Dynamic tail risk forecasting: what do realized skewness and kurtosis add?

Giampiero M. Gallo[a,1], Ostap Okhrin[b,c], Giuseppe Storti[d,1,*]

[a]*Corte dei conti, New York University in Florence, and CRENoS*
[b]*Technische Universität Dresden, 01062 Dresden, Germany*
[c]*Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Germany*
[d]*Università di Salerno, Department of Economics and Statistics, Fisciano, Italy*

Abstract

This paper compares the accuracy of tail risk forecasts with a focus on including realized skewness and kurtosis in "additive" and "multiplicative" models. Utilizing a panel of 960 US stocks, we conduct diagnostic tests, employ scoring functions, and implement rolling window forecasting to evaluate the performance of Value at Risk (VaR) and Expected Shortfall (ES) forecasts. Additionally, we examine the impact of the window length on forecast accuracy. We propose model specifications that incorporate realized skewness and kurtosis for enhanced precision. Our findings provide insights into the importance of considering skewness and kurtosis in tail risk modeling, contributing to the existing literature and offering practical implications for risk practitioners and researchers.

Keywords: Value at Risk, CAViaR, Expected Shortfall, Realized Skewness, Realized Kurtosis.

1. Introduction

Starting approximately thirty years ago, the issue of capital adequacy has received increased attention, with significant impetus given to supervisory and regulatory functions to closely monitor the impact of volatility and interconnectedness on financial institution portfolios. Modern risk management is based on the principle that increased risks must be adequately covered with sufficient resources to avoid liquidity crises or defaults that could affect other institutions and the financial system as a whole. The consequences of the 2007-2008 financial crisis underscored the need for suitable capital risk measures exhibiting forecastability over relevant time horizons.

The various recommendations of the Basel Committee on Banking Supervision regarding capital risk regulations emphasize that the main parameters of a conditional distribution of returns to be

---

*Corresponding author

Email addresses: `giampiero.gallo@nyu.edu` (Giampiero M. Gallo), `ostap.okhrin@tu-dresden.de` (Ostap Okhrin), `storti@unisa.it` (Giuseppe Storti)

[1]Opinions expressed here are personal and do not involve the Corte dei conti.

monitored are some position index, specifically the threshold (Value at Risk, VaR, Jorion, 1997) corresponding to a certain probability in the tail where losses occur, and the average value of the loss once that threshold has been surpassed (Expected Shortfall, ES, Artzner et al., 1999). In this context, without loss of generality, we assume that the tail in question is the left tail, representing losses in long positions.

Market activity, characterized by price and volume movements in response to news, renders the conditional distribution of returns non-constant over time. Consequently, both Value at Risk (VaR) and Expected Shortfall (ES) become time-varying risk measures. Moreover, observed persistence in market behavior suggests dynamics that leverage valuable past information. From an econometric perspective, it is challenging to determine which features of past market behavior are relevant for predicting VaR and ES, as these measures represent conditional quantiles and expectations, respectively, in the tail of the asset return distribution.

Approaches to address this issue can broadly be categorized into three main groups. The first category assumes a known parametric distribution for returns, typically a Student-t distribution, and focuses on the dynamic evolution of the conditional variance of returns. This approach augments the fixed quantile identification with a GARCH process that models the dependence of conditional variance on recent returns and past estimates. Parameters are estimated using (Quasi) Maximum Likelihood (QML) methods. At the opposite end of the spectrum, parametric assumptions about the return distribution or its dynamics are entirely discarded. So-called historical simulation methods are employed, where future outcomes are simulated by repeating observed past behaviors.

A third stream adopts an intermediate stance, focusing on the dynamics of the risk measure of interest while limiting or avoiding reliance on parametric assumptions about the shape of the conditional return distribution. This semi-parametric approach to financial risk modeling is gaining popularity due to its flexibility and often demonstrates competitive performance compared to more complex parametric models. In what follows, we will position ourselves in this stream of literature, addressing, in particular, the role that higher-order conditional moments, notably skewness and kurtosis, have on the refinement of predictions, hence highlighting the role of the time-varying evolution of asymmetry and tail density of the return distribution in sharpening the projections of VaR and ES.

