Regression-Based Expected Shortfall Backtesting*

Sebastian Bayer¹ and Timo Dimitriadis²,³

¹University of Konstanz, ²University of Hohenheim and ³Heidelberg Institute for Theoretical Studies (HITS)

Address correspondence to Timo Dimitriadis, Heidelberg Institute for Theoretical Studies HITS gGmbH, Schloß-Wolfsbrunnenweg 35, 69118 Heidelberg, or e-mail: timo.dimitriadis@h-its.org.

Received 12 January 2018; revised 22 April 2020; editorial decision 22 April 2020; accepted 27 April 2020

Abstract

This article introduces novel backtests for the risk measure Expected Shortfall (ES) following the testing idea of Mincer and Zarnowitz (1969). Estimating a regression model for the ES stand-alone is infeasible and thus, our tests are based on a joint regression model for the Value at Risk (VaR) and the ES, which allows for different test specifications. These ES backtests are the first which solely backtest the ES in the sense that they only require ES forecasts as input variables. As the tests are potentially subject to model misspecification, we provide asymptotic theory under misspecification for the underlying joint regression. We find that employing a misspecification robust covariance estimator substantially improves the tests’ performance. We compare our backtests to existing joint VaR and ES backtests and find that our tests outperform the existing alternatives throughout all considered simulations. In an empirical illustration, we apply our backtests to ES forecasts for 200 stocks of the S&P 500 index.

Key words: asymptotic theory, backtesting, expected shortfall, forecast evaluation, Mincer–Zarnowitz regression, model misspecification

JEL classification: C12, C32, C52, C53, C58, G32

* We thank the editor Andrew Patton, an anonymous associate editor, and two referees for very helpful comments. We further thank Tobias Fissler, Lyudmila Grigoryeva, Roxana Halbleib, Phillip Heiler, Ekaterina Kazak, Winfried Pohlmeyer, James Taylor, and Johanna Ziegel for suggestions which inspired some results of this paper. Financial support by the Heidelberg Academy of Sciences and Humanities (HAW) within the project “Analyzing, Measuring and Forecasting Financial Risks by means of High-Frequency Data”, the Klaus Tschira Foundation, the University of Hohenheim, and by the German Research Foundation (DFG) within the research group “Robust Risk Measures in Real Time Settings” is gratefully acknowledged. The authors acknowledge support by the state of Baden-Württemberg through bwHPC. The majority of the work on this paper was conducted while both authors were at the University of Konstanz.

© The Author(s) 2020. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Through the transition from Value at Risk (VaR) to Expected Shortfall (ES) as the primary market risk measure in the Basel Accords (Basel Committee, 2016, 2017), 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 $\tau$-quantile (the VaR), where $\tau$ is usually chosen to be 2.5% as stipulated by the Basel Accords. The ES is introduced into the banking regulation because it overcomes several shortcomings of the VaR, such as being not coherent and its inability to capture tail risks beyond the $\tau$-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 not as straight-forward for backtesting ES forecasts (Emmer, Kratz, and Tasche, 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 difficulty in directly backtesting ES is its non-elicitability and non-identifiability (Weber, 2006; Gneiting, 2011; Fissler and Ziegel, 2016; Fissler, Ziegel, and Gneiting, 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 almost all VaR backtests.¹

As a consequence, most of the proposed procedures in the growing literature on backtesting ES use indirect approaches by formally backtesting some quantity which is closely related to the ES. Examples include tests based on the entire tail distribution, a linear approximation of the ES through several quantiles, or the pair consisting of the VaR and the ES.² We argue that formally, these approaches are backtests for the auxiliary quantities rather than for the ES itself, see also Nolde and Ziegel (2017). This distinction is particularly important as these backtests require further input variables such as forecasts for the VaR at multiple levels, the tail distribution beyond some quantile, or even the entire distribution. The regulatory authorities, however, do not have this additional information at hand as it is not mandatorily reported by the financial institutions (Aramonte et al., 2011; Basel Committee, 2016, 2017). As a consequence, the existing, so-called ES backtests are not applicable where they are most needed.

In this article, we propose novel backtests for ES forecasts which are the first strict ES backtests in the literature in the sense that besides the realized returns, they only require ES forecasts as input variables. Our tests follow the general regression-based testing idea of Mincer and Zarnowitz (1969). For this, we estimate a regression model that models the


² In particular, several tests require the whole or tail distribution of the returns or equivalently the cumulative violation process (Kerkhof and Melenberg, 2004; Wong, 2008; Graham and Pál, 2014; Acerbi and Szekely, 2014; Du and Escanciano, 2017; Löser, Wied, and Ziggel, 2018; Costanzino and Curran, 2018), multiple quantiles at different levels (Emmer, Kratz, and Tasche, 2015; Costanzino and Curran, 2015; Kratz, Lok, and McNeil, 2018; Couperier and Leymarie, 2019), the VaR and the volatility (McNeil and Frey, 2000; Nolde and Ziegel, 2017; Righi and Ceretara, 2013, 2015), or the VaR (McNeil and Frey, 2000; Nolde and Ziegel, 2017) in addition to the ES forecasts. See Section S.1.2 in the Supplementary Appendix for an overview over the existing backtesting approaches.
conditional ES at level $\tau$ as a linear function $\text{ES}_s(Y_t|\mathcal{F}_{t-1}) = \gamma_1 + \gamma_2 \hat{e}_t$, where we use financial returns $Y_t$ as the response variable and the given ES forecasts $\hat{e}_t$ as the explanatory variable including an intercept term. For correctly specified ES forecasts, the intercept and slope parameters equal zero and one, which we test for by using a Wald statistic. As the ES is not elicitable (Gneiting, 2011), we face the methodological difficulty that we cannot estimate such a regression model for the ES stand-alone as neither loss nor identification functions are available for the ES, which could be used as objective functions for maximum (M) or generalized method of moment (GMM) estimation (Dimitriadis and Bayer, 2019). Recently, Patton, Ziegel, and Chen (2019) and Dimitriadis and Bayer (2019) propose a feasible alternative by specifying an auxiliary quantile regression equation $Q_s(Y_t|\mathcal{F}_{t-1}) = \beta_1 + \beta_2 \hat{q}_t$ (with explanatory variable $\hat{q}_t$) and by jointly estimating the regression parameters $(\beta, \gamma)$ by employing a joint loss function for the quantile and the ES from Fissler and Ziegel (2016).

The specification of the quantile equation allows for different testing approaches. First, we employ auxiliary VaR forecasts $\hat{v}_t$ as the explanatory variable in the quantile equation, but only test the ES-specific parameters $\gamma$. We refer to this test as the Auxiliary ESR (ES regression) backtest. The main drawback of this test is that it requires auxiliary VaR forecasts and consequently, it is formally a joint backtest for the VaR and ES which, however, mainly focuses on the ES by only testing the ES-specific regression parameters. Second, we use the ES forecasts $\hat{e}_t$ as the explanatory variable in both, the quantile and the ES equation and again only test on the ES-specific parameters $\gamma$. We refer to this test as the Strict ESR backtest as it only requires ES forecasts as input variables and thus, it is the first test in the literature which solely backtests ES forecasts. This testing idea comes at the drawback of a potential model misspecification in the quantile equation if the underlying data go beyond a pure scale (volatility) model. Therefore, we provide asymptotic theory for the joint quantile and ES regression framework under model misspecification, which generalizes the asymptotic theory introduced in Dimitriadis and Bayer (2019) and Patton, Ziegel, and Chen (2019). The potential model misspecification results in a more complex and usually inflated asymptotic covariance matrix. We account for this in the implementation of our tests by employing a covariance estimation technique which explicitly estimates these additional covariance terms.

We further introduce an intercept variant of the Strict ESR backtest by fixing the slope parameter in the ES regression to one, and by only estimating and testing the intercept term. We refer to this backtest as the Intercept ESR backtest. This test allows for both, testing against one-sided and two-sided alternatives. In contrast, the other two proposed ESR backtests only allow for testing against two-sided alternatives as it is generally unclear how underestimated and overestimated ES forecasts influence the intercept and slope 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 report risk forecasts which are too risky in order to minimize the expensive capital requirements. In contrast, issuing too conservative risk forecasts results in holding costly capital reserves for the financial institutions but poses no risk to the society as a whole. Thus, the regulators only have to prevent and penalize the underestimation of the financial risks, 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. Like the Strict ESR backtest, the Intercept ESR backtest also has the desired characteristic to only
require ES forecasts as input variables and consequently is the first procedure that solely backtests the ES against a one-sided alternative. We provide implementations of the three ESR backtests proposed in this article in the R package esback (Bayer and Dimitriadis, 2019a).

Such regression-based forecast evaluation approaches are already used for testing mean forecasts (Mincer and Zarnowitz, 1969), quantile forecasts (Gaglianone et al., 2011; Guler, Ng, and Xiao, 2017), and expectile forecasts (Guler, Ng, and Xiao, 2017). In contrast to these functionals where regression techniques are easily available (see e.g. Koenker and Bassett, 1978, Efron, 1991), the non-elicitability of the ES makes our approach more involved but also opens up the possibility for the different testing specifications we introduce. Our multivariate generalization approach of the Mincer and Zarnowitz (1969) testing idea can be applied equivalently to other higher-order elicitable functionals (Fissler and Ziegel, 2016) such as, for example, the variance (in the presence of a non-zero mean) and the Range VaR (Cont, Deguest, and Scandolo, 2010; Embrechts, Liu, and Wang, 2018).

We evaluate the empirical properties of our ESR backtests and compare them to the existing joint VaR and ES backtests of McNeil and Frey (2000) and Nolde and Ziegel (2017) in several simulation designs. In the first setup, we implement the classical size and power analysis for backtesting risk measures, where we simulate data stemming from several realistic data generating processes (DGPs) and evaluate the empirical rejection frequencies of the backtests for forecasts stemming from the true and from some misspecified forecasting model. In order to assess how the potential model misspecification affects the Strict and the Intercept ESR backtests, we utilize DGPs which go beyond the class of pure scale (volatility) processes. For this, we implement two different Student’s-t generalized autoregressive score (GAS) models with time-varying higher moments (Creal, Koopman, and Lucas, 2013) and furthermore use an autoregressive (AR) GARCH (generalized autoregressive conditional heteroskedasticity) model which allows for gradually increasing the degree of misspecification through the AR parameter. In the second setup, we introduce a new technique for evaluating the power of backtests for financial risk measures, where we continuously misspecify certain model parameters of the DGP 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.

The simulations show that all three ESR backtests we propose in this article are well-sized, especially when the tests are applied using the covariance estimation method which accounts for possible model misspecification. We further find that the performance of our testing procedures is almost unaffected by the DGPs which cause model misspecification in the Strict and the Intercept ESR tests. Moreover, our tests 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 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 article is organized as follows. Section 1 introduces the ESR backtests and presents asymptotic theory under model misspecification. Section 2 contains several
simulation studies and Section 3 applies the backtests to ES forecasts for a large amount of stocks from the S&P 500 index. Section 4 concludes. The proofs are deferred to Appendix A and a Supplementary Appendix contains further details of the proofs.

1 Theory

1.1 Setup and Notation

We consider a stochastic process

\[ Z = \{Z_t : \Omega \to \mathbb{R}^{l+1}, l \in \mathbb{N}, t = 1, \ldots, T\}, \quad (1.1) \]

defined on some complete probability space \((\Omega, \mathcal{F}, \mathbb{P})\), with the filtration \( \mathcal{F} = \{\mathcal{F}_t, t = 1, \ldots, T\} \) and \( \mathcal{F}_t = \sigma\{Z_s, s \leq t\} \) for all \( t = 1, \ldots, T \), where \( T \in \mathbb{N} \). We partition the stochastic process \( Z_t = (Y_t, U_t) \), where \( Y_t \) is an absolutely continuous random variable of interest and \( U_t \) is an \( l \)-dimensional vector of explanatory variables. We denote the conditional cumulative distribution function of \( Y_t \) given the past information \( \mathcal{F}_{t-1} \) by \( F_t(y) = \mathbb{P}(Y_t \leq y | \mathcal{F}_{t-1}) \) 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 \( \text{Var}_t(\cdot) \).

