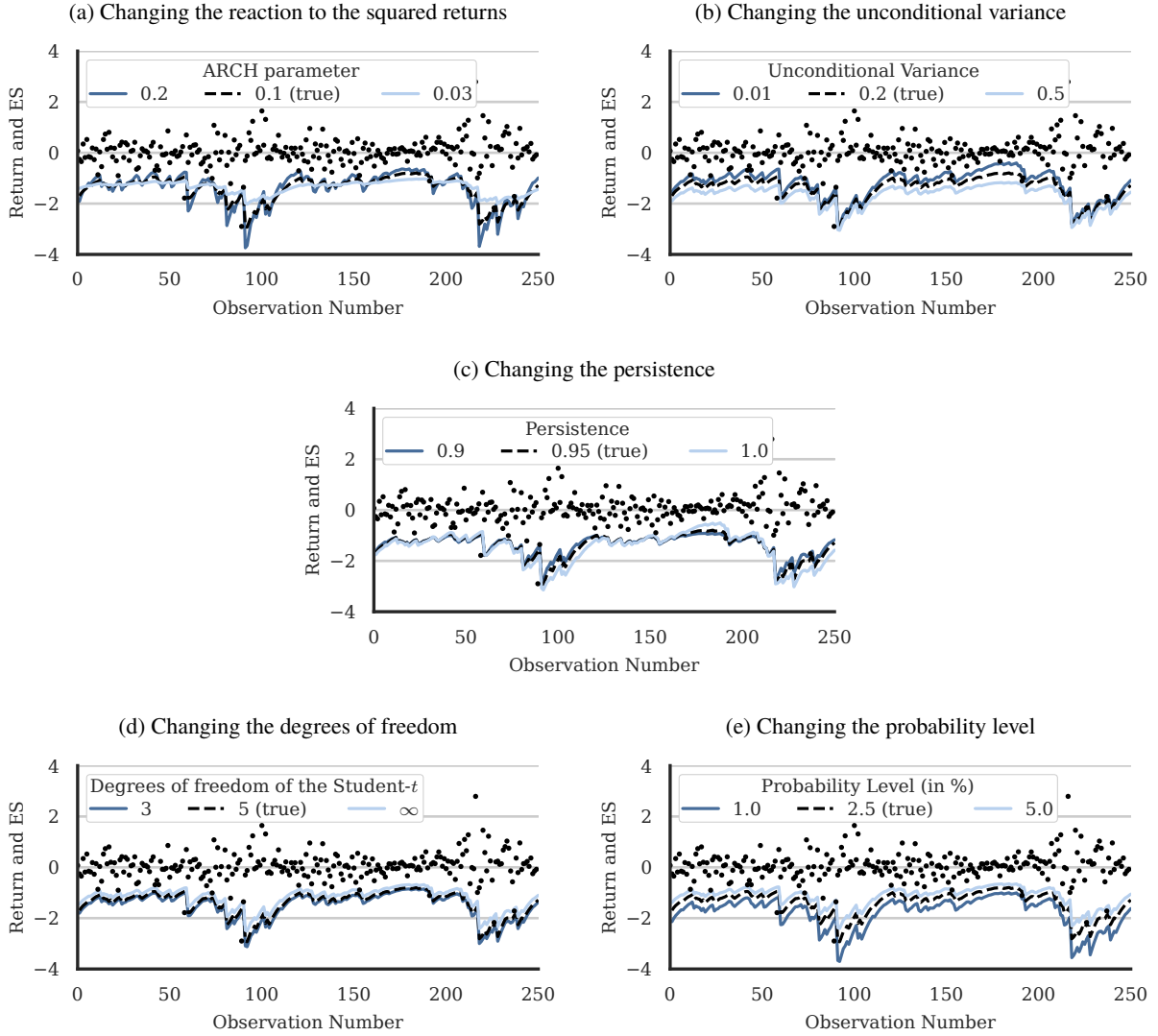(e) Changing the probability level

Figure 2: These plots show exemplary simulated return series with 250 observations for the DGP given in (4.4) and for the five parameter misspecifications illustrated in the points (a) - (e) in Section 4.2. In each of the subfigures, the black dashed line corresponds to the true model parameters.

case of misspecified tails of the residual distribution, which affects the relative distance between the VaR and ES forecasts (Figure 3d). If these backtests would be used by the regulatory authorities, banks could submit joint VaR and ES forecasts for some level $\tilde{\tau} > \tau$ or some (too small) volatility process in order to minimize their capital requirements without facing the risk of being detected by these backtests. In comparison, our intercept ESR backtest which is similar to the ER backtests by construction is clearly able to identify these misspecified probability levels.

(4) Throughout all five misspecifications, the simple CC backtest also exhibits good power properties, similar to our proposed backtests. However, our two ESR backtests exhibit much better size properties (see Tables 1 and 2) and in contrast to the simple CC test, they do not fail to reject the Historical Simulation forecasts in the first simulation study (see Figure 1).