Our synthesis in this field is to identify two main categories of semi-parametric modeling approaches for tail risk measures. The first is called the "additive" approach, which utilizes linearized representations of GARCH models, such as in CAViaR models. The second approach, referred to as the "multiplicative" approach, involves estimating GARCH-type models via the minimization of a properly defined strictly consistent scoring function. We consider the recent literature (e.g., Neuberger, 2012; Neuberger and Payne, 2021; Bae and Lee, 2021) on the derivation of realized measures of skewness and kurtosis as consistent estimates of the conditional skewness and kurtosis of daily returns. For our purposes, these additional features of the conditional distributions may be relevant when included in the specifications. Given that Amaya et al. (2015) provides evidence that realized skewness and kurtosis are useful when forecasting the cross-section distribution of equity returns, our interest here is to assess whether these benefits extend to risk forecasting as well.

Although the additive approach has gained popularity, there is still a lack of extensive forecast-

ing comparison between these two methodologies. Hence, we aim to bridge a gap in the literature by proposing an application that evaluates the accuracy of forecasts generated by additive and multiplicative modeling strategies for a panel of 960 US stocks. To achieve this, we employ various diagnostic tests and scoring functions for both VaR and ES forecasts. Additionally, we investigate the impact of window length on forecasting accuracy, a critical issue for practitioners. Short windows tend to minimize bias but increase variability in risk forecasts, while long windows have the opposite effect: hence, we conduct a rolling window estimation/forecasting exercise and evaluate the performance of three different window lengths, 500, 1000, and 2000 days.

Our novel model specifications using information on realized higher-order moments to forecast tail risk measures are both regression quantile time series models for forecasting VaR, as well as bivariate semi-parametric models for joint VaR and ES forecasting: we are interested in providing specific evidence on the relevance of the realized skewness and kurtosis via Wald-type tests, but also on their contribution in improving the forecast performance, assessed with standard backtesting procedures. Their predictive performances are compared to those of competitors that do not include such information.

In a nutshell, the evidence on the vast panel of stock indicates that multiplicative models are preferred to additive ones, and that the extension to higher moments does not buy a generalized relevant improvement in the outcome. In general, simpler models are to be preferred to more complex ones.

The structure of the paper is as follows. In Section 2, we propose our models in Subsection 2.1, while the related estimation procedures and some properties of the estimators are illustrated in Subsection 2.2. In Section 3, we present a recent literature review on realized estimators of skewness and kurtosis of financial returns. Section 4 is dedicated to the empirical application, while Section 5 concludes.

2. Semi-parametric risk modeling

The literature on semi-parametric risk modeling features a seminal paper by Engle and Manganelli (2004), who introduced the Conditional Autoregressive Value-at-Risk (CAViaR) model for forecasting VaR. This model has interesting connections with both quantile regression and GARCH models, in that the CAViaR model can be viewed as a quantile autoregression with a recursive term. By the same token, a linear GARCH model of a given order can be represented as a CAViaR model of the same order. Building on the duality between GARCH and CAViaR, Xiao and Koenker (2009) present an original approach to estimating parameters of a GARCH model, proposing to minimize the typical quantile loss function used in quantile regression models.

Direct semi-parametric modeling of ES is not feasible because, unlike VaR, ES is not elicitable relative to a given loss function. However, Fissler et al. (2015) have derived a class of loss functions that are strictly consistent for the pair (VaR, ES), in the sense that the expected loss is minimized by the true (VaR, ES). Within this framework Taylor (2020) proposes a class of semi-parametric models for (VaR, ES), augmenting the standard CAViaR setup with an additional dynamic equation for ES, and replacing the usual quantile loss with a member of the Fissler-Ziegel (FZ) class. In particular, among the available choices, Taylor (2019) considers a loss, or

scoring, function based on the Asymmetric Laplace quasi-likelihood function, AL for short. Patton et al. (2019) extend the work by Taylor (2020) in two different directions. First, they consider time-varying semi-parametric (VaR, ES) models based on the Generalized Autoregressive Score (GAS) framework (Creal et al., 2013). Second, as done by Xiao and Koenker (2009) for VaR, they consider directly estimating GARCH models minimizing a specific strictly consistent loss function in the FZ class called FZ0 (owing its denomination to the fact that, when using this loss to compare two models, it yields loss differentials that are homogeneous of degree zero). This property can lead to a higher power in Diebold-Mariano tests (Diebold and Mariano, 2002).

2.1. The model setup

Let r_t be the log-return for the day t, for $t = 1, \ldots, T$, and $Q_{\alpha,t} = F_r^{-1}(\alpha | \mathcal{I}_{t-1})$ indicate the conditional α-quantile of r_t (level-α Value-at-Risk –VaR), with F_r being the cdf of r_t; correspondingly, $ES_{\alpha,t} = E(r_t | r_t < Q_{\alpha,t}, \mathcal{I}_{t-1})$ indicates the conditional α-tail expectation of r_t, given past information $\mathcal{I}_{t-1}$ (level-α Expected Shortfall – ES).