For financial applications, the variable \( Y_t \) denotes the daily log returns of a financial asset (for instance, a stock or a portfolio), that is, \( Y_t = \log P_t - \log P_{t-1} \), where \( P_t \) denotes the price of the asset at day \( t = 1, \ldots, T \). This means that throughout this article, we use the sign convention that positive returns denote profits, and negative returns denote losses. The vector \( U_t \) 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 \( Q_t(Y_t | \mathcal{F}_{t-1}) = F_t^{-1}(\tau) = \inf\{y \in \mathbb{R} : F_t(y) \geq \tau\} \), which is called the VaR at level \( \tau \) in financial applications. Furthermore, we define the functional ES at level \( \tau \) of \( Y_t \) given \( \mathcal{F}_{t-1} \) as

\[ \text{ES}_t(Y_t | \mathcal{F}_{t-1}) = \frac{1}{\tau} \int_{0}^{\tau} F_t^{-1}(s) ds. \]

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

\[ \text{ES}_t(Y_t | \mathcal{F}_{t-1}) = \mathbb{E}_t[Y_t | Y_t \leq Q_t(Y_t | \mathcal{F}_{t-1})]. \quad (1.2) \]

We denote an \( \mathcal{F}_{t-1} \)-measurable one-step-ahead forecast for day \( t \) for the risk measure \( \rho \) of the distribution \( F_t \), stemming from some external forecaster or from some given forecasting model\(^3\) by \( \hat{\rho}_t = \hat{\rho}_t(\mathcal{F}_{t-1}) \). Following this notation, we denote forecasts for the \( \tau \)-VaR by \( \hat{\nu}_t \) and for the \( \tau \)-ES by \( \hat{\varepsilon}_t \) for some fixed level \( \tau \in (0, 1) \). For simplicity of notation, we drop the dependence on \( \tau \) 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 article is to develop statistical tests for correctness of a given series of forecasts \( \{\hat{\rho}_t, t = 1, \ldots, T\} \) for the risk measure \( \rho \) relative to the realized return series

---

\(3\) For recent overviews on VaR and ES forecasting approaches, see Komunjer (2004) and Nadarajah, Zhang, and Chan (2014).
\[ \{Y_t, t = 1, \ldots, T\} \]. This is in the literature usually referred to as backtesting of the risk measure \( \rho \) without strictly defining this terminology. We provide such a definition in the following.

**Definition 1.1** A backtest for the series of forecasts \( \{\hat{\rho}_t, t = 1, \ldots, T\} \) for the \( d \)-dimensional risk measure (functional) \( \rho \) relative to the realized return series \( \{Y_t, t = 1, \ldots, T\} \) is a function

\[
 f : \mathbb{R}^T \times \mathbb{R}^{T \times d} \to \{0, 1\},
\]

which maps the return and forecast series onto the respective test decision.

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 variables. 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 variables 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 some vector of risk measures (or the entire tail distribution).

### 1.2 The ESR Backtests

We propose backtests for the risk measure ES that test 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 realized returns \( \{Y_t, t = 1, \ldots, T\} \). We follow the general testing idea of Mincer and Zarnowitz (1969) and regress the returns \( Y_t \) on the forecasts \( \hat{e}_t \) and an intercept term by using a regression equation designed specifically for the functional ES,

\[
 Y_t = \gamma_1 + \gamma_2 \hat{e}_t + \epsilon_t,
\]

where \( \text{ES}_t(\epsilon_t|\mathcal{F}_{t-1}) = 0 \) almost surely. Given the structure in Equation (1.4) and since the forecasts \( \hat{e}_t \) are generated by using the information set \( \mathcal{F}_{t-1} \), this condition on the error term is equivalent to

\[
 \text{ES}_t(Y_t|\mathcal{F}_{t-1}) = \gamma_1 + \gamma_2 \hat{e}_t.
\]

We then test the hypothesis

\[
 \mathbb{H}_0 : (\gamma_1, \gamma_2) = (0, 1) \quad \text{against} \quad \mathbb{H}_A : (\gamma_1, \gamma_2) \neq (0, 1).
\]

Under \( \mathbb{H}_0 \), the ES forecasts are correctly specified as it holds that \( \hat{e}_t = \text{ES}_t(Y_t|\mathcal{F}_{t-1}) \) almost surely.\(^4\) In general, Equation (1.4) is an example of a linear regression equation for

---

4 Given that the ES forecasts are correctly specified, that is, \( \hat{e}_t = \text{ES}_t(Y_t|\mathcal{F}_{t-1}) \), the correct specification condition in Equation (1.5) is equivalent to \( \gamma_1 = (1 - \gamma_2) \hat{e}_t \). This results in the remark of
the ES of the form $Y_t = W_t^T \gamma + \nu_t^e$, for some general vector of covariates $W_t$. As outlined in Dimitriadis and Bayer (2019) and Patton, Ziegel, and Chen (2019), estimating the parameters $\gamma$ by M- or GMM estimation stand-alone is not possible since there do not exist strictly consistent loss and identification functions for the functional ES (Gneiting, 2011). Based on the seminal work of Fissler and Ziegel (2016) who introduce joint loss and identification functions for the VaR and ES, Dimitriadis and Bayer (2019), Patton, Ziegel, and Chen (2019), and Barendse (2020) propose the joint regression technique,

$$Y_t = V_t^T \beta + \nu_t^d, \quad \text{and} \quad Y_t = W_t^T \gamma + \nu_t^e, \quad (1.7)$$

where $V_t$ and $W_t$ are $k$-dimensional, $\mathcal{F}_{t-1}$-measurable covariate vectors, and where $Q_t(\nu_t^d|\mathcal{F}_{t-1}) = 0$ and $\text{ES}(\nu_t^e|\mathcal{F}_{t-1}) = 0$ almost surely. Setting up this joint regression model facilitates the estimation of the joint parameters $(\beta, \gamma)$, whereas stand-alone estimation of $\gamma$ is infeasible. We use this joint setup to propose the following regression based backtests for the ES:

**The Auxiliary ESR Backtest:** We choose $V_t = (1, \hat{v}_t)$ and $W_t = (1, \hat{e}_t)$, i.e. we set up the regression system

$$Y_t = \beta_1 + \beta_2 \hat{v}_t + \nu_t^d, \quad \text{and} \quad Y_t = \gamma_1 + \gamma_2 \hat{e}_t + \nu_t^e, \quad (1.8)$$

and test

$$H_0 : (\gamma_1, \gamma_2) = (0, 1) \quad \text{against} \quad H_A : (\gamma_1, \gamma_2) \neq (0, 1), \quad (1.9)$$

using the Wald-type test statistic

$$T_{A-ESR} = T(\hat{\gamma}_T - (0, 1)) \hat{\Omega}_\gamma^{-1}(\hat{\gamma}_T - (0, 1))^T, \quad (1.10)$$

based on some (consistent) covariance estimator $\hat{\Omega}_\gamma$ for the covariance of the subvector $\gamma$.

**The Strict ESR Backtest:** We choose $V_t = W_t = (1, \hat{e}_t)$, i.e. we set up the regression system

$$Y_t = \beta_1 + \beta_2 \hat{e}_t + \nu_t^d, \quad \text{and} \quad Y_t = \gamma_1 + \gamma_2 \hat{e}_t + \nu_t^e, \quad (1.11)$$

and test

$$H_0 : (\gamma_1, \gamma_2) = (0, 1) \quad \text{against} \quad H_A : (\gamma_1, \gamma_2) \neq (0, 1), \quad (1.12)$$

using the Wald-type test statistic

$$T_{S-ESR} = T(\hat{\gamma}_T - (0, 1)) \hat{\Omega}_\gamma^{-1}(\hat{\gamma}_T - (0, 1))^T, \quad (1.13)$$

based on some (consistent) covariance estimator $\hat{\Omega}_\gamma$ for the covariance of the subvector $\gamma$.

We discuss the employed covariance estimators $\hat{\Omega}_\gamma$ in Section 1.5. Whereas setting up Mincer–Zarnowitz tests for classical elicitable functionals such as the mean, quantiles, and expectiles is straightforward (see Mincer and Zarnowitz (1969), Gaglianone et al. (2011), Holden and Peel (1990), who claim that the null hypothesis, given in Equation (1.6) is only a sufficient, but not a necessary condition for correctly specified forecasts as $\gamma_t = (1 - \gamma_2) \hat{e}_t$ 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 Equation (1.6) for our backtesting procedure.
Guler, Ng, and Xiao (2017), in the case of higher-order elicitable functionals such as the ES, we have several choices as illustrated above. The Auxiliary ESR backtest is based on the regression specification in Equation (1.8) and requires both, VaR and ES forecasts as input variables. Thus, following Definition 1.1, this backtest is formally a joint VaR and ES backtest, however, with a strong emphasis on backtesting ES forecasts. In contrast, the Strict ESR backtest only incorporates ES forecasts and consequently is the first backtest for the ES stand-alone.

The Strict ESR test, however, comes at the cost of a potential model misspecification. Given that the financial returns $Y_t$ follow some pure scale (volatility) process, it holds that the VaR and ES forecasts are perfectly colinear, $\hat{e}_t = c \hat{v}_t$ for some $c \in \mathbb{R}$. Consequently, if $\hat{v}_t$ equals the true conditional VaR, the first equation in (1.11) is correctly specified for the true parameter values $(\beta_1, \beta_2) = (0, 1/c)$. Most of the financial econometrics literature (almost the entire GARCH, stochastic volatility, and Realized Volatility literature) is based on such an assumption for daily returns, which motivates the applicability of this Strict ESR backtest. However, this backtest is also applicable in the general case where the true VaR and ES forecasts are not necessarily colinear. For this, we provide asymptotic theory for M-estimation of the joint VaR and ES regression under potential model misspecification in Section 1.4.

1.3 The One-Sided Intercept ESR Backtest
The two ESR backtests introduced in the previous section only allow for testing two-sided hypotheses as specified in Equations (1.9) and (1.12), as it is generally unclear how too risky (or too conservative) forecasts influence the parameters $\gamma_1$ and $\gamma_2$. Because the capital requirements that the financial institutions have to keep as a reserve depend on the reported risk forecasts, the market participants have an incentive to report too risky forecasts for the ES in order to keep as little capital requirements as possible. In contrast, issuing too conservative risk forecasts and facing higher capital requirements do not have to be punished by the regulatory authorities. Thus, the regulators only have to prevent and consequently penalize the underestimation of financial risks, 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. Hence, we also introduce a regression-based backtesting procedure for the ES that allows for testing one-sided hypotheses.

The Intercept ESR Backtest: This backtest is based on a regression setup similar to the Strict ESR backtest by regressing the forecast errors, $Y_t - \hat{e}_t$, only on an intercept term in the ES-specific regression equation,

\[
Y_t - \hat{e}_t = \beta_1 + \beta_2 \hat{v}_t + u_t^e, \quad \text{and} \quad Y_t - \hat{e}_t = \gamma_1 + u_t^e, \quad (1.14)
\]

where $Q_e(u_t^e|\mathcal{F}_{t-1}) = 0$ and $\text{ES}_e(u_t^e|\mathcal{F}_{t-1}) = 0$ almost surely. By using this restricted regression equation, we can define a one-sided and a two-sided alternative,
$H^0_{2s} : \gamma_1 = 0$ against $H^2_A : \gamma_1 \neq 0$, and

$H^1_{0s} : \gamma_1 \geq 0$ against $H^1_A : \gamma_1 < 0$,

which we test by using a $t$-test based on the estimated asymptotic covariance described in Section 1.5.