Together with the results from the first simulation study, these findings show that our proposed ESR backtests are a powerful choice for backtesting ES forecasts. They are reasonably sized and exhibit good power properties against a variety of misspecifications. Notably, in contrast to the existing backtests, there is no single type of misspecification where our ESR tests are unable to discriminate between forecasts of the true and the misspecified models.
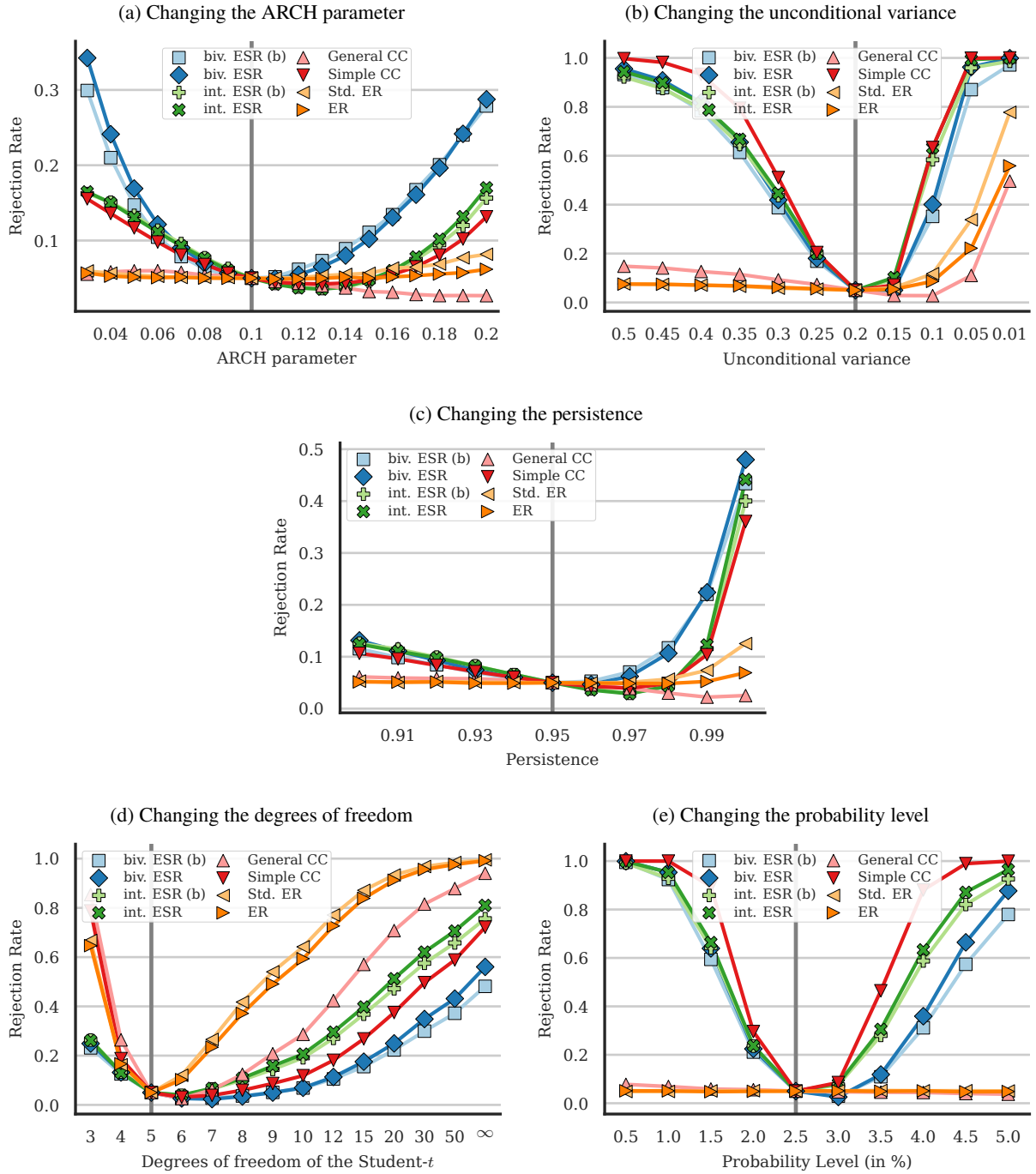
13

Figure 3: Size-adjusted rejection rates for various types of misspecification. The gray vertical line depicts the true model. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

4.3. Testing one-sided hypotheses

For the regulatory authorities, testing against a one-sided alternative might be more meaningful than the two-sided version we consider in the previous section. Holding more money than stipulated in the Basel accords is no concern for regulators as it is only important that banks keep enough monetary reserves to cover the risk from their market activities. As all backtests (with exception of the bivariate ESR test) allow

for testing against one-sided alternatives, we assess their ability to reject the null hypothesis that the issued ES forecasts are smaller or equal to the true ES, i.e. that the associated market risk is not underestimated.

In Figures 4a to 4e, we present the size-adjusted rejection rates for the one-sided versions of the considered backtests and for the five continuous parameter misspecifications described in the points (a) - (e) from the previous section. The structure of these figures is analog to the two-sided case where the x-axis is oriented such that too small (too risky) ES forecasts are on the left side of the true model (vertical gray line). As it can be seen in Figures 2a to 2e, the five modifications of the true model exhibit clear patterns when they are over-, respectively underestimating the true ES, where the overestimation holds strictly for the cases (b), (d) and (e) and on average for the cases (a) and (c). Thus, the one-sided backtests should only reject the null hypothesis for ES forecasts that overestimate the true ES, i.e. which are on the right side of the true model in Figures 4a to 4e.

We find that our intercept ESR backtest (in the asymptotic and the bootstrap version) is reasonably sized (compare Table 2) and clearly dominates the ER and the CC tests in terms of their power in four out of five misspecification designs. Only when changing the degrees of freedom, the ER test is slightly more powerful than the intercept ESR test. Surprisingly, we see that in four out of the five cases, the one-sided CC tests (both, the simple and the general version) also reject too small (too risky) ES forecasts, even though these should not be rejected by the specifications of the one-sided tests.[12] Furthermore, as for the two-sided tests, both ER backtests fail to detect misspecifications of the underlying volatility process and of the underlying probability level. Summarizing these results, the proposed intercept ESR backtest is a powerful backtest with good size properties for testing one-sided hypotheses which clearly dominates the existing one-sided backtesting techniques in the literature.

5. Empirical application

In the empirical application, we predict the market risk for the daily close-to-close log-returns of the S&P500 index for the time period from January 3, 2000 to October 18, 2017, totaling up to 4478 days. We predict the ES (and the VaR for application of the existing tests) for this return series using 10 different risk models. The first two are the Historical Simulation estimated with a rolling window of 250 days and RiskMetrics. The other 8 models follow the volatility specifications of the GARCH(1,1) model and the asymmetric GJR-GARCH(1,1) model of Glosten et al. (1993) and use four different assumptions on the conditional distribution of the innovations. These are the standard normal distribution (abbreviated by N), the standardized Student-t (t), the standardized skewed Student-t (skew-t) and the semi-parametric filtered historical simulation approach (FHS) of Barone-Adesi et al. (1999), where the quantile, respectively the ES of the innovations is estimated from the standardized returns. We estimate these 8 models on a rolling window of 1000 days.