We let RV_t, Sk_t, and Ku_t denote, respectively, the conditional variance, skewness, and kurtosis of daily returns r_t as follows

$$RV_t = E_0\{(r_t - \mu_{1t})^2 | \mathcal{I}_{t-1}\} = \mu_{2t} - \mu_{1t}^2,$$

$$Sk_t = E_0\left\{\left(\frac{r_t - \mu_{1t}}{RV_t^{1/2}}\right)^3 | \mathcal{I}_{t-1}\right\} = \frac{\mu_{3t} - \mu_{1t}\mu_{2t} + 2\mu_{1t}^3}{(\mu_{2t} - \mu_{1t}^2)^{3/2}},$$

$$Ku_t = E_0\left\{\left(\frac{r_t - \mu_{1t}}{RV_t^{1/2}}\right)^4 | \mathcal{I}_{t-1}\right\} = \frac{\mu_{4t} - 4\mu_{3t}\mu_{1t} + 6\mu_{2t}\mu_{1t}^2 - 3\mu_{1t}^4}{(\mu_{2t} - \mu_{1t}^2)^2},$$

where $\mu_{kt} = E_0(r_t^k | \mathcal{I}_{t-1})$ indicates the k-th conditional noncentered moment under the true measure. Estimates of these quantities can be readily obtained by replacing the involved conditional moments μ_{kt} with their estimated counterparts, at least using daily observations. In Section 3, we will formally address the estimation of μ_{kt} for $1 \leq k \leq 4$.

We can now present the two alternative modeling frameworks under which VaR and ES forecasts are generated, denoted as, for ease of reference, the *additive* and the *multiplicative* models, respectively. We simplify the notation by defining $v_t \equiv Q_{\alpha,t}$ and $e_t \equiv ES_{\alpha,t}$. Thus, the *additive modeling framework* can be represented as a regression model for the 1-step ahead expected α-level of VaR, v_t:

$$r_t = v_t + \eta_t,$$

where the error term η_t is controlling the left tail of the conditional distribution of returns, so that, under correct specification of v_t, the error term η_t is such that $F_\eta^{-1}(\alpha | \mathcal{I}_{t-1}) = 0$.

This is a general framework since several models can be derived as special cases by varying the dynamic specifications for v_t. Noting that a $\widehat{\cdot}$ is used to indicate an estimate, $\bar{r} = T^{-1} \sum_{t=1}^{T} r_t$ and $\widehat{Sk}_t^+$ and $\widehat{Sk}_t^-$, represent negative and positive skewness:

$$\widehat{Sk}_{1t}^+ = \widehat{Sk}_{1t} \cdot I\{\widehat{Sk}_{1t} > 0\}, \qquad \widehat{Sk}_{1t}^- = -\widehat{Sk}_{1t} \cdot I\{\widehat{Sk}_{1t} < 0\},$$

4

in what follows, we investigate three specifications.

The first is the simple additive form of the VaR being driven only by the lagged observation of an estimator of the integrated volatility $(\widehat{RV}_{t-1})$[2].

$$\texttt{add\_sim:}\ v_t = d_0 + d_1\widehat{RV}_{t-1}^{1/2} + d_2 v_{t-1}, \tag{1}$$

To account for the potential misspecification in $\texttt{add\_sim}$, we can resort to the Cornish-Fisher (CF) expansion (Hill and Davis, 1968), which approximates the quantiles of an unknown non-Gaussian distribution using the information on sample skewness and kurtosis to adjust the value of the corresponding Gaussian quantiles. Considering as an illustration a random variable $X \sim (0, 1)$, the CF approximation for the α-quantile of X reads as

$$X_\alpha^{CF} = z_\alpha + \frac{z_\alpha^2 - 1}{6}Sk + \frac{z_\alpha^3 - 3z_\alpha}{2}Ku - \frac{2z_\alpha^3 - 5z_\alpha}{36}Sk^2,$$

where Sk and Ku are the usual moment-based sample skewness and kurtosis coefficients of X respectively, and $z_\alpha = \Phi^{-1}(\alpha)$ is the α-quantile of a $N(0, 1)$ random variable.