Note that this testing procedure is equivalent to fixing the slope parameter of the ES equation in the Strict ESR test given in Equation (1.11) to one and only estimating and testing the intercept term. Therefore, we call this backtest the *Intercept ESR* backtest. We keep the slope parameter in the quantile regression equation, as for pure scale models where $\hat{\epsilon}_t = c\hat{v}_t$, it holds that $\beta_1 = 0$ and $\beta_2 = (1 - c)/c$ under the null hypothesis.

### 1.4 Asymptotic Theory under Model Misspecification

In this section, we consider the asymptotic properties of the M-estimator of the joint VaR and ES regression framework given in Equation (1.7) under potential model misspecification. In the following, we write $X_t = (V_t, W_t)$ for the compound vector of covariates. Following Dimitriadis and Bayer (2019) and Patton, Ziegel, and Chen (2019), the M-estimator of the regression parameters $\theta = (\beta, \gamma)$ is defined by:

$$
\hat{\theta}_T = \arg\min_{\theta \in \Theta} Q_T(\theta), \quad \text{where}
$$

$$
Q_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \rho(Y_t, X_t, \theta) \quad \text{and}
$$

$$
\rho(Y_t, X_t, \theta) = \frac{1}{W_t^\top \gamma - V_t^\top \beta + \frac{(V_t^\top \beta - Y_t)1_{\{Y_t \leq V_t^\top \beta\}}}{\tau}} + \log(-W_t^\top \gamma),
$$

where the loss function in Equation (1.18) is a strictly consistent loss function for the pair quantile and ES (Fissler and Ziegel, 2016). Dimitriadis and Bayer (2019) and Patton, Ziegel, and Chen (2019) show consistency and asymptotic normality for the M-estimator in the case of a correctly specified parametric model, that is, under the assumption that there exists a true parameter $\theta_0 \in \Theta$ such that $Q(u_t^s | \mathcal{F}_{t-1}) = 0$ and $ES(u_t^s | \mathcal{F}_{t-1}) = 0$, almost surely. In the following, we extend this theory by relaxing these assumptions which allows for the general case of misspecified models. For this, we define the pseudo-true parameter

$$
\theta^*_T = \arg\min_{\theta \in \Theta} Q^0_T(\theta), \quad \text{where} \quad Q^0_T(\theta) = \mathbb{E}[Q_T(\theta)].
$$

For the classical case of a correctly specified model, the pseudo-true parameter coincides with the true regression parameter $\theta^*_T = \theta_0$ and is independent of $T$. In the following, we restrict our attention to processes and models for the conditional quantile and ES which follow the following conditions.

**Assumption 1.2** (A1) The distribution $F_t$ is absolutely continuous with density function $f_t$, which is bounded from above, that is, there exists a constant $c > 0$ s.t. $\sup_{y \in \mathbb{R}} f_t(y) \leq c$ and $\sup_{y \in \mathbb{R}} f^*_t(y) \leq c$.

(A2) The parameter space $\Theta \subseteq \mathbb{R}^{2k}$ is compact, convex, and has nonempty interior.
We assume that the pseudo-true parameter \( \theta^*_T \) defined in Equation (1.19) is in the interior of \( \Theta \) and is the unique minimizer of the objective function \( Q_0^T(\theta) \) and that the sequence \( \nabla_\theta Q_0^T(\theta) \) is uncorrelated.

(A4) \( V_t, W_t \in F_{t-1} \) and the matrices \( \mathbb{E}[V_t V_t^\top] \) and \( \mathbb{E}[W_t W_t^\top] \) have full rank.

(A5) The matrix \( \Lambda_T \), defined in Theorem 1.4 has strictly positive Eigenvalues for all \( T \) sufficiently large enough.

(A6) The stochastic process \( \{Y_t, V_t, W_t\} \) is strong mixing of size \( -r/(r-2) \) for some \( r > 2 \).

(A7) For all \( h \in H \), it holds that \( \|W_t\|_{T,1}^2 \) a.s. for some constant \( K > 0 \).

(A8) It holds that \( \mathbb{E}[||V_t||^{r+1}] < \infty, \mathbb{E}[||W_t||^{r+1}] < \infty, \mathbb{E}[||V_t||^{r+1}||W_t||^r] < \infty \) and \( \mathbb{E}[||W_t||^{r+1}|Y_t^r] < \infty \) for the \( r > 2 \) from condition (A6).

(A9) For any \( T \in \mathbb{N} \), \( \sup_{\theta \in \Theta} \sum_{t=1}^T 1_{\{Y_t = V_t^\top \beta\}} \leq K \) a.s. for some constant \( K > 0 \).

The conditions in Assumption 1.2 mainly resemble the regularity conditions for asymptotic normality for correctly specified models from Patton, Ziegel, and Chen (2019) and we refer to Patton, Ziegel, and Chen (2019) for a discussion of these conditions. The key condition that allows for misspecified models is the unique minimization condition of the pseudo-true parameter \( \theta^*_T \) in condition (A3). The above assumptions contain the case of correctly specified models as then, the condition (A3) is naturally fulfilled as the utilized loss function is a strictly consistent loss function for the VaR and the ES (Fissler and Ziegel, 2016).

We connect this weaker condition (A3) to classical misspecified regression models for the mean and for quantiles of White (1980), Gourieroux, Monfort, and Trognon (1984), Kim and White (2003), Komunjer (2005), and Angrist, Chernozhukov, and Fernandez-Val (2006). For correctly specified models, we usually impose the strong condition that for all \( t = 1, \ldots, T \),

\[
\mathbb{E}[\psi(Y_t, X_t, \theta)] = 0 \quad \text{a.s.} \iff \theta = \theta^*_T, \tag{1.20}
\]

where \( \psi(Y_t, X_t, \theta) \) is almost surely the derivative of \( \rho(Y_t, X_t, \theta) \) and corresponds to the identification functions of the model (Gneiting, 2011). The weaker condition (A3) is essentially equivalent to the unconditional moment condition

\[
\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \psi(Y_t, X_t, \theta)\right] = 0 \iff \theta = \theta^*_T. \tag{1.21}
\]

Thus, the condition (1.21) can be interpreted as an average identification condition, that is, \( V_t^\top \beta^*_T \) and \( W_t^\top \gamma^*_T \) are some best averaged linear approximations of the true unknown conditional quantile and ES models.

**Theorem 1.3** (Consistency Misspecified Model). Given the conditions from Assumption 1.2, it holds that \( \theta_T - \theta^*_T \to 0 \) as \( T \to \infty \), where \( \theta^*_T \) is the pseudo-true parameter as defined in Equation (1.19).

The proof of Theorem 1.3 is given in Appendix A.

**Theorem 1.4** (Asymptotic Normality Misspecified Model). Given the conditions of Assumption 1.2, it holds that
\[\Sigma_T(\theta_T^*)^{-1/2} A_T(\theta_T^*) \sqrt{T} (\hat{\theta}_T - \theta_T^*) \overset{d}{\rightarrow} N(0, I_{2k}), \quad (1.22)\]

where

\[A_T(\theta_T^*) = \begin{pmatrix} A_{11,T}(\theta_T^*) & A_{12,T}(\theta_T^*) \\ A_{21,T}(\theta_T^*) & A_{22,T}(\theta_T^*) \end{pmatrix} \quad \text{and} \quad \Sigma_T(\theta_T^*) = \begin{pmatrix} \Sigma_{11,T}(\theta_T^*) & \Sigma_{12,T}(\theta_T^*) \\ \Sigma_{21,T}(\theta_T^*) & \Sigma_{22,T}(\theta_T^*) \end{pmatrix} \quad (1.23)\]

with

\[A_{11,T}(\theta_T^*) = -\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ V_t V_t^T f_t(\gamma_t) - \frac{1}{\tau} W_t^T \gamma_t \right], \quad (1.24)\]

\[A_{12,T}(\theta_T^*) = A_{21,T}(\theta_T^*) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ V_t W_t^T \left( \frac{F_t(\gamma_t^T - \frac{\tau}{\tau})}{\tau} \right) \right], \quad (1.25)\]

\[A_{22,T}(\theta_T^*) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ W_t W_t^T \left( \frac{1}{\left( W_t^T \gamma_t \right)^2} \right) \right], \quad (1.26)\]

\[-2 \frac{2}{T} \sum_{t=1}^{T} \mathbb{E} \left[ W_t W_t^T \left( \frac{1}{W_t^T \gamma_t} \right) \left( \frac{1}{W_t^T \gamma_t} \right)^{-1} \right] \left[ Y_t 1_{\{Y_t \leq V_t^T \beta_t^*\}} + V_t^T \beta_t^* \left( \frac{F_t(\gamma_t^T - \frac{\tau}{\tau})}{\tau} \right) \right], \quad (1.27)\]

and

\[\Sigma_{11,T}(\theta_T^*) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ V_t V_t^T \left( \frac{1}{(V_t^T \gamma_t)^2} \right) \left( 1 - \frac{1}{\tau} + \frac{(1 - 2\tau)(F_t(\gamma_t^T - \frac{\tau}{\tau})}{\tau^2} \right) \right], \quad (1.28)\]

\[\Sigma_{12,T}(\theta_T^*) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ V_t W_t^T \left( \frac{1}{-W_t^T \gamma_t} \right) \left( 1 - \frac{1}{\tau} \right) \right] \left( V_t^T \beta_t^* - W_t^T \gamma_t \right), \quad (1.29)\]

\[+ \frac{1}{\tau} \left( V_t^T \beta_t^* - \frac{F_t(\gamma_t^T - \frac{\tau}{\tau})}{\tau} \right) + W_t^T \gamma_t - \frac{1}{\tau} \mathbb{E} \left[ Y_t 1_{\{Y_t \leq V_t^T \beta_t^*\}} \right], \quad (1.30)\]

\[-\frac{F_t(\gamma_t^T - \frac{\tau}{\tau})}{\tau} \left( V_t^T \beta_t^* - W_t^T \gamma_t \right), \quad (1.31)\]

\[\Sigma_{22,T}(\theta_T^*) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ W_t W_t^T \left( \frac{1}{W_t^T \gamma_t} \right)^2 \left( \frac{1}{\tau} \text{Var}_t(V_t^T \beta_t^* - Y_t | Y_t \leq V_t^T \beta_t^*) + \frac{1}{\tau} (V_t^T \beta_t^* - W_t^T \gamma_t)^2 \right) \right], \quad (1.32)\]

\[+ 2 \left( V_t^T \beta_t^* - W_t^T \gamma_t \right) V_t^T \beta_t^* \frac{\tau - F_t(\gamma_t^T - \frac{\tau}{\tau})}{\tau}, \quad (1.33)\]

The proof of Theorem 1.4 is given in Appendix A. The asymptotic theory derived here embeds the asymptotic theory of Patton, Ziegel, and Chen (2019) and Dimitriadis and Bayer (2019) in the simplified case of correctly specified models. Correct specification implies that \( F_t(\gamma_t^T - \frac{\tau}{\tau}) = \tau \) and \( W_t^T \gamma_t = \frac{1}{\tau} \mathbb{E} \left[ Y_t 1_{\{Y_t \leq V_t^T \beta_t^*\}} \right] \) almost surely for all \( t = 1, \ldots, T \). Imposing these two conditions simplifies the asymptotic covariance matrix of Theorem 1.4 to the asymptotic covariances from Patton, Ziegel, and Chen (2019) and Dimitriadis and Bayer (2019). In general, allowing for model misspecification in regression models comes at the cost of an inflated and more complicated asymptotic covariance matrix, see White (1980), White (1994), Kim and White (2003), Komunjer (2005), and...
Angrist, Chernozhukov, and Fernandez-Val (2006) for examples of semiparametric models for the mean and quantiles.