Table 3 presents the p-values of the different ES backtests (for the two-sided hypothesis), the average losses of the strictly consistent 0-homogeneous loss function for the pair VaR and ES[13] (Fissler and Ziegel, 2016), and the p-value of the Model Confidence Set (MCS) of Hansen et al. (2011) applied to this loss function. With the MCS p-values, we can determine a set of models having equal predictive ability at a certain significance level with respect to the losses. The models are sorted according to the average loss.

From this table we can draw several conclusions. First, the MCS rejects 7 out of 10 models at the 5% significance level, i.e. only 3 models have equal predictive power with respect to the joint loss function. These three (GJR-GARCH-skew-t, -FHS, -t) share the same assumption on the volatility process and only differ with respect to the assumption on the innovations. Moreover, for these three models the null hypothesis of correct forecasts is not rejected by almost all backtests at the 5% significance level. Thus, the backtests and the MCS agree on which models predict the ES (and the VaR) well. Second, the CC and

---

[12]We verified our implementation of the CC tests with the codes provided by Nolde and Ziegel (2017) at https://github.com/nnolde/Elicitability-and-Backtesting/.

[13]This is in fact the loss function given in (A.5), applied to a scenario of forecast comparison.

(a) Changing the reaction to the squared returns

(b) Changing the unconditional variance

(c) Changing the persistence

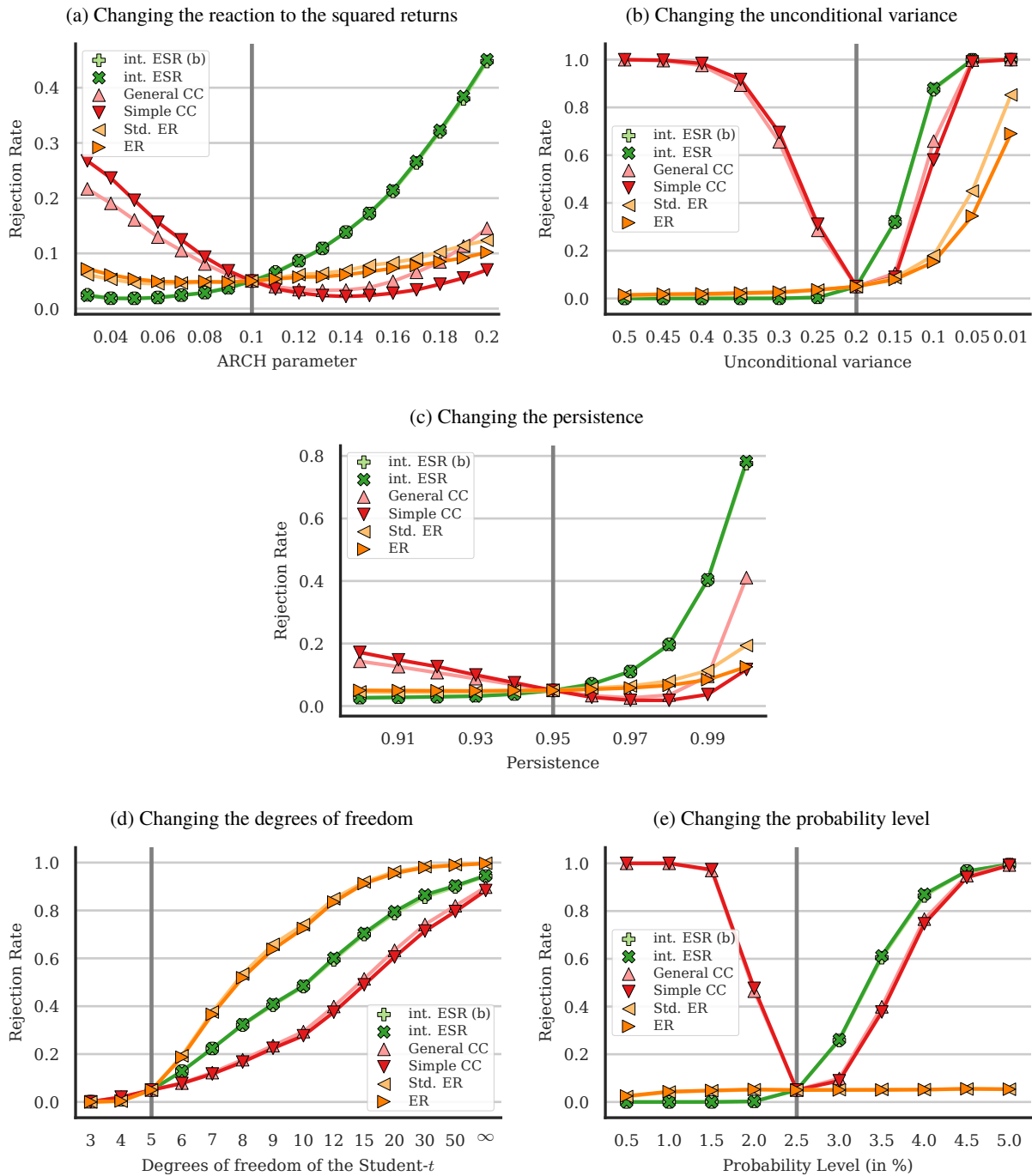(d) Changing the degrees of freedom

(e) Changing the probability level

Figure 4: Size-adjusted rejection rates for various types of misspecification with a one-sided hypothesis. The gray vertical line depicts the true model. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

ER tests reject less forecasts at the 5% significance level than the two ESR backtests, which reflects the findings of the simulation studies where these backtests are often less powerful than our ESR tests. In particular, the null hypothesis is not rejected for the Historical Simulation model, although this approach yields large losses. Third, incorporating leverage into the volatility dynamics appears to be important, since mainly the models using the GJR-GARCH are not rejected by the backtests. Additionally, it is crucial to consider models with flexible tails, e.g. by using the skewed Student-t or the FHS approach, since the

Table 3: Results of the empirical application.