Therefore, the second model adds realized negative and positive skewnesses (Sk_{t-1}^- and Sk_{t-1}^+) and kurtosis (Ku_{t-1}) to the $\texttt{add\_sim}$[3]:

$$\texttt{add\_skk:}\ v_t = d_0 + d_1\widehat{RV}_{t-1}^{1/2} + d_2 v_{t-1} + (a_1\widehat{Sk}_{t-1}^- + a_2\widehat{Sk}_{t-1}^+ + a_3\widehat{Ku}_{t-1}). \tag{2}$$

The inclusion of the skewness and kurtosis terms are thus motivated by a data-driven CF expansion, whose coefficients, as it will be later illustrated, can be estimated in a semi-parametric fashion by minimizing a strictly consistent loss function.

The third model further extends the $\texttt{add\_skk}$ with an asymmetric impact of the integrated volatility in correspondence with returns smaller than their average (leverage effect):

$$\texttt{add\_lev:}\ v_t = d_0 + d_1\widehat{RV}_{t-1}^{1/2} + d_2 v_{t-1} + (a_1\widehat{Sk}_{t-1}^- + a_2\widehat{Sk}_{t-1}^+ + a_3\widehat{Ku}_{t-1}) + d_3\widehat{RV}_{t-1}^{1/2}I\{r_{t-1} \le \bar{r}\}. \tag{3}$$

By contrast, a *multiplicative modeling framework* can be represented in terms of the following nonlinear regression model

$$r_t = v_t \eta_t,$$

where, under correct specification of v_t, η_t is such that $F_\eta^{-1}(\alpha|\mathcal{I}_{t-1}) = 1$.

This framework can also be motivated by a simple location-scale representation of the returns process

$$r_t = h_t z_t, \qquad z_t \stackrel{iid}{\sim} (0, 1),$$

---

[2]In the absence of jumps, the integrated variance coincides with the quadratic variation that, in turn, diverges from the conditional variance RV_t by a zero mean error, thus motivating our notation (Andersen et al., 2001). A set of alternative choices for $\widehat{RV}_{t-1}$ will be presented and discussed in Section 3.

[3]In our approach, we focus on the conditional distribution of returns rather than on their unconditional distribution, as would happen when using the standard CF expansion. Hence, the sample skewness and kurtosis coefficients are replaced by their realized counterparts that provide point estimates of daily conditional skewness and kurtosis.

where the dynamics of $h_t^2 = \mathsf{Var}(r_t|\mathcal{I}_{t-1})$ can be modelled by means of GARCH type models. Under the iid assumption for z_t the 1-step ahead α-level VaR of r_t is given by $v_t = h_t z_\alpha$ where $z_\alpha = F_z^{-1}(\alpha)$. Thus, the first multiplicative model is for the α-level 1-step ahead VaRis

$$\texttt{mlt\_sim:}\ v_t = h_t, \quad h_t^2 = d_0 + d_1 \widehat{\mathrm{RV}}_{t-1} + d_2 h_{t|t-1}^2. \tag{4}$$

When we allow for a time-varying conditional skewness and kurtosis in the returns distribution, this assumption must be generalized to read

$$v_t = h_t z_{\alpha,t},$$

where the time variation in the conditional error quantile $z_{\alpha,t}$ is driven by the time-varying conditional skewness and kurtosis values as in the $\texttt{mlt\_lev}$ and $\texttt{mlt\_skk}$ specifications introduced below. Thus, within the multiplicative modeling framework, we consider the following alternative specifications for the α-level 1-step ahead VaR

$$\texttt{mlt\_skk:}\ v_t = h_t(a_1 \mathrm{Sk}_{t-1}^- + a_2 \mathrm{Sk}_{t-1}^+ + a_3 \mathrm{Ku}_{t-1}), \quad h_t^2 = d_0 + d_1 \widehat{\mathrm{RV}}_{t-1} + d_2 h_{t|t-1}^2, \tag{5}$$

$$\texttt{mlt\_lev:}\ v_t = h_t(a_1 \mathrm{Sk}_{t-1}^- + a_2 \mathrm{Sk}_{t-1}^+ + a_3 \mathrm{Ku}_{t-1}), \quad h_t^2 = d_0 + d_1 \widehat{\mathrm{RV}}_{t-1} + d_2 h_{t|t-1}^2 + d_3 \widehat{\mathrm{RV}}_{t-1} I\{r_{t-1} \le \bar{r}\}. \tag{6}$$

Multiplicative models closely mirror additive models (1), (2) and (3). The $\texttt{mlt\_sim}$ is similar to (1) and is the simplest specification with only the integrated volatility driving the dynamics of the scale. Further $\texttt{mlt\_skk}$ assumes similar to (2) in the additional information incorporated in realized skewness and kurtosis that drives the dynamics of the scale. The most complex model $\texttt{mlt\_lev}$ also controls for the leverage in h_t in the same fashion as in the model (3).