Given consistency and asymptotic normality, we can derive the asymptotic distribution of the test statistics of the regression-based ESR backtests. Henceforth, we use the short notation $\Omega_T = \Lambda_T(\theta_T)^{-1}\Sigma_T(\theta_T')\Lambda_T(\theta_T)^{-1}$ for the asymptotic covariance. As the Auxiliary ESR backtest is not subject to model misspecification, under the null hypothesis it holds that $\gamma_T = (0,1)$ for all $T \in \mathbb{N}$. However, this does not necessarily hold for the Strict ESR and the Intercept ESR backtests and we define the following modified test statistics for these backtests,

$$\tilde{T}_{S-\text{ESR}}(\gamma_0^c) = T(\gamma_T - \gamma_0)\tilde{\Omega}_{T,c}^{-1}(\gamma_T - \gamma_0)^T, \quad (1.34)$$

$$\tilde{T}_{I-\text{ESR}}(\gamma_0^c) = T(\gamma_{1,T} - \gamma_0^c)\Omega_{T,1}^{-1}(\gamma_{1,T} - \gamma_0^c)^T, \quad (1.35)$$

depending on the parameter $\gamma_0^c$, which we test for. The matrices $\tilde{\Omega}_{T,c}$ and $\Omega_{T,1}$ are the ES-specific parts of the estimators for the asymptotic covariance matrix and $\gamma_{1,T}$ and $\gamma_0^c$ refer to the intercept components of the ES-specific parameter vectors $\gamma_T^c$ and $\gamma_0$. Given these modified test statistics, we can state the following corollary.

**Corollary 1.5** Given the conditions of Assumption 1.2 and given that $\hat{\Omega}_T - \Omega_T \overset{p}{\rightarrow} 0$, it holds that

$$T_{A-\text{ESR}} \overset{d}{\rightarrow} \chi^2_2, \quad \tilde{T}_{S-\text{ESR}}(\gamma_T^c) \overset{d}{\rightarrow} \chi^2_2, \quad \text{and} \quad \tilde{T}_{I-\text{ESR}}(\gamma_{1,T}^c) \overset{d}{\rightarrow} \chi^2_1. \quad (1.36)$$

The proof of Corollary 1.5 is given in Appendix A. For the Auxiliary ESR test, we can simply use the test statistic $T_{A-\text{ESR}}$ in order to test whether $\gamma_T = (0,1)$. However, for the Strict and Intercept ESR tests, we do not know the exact form of the pseudo-true parameter $\gamma_T$ under the null hypothesis due to the potential model misspecification. Consequently, we derive the distribution of the test statistic $\tilde{T}_{S-\text{ESR}}(\gamma_T^c)$ at the pseudo-true parameter $\gamma_T$. In the following, we argue that in realistic financial settings, it holds that $\gamma_T^c \approx (0,1)$ and thus, $\tilde{T}_{S-\text{ESR}}(\gamma_T^c) \approx \tilde{T}_{S-\text{ESR}}((0,1)) = T_{S-\text{ESR}}$ holds approximately, and equivalently for the Intercept ESR test. This implies that these tests still have approximately correct size under these slightly misspecified null hypotheses.

First, the majority of the literature in financial econometrics finds that pure scale processes (e.g., GARCH and stochastic volatility models) approximate the true underlying daily financial data well enough. Thus, $\hat{e}_t \approx c \omega t$, for some $c > 0$ and we find that under the null hypothesis, the regression model in Equation (1.11) is only subject to slight model misspecification. Second, the misspecification is in the auxiliary quantile equation, while we test the parameters of the ES equation in Equation (1.11), which is correctly specified under the null. Thus, the model misspecification enters our test statistic only indirectly through the auxiliary effect of the joint parameter estimation. We confirm that these approximations are very precise in the simulation setup of Appendix B, even for cases of unrealistically strong model misspecification. Furthermore, the simulation study in Section 2 confirms these results by showing that the Strict ESR backtest based on $T_{S-\text{ESR}}$ exhibits correct size and performs almost indistinguishably to the Auxiliary ESR backtest, also in the simulation setups where the underlying data do not follow a pure scale processes. This shows that the
approximation error is negligible under the null hypothesis in realistic financial settings and that the Strict and Intercept ESR backtests can indeed be applied in practice.

The following corollary specifies the behavior of our tests under alternative hypotheses. For this, we define the hypothetical parameter $\gamma_T^{H_0}$ as the ES-specific pseudo-true parameter of the model for correctly specified ES forecasts. While for correctly specified regression equations it holds that $\gamma_T^{H_0} = (0, 1)$, its exact form is unknown in the general case.

**Corollary 1.6** Under the alternative hypotheses,

\[
\mathbb{H}_A^{ESR} : \gamma_T^* \neq (0, 1), \quad \mathbb{H}_A^{S-ESR} : \gamma_T^* \neq \gamma_T^{H_0}, \quad \text{and} \quad \mathbb{H}_A^{I-ESR} : \gamma_{1,T}^* \neq \gamma_{1,T}^{H_0},
\]

for all $T \geq T_A$ for some $T_A \in \mathbb{N}$, and given the conditions of Assumption 1.2 and given that $\Omega_T - \Omega^2 \sim 0$, it holds that for all $c > 0$,

\[
P(T_{A-ESR} \geq c) \xrightarrow{P} 1, \quad P(\tilde{T}_{S-ESR}(\gamma_T^{H_0}) \geq c) \xrightarrow{P} 1, \quad \text{and} \quad P(\tilde{T}_{1-ESR}(\gamma_{1,T}^{H_0}) \geq c) \xrightarrow{P} 1.
\]  

(1.38)

The proof of Corollary 1.6 is given in Appendix A. While the parameter $\gamma_{1,T}^{H_0}$ is unknown in the general case, we argue above that $\gamma_T^{H_0} \approx (0, 1)$ still holds approximately in realistic settings and consequently $\tilde{T}_{S-ESR}(\gamma_T^{H_0}) \approx \tilde{T}_{S-ESR}((0, 1)) = T_{S-ESR}$. Corollary 1.6 theoretically implies diverging power for any case where $\gamma_T^* \neq \gamma_T^{H_0}$, that is, also in misspecified cases when $\gamma_T^{H_0} \neq (0, 1)$, and thus, diverging power by employing the approximated test statistic $T_{S-ESR} = \tilde{T}_{S-ESR}((0,1))$. While this holds theoretically, the empirical performance of the Strict and Intercept ESR tests is almost entirely unaffected by the small approximation error stemming from the indirect misspecification of the quantile model, as can be seen in the simulation results in Appendix B and in Section 2.

### 1.5 Implementation of the Tests

The M-estimation of the parameters $\theta_T$ is carried out by using the R package esreg (Bayer and Dimitriadis, 2019b). The main difficulty in the implementation of the backtests is estimation of the asymptotic covariance matrix $\Omega_T = \Lambda_T(\theta_T^*)^{-1} \Sigma_T(\theta_T^*) \Lambda_T(\theta_T^*)^{-1}$. Generally, this is implemented by using the sample counterparts of the expectation of the components given in Equations (1.24)–(1.33) in Theorem 1.4, which are however subject to the following four nuisance quantities:

(a) the conditional density function, evaluated at the conditional quantile, $\hat{f}_t(V_i^T \hat{\beta}_T)$,

(b) the conditional truncated variance, $\hat{\text{Var}}_t(V_i^T \hat{\beta}_T - Y_i | Y_i \leq V_i^T \hat{\beta}_T)$,

(c) the conditional distribution function, $\hat{F}_t(V_i^T \hat{\beta}_T)$, and

(d) the conditional truncated expectation $\frac{1}{\tau} \hat{\mathbb{E}}_{t|Y_i \leq V_i^T \hat{\beta}_T}[Y_i | Y_i \leq V_i^T \hat{\beta}_T]$.

We implement a novel and misspecification robust covariance estimator by estimating the four nuisance quantities above in the following way. The terms (a) and (b) are subject to the asymptotic covariance of correctly specified models for the quantile and the ES of Dimitriadis and Bayer (2019), Patton, Ziegel, and Chen (2019), and Barendse (2020). Thus, we follow the approach of Dimitriadis and Bayer (2019) and apply the *nid* estimator of Hendricks and Koenker (1992) for (a), the conditional density and the flexible *scl-sp* estimator of Dimitriadis and Bayer (2019) for (b), the conditional truncated variance.
In order to estimate (c), the conditional distribution function \( \hat{F}_t(V_t^T \hat{\beta}_T) \), we follow the general approach of the scl-sp estimator of Dimitriadis and Bayer (2019), that is, we assume that \( F_t \) follows a conditional location-scale model with innovations \( \epsilon_t \) with a flexible zero mean and unit variance distribution. We standardize \( Y_t \) by the estimates of the conditional mean and variance, estimated by pseudo-maximum likelihood and apply a kernel density estimator in order to obtain the distribution function of \( \epsilon_t \). Hence, we can recover the distribution of \( Y_t \) given \( F_t \). Notice that for the minor degree of misspecification we are subject to in our backtesting approach, it approximately holds that \( \hat{F}_t(V_t^T \hat{\beta}_T) \approx \tau \) for all \( t \). We find that this semiparametric estimation approach, which is subject to the location-scale assumption, performs better than pure nonparametric alternatives as we are estimating the conditional distribution evaluated at rather extreme quantiles such as at \( \tau = 2.5\% \).

The last nuisance quantity, \( \frac{1}{\tau} E_t[Y_t 1\{Y_t \leq V_t^T \hat{\beta}_T\}] \), is the mean, given the observations are smaller than the possibly misspecified linear quantile model. This quantity is closely related to the conditional ES, which is assumed to be a linear function in our approach. As for realistic financial data, we only face a minor degree of misspecification in the quantile model, this nuisance quantity is assumed to still be approximately linear, and thus, we obtain that \( \frac{1}{\tau} E_t[Y_t 1\{Y_t \leq V_t^T \hat{\beta}_T\}] = W_t^\tau \hat{\beta}_T \) for all \( t \). Nonparametric estimation of this nuisance quantity again introduces too much estimation noise.

We further implement our backtests based on a covariance estimator from Dimitriadis and Bayer (2019) and Patton, Ziegel, and Chen (2019), which does not account for possible model misspecification. This estimator is based on the simplified covariance structure given in Dimitriadis and Bayer (2019) and Patton, Ziegel, and Chen (2019), where the correct model specification assumption implies that \( F_t(V_t^T \hat{\beta}_T) = s_t \) and \( \frac{1}{\tau} E_t[Y_t 1\{Y_t \leq V_t^T \hat{\beta}_T\}] = W_t^\tau \hat{\beta}_T \), almost surely. Thus, we only estimate the nuisance quantities (a) and (b) in this approach.