Model	biv. ESR (b)	biv ESR	int. ESR (b)	int. ESR	General CC	Simple CC	Std. ER	ER	Mean Loss	MCS p-value
GJR-GARCH-skew-t	0.07	0.12	0.78	0.77	0.85	0.08	0.87	0.14	0.951	1.00
GJR-GARCH-FHS	0.13	0.20	0.30	0.30	0.39	0.69	0.34	0.55	0.953	0.70
GJR-GARCH-t	0.06	0.10	0.06	**0.04**	0.28	0.11	0.23	0.90	0.963	0.28
GJR-GARCH-N	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.982	**0.02**
GARCH-skew-t	0.12	0.19	0.92	0.92	0.33	**0.05**	0.38	0.10	0.986	**0.05**
GARCH-FHS	0.09	0.14	0.19	0.17	0.67	0.31	0.68	0.69	0.993	**0.03**
GARCH-t	**0.03**	**0.05**	0.05	**0.03**	0.57	**0.02**	0.58	0.60	1.000	**0.03**
GARCH-N	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	1.030	**0.01**
RiskMetrics	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	1.075	**0.00**
Historical Simulation	**0.01**	**0.00**	**0.01**	**0.00**	0.11	**0.01**	0.06	0.06	1.132	**0.01**

Notes: In this table, *p*-values smaller than 5% are printed bold-faced and the models are sorted by the average loss. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000). We compute the MCS *p*-values using the *R*-statistic of Hansen et al. (2011).

models based on conditionally normally distributed returns are collectively rejected by the backtests and the MCS.

6. Conclusion

In this paper, we introduce two novel backtests for ES forecasts which regress the realized returns on the issued ES forecasts using an appropriate regression method for the ES introduced in Dimitriadis and Bayer (2017); Patton et al. (2017); Barendse (2017) and test the resulting parameter estimates. We introduce a bivariate version, denoted *bivariate ESR backtest*, where we test the intercept and the slope parameters for zero and one, and an intercept version, denoted *intercept ESR backtest*, that only incorporates an intercept term being estimated and tested for zero. The motivation for the latter test is the possibility to specify a one-sided hypothesis that is particularly relevant for the regulatory authorities. These backtests can be interpreted as ES-specific versions of the classical Mincer and Zarnowitz (1969) test for evaluating mean forecasts.

A unique feature of the backtests proposed in this paper is that they solely require forecasts for the ES and are consequently the first backtests for the ES stand-alone. In contrast to that, a common drawback of the existing backtests is that they need forecasts of further input parameters, such as the VaR, the volatility, the tail distribution or even the whole return distribution. Using more information than the ES forecasts is problematic for two reasons. First, these tests are not applicable for the regulatory authorities, who receive forecasts of the ES, but not of the additional information required by these tests. Second, rejecting the null hypothesis does not necessarily imply that the ES forecasts are incorrect as the rejection can be a result of a false prediction of any of the input parameters.

In several simulation studies, we assess the empirical size and power properties of our proposed backtests and compare them to the approaches of McNeil and Frey (2000) and Nolde and Ziegel (2017), which jointly backtest the VaR and the ES. We find that our regression-based tests are reasonably sized, especially when they are applied using the bootstrap. Moreover, in almost all simulation designs our two proposed backtests are more powerful than the existing tests. The backtests from the literature are often not able to distinguish between forecasts stemming from the true model and some misspecified model, for instance when we consider a misspecified volatility process or a wrong probability level of the ES. In contrast to that, our two ESR backtests detect the misspecifications in all considered simulation experiments. We provide an implementation of our backtests and of several approaches from the literature in the esback package for R (Bayer and Dimitriadis, 2017).

This paper contributes to the ongoing discussion about which risk measure is the best in practice in the following way. As the VaR is criticized for not being subadditive and for not capturing tail risks beyond itself, the recent literature proposes both, the ES and expectiles as alternative risk measures. Expectiles are suggested as they are coherent, elicitable and are able to capture extreme risks beyond the VaR and thus, they simultaneously overcome the drawbacks of the VaR and the ES (Bellini et al., 2014; Ziegel, 2016). Unfortunately, as opposed to the VaR and ES, they lack a visual and intuitive interpretation (Emmer et al., 2015). In contrast, the ES is mainly criticized for its theoretical deficiencies of being not elicitable and not (only with difficulties) backtestable. However, starting with the joint elicitability result of VaR and ES of Fissler and Ziegel (2016), there is a growing body of literature using this result for a regression procedure (Dimitriadis and Bayer, 2017; Barendse, 2017; Patton et al., 2017) and for relative forecast comparison (Fissler et al., 2016; Nolde and Ziegel, 2017), which is extended by this paper through introducing the ESR backtests, which are the first sensible backtests for the ES stand-alone. This shows that, even though technically more demanding, the ES can be modeled, evaluated and backtested in the same way as quantiles and expectiles. Combining this with its ability to capture extreme tail risks and its intuitive visual interpretation, the ES is an appropriate candidate for being the standard risk measure in practice.

Acknowledgments

Appendix A   The Joint Quantile and ES Regression Technique

Assume we have a response variable Y_t and a k-dimensional vector of covariates X_t following the definition of the stochastic process in (2.1) and assume that we are interested in a linear regression technique for the ES at level $\tau \in (0, 1)$,

$$Y_t = X_t'\theta^e + u_t^e, \tag{A.1}$$

where $\text{ES}_\tau(u_t^e \mid \mathcal{F}_{t-1}) = 0$ for all $t = 1, \ldots, T$, $T \geq 1$. This means that we model the conditional ES through a linear function, $\text{ES}_\tau(Y_t \mid \mathcal{F}_{t-1}) = X_t'\theta^e$. As outlined in Dimitriadis and Bayer (2017), Patton et al. (2017) and Barendse (2017), it is infeasible to estimate the parameter vector θ^e by M- or Z-/GMM-estimation using a semiparametric method without specifying the full conditional distribution of the error term u_t^e. The underlying reason for this is that the functional ES is not elicitable (Gneiting, 2011), i.e. there exists no strictly consistent loss function for the ES, which could be minimized in M-estimation of the regression parameters.[14]