For both additive and multiplicative frameworks, the ES can be modeled according to two different alternative specifications:

$$\texttt{ES\_sim:}\quad e_t = \{1 + \exp(b_0)\}v_t, \tag{7}$$

$$\texttt{ES\_skk:}\quad e_t = \{1 + \exp(b_0 + b_1 \mathrm{Sk}_{t-1} + b_2 \mathrm{Ku}_{t-1})\}v_t. \tag{8}$$

Here $\texttt{ES\_sim}$ is the simple specification assuming that the ES is a rescaling of VaR. Taylor (2020) shows that this simple specification provides competitive VaR forecasts. More recently, Wang et al. (2023) have extended the framework proposed in Taylor (2020) to allow for separate VaR and ES dynamics as well as for the incorporation of realized measures.

The more complex $\texttt{ES\_skk}$ brings the dynamics of the ES to be also driven by the skewness and kurtosis, possibly accounting for the misspecification of $\texttt{ES\_sim}$. Differently from VaR, in this case, we did not split the skewness into negative and positive. By construction, both $\texttt{ES\_skk}$ and $\texttt{ES\_sim}$ specifications avoid the crossing of VaR and ES forecasts.

In the additive case, neglecting to model the ES dynamics leads to a pure VaR model. This case, labeled as $\texttt{ES\_no}$, corresponds to a model specification that is close in spirit to a Engle and Manganelli (2004) CAViaR type one where some realized estimator of the integrated variance replaces the volatility measure based on lagged daily returns.

2.2. Estimation

Estimation of the vector $\boldsymbol{\theta}$ of unknown parameters describing the models for v_t and e_t both in the additive and multiplicative models (1)-(8) is done semi-parametrically by minimizing a strictly consistent scoring rule,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} S_t^{(\alpha)}, \tag{9}$$

where $S_t^{(\alpha)}$ is a member of the general class presented by Fissler and Ziegel (2015), i.e.

$$S_t^{(\alpha)} \equiv S(v_t, e_t | r_t; \alpha) = \{I(r_t \le v_t) - \alpha\}G_1(v_t) - I(r_t \le v_t)G_1(r_t) + G_2(e_t)$$
$$\left\{e_t - v_t + I(r_t \le v_t)\frac{v_t - r_t}{\alpha}\right\} - \zeta_2(e_t) + a(r_t).$$

In the definition of $S_t^{(\alpha)}$, the functions G_1, ζ_2, and G_2 satisfy the following conditions: G_1 is increasing, ζ_2 is increasing and convex, and $G_2 = \zeta_2'$. In particular, setting $G_1(\cdot) = 0$, $G_2(x) = -1/x$, $\zeta_2(x) = -\log(-x)$, $a(r_t) = 1 - \log(1 - \alpha)$, leads to the following scoring rule (Taylor, 2020)

$$AL_t^{(\alpha)} = \frac{I(r_t \le v_t)r_t + v_t\{\alpha - I(r_t \le v_t)\}}{\alpha e_t} + \log(-e_t) - \log(1 - \alpha)$$
$$= \frac{I(r_t \le v_t)r_t + v_t\{\alpha - I(r_t \le v_t)\}}{\alpha e_t} - \log\left(\frac{1 - \alpha}{e_t}\right).$$

Adding and subtracting αr_t to the numerator of the second term on the right-hand side of the previous equation, we get

$$AL_t^{(\alpha)} = -\log\left(\frac{\alpha - 1}{e_t}\right) - \frac{(r_t - v_t)\{\alpha - I(r_t \le v_t)\}}{\alpha e_t} + \frac{r_t}{e_t}.$$

In the simplified expression of $AL_t^{(\alpha)}$ obtained by Taylor (2020), the last term on the right-hand side is dropped. This simplification arises from the assumption that the conditional mean of returns is zero, as shown in their equation (19). Notably, it can be demonstrated that the negative value of $AL_t^{(\alpha)}$ quantifies the contribution of the t-observation to a quasi-likelihood function, which is constructed based on the Asymmetric Laplace distribution (Taylor, 2020).