2 Monte-Carlo Simulations

In this section, we evaluate the empirical performance of our proposed ESR backtests and compare them to the tests of McNeil and Frey (2000) and Nolde and Ziegel (2017). For this, we assess the empirical size and power of the tests, which are defined as the rejection frequency of the tests under the null and alternative hypothesis, respectively. This comparison is conducted using two different approaches. The first, presented in Section 2.1, follows the typical strategy in the related literature of first assessing the size of the backtests with several realistic DGPs, followed by an evaluation of the power by backtesting forecasts stemming from an overly simplified model, in this case the Historical Simulation (HS) model. In the second setup, presented in Section 2.2, 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 advantages. First, we obtain power curves which can be used to draw conclusions of how an increasing model misspecification influences the test decisions. Second, misspecifying the different model parameters in isolation allows us to misspecify certain model characteristics while leaving the remaining model unchanged.
2.1 Traditional Size and Power Comparisons

In order to compare the proposed backtests from the previous sections, we simulate data from several DGPs. Besides pure scale (volatility) model specifications, under which the Strict and Intercept ESR backtests are correctly specified, we also consider more general Student’s-t GAS models (Creal, Koopman, and Lucas 2013) with time-varying higher moments and AR-GARCH specifications where our ESR backtests are subject to model misspecification under the null hypothesis.

**EGARCH:** The first DGP is an EGARCH(1,1) model (Nelson, 1991) with t-distributed innovations, where the parameter values are calibrated to daily returns of the S&P 500 index,

\[ Y_t = \sigma_t z_t, \quad \text{where} \quad z_t \sim t_{7.39}, \quad \text{and} \]

\[
\log(\sigma^2_t) = -0.0012 - 0.161 z_{t-1} + 0.136(|z_{t-1}| - E[|z_{t-1}|]) + 0.978 \log(\sigma^2_{t-1}).
\]

(2.1)

This model represents a highly flexible GARCH specification and due to its calibrated parameter values, this DGP accurately replicates the distributional properties of daily financial returns. As we assume zero mean for this model, the true VaR and ES forecasts are perfectly colinear and consequently, the regression equations for the Strict and the Intercept ESR backtests are correctly specified under the null hypothesis.

**AR-GARCH:** The next specification is an AR(1)-GARCH(1,1) model with Gaussian innovations,

\[ Y_t = \phi Y_{t-1} + \sigma_t z_t, \quad \text{where} \quad z_t \sim \mathcal{N}(0, 1), \quad \text{and} \]

\[
\sigma^2_t = 0.01 + 0.1 Y^2_{t-1} + 0.85 \sigma^2_{t-1},
\]

(2.2)

where we consider the three specifications \( \phi \in \{0, 0.1, 0.5\} \) for the AR parameter. This DGP introduces model misspecification for the Strict and Intercept ESR backtests through the non-zero conditional mean specification, while leaving the realistic volatility structure of the financial returns unchanged. For this DGP, the ratio between true ES and VaR is given by:

\[
\frac{\hat{t}_t}{\hat{v}_t} = \frac{\mu_t + \sigma_t q_z(\tau)}{\mu_t + \sigma_t \xi_z(\tau)},
\]

(2.3)

where \( \mu_t \) is the conditional mean of \( Y_t \) given \( \mathcal{F}_{t-1} \) and \( q_z(\tau) \) and \( \xi_z(\tau) \) are the \( \tau \)-quantile, respectively the \( \tau \)-ES of the innovations \( z_t \). If \( \mu_t \) equals zero, the ratio is constant and thus, the regression equations in (1.11) are correctly specified under the null. By increasing the time-dependence of the conditional mean model through the AR parameter, we can monotonically strengthen the model misspecification in this DGP.

**GAS-STD:** We use a 3-factor Student’s-t GAS model with time-varying location \( \mu_t \), scale \( \sigma_t \), and degrees of freedom \( \xi_t \), with parameters calibrated to daily returns of the S&P 500 index. This model is estimated and simulated by using the R package GAS (Ardia, Boudt, and Catania, 2019) and is based on the following model specification

\[ Y_t|Y_1, \ldots, Y_{t-1} \sim t(\mu_t, \sigma_t, \xi_t), \]

(2.4)

where the vector \((\mu_t, \sigma_t, \xi_t)\) follows an autoregressive specification, driven by the lagged
score of the log-likelihood of the distributional specification in Equation (2.4). Creal, Koopman, and Lucas (2013) and Harvey (2013) introduce the general GAS specification, which nests many well known models, including ARMA, GARCH (Bollerslev, 1986), and ACD (Engle and Russell, 1998) models. Koopman, Lucas, A., and Scharth (2016) provides an overview of GAS and related models. We refer to Appendix A of Ardia, Boudt, and Catania (2019) for the exact parametric specification of this Student’s-\( t \) GAS model.

GAS-SSTD: We generalize the previous GAS model to a 4-factor asymmetric Student’s-\( t \) GAS model with time-varying location \( \mu_t \), scale \( \sigma_t \), skewness \( \lambda_t \), and degrees of freedom \( n_t \),

\[
Y_t(Y_1, \ldots, Y_{t-1}) \sim t(\mu_t, \sigma_t, \lambda_t, n_t).
\]  

(2.5)

Compared to the previous 3-factor GAS specification, this model further allows for asymmetries in the conditional return distribution through allowing for an additional time-varying skewness parameter with an autoregressive GAS specification.

For the two location-scale DGPs, we obtain VaR and ES forecasts at level \( \tau \) by:

\[
\hat{\nu}_t = \hat{\mu}_t + \hat{\sigma}_t q_z(\tau) \quad \text{and} \quad \hat{e}_t = \hat{\mu}_t + \hat{\sigma}_t \xi_z(\tau),
\]  

(2.6)

where \( \hat{\mu}_t \) and \( \hat{\sigma}_t \) are the respective location and volatility forecasts generated by the location and scale models and \( q_z(\tau) \) and \( \xi_z(\tau) \) are the \( \tau \)-quantile, respectively the \( \tau \)-ES of the innovations \( z_t \). For the \( t \)-distributions of the two GAS models, we obtain the ES forecasts through numerical integration. For the following size and power analysis of the backtests, we simulate data from the DGPs given above with varying sample sizes of 250, 500, 1000, 2500, and 5000 observations and 250 additional pre-sample values required for the power analysis. We run 10,000 Monte-Carlo replications for each of the DGPs. As stipulated by the Basel Accords, we fix the probability level to \( \tau = 2.5\% \) for the VaR and ES forecasts for each of the DGPs. In this part of the study, we focus on two-sided hypotheses and defer the one-sided case to Section 2.3. We compare our three ESR backtests to two specifications of the conditional calibration (CC) backtest of Nolde and Ziegel (2017) and to two specifications of the exceedance residual (ER) backtests of McNeil and Frey (2000), which are further described in Section S.1.2.1 and Section S.1.2.2 in the Supplementary Appendix.

Table 1 presents the empirical sizes of the considered backtests for the different DGPs introduced above and for the different sample sizes and a nominal test size of 5%. Table S.1 and Table S.2 in Section S.1.3 in the Supplementary Appendix show equivalent results for nominal significance levels of 1% and 10%. We find that in large samples, all backtests display rejection rates close to the respective nominal size for all considered DGPs. However, in small samples, the ESR tests based on the misspecification covariance estimator exhibit much better sizes compared to the equivalent ESR tests which do not account for the potential misspecification. As this holds for both, DGPs which do and do not generate misspecification under the null, this indicates that the misspecification covariance estimator better approximates the finite sample distribution and should consequently be applied in empirical applications.

We further find that the Strict ESR test and the Auxiliary ESR test perform almost identical throughout all considered DGPs. This implies that the indirect misspecification the Strict ESR test introduces is negligible for realistic financial data. Even for the AR-GARCH model with increasing AR parameter \( \phi \), the size properties of the Strict and the Intercept ESR tests are not adversely affected by the increasing degree of misspecification, see the
results of Appendix B for further details on this. From the four backtests from the literature, the general CC and the ER and its standardized version exhibit satisfactory sizes whereas the Simple CC test is severely oversized, especially in small samples.

<table>
<thead>
<tr>
<th>DGP</th>
<th>Sample size</th>
<th>Str. ESR</th>
<th>Aux. ESR</th>
<th>Int. ESR</th>
<th>Str. ESR</th>
<th>Aux. ESR</th>
<th>Int. ESR</th>
<th>Gen. CC</th>
<th>Sim. CC</th>
<th>Std. ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGARCH-STD</td>
<td>250</td>
<td>0.09</td>
<td>0.09</td>
<td>0.04</td>
<td>0.24</td>
<td>0.25</td>
<td>0.15</td>
<td>0.08</td>
<td>0.29</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.15</td>
<td>0.15</td>
<td>0.11</td>
<td>0.10</td>
<td>0.20</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>0.11</td>
<td>0.11</td>
<td>0.09</td>
<td>0.09</td>
<td>0.14</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>GAS-STD</td>
<td>250</td>
<td>0.10</td>
<td>0.10</td>
<td>0.05</td>
<td>0.26</td>
<td>0.26</td>
<td>0.15</td>
<td>0.07</td>
<td>0.28</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.08</td>
<td>0.08</td>
<td>0.05</td>
<td>0.16</td>
<td>0.16</td>
<td>0.11</td>
<td>0.10</td>
<td>0.20</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.11</td>
<td>0.11</td>
<td>0.09</td>
<td>0.09</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.07</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>0.04</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>GAS-SSTD</td>
<td>250</td>
<td>0.09</td>
<td>0.09</td>
<td>0.05</td>
<td>0.25</td>
<td>0.25</td>
<td>0.15</td>
<td>0.07</td>
<td>0.26</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.06</td>
<td>0.06</td>
<td>0.04</td>
<td>0.15</td>
<td>0.15</td>
<td>0.10</td>
<td>0.09</td>
<td>0.18</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>0.10</td>
<td>0.10</td>
<td>0.08</td>
<td>0.08</td>
<td>0.13</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>AR-GARCH, φ = 0.0</td>
<td>250</td>
<td>0.05</td>
<td>0.05</td>
<td>0.03</td>
<td>0.18</td>
<td>0.18</td>
<td>0.11</td>
<td>0.06</td>
<td>0.22</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.12</td>
<td>0.12</td>
<td>0.09</td>
<td>0.07</td>
<td>0.14</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.09</td>
<td>0.09</td>
<td>0.07</td>
<td>0.07</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>AR-GARCH, φ = 0.1</td>
<td>250</td>
<td>0.05</td>
<td>0.05</td>
<td>0.03</td>
<td>0.18</td>
<td>0.18</td>
<td>0.11</td>
<td>0.06</td>
<td>0.22</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.12</td>
<td>0.12</td>
<td>0.08</td>
<td>0.07</td>
<td>0.14</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.09</td>
<td>0.09</td>
<td>0.07</td>
<td>0.07</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>AR-GARCH, φ = 0.5</td>
<td>250</td>
<td>0.04</td>
<td>0.04</td>
<td>0.02</td>
<td>0.17</td>
<td>0.17</td>
<td>0.11</td>
<td>0.06</td>
<td>0.22</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.12</td>
<td>0.12</td>
<td>0.09</td>
<td>0.07</td>
<td>0.14</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.09</td>
<td>0.09</td>
<td>0.07</td>
<td>0.07</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Notes: This table reports the empirical sizes of the backtests for the different DGPs described in Section 2.1 and for a nominal test size of 5%. 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 three backtests introduced in this article and we consider versions with covariance estimation with and without model misspecification. CC refers to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).
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 HS approach which forecasts the VaR and ES by using their empirical counterparts from previous trading days,

\[
\hat{v}_t = \hat{Q}_\tau(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-w}) \quad \text{and} \quad \hat{e}_t = \frac{1}{w} \sum_{i=1}^{w} Y_{t-i} \cdot 1_{\{Y_{t-i} \leq \hat{v}_t\}},
\]

where \(\hat{Q}_\tau\) is the empirical \(\tau\)-quantile and \(w\) is the length of a rolling window, that we set to 250, that is, one year of data. Since the standardized ER and the general CC backtests 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\(^6\) of the backtests (Lloyd, 2005). For this, the original critical values of the tests are modified 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.