However, as the ES and the quantile at common probability level τ are jointly elicitable (Fissler and Ziegel, 2016), the parameters θ^e in (A.1) can be estimated by jointly modeling a regression equation for the quantile and for the ES,

$$Y_t = X_t'\theta^q + u_t^q \qquad \text{and} \tag{A.2}$$

$$Y_t = X_t'\theta^e + u_t^e, \tag{A.3}$$

---

[14]In contrast, classical mean regression is built on minimizing the squared loss function, which is a strictly consistent loss function for the mean. Furthermore, quantile regression (Koenker, 2005) is estimated by minimizing the asymmetric linear loss function, a strictly consistent loss function for quantiles and expectile regression (Efron, 1991) by the asymmetric squared error loss, a strictly consistent loss function for expectiles.

where $Q_\tau(u_t^q \mid \mathcal{F}_{t-1}) = 0$ and $\mathrm{ES}_\tau(u_t^e \mid \mathcal{F}_{t-1}) = 0$ for all $t = 1, \ldots, T$, $T \geq 1$. Here, $\theta = (\theta^q, \theta^e)$ denotes the $2k$-dimensional vector of regression parameters of the joint model and the quantile and ES equations are modelled through the separate $k-$dimensional parameter vectors θ^q and θ^e. The M-estimator of the regression parameters θ is obtained by

$$\widehat{\theta}_T = \arg\min_\theta \frac{1}{T} \sum_{t=1}^{T} \rho(Y_t, X_t, \theta), \tag{A.4}$$

where the loss function[15] is given by

$$\rho(Y_t, X_t, \theta) = \frac{1}{-X_t'\theta^e} \left(X_t'\theta^e - X_t'\theta^q + \frac{(X_t'\theta^q - Y_t)\mathbb{1}_{\{Y_t \leq X_t'\theta^q\}}}{\tau} \right) + \log(-X_t'\theta^e). \tag{A.5}$$

Consistency and the asymptotic normality of the M-estimator of θ is shown by Patton et al. (2017) for an α-mixing stochastic process $Z_t = (Y_t, X_t)$. Under the further technical conditions in Assumption 1 and 2 in Patton et al. (2017), it holds that

$$\sqrt{T} C_T^{-1/2} \Lambda_T (\hat{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}(0, I), \tag{A.6}$$

where θ_0 denotes the unknown true parameter value and where

$$\Lambda_T = \begin{pmatrix} \Lambda_{11,T} & 0 \\ 0 & \Lambda_{22,T} \end{pmatrix} \quad \text{and} \quad C_T = \begin{pmatrix} C_{11,T} & C_{12,T} \\ C_{21,T} & C_{22,T} \end{pmatrix}, \tag{A.7}$$

with

$$\Lambda_{11,T} = -\frac{1}{T} \sum_{t=1}^{T} \frac{1}{\alpha} \mathbb{E}\left[(X_t X_t') f_t(X_t'\theta_0^q)/(X_t'\theta_0^e) \right], \tag{A.8}$$

$$\Lambda_{22,T} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[(X_t X_t')/(X_t'\theta_0^e)^2 \right], \tag{A.9}$$

$$C_{11,T} = \frac{1}{T} \sum_{t=1}^{T} \frac{1-\alpha}{\alpha} \mathbb{E}\left[(X_t X_t')/(X_t'\theta_0^e)^2 \right], \tag{A.10}$$

$$C_{12,T} = C_{21,T} = -\frac{1}{T} \sum_{t=1}^{T} \frac{1-\alpha}{\alpha} \mathbb{E}\left[(X_t X_t')(X_t'\theta_0^q - X_t'\theta_0^e)/(X_t'\theta_0^e)^3 \right], \tag{A.11}$$

$$C_{22,T} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[(X_t X_t')/(X_t'\theta_0^e)^4 \left(\frac{1}{\alpha} \mathrm{Var}_t\left(Y_t - X_t'\theta_0^q \mid Y_t \leq X_t'\theta_0^q \right) + \frac{1-\alpha}{\alpha} \left(X_t'\theta_0^q - X_t'\theta_0^e \right)^2 \right) \right]. \tag{A.12}$$

Appendix B Robustness Check

The DGP used by Gaglianone et al. (2011) is a GARCH(1,1) model with standard normally distributed innovations,

$$\begin{aligned} Y_t &= \sigma_t z_t, \\ \sigma_t^2 &= 0.05 + 0.05 Y_{t-1}^2 + 0.90 \sigma_{t-1}^2, \\ z_t &\sim \mathcal{N}(0, 1). \end{aligned} \tag{B.1}$$

For this DGP, Table B.4 and Figure B.5 present the empirical sizes and the PAUC analog to the results provided in Section 4.1.

---

[15]As shown by Dimitriadis and Bayer (2017), consistent and asymptotically normal M-estimation of these regression parameters can be obtained by employing loss functions from a whole class of functions, originally introduced by Fissler and Ziegel (2016) in the context of forecast evaluation. However, consensus seems to emerge on the 0-homogeneous loss function presented in (A.5), see e.g. Dimitriadis and Bayer (2017); Taylor (2017); Patton et al. (2017); Barendse (2017) and Nolde and Ziegel (2017).