With the same choices for G_1, G_2, and ζ_2 as above, but setting $a(r_t) = 0$, leads to the zero-degree homogeneous loss used by Patton et al. (2019)

$$FZ0_t^{(\alpha)} = \frac{I(r_t \le v_t)r_t + v_t\{\alpha - I(r_t \le v_t)\}}{\alpha e_t} + \log(-e_t) - 1 = -\frac{I(r_t \le v_t)(v_t - r_t)}{\alpha e_t} + \frac{v_t}{e_t} + \log(-e_t) - 1.$$

Therefore, in the case of the model EM_no, where only the VaR is estimated, the objective function is given by the quantile loss:

$$EM_t^{(\alpha)} = \{\alpha - I(r_t < v_t)\} \cdot (r_t - v_t). \tag{10}$$

In what follows, models estimated relying on loss functions $AL_t^{(\alpha)}$, $FZ0_t^{(\alpha)}$ and $EM_t^{(\alpha)}$ are labeled as `Loss=ALS`, `Loss=FZ0` and `Loss=EM`, respectively. Standard errors are computed using the asymptotic theory developed by Engle and Manganelli (2004), for pure VaR models, and Patton et al. (2019), for joint VaR-ES models. Technical details are provided in Section 5.

It is worth noting that optimization in the (9) is a challenging task irrespective of what function $S_t^{(\alpha)}$ is chosen, be it either $AL_t^{(\alpha)}$, $FZ0_t^{(\alpha)}$ or $EM_t^{(\alpha)}$. In particular, the optimization of these loss functions is typically strongly dependent upon the chosen initial values. For this reason, we implemented an optimization technique similar to Engle and Manganelli (2004) which, for ease of reference, we call *complete estimation*. Namely, for each model, we evaluated the objective function on $\mathfrak{n} = 5 \cdot 10^4$ uniformly sampled possible parameter constellations, and among them, we selected the $\mathfrak{m} = 10$ parameter vectors that lead to the smallest objective function values. Selecting each of these $\mathfrak{m}$ vectors as a starting point, we re-estimated the model $\mathfrak{m}$-times iterating between a Nelder-Mead and a BFGS optimizer until convergence is achieved, and the final estimates are those delivering the smallest value of the objective function. In a rolling window forecasting exercise, one may be advised to follow a parsimonious estimation strategy, by using the most recent estimates as the starting point for the next estimation round, at regular intervals.

3. The underlying process and the derived realized measures

Having developed a setup where the theoretical estimators of conditional moments such as RV, Sk, and Ku are considered within suitable models, we are left with the delicate phase to choose which operational counterparts we can count on at daily frequencies, employing rolling windows and sample statistics. The standard framework starts from a true underlying continuous log-price following a diffusion process, disregarding, for example, the presence of structural breaks:

$$\mathrm{d}X_\mathfrak{t} = \mu(X_\mathfrak{t})\mathrm{d}\mathfrak{t} + \sigma(X_\mathfrak{t})\mathrm{d}W_\mathfrak{t},$$

where, $W_\mathfrak{t}$ represents the standard Brownian motion, $\mu(X_\mathfrak{t})$ is the drift càdlàg finite variation process, and $\sigma(X_\mathfrak{t})$ is the time-varying càdlàg volatility function. It is important to note that $\sigma(X_\mathfrak{t})$ may depend on a separate Brownian motion, which could potentially be correlated with $W_\mathfrak{t}$. This general family encompasses well-known processes such as the Heston or Bates processes (see Heston (1993); Bates (1996)). In this context, the parameter $\mathfrak{t}$ represents the continuous temporal component that spans within and across days.

The second moment μ_{2t} is known as the *integrated variance*, an object of paramount importance to researchers and practitioners. By utilizing the aforementioned process over a one-day interval $[t - 1d, t]$, the integrated variance can be computed as $\int_{t-1d}^{t} \sigma^2(u)\mathrm{d}u$.

The temporal component then needs to be somehow aggregated to get the daily estimators for the relevant moments: in this respect, we ground ourselves in the massive literature that considers the market activity of a day (using the same index $t \in \{1, \dots, T\}$) between opening and closing to be divided into regularly spaced intervals $i \in \{0, \dots, N\}$. We then take the high-frequency log-prices $x_{t,i}$ as the elementary information, to be converted into $r_{t,i} = x_{t,i} - x_{t,i-1}$, the corresponding *intraday* log-returns, $i = 1, \dots, N$.