The left panels in Figure 1 and Figure 2 contain the size-adjusted power of the backtests for all empirical sizes in the unit interval for a sample size of 1000 and for the different DGPs.\(^7\) 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. For the three ESR backtests, we only report power for the tests relying on the misspecification robust covariance estimator as these versions of the tests exhibit superior size properties for all considered DGPs. We observe that throughout all six considered DGPs, the Strict and the Auxiliary ESR backtests clearly dominate the other tests in terms of power at almost all empirical sizes, including the most relevant region of test sizes between 1% and 10%. The Intercept ESR test is not as powerful, which is not unexpected as due to its unity restriction in the slope coefficient, it cannot account for misspecifications in the dynamics as precisely as the Strict and Auxiliary ESR tests.

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 can be interpreted as the test power averaged over the different test sizes. In the right-hand panels of Figure 1 and Figure 2, we present the PAUC for all backtests, DGPs, and sample sizes. As expected, the average power increases with the sample size, so that using more information leads to more reliable decisions about the quality of a forecast.

\(^6\) A comparison of the raw power, that is, 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.

\(^7\) These plots are known as the receiver operating characteristic 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.
We find that for all considered sample sizes, the Strict and Auxiliary ESR backtests dominate the other testing approaches. The almost identical performance of the Strict and the Auxiliary ESR tests throughout all simulation designs in Figure 1 and Figure 2 emphasizes that the misspecification introduced by the Strict ESR test seems to be unproblematic for realistic financial data.

2.2 Continuous Model Misspecification
In the second simulation study, we use a GARCH(1,1) model with standardized Student-\(t\) distributed innovations,
$Y_t = \sigma_t z_t$, where $z_t \sim t_\nu$, and

$$\sigma_t^2 = \eta_0 + \eta_1 Y_{t-1}^2 + \eta_2 \sigma_{t-1}^2,$$

(2.8)

with the parameter values $\eta_0 = 0.01, \eta_1 = 0.1, \eta_2 = 0.85$, 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 presents the empirical sizes of the backtests for a nominal size of 5% for both, the two- and one-sided hypothesis. As in the first simulation study, we find that most of the backtests are reasonably sized with rejection frequencies close to the nominal value.

Figure 2. Size-adjusted power and PAUC plots against HS 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 article with (m) indicating the version which account for the additional covariance terms induced by the misspecified model. CC refers to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000). (a) AR-GARCH $\phi = 0$: Size-adjusted Power; (b) AR-GARCH $\phi = 0$: PAUC; (c) AR-GARCH $\phi = 0.1$: Size-adjusted Power; (d) AR-GARCH $\phi = 0.1$: PAUC; (e) AR-GARCH $\phi = 0.5$: Size-adjusted Power; (f) AR-GARCH $\phi = 0.5$: PAUC.
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 $g_1$. We choose $\tilde{g}_1$ between 0.03 and 0.2 and let $\tilde{g}_2 = 0.95 - \tilde{g}_1$, such that the persistence of the GARCH process remains constant. When $\tilde{g}_1 < g_1$, 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 $E[\sigma^2_t] = g_0/(1 - \tilde{g}_1 - \tilde{g}_2)$ between 0.5 and 0.01 by varying the parameter $g_0$ while holding $\tilde{g}_1$ and $\tilde{g}_2$ 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 forecasts are too conservative when the unconditional variance is larger than its true value, and vice versa.

(c) We vary the persistence of shocks between 0.9 and 0.999 by setting $\tilde{\eta}_1 = d \cdot \eta_1$ and $\tilde{\eta}_2 = d \cdot \eta_2$ for a varying constant $d > 0$ and by setting $\tilde{g}_0 = E[\sigma^2_t](1 - \tilde{\eta}_1 - \tilde{\eta}_2)$ in order to stabilize the unconditional variance. 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 $\infty$. Since the conditional variance is unaffected, this modification implies a relative horizontal shift of the ES forecasts.

(e) We misspecify the probability level $\tilde{s}$ 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 $\tilde{s} \neq s$. Similar to changing the degrees of freedom, this modification implies a relative horizontal shift of the ES forecasts.

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 $g_1$. We choose $\tilde{g}_1$ between 0.03 and 0.2 and let $\tilde{g}_2 = 0.95 - \tilde{g}_1$, such that the persistence of the GARCH process remains constant. When $\tilde{g}_1 < g_1$, 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 $E[\sigma^2_t] = g_0/(1 - \tilde{g}_1 - \tilde{g}_2)$ between 0.5 and 0.01 by varying the parameter $g_0$ while holding $\tilde{g}_1$ and $\tilde{g}_2$ 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 forecasts are too conservative when the unconditional variance is larger than its true value, and vice versa.

(c) We vary the persistence of shocks between 0.9 and 0.999 by setting $\tilde{\eta}_1 = d \cdot \eta_1$ and $\tilde{\eta}_2 = d \cdot \eta_2$ for a varying constant $d > 0$ and by setting $\tilde{g}_0 = E[\sigma^2_t](1 - \tilde{\eta}_1 - \tilde{\eta}_2)$ in order to stabilize the unconditional variance. 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 $\infty$. Since the conditional variance is unaffected, this modification implies a relative horizontal shift of the ES forecasts.

(e) We misspecify the probability level $\tilde{s}$ 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 $\tilde{s} \neq s$. 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 S.1a to S.1e in the Supplementary Appendix depict 250 realizations of the returns of the true DGP in Equation (2.8), together with the corresponding ES forecasts of the true model (black dashed line) and of two exemplary 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 Figure 3a–e. The true model is indicated by the gray
vertical line and, induced by the results of Figure S.1 in the Supplementary Appendix; the X-axis is oriented such that too risky (too small in absolute value) ES forecasts are on the right side of the true model. Even though there is no backtest that dominates the others throughout all considered designs, several conclusions can be drawn from this figure.

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 article with (m) indicating the version which account for the additional covariance terms induced by the misspecified model. CC refers 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.

vertical line and, induced by the results of Figure S.1 in the Supplementary Appendix; the X-axis is oriented such that too risky (too small in absolute value) ES forecasts are on the right side of the true model. Even though there is no backtest that dominates the others throughout all considered designs, several conclusions can be drawn from this figure.

Notice that this inequality of the forecast magnitude only holds on average in the cases of Figure 3a and c whereas it holds strictly for Figure 3b, d and e.
1. Overall, the Strict and Auxiliary ESR tests perform almost indistinguishably and in four out of the five considered designs, their performance is superior compared to the general CC and both ER backtesting approaches (Figure 3a–c). The ESR backtests outperform the competitors especially when we misspecify the volatility dynamics of the underlying GARCH process (Figure 3a–c). This shows that, in contrast to the existing

**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 article with (m) indicating the version which account for the additional covariance terms induced by the misspecified model. CC refers 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.
approaches, our ESR backtests can be used to detect misspecifications in the dynamics used to construct the ES forecasts which go beyond level shifts.

2. The two ER tests (and the general CC test that is constructed to be similar to the ER backtest) can hardly discriminate between forecasts for the VaR and ES issued through misspecified volatility processes (Figure 3a–c) and through misspecified probability levels $\tilde{\tau} \neq \tau$ (Figure 3e). This confirms the theoretical results discussed in Section S.1.2.1 in the Supplementary Appendix 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 case of misspecified tails of the residual distribution, which particularly 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.

3. Throughout all five misspecifications, the simple CC backtest also exhibits good power properties, similar to our proposed backtests. However, our three 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 HS forecasts in the first simulation study (see Figure 1).

4. The Intercept ESR test performs well for misspecifications in the residual distribution, while it exhibits lower power against misspecifications in the dynamics of the model when we alter the ARCH parameter and the persistence of the process. This confirms the theoretical considerations that the Strict and Auxiliary test have a greater ability to reject these misspecifications.

Together with the results from the first simulation study, these findings demonstrate 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.

2.3 Testing One-Sided Hypotheses

For the regulatory authorities, testing against a one-sided alternative might be more meaningful than the two-sided versions of the tests we consider in the previous sections. Holding more money than stipulated by the Basel Accords is no concern for regulators as it is only important that banks keep enough monetary reserves to cover the risks from their market activities. In the following, we assess the performance of the Intercept ESR backtest and the one-sided versions of the four competitor backtests in rejecting the null hypothesis that the issued ES forecasts are at least as conservative (not smaller in absolute value) as the true ES, that is, that the associated market risk is not underestimated.

In Figure 4a–e, 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 risky ES forecasts are on the right side of the true model (vertical gray line). As it can be seen in Figures S.1a to S.1e in the
Supplementary Appendix, the five modifications of the true model exhibit clear patterns when they issue too risky, respectively too conservative forecasts for the true ES, where this finding 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 issue too risky (too small in absolute value) forecasts, that is, which are on the right side of the true model in Figure 4a–e.

We find that our Intercept ESR backtest is reasonably sized (compare Table 2) and dominates the ER and the CC tests in terms of their power in three out of the five misspecification designs. Only when altering the degrees of freedom, the ER tests are more powerful than the Intercept ESR test. When changing the persistence of the process, the Intercept ESR test performs overall comparably to its competitors throughout the different degrees of misspecification. 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 conservative ES forecasts, even though these should not be rejected by the specifications of the one-sided tests. Furthermore, as for the two-sided tests, both ER backtests fail to detect missspecifications 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 dominates the existing one-sided (joint VaR and ES) backtesting techniques in the literature.

3 Empirical Application

In the empirical application, we apply our backtests to compare ES forecasts along three dimensions: the complexity of the risk model, the length of the estimation window, and the model refitting frequency. From a practitioners point of view, it would be desirable to have a parsimonious model that can be estimated with few observations and is valid over a long period of time, for reasons of low engineering effort, data storage, and human and computational effort for updating the model. To assess whether such a setup is reasonable, and if not, which dimensions are crucial for a good performance, we compare rejection rates of ES forecasts using our backtests.

For this application, we use daily log returns of the 200 most highly capitalized stocks of the S&P 500 index (as of September 1, 2019), with a sufficiently long history of stock prices. We consider four different risk models: the standard GARCH(1,1) of Bollerslev (1986) and the GJR-GARCH(1,1) model of Glosten, Jagannathan, and Runkle (1993), both coupled with Gaussian and Student-t distributed innovations. For all four models and 200 stocks, we compare the same evaluation horizon, the period from January 2010 to August 2019 with a total of 2432 daily observations. We furthermore consider five different lengths of the rolling estimation window ranging from one year (250 trading days) up to eight years (2000 trading days) and refitting horizons of 5, 21, 62, 125, and 250 days, corresponding to weekly, monthly, quarterly, bi-yearly, and yearly updating of the models.

Table 3 presents the rejection rates of the one-sided Intercept ESR backtest with a nominal size of 5% for the 200 stocks under investigation, for the four GARCH specifications, the five estimation window sizes, and the five refitting frequencies. We choose to use the one-sided Intercept ESR test as this is the only one-sided and strict ES backtest in the literature. Given the currently implemented traffic light system of the Basel Committee,
The results show that both, the GARCH-N and GJR-GARCH-N are rejected for almost all the stocks (in more than 80% of the cases) uniformly over the different estimation sample sizes and refitting frequencies. Independent of the sample length and refitting frequency, this supports the well-known finding that Gaussian residuals generally fail to capture the riskiness of financial assets, especially in the tails of the distribution. In contrast, for the two GARCH specifications with Student-\(t\) distributed innovations, the rejection frequencies are considerably lower and for almost all choices of the refitting frequency and the estimation window length, they are below the nominal significance level of 5%. Furthermore, refitting the models more frequently tends to slightly decrease the rejection frequency for the models with Student-\(t\) distributed innovations, however, it tends to increase the rejection frequency for the Gaussian models. Overall, this implies that the refitting frequency is not a key factor in the model performance. Increasing the size of the estimation window tends to decrease the rejection frequency, whereas the results stabilize for lengths above 1000 days. Interestingly, employing the GJR-GARCH model, which accounts for a potential leverage effect in the volatility process, does not perform better than the standard GARCH model.