Table B.4: Empirical sizes of the backtests

Nominal Size	Sample Size	bivariate ESR (b)	bivariate ESR	intercept ESR (b)	intercept ESR	General CC	Simple CC	Std. ER	ER
	250	0.02	0.10	0.01	0.07	0.01	0.17	0.04	0.04
	500	0.02	0.05	0.01	0.05	0.02	0.08	0.00	0.00
1%	1000	0.01	0.04	0.01	0.03	0.02	0.05	0.00	0.01
	2500	0.01	0.02	0.01	0.02	0.02	0.03	0.01	0.01
	5000	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.01
	250	0.08	0.18	0.05	0.13	0.06	0.22	0.06	0.07
	500	0.06	0.12	0.05	0.10	0.07	0.14	0.04	0.04
5%	1000	0.06	0.09	0.05	0.07	0.07	0.10	0.04	0.04
	2500	0.06	0.07	0.05	0.06	0.06	0.07	0.05	0.05
	5000	0.05	0.06	0.05	0.06	0.05	0.06	0.05	0.05
	250	0.14	0.24	0.10	0.18	0.13	0.26	0.11	0.11
	500	0.12	0.18	0.10	0.14	0.13	0.19	0.08	0.08
10%	1000	0.11	0.14	0.10	0.12	0.12	0.15	0.09	0.09
	2500	0.10	0.12	0.10	0.11	0.11	0.12	0.10	0.10
	5000	0.10	0.11	0.10	0.11	0.11	0.11	0.10	0.10

Notes: The table reports the empirical sizes of the backtests for a GARCH(1,1)-N process. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

(a) Size-adjusted Power
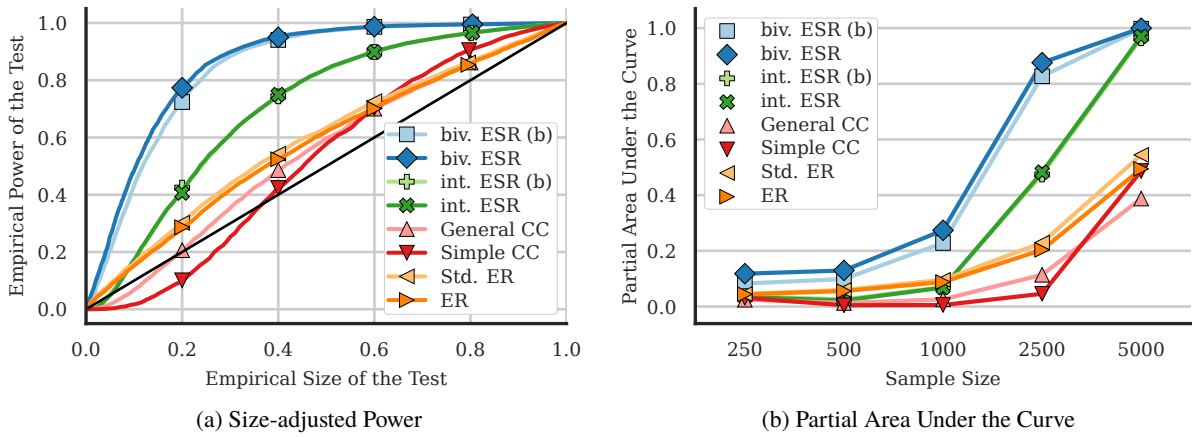
(b) Partial Area Under the Curve

Figure B.5: Size-adjusted power and Partial Area Under the Curve plots against Historical Simulation for a sample size of 1000 days. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

Appendix C   Raw Power

Table C.5: Raw empirical power of the backtests against Historical Simulation

Nominal Size	Sample Size	bivariate ESR (b)	bivariate ESR	intercept ESR (b)	intercept ESR	General CC	Simple CC	Std. ER	ER
1%	250	0.13	0.46	0.04	0.18	0.02	0.23	0.07	0.07
	500	0.08	0.29	0.02	0.11	0.01	0.09	0.03	0.03
	1000	0.07	0.32	0.05	0.22	0.02	0.09	0.09	0.07
	2500	0.46	0.83	0.52	0.81	0.17	0.34	0.37	0.25
	5000	0.96	1.00	0.98	1.00	0.61	0.90	0.81	0.61
5%	250	0.25	0.58	0.12	0.31	0.06	0.34	0.13	0.12
	500	0.18	0.48	0.11	0.29	0.06	0.20	0.13	0.11
	1000	0.28	0.61	0.31	0.51	0.15	0.23	0.28	0.21
	2500	0.89	0.97	0.91	0.97	0.50	0.65	0.67	0.52
	5000	1.00	1.00	1.00	1.00	0.90	0.99	0.95	0.85
10%	250	0.35	0.66	0.21	0.41	0.14	0.42	0.21	0.19
	500	0.29	0.62	0.26	0.41	0.15	0.30	0.23	0.20
	1000	0.53	0.77	0.54	0.65	0.28	0.36	0.41	0.33
	2500	0.98	0.99	0.98	0.99	0.70	0.80	0.80	0.67
	5000	1.00	1.00	1.00	1.00	0.96	1.00	0.98	0.92