The overwhelming attention of the literature was devoted to the design of consistent estimators of the integrated variance μ_{2t} of the continuous process over a discrete interval (Andersen

8

et al., 2010), with specific care devoted to departures from the standard framework (e.g. jumps) or to the nature of observed prices which are affected by trading mechanisms (so-called market microstructure). There exists a range of options for researchers and practitioners alike seeking to estimate these quantities accurately and efficiently. Starting from the realized variance (Andersen and Bollerslev, 1998),

$$\widehat{\mu}_{2t}^{RV} = \sum_{i=1}^{N} r_{t,i}^2,$$

other widely used estimators of the integrated volatility are the, proposed in Barndorff-Nielsen and Shephard (2004) and Andersen et al. (2012), bipower variation $\widehat{\mu}_{2t}^{BPV} = \frac{\pi}{2} \frac{N}{N-1} \sum_{i=1}^{N-1} |r_{t,i}||r_{t,i+1}|$, or the upside and downside semivariances $\widehat{\mu}_{2t}^{SVPOS} = \sum_{i=1}^{N} r_{t,i}^2 \cdot I\{r_{t,i} > 0\}$ and $\widehat{\mu}_{2t}^{SVNEG} = \sum_{i=1}^{N} r_{t,i}^2 \cdot I\{r_{t,i} < 0\}$ developed in Barndorff-Nielsen et al. (2008) and Bollerslev et al. (2020).

Barring a horse race among the many estimators of μ_{2t} available, we limit ourselves to a single choice, and our preference goes to the median estimator

$$\widehat{\mu}_{2t}^{MED} = \frac{\pi}{6 - 4\sqrt{3} + \pi} \frac{N}{N-2} \sum_{i=2}^{N-1} \mathrm{med}(|r_{t,i-1}|, |r_{t,i}|, |r_{t,i+1}|)^2,$$

proposed by Andersen et al. (2012), because of its documented robustness properties.

Several new estimators for *higher-order* moments have emerged in recent years. While these estimators do not directly estimate daily skewness or kurtosis, they instead estimate the integrated third or fourth power of intraday returns or the averaged jump component. Empirical evidence suggests that these estimators can be informative in predicting cross-sectional next week's stock returns or in forecasting RV at medium- to long-term horizons, as demonstrated by Mei et al. (2017). The simplest estimator of the integrated k-order moments was proposed by Amaya et al. (2015), shadowing the relationship between the realized variance and the integrated volatility (case $k = 2$).

$$\widehat{\mu}_{kt}^{ACJV} = \sum_{i=1}^{N} r_{t,i}^k.$$

Amaya et al. (2015) demonstrate the estimator's consistency, which asymptotically captures only the jump component and the average jump size but does not capture skewness arising from the leverage effect and heavily depends on the sampling frequency. Later Liu et al. (2014) derived asymptotic properties of the Amaya et al. (2015) estimator and developed their own measures of realized skewness accounting for market microstructure noise. Another extension was provided by Choe and Lee (2014), who showed that the daily third moment is proportional to the quadratic co-variation between the squared return and the return process, and the fourth moment is proportional to the quadratic variation of the squared return process with some additional cross-terms.

Based on some preliminary analysis, our choice for the realized skewness and kurtosis falls on the estimators by Neuberger (2012) and Neuberger and Payne (2021):

$$\widehat{\mu}_{3t}^{NP} = \frac{1}{\tau} \sum_{j=0}^{\tau-1} \sum_{i=1}^{N} \left(r_{t-j,i}^3 + 3y_{t-j,i-1}^* r_{t,i}^2 \right), \quad \widehat{\mu}_{4t}^{NP} = \frac{1}{\tau} \sum_{j=0}^{\tau-1} \sum_{i=1}^{N} \left(r_{t-j,i}^4 + 4y_{t-j,i-1}^* r_{t-j,i}^3 + 6z_{t-j,i-1}^* r_{t-j,i}^2 \right),$$

where $y_{t,i-1}^* = \frac{1}{N} \sum_{j=1}^{N} (x_{t,i-1} - x_{t,i-j})$ and $z_{t,i-1}^* = \frac{1}{N} \sum_{j=1}^{N} (x_{t,i-1} - x_{t,i-j})^2$ measure local (daily) trends in simple and squared log-prices. Similar to Choe and Lee (2014), they assume that the conditional mean of the returns is zero. Also, in what follows, we choose $\tau = 5$.