Overall, the results of this application which are diversified over 200 individual stocks, imply that using a fat-tailed residual distribution and an estimation window above 1000 days (roughly four years) suffices to obtain rejection rates uniformly below 1%.

### Table 3. Results of the empirical application

<table>
<thead>
<tr>
<th>Rolling window</th>
<th>Refitting frequency</th>
<th>Refitting frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>250</td>
<td>GARCH-N</td>
<td>0.96</td>
</tr>
<tr>
<td>500</td>
<td>GARCH-N</td>
<td>0.94</td>
</tr>
<tr>
<td>1000</td>
<td>GARCH-N</td>
<td>0.86</td>
</tr>
<tr>
<td>1500</td>
<td>GARCH-N</td>
<td>0.85</td>
</tr>
<tr>
<td>2000</td>
<td>GARCH-N</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>GARCH-(t)</td>
<td>0.00</td>
</tr>
<tr>
<td>250</td>
<td>GARCH-(t)</td>
<td>0.00</td>
</tr>
<tr>
<td>500</td>
<td>GARCH-(t)</td>
<td>0.00</td>
</tr>
<tr>
<td>1000</td>
<td>GARCH-(t)</td>
<td>0.00</td>
</tr>
<tr>
<td>1500</td>
<td>GARCH-(t)</td>
<td>0.00</td>
</tr>
<tr>
<td>2000</td>
<td>GARCH-(t)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Notes: This table shows the rejection rates of the one-sided ESR backtest for ES forecasts stemming from the two GARCH-type models with Student’s \(t\) and Gaussian residuals, different rolling window sizes, and model refitting lengths (in days). The rejection frequencies are averaged over the analyzed 200 most capitalized stocks of the S&P 500 index. The out-of-sample window covers the time from January 2010 to August 2019 resulting in a sample size of 2432 days.
4 Conclusion

With the upcoming implementation of the third Basel Accords, risk managers and regulators will shift attention to the risk measure ES for the forecasting and evaluation of financial risks. In this article, we introduce regression based ESR backtests for ES forecasts, which extend the classical Mincer and Zarnowitz (1969) test to ES-specific versions. As estimation of regression parameters for the ES stand-alone is infeasible, our tests build on a recently developed joint VaR and ES regression, which allows for different specifications of our tests, titled the Auxiliary, Strict, and Intercept ESR backtests. As these tests are potentially subject to model misspecification, we extend the asymptotic theory for the joint VaR and ES regression model to possibly misspecified models and verify the tests’ performance in finite samples through an extensive simulation study. We apply our tests to 200 stocks from the S&P 500 index in order to analyze the performance of ES forecasts stemming from the GARCH model family. We find that using fat-tailed (Student’s $t$) residual distributions and more than four years of data yield satisfactory ES forecasts.

A unique and essential feature of the Strict and Intercept ESR backtests is that they solely require forecasts for the ES and are consequently the first backtests for the ES stand-alone. In contrast, a common drawback of the existing backtests in the literature is that they need forecasts of further input variables, 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.

This article 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, Kratz, and Tasche, 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, 2019; Patton, Ziegel, and Chen, 2019; Barendse, 2020) and for relative forecast comparison (Fissler, Ziegel, and Gneiting, 2016; Nolde and Ziegel, 2017), which is extended by this article by 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.

Supplementary Data

Supplementary data are available at Journal of Financial Econometrics online.
Appendix A: Proofs

Proof of Theorem 1.3: We check that the necessary conditions (i)–(iv) of the basic consistency theorem, given in Theorem 2.1 in Newey and McFadden (1994), p. 2121 hold, where we consider the objective functions \( Q_T(\theta) \) and \( Q_T^0(\theta) \) as defined in Equations (1.17) and (1.19). First, notice that condition (ii) holds by imposing condition (A2). The unique identification condition (i) holds by assumption (A3). Next, we verify the uniform convergence condition (iv) by applying the uniform weak law of large numbers given in Theorem A.2.5. in White (1994). For that, we have to show that

(A). the map \( \theta \mapsto \rho(Y_t, X_t, \theta) \) is Lipschitz-\( L_1 \) on \( \Theta \), see Definition A.2.3 in White (1994),

(B). For all \( \theta^0 \in \Theta \), there exists \( \delta^0 > 0 \), such that for all \( \delta, 0 < \delta \leq \delta^0 \), the sequences

\[
\bar{\rho}_t(\theta^0, \delta) := \sup_{\theta \in \Theta} \{ \rho(Y_t, X_t, \theta) | \| \theta - \theta^0 \| < \delta \} \quad \text{and} \quad (A.1) \\
\rho_t(\theta^0, \delta) := \inf_{\theta \in \Theta} \{ \rho(Y_t, X_t, \theta) | \| \theta - \theta^0 \| < \delta \} \quad (A.2)
\]

obey a weak law of large numbers.

Condition (A) follows directly from Lemma S.1.1 and we turn to condition (B). As the processes \( Y_t, V_t, W_t \) is strong mixing of size \(-r/(r-2)\) for some \( r > 2 \) by condition (A6), the processes \( V_t \) and \( W_t \) are strong mixing of the same size by Theorem 3.49 in White (2001), p. 50. As the functions \( \rho(Y_t, X_t, \theta) \) and the supremum/infimum functions are \( \mathcal{F}_t \)-measurable for all \( t \in \mathbb{N} \), we can conclude that the sequences \( \bar{\rho}_t(\theta^0, \delta) \) and \( \rho_t(\theta^0, \delta) \) are also strong mixing of the same size by applying the same theorem.

Furthermore, for \( \bar{r} > 1 \) and for some \( \delta > 0 \) sufficiently small enough, \( r \geq \bar{r} + \bar{\delta} \) and thus

\[
\mathbb{E}[|\bar{\rho}_t(\theta^0, \delta)|^{\bar{r} + \bar{\delta}}] \leq \sup_{1 \leq t \leq T} \mathbb{E}[\sup_{\theta \in \Theta} |\rho(Y_t, X_t, \theta)|^{\bar{r}}] \quad \text{for all} \quad t, 1 \leq t \leq T, T \geq 1.
\]

As \( \Theta \) is compact, there exists some \( c > 0 \) such that \( \sup_{\theta \in \Theta} |\theta|| < c \) and thus, for all \( t = 1, \ldots, T \), it holds that

\[
\mathbb{E}[\sup_{\theta \in \Theta} |\rho(Y_t, X_t, \theta)|^{\bar{r}}] \leq 4^{\bar{r}-1}\left\{ 1 + \left( \frac{c}{K} \right) \left( 1 + \frac{1}{\bar{r}} \right) \mathbb{E}[|V_t||^{\bar{r}} + \frac{1}{\bar{r}K} \mathbb{E}[|Y_t||^{\bar{r}} + \sup_{\theta \in \Theta} \mathbb{E}[|\log(W_t^\gamma)||^{\bar{r}}] \right)^{\bar{r}}. \quad (A.3)
\]

which is bounded by condition (A8) and as \( \log(z) \leq z \) for \( z \) large enough. The same inequality holds for \( |\rho_t(\theta^0, \delta)| \). Thus, we can apply the weak law of large numbers for strong mixing sequences in Corollary 3.48 in White (2001), p. 49 in order to conclude that for all \( \theta^0 \in \Theta \) such that \( |\theta^0 - \theta|| \leq \delta \), it holds that

\[
\frac{1}{T} \sum_{t=1}^{T} (\bar{\rho}_t(\theta^0, \delta) - \mathbb{E}[\bar{\rho}_t(\theta^0, \delta)]) \xrightarrow{P} 0 \quad \text{and} \\
\frac{1}{T} \sum_{t=1}^{T} (\rho_t(\theta^0, \delta) - \mathbb{E}[\rho_t(\theta^0, \delta)]) \xrightarrow{P} 0,
\]

which shows condition (B). Consequently, the uniform convergence condition (iv) holds by applying the uniform weak law of large numbers given in Theorem A.2.5. in White (1994).

As we have shown that the map \( \theta \mapsto \rho(Y_t, X_t, \theta) \) is Lipschitz-\( L_1 \) in Lemma S.1.1, the map

\[
\theta \mapsto Q_T^0 = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\rho(Y_t, X_t, \theta)]
\]

is also continuous which shows condition (iii). Thus, we can

---

9 Notice that we do not have a double index and thus we suppress the \( n \) in the notation of White (1994). Furthermore, we apply the definition by using the identify function for \( a^3 \).
apply Theorem 2.1. of Newey and McFadden (1994) which concludes the proof of this theorem.

Proof of Theorem 1.4: Let

\[ \psi(Y_t, X_t, \theta) = \left( \begin{array}{c}
-\frac{V_t}{\tau W_t \gamma} (1(Y_t \leq \beta) - \tau) \\
(W_t^{T} \gamma - V_t^{T} \beta + \frac{1}{\tau}(V_t^{T} \beta - Y_t)1(Y_t \leq \beta))
\end{array} \right), \tag{A.4} \]

which is almost surely the derivative of \( \rho(Y_t, X_t, \theta) \) with respect to \( \theta \). We further define \( \Psi_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \psi(Y_t, X_t, \theta) \) and \( \Psi_T^2(\theta) = \mathbb{E}[\psi^2(Y_t, X_t, \theta)] \). From the proof of Lemma S.1.2, we get the mean value expansion (for \( \hat{\theta}_T \) close to \( \theta_T^* \)),

\[ \Psi_T^0(\hat{\theta}_T) - \Psi_T^0(\theta_T^*) = \Delta_T(\hat{\theta}_1, \hat{\theta}_2)(\hat{\theta}_T - \theta_T^*), \tag{A.5} \]

for some values \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) somewhere on the line between \( \hat{\theta}_T \) and \( \theta_T^* \), where the components of \( \Delta_T(\hat{\theta}_1, \hat{\theta}_2) \) are given in Equation (S.1.8) and Equation (S.1.9), and where \( \Psi_T^0(\theta_T^*) = 0. \)

Furthermore, it holds that \( \Delta_T(\theta_T^*, \theta_T^*) = \Delta_T(\hat{\theta}_1, \hat{\theta}_2) \) is a continuous function in its arguments \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \). Using that \( \Delta_T(\hat{\theta}_1, \hat{\theta}_2) \) has Eigenvalues bounded away from zero (for \( T \) large enough), we also get that \( \Delta_T(\hat{\theta}_1, \hat{\theta}_2) \) is non-singular in a neighborhood around \( \theta_T^* \) (for all arguments) for \( T \) large enough as the map which maps the matrix onto its Eigenvalues is continuous. As we further know that \( \hat{\theta}_T - \theta_T^* \to 0 \) and \( ||\hat{\theta}_j - \theta_T^*|| \leq ||\hat{\theta}_T - \theta_T^*|| \) for all \( j = 1, 2 \), we get from the continuous mapping theorem that

\[ \Delta_T^{-1}(\hat{\theta}_1, \hat{\theta}_2) - \Delta_T^{-1}(\theta_T^*) \to 0. \tag{A.6} \]