Notes: The table reports the raw empirical power of the backtests against the Historical Simulation for the EGARCH(1,1) process given in (4.1). The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).
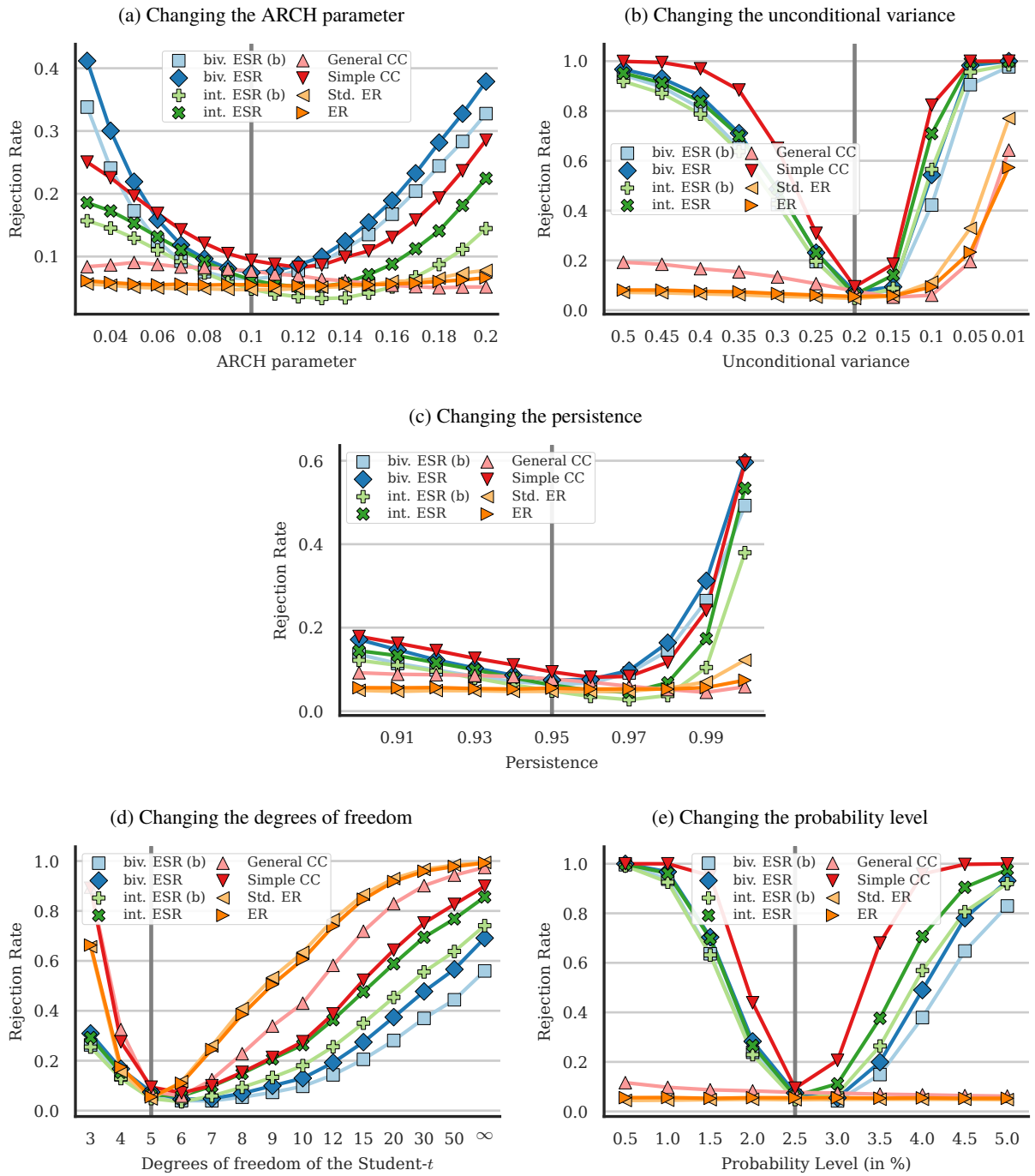
Figure C.6: Raw rejection rates for various types of misspecification. The gray vertical line depicts the true model. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

(a) Changing the reaction to the squared returns

(b) Changing the unconditional variance

(c) Changing the persistence

(d) Changing the degrees of freedom
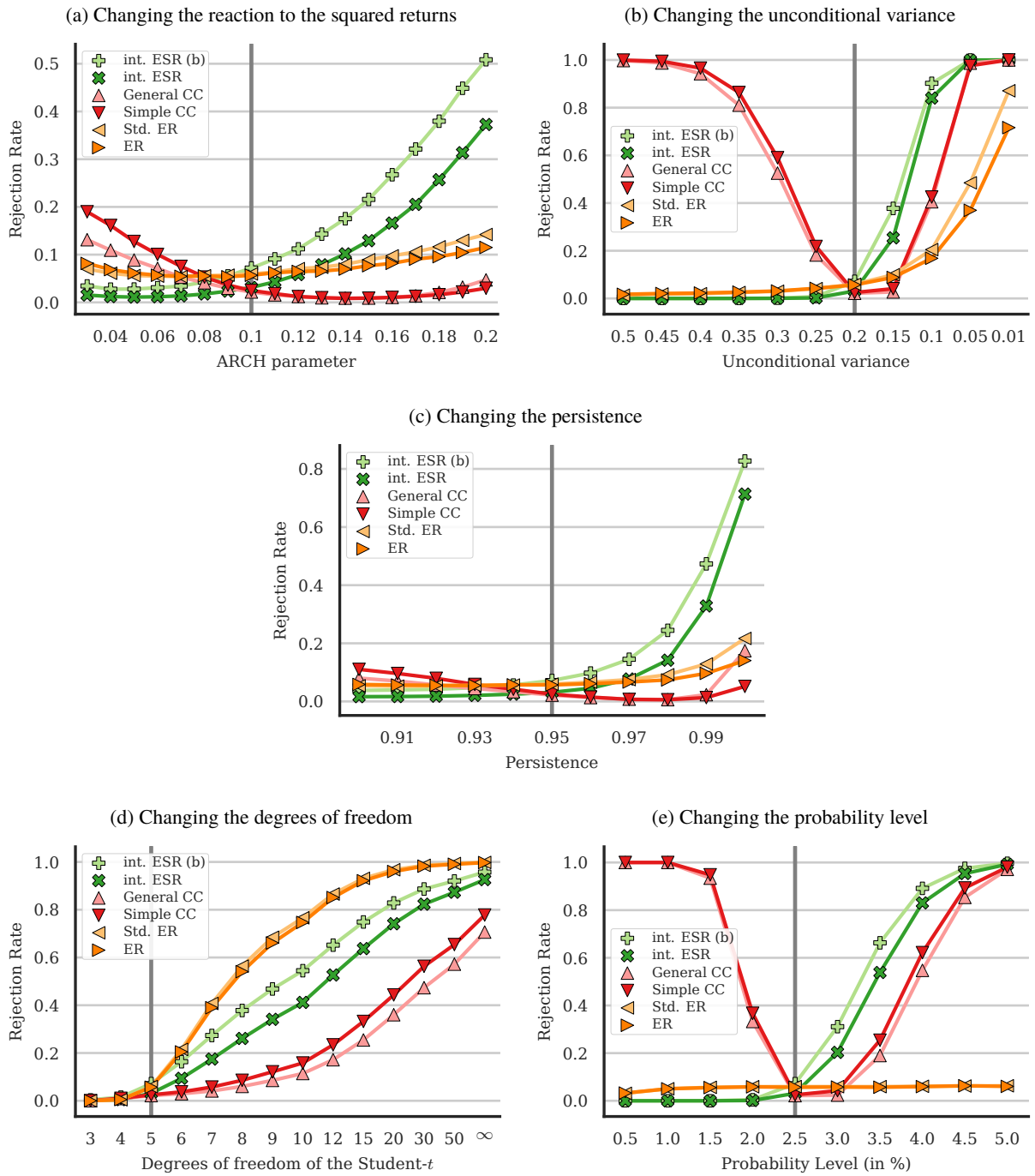
(e) Changing the probability level

Figure C.7: Raw rejection rates for various types of misspecification with a one-sided hypothesis. The gray vertical line depicts the true model. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

References

Acerbi, C. and Szekely, B. (2014). Backtesting Expected Shortfall. *Risk Magazine*, December:76–81.