In the following, we apply Lemma A.1 in Weiss (1991) (by verifying its assumptions), which extends the i.i.d. results of Huber (1967) to strong mixing sequences. Assumption (N1) of Lemma A.1 in Weiss (1991) is satisfied as every almost surely continuous stochastic process is separable in the sense of Doob (Gikhman and Skorokhod, 2004) and the functions \( \psi(Y_t, X_t, \theta) \) are almost surely continuous for all \( t \in \mathbb{N} \). Assumption (N2) is satisfied as shown in the proof of Theorem 1.3. Assumption (N3)(i) is shown in Lemma S.1.2. The technical Assumptions (N3)(ii) and (N3)(iii) follow from Lemma 4 and Lemma 5 in the Supplementary Appendix of Patton, Ziegel, and Chen (2019). For this, notice that the moment conditions in Assumption 2 (C) and (D) of Patton, Ziegel, and Chen (2019) are implied by the condition (A8) in Assumption 1.2 for the simplified case of linear models. Assumption (N4) follows from the moment conditions (A8) in Assumption 1.2 and Assumption (N5) from the strong mixing condition (A6). Furthermore, Lemma 2 in the Supplementary Appendix of Patton, Ziegel, and Chen (2019) implies that \( \sqrt{T}\Psi_T(\hat{\theta}_T) \to 0 \). Thus, we can applyLemma A.1 in Weiss (1991) and get that

10 The mean-value theorem cannot be generalized in a straight-forward fashion to vector-valued functions. Thus, we have to consider the mean value expansion in each component separately which gives this more complicated expression.
\[ \sqrt{T} \Psi_T^0(\hat{\theta}_T) - \sqrt{T} \Psi_T^0(\theta_T^*) \xrightarrow{p} 0. \] (A.7)

Combining Equations (A.5), (A.6), and (A.7), we get that
\[ \sqrt{T}(\hat{\theta}_T - \theta_T^*) = -\Delta_T(\hat{\theta}_1, \hat{\theta}_2)^{-1} \sqrt{T} \Psi_T^0(\hat{\theta}_T) \]
\[ = -\left( \Lambda^{-1}(\theta_T^*) + o_p(1) \right) \cdot \left( \sqrt{T} \Psi_T(\theta_T^*) + o_p(1) \right) \]
\[ = -\Lambda^{-1}(\theta_T^*) \cdot \sqrt{T} \Psi_T(\theta_T^*) + o_p(1). \] (A.8)

Furthermore,
\[ \Sigma_T^{-1/2}(\theta_T^*) \sqrt{T} \Psi_T(\theta_T^*) = \Sigma_T^{-1/2}(\theta_T^*) \sqrt{T} \left( \Psi_T(\theta_T^*) - \Psi_T^0(\theta_T^*) \right) \xrightarrow{d} N(0, I_{2k}), \] (A.11)

by Lemma S.1.3 and thus,
\[ \Sigma_T^{-1/2}(\theta_T^*) \Lambda_T(\theta_T^*) \sqrt{T}(\hat{\theta}_T - \theta_T^*) \xrightarrow{d} N(0, I_{2k}), \] (A.12)

which concludes the proof of this theorem. \( \square \)

Proof of Corollary 1.5: We first notice that
\[ \hat{\Omega}_T^{-1/2} \sqrt{T}(\hat{\theta}_T - \theta_T^*) = \Omega_T^{-1/2} \sqrt{T}(\hat{\theta}_T - \theta_T^*) + (\hat{\Omega}_T^{-1/2} - \Omega_T^{-1/2}) \sqrt{T}(\hat{\theta}_T - \theta_T^*). \] (A.13)

From Theorem 1.4, we get that \( \Omega_T^{-1/2} \sqrt{T}(\hat{\theta}_T - \theta_T^*) \xrightarrow{d} N(0, I_k) \). Furthermore, as \( \hat{\Omega}_T^{-1/2} - \Omega_T^{-1/2} = o_p(1) \) it holds by Slutzky’s theorem, that \( \hat{\Omega}_T^{-1/2} - \Omega_T^{-1/2} \sqrt{T}(\hat{\theta}_T - \theta_T^*) = o_p(1) \) and consequently,
\[ \hat{\Omega}_T^{-1/2} \sqrt{T}(\hat{\theta}_T - \theta_T^*) \xrightarrow{d} N(0, I_k). \] (A.14)

Thus,
\[ T_{A-ESR} = \left( \Omega_T^{-1/2} \sqrt{T}(\hat{\gamma}_T - \gamma_T^*) \right)^\top \left( \Omega_T^{-1/2} \sqrt{T}(\hat{\gamma}_T - \gamma_T^*) \right) \xrightarrow{d} \chi^2_{2}, \] (A.15)

\[ \hat{T}_{S-ESR}(\hat{\gamma}_T^*) = \left( \hat{\Omega}_T^{-1/2} \sqrt{T}(\hat{\gamma}_T^* - \gamma_T^*) \right)^\top \left( \hat{\Omega}_T^{-1/2} \sqrt{T}(\hat{\gamma}_T^* - \gamma_T^*) \right) \xrightarrow{d} \gamma^2_{2}, \] and
\[ \hat{T}_{1-ESR}(\hat{\gamma}_1^*) = \left( \hat{\Omega}_{T,1}^{-1/2} \sqrt{T}(\hat{\gamma}_1^* - \gamma_1^*) \right)^\top \left( \hat{\Omega}_{T,1}^{-1/2} \sqrt{T}(\hat{\gamma}_1^* - \gamma_1^*) \right) \xrightarrow{d} \chi^2_{1}. \] (A.16)

Proof of Corollary 1.6: In the following, we show the result for the Strict ESR test statistic, while equivalent results for the other two ESR tests follow from straightforward simplifications of this proof. Given the alternative hypothesis
\[ H_A^{S-ESR} : \gamma_T^* \neq \gamma_T^{\text{Es}} \quad \forall T \geq T_A \quad \text{for some } T_A \in \mathbb{N}, \] (A.18)

it holds that \( \|\hat{\gamma}_T^* - \gamma_T^{\text{Es}}\| \geq 2\epsilon_A \) for all \( T \geq T_A \) and for some \( \epsilon_A > 0 \). Thus,
\[ \|\hat{\gamma}_T - \gamma_T^{\text{Es}}\| = \|\hat{\gamma}_T - \gamma_T^* + \gamma_T^* - \gamma_T^{\text{Es}}\| \geq \|\gamma_T^* - \gamma_T^{\text{Es}}\| - \|\hat{\gamma}_T - \gamma_T^*\| \geq \epsilon_A > 0, \] (A.19)

with probability approaching one by the inverse triangle inequality and as \( \|\hat{\gamma}_T - \gamma_T^*\| \xrightarrow{p} 0 \) and thus, \( \|\hat{\gamma}_T - \gamma_T^*\| \leq \epsilon_A \) with probability approaching one as \( T \to \infty \). Consequently, for all \( \epsilon \in \mathbb{R} \), it holds that
\[ \mathbb{P}\left( \left\| \sqrt{T_{2,T}^{-1/2} (\hat{\gamma}_T - \gamma_{T0})} \right\| \geq c \right) \to 1, \quad (A.20) \]

and thus,
\[ \mathbb{P}\left( \hat{T}_{S-ESR}(\gamma_{T0}^{E}) \geq c \right) \to 1. \quad (A.21) \]

The proof for \( \hat{T}_{1-ESR} \) follows along the lines and the one of \( T_{A-ESR} \) is a simplified version as we can consider the test parameters (0, 1) instead of the hypothetical pseudo-true parameters under the null \( \gamma_{T0}^{E} \).

Appendix B: Approximation Accuracy of the Misspecified Parameters

In this section, we present a simulation study in order to analyze the accuracy of the approximations of the pseudo-true parameter \( \gamma_{T0}^{E} \) by the tested restriction \( \gamma^{0} = (0, 1) \) under the null. Subsequently, we analyze the approximation of the test statistic \( T_{A-ESR} \) by

\[ c_{t} = \hat{\delta}_t / \hat{\nu}_t, \phi = 0 \]
\[ c_{t} = \hat{\delta}_t / \hat{\nu}_t, \phi = 0.1 \]
\[ c_{t} = \hat{\delta}_t / \hat{\nu}_t, \phi = 0.5 \]

Figure 5. This figure illustrates the effect the misspecified regression equations of the Strict ESR test have on the respective (average) parameter estimates and associated test statistics. The three plot columns of the figure correspond to the different values of the AR-parameter \( \phi \in \{0, 0.1, 0.5\} \), which governs the degree of misspecification. For each column, the first row illustrates the empirical degree of misspecification through a plot of \( c_t = \hat{\delta}_t / \hat{\nu}_t \). The subsequent rows show the estimated densities of the Strict ESR test (solid black lines) and the Auxiliary ESR test (dashed green lines) for the quantile-specific parameters \( \hat{\beta}_T \), the ES-specific parameter \( \hat{\gamma}_T \) and the respective test statistics \( T_{S-ESR} \) and \( T_{A-ESR} \).
For this, we generate 1000 Monte-Carlo replications of the AR-GARCH DGP, given in Equation (2.2) of length \( T = 1000 \) with varying AR parameter \( \phi \in \{0, 0.1, 0.5\} \), and generate optimal VaR and ES forecasts \( \hat{v}_t \) and \( \hat{e}_t \). As already specified in Equation (2.3), for this DGP, the degree of misspecification in the quantile model of the Strict and the Intercept ESR tests is analytically tractable through the magnitude of \( \phi \), where \( \phi = 0 \) represents the case of a correct specification. In contrast, the case \( \phi = 0.5 \), which is highly unrealistic for financial returns, can be seen as a worst-case scenario. For each simulated time series, we estimate the underlying regression equations of the Auxiliary and the Strict ESR tests, given in Equations (1.8) and (1.11) and denote the respective parameter estimates through the superscripts, \( \gamma_{T}^{\text{A-ESR}} \) and \( \gamma_{T}^{\text{S-ESR}} \). Under the null hypothesis, the (average) difference of the estimated parameters \( \gamma_{T}^{\text{A-ESR}} \) and \( \gamma_{T}^{\text{S-ESR}} \) governs the (average) effect the misspecification has on the ES-specific parameters.

Figure 5 illustrates the results, where the different columns of the plots correspond to the different values of \( \phi \). In each column, the first row illustrates the degree of misspecification through a plot of \( c_{t} = \hat{e}_{t}/\hat{v}_{t} \). The subsequent rows illustrate the estimated densities of the Strict and Auxiliary ESR tests for the quantile-specific parameters \( \beta_{T} \), the ES-specific parameter \( \gamma_{T} \) and the respective test statistics \( T_{S-ESR} \) and \( T_{A-ESR} \).

For the quantile-specific parameters \( \beta_{T} \), we can verify the expected misspecified behavior of the quantile regression equation through differing parameter estimates between the Strict and Auxiliary ESR tests. For \( \phi = 0 \), \( c_{t} \) is constant and the quantile regression is correctly specified, however with a slope coefficient generally unequal to 1. For an increasing degree of misspecification, we expectedly find an increasing degree of misspecification in the estimates \( \beta_{T}^{\text{S-ESR}} \) compared to \( \beta_{T}^{\text{A-ESR}} \). In contrast, this effect is almost negligible for the ES-specific parameter estimates \( \gamma_{T}^{\text{S-ESR}} \) compared to \( \gamma_{T}^{\text{S-ESR}} \). This illustrates that the misspecification mainly affects the quantile parameters while leaving the ES-specific parameters almost unchanged, even for the worst-case scenario of \( \phi = 0.5 \). The same approximation behavior can be observed for the associated test statistics reported in the last row of plots. This finding explains the almost identical behavior of the Auxiliary and the Strict ESR tests in the simulation exercises in Section 2.

References


---

11 We use this as a proxy for the approximation of \( T_{S-ESR} = \bar{T}_{S-ESR}(\phi) \) by \( \bar{T}_{S-ESR}(\gamma^{\phi}_{T}) \) as \( \bar{T}_{S-ESR}(\gamma^{\phi}_{T}) \) is unfortunately not observable.