Aramonte, S., Durand, P., Kobayashi, S., Kwast, M., Lopez, J. A., Mazzoni, G., Raupach, P., Summer, M., and Wu, J. (2011). Messages from the academic literature on risk measurement for the trading book. Technical report, Bank for International Settlements. Working Paper No. 19, available at http://www.bis.org/publ/bcbs_wp19.pdf.

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

Barendse, S. (2017). Interquantile Expectation Regression. Tinbergen Institute Discussion Paper 2017-034/III. Available at https://ssrn.com/abstract=2937665.

Barone-Adesi, G., Giannopoulos, K., and Vosper, L. (1999). VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets*, 19(5):583–602.

Basel Committee (1996). Overview of the Amendment to the Capital Accord to Incorporate Market Risks. Technical report, Bank for International Settlements. Available at http://www.bis.org/publ/bcbs23.pdf.

Basel Committee (2013). Fundamental review of the trading book: A revised market risk framework. Technical report, Bank for International Settlements. Available at http://www.bis.org/publ/bcbs265.pdf.

Basel Committee (2016). Minimum capital requirements for Market Risk. Technical report, Bank for International Settlements. Available at http://www.bis.org/bcbs/publ/d352.pdf.

Basel Committee (2017). Pillar 3 disclosure requirements – consolidated and enhanced framework. Technical report, Basel Committee on Banking Supervision. Available at http://www.bis.org/bcbs/publ/d400.pdf.

Bayer, S. and Dimitriadis, T. (2017). *esback: Expected Shortfall Backtesting*. R package version 0.1.1, available at https://github.com/BayerSe/esback.

Bellini, F., Klar, B., Müller, A., and Gianin, E. R. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54(Supplement C):41 – 48.

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

Carver, L. (2013). Mooted VAR substitute cannot be back-tested, says top quant. *Risk Magazine*, March.

Costanzino, N. and Curran, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *Risk Magazine*, March.

Danielsson, J., Embrechts, P., Goodhart, C., Keating, C., Muennich, F., Renault, O., and Shin, H. S. (2001). An Academic Response to Basel II. Financial Markets Group Special Papers, available at https://EconPapers.repec.org/RePEc:fmg:fmgsps:sp130.

Dimitriadis, T. and Bayer, S. (2017). A Joint Quantile and Expected Shortfall Regression Framework. arXiv:1704.02213 [math.ST].

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

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1(1):93–125.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

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(2):31–60.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4):1680–1707.

Fissler, T., Ziegel, J. F., and Gneiting, T. (2016). Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting. *Risk Magazine*, January:58–61.

Gaglianone, W. P., Lima, L. R., Linton, O., and Smith, D. R. (2011). Evaluating Value-at-Risk Models via Quantile Regression. *Journal of Business & Economic Statistics*, 29(1):150–160.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance*, 48(5):1779–1801.

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Graham, A. and Pál, J. (2014). Backtesting value-at-risk tail losses on a dynamic portfolio. *The Journal of Risk Model Validation*, 8(2):59.

Guler, K., Ng, P. T., and Xiao, Z. (2017). Mincer–Zarnowitz quantile and expectile regressions for forecast evaluations under aysmmetric loss functions. *Journal of Forecasting*, 36(6):651–679.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The Model Confidence Set. *Econometrica*, 79(2):453–497.

Holden, K. and Peel, D. A. (1990). On Testing For Unbiasedness And Efficiency Of Forecasts. *The Manchester School*, 58(2):120–127.

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

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, R. W. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Komunjer, I. (2004). Quantile Prediction. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, chapter 17, pages 961–994. Elsevier.

Kratz, M., Lok, Y. H., and McNeil, A. J. (2016). Multinomial VaR Backtests: A simple implicit approach to backtesting expected shortfall. arXiv:1611.04851 [q-fin.RM].

Lloyd, C. J. (2005). Estimating test power adjusted for size. *Journal of Statistical Computation and Simulation*, 75(11):921–933.

MacKinnon, J. G. (2009). Bootstrap Hypothesis Testing. In Belsley, D. A. and Kontoghiorghes, E. J., editors, *Handbook of Computational Econometrics*, chapter 6, pages 183–213. John Wiley & Sons, Ltd.

McNeil, A. J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3–4):271–300.

Mincer, J. and Zarnowitz, V. (1969). The Evaluation of Economic Forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pages 3–46. National Bureau of Economic Research, Inc.

Nadarajah, S., Zhang, B., and Chan, S. (2014). Estimation methods for expected shortfall. *Quantitative Finance*, 14(2):271–291.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, 59(2):347–370.

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

Patton, A. J., Ziegel, J. F., and Chen, R. (2017). Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk). arXiv:1707.05108 [q-fin.EC].

Righi, M. B. and Ceretta, P. S. (2013). Individual and flexible expected shortfall backtesting. *Journal of Risk Model Validation*, 7(3):3–20.

Righi, M. B. and Ceretta, P. S. (2015). A comparison of Expected Shortfall estimation models. *Journal of Economics and Business*, 78:14–47.

Taylor, J. W. (2017). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics*, 0(0):1–13.

Weber, S. (2006). Distribution Invariant Risk Measures, Information, and Dynamic Consistency. *Mathematical Finance*, 16(2):419–441.

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

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

Ziegel, J. F. (2016). Coherence and elicitability. *Mathematical Finance*, 26(4):901–918.