We note that the estimators for realized skewness and kurtosis sometimes produce outliers that can significantly affect the performance of VaR and ES models. To address this issue, we applied a filter that removes estimated skewness and kurtosis values falling outside the ranges of $(-15; 15)$ and $(0; 20)$, respectively. Outliers excluded from our analysis are then smoothed out using interpolation techniques accounting for autocorrelation.

4. Empirical evidence

4.1. Data and forecasting design

In this section, we present the results of our setup to a very large panel of 960 U.S. stocks traded on the New York Stock Exchange (NYSE), included in the S&P500 index at various times over the considered period. The list of stocks can be found in Web Appendix *List of Tickers*. The original dataset for each stock consists of intra-daily prices adjusted for stock splits and dividends sampled every 5 minutes. We focus only on regular trading hours, from 9:30 am to 4:00 pm, resulting in 78 observations for each trading day. The stocks have different timespans, starting within a range between 1998-01-02 and 2016-10-11, and ending between 1998-01-09 and 2017-02-09. Furthermore, to ensure an adequate sample size, we limit our analysis to assets with a continuous record of at least 500 daily observations. This reduces the cross-sectional size of our sample to 823 assets (marked in the Web Appendix *List of Tickers* in italics).

Our empirical strategy consists of two main steps. In the first, we conduct a full-sample analysis to assess the performance of various models in fitting VaR and ES. In the second step, we focus on the out-of-sample forecasting performance using a rolling window approach. We consider three estimation windows: 500, 1000, and 2000 days. For the out-of-sample analysis, we include assets with a continuous record of daily pricing observations from the start date of our sample to its end, 2017-02-09. In this case, we were able to obtain one-step ahead predictions for the dates 2000-01-04 – 2017-02-09, for $w = 500$, 2002-01-03 – 2017-02-09 for $w = 1000$, and 2005-12-21 – 2017-02-09 for $w = 2000$. 406 of the original 960 stocks meet this criterion and are included in the out-of-sample analysis (marked in boldface in the Web Appendix *List of Tickers*).

The model universe considered for both the full-sample and out-of-sample analysis includes all the specifications presented in Section 2. As discussed, each of these is coupled with three different ES specifications, ES_sim, ES_sk, and ES_no, for a total of 18 different models. Implementing the procedures discussed in Section 2.2, each model is estimated for three different risk levels, $\alpha \in \{0.01, 0.025, 0.05\}$.

4.2. Analysis of the in- and out-of-sample losses and coverage

Before delving into the assessment of the model performances through the various tests conducted on the extensive dataset, it is first useful to visually assess their in- and out-of-sample coverage. Figure D.1 presents a comprehensive overview of the aggregated information across all datasets, three coverage levels, and all models for in-sample performance. Each model is estimated

10

for every dataset, and the in-sample empirical coverage ($\hat{\alpha}$) is calculated. The models are represented by row blocks in the figure, with corresponding names on the y-axis, such as "VaR=m_lev, ES=no, Loss=EM". Within each block, three box plots display the coverage for all datasets, with blue indicating $\alpha = 0.01$, green representing $\alpha = 0.025$, and red representing $\alpha = 0.05$. These levels are also depicted by vertical dashed lines. All the models exhibit similar behavior and, on average, achieve the desired coverage level, albeit with slight variations. Simpler models generally exhibit less variability. Some cases encountered convergence difficulties, leading to the inability to estimate certain models. The right panel of Figure D.1 shows the fraction of such problematic cases, consistently below 3%.

A similar analysis has been conducted for out-of-sample coverage, utilizing three different window sizes of 500, 1000, and 2000 days. Aggregated results are presented in Figure D.2. In addition to the three colors representing coverage levels ($\alpha = 0.01$, $\alpha = 0.025$, and $\alpha = 0.05$), varying color intensities indicate window size (lightest shade = 2000 days; darkest = 500 days).[4] The out-of-sample results reveal a less favorable situation than the in-sample, as all models tend to overestimate the coverage on average, less severely so with a wider rolling window. Surprisingly, the variance also increases in this scenario. This can be attributed to longer intervals containing more diverse data from potentially different underlying models, thus imperfectly capturing future behavior. Despite these nuances, all models demonstrate similar behavior based on simple visual inspection. Additionally, Figure D.3 provides aggregated loss information. It is evident that both the values and spreads of the loss function decrease with larger sample sizes.

4.3. Evaluation metrics

Following the practice by researchers and risk managers, the in- and out-of-sample performances of the dynamic models[5] for VaR and ES presented in Section 2 are firstly assessed using some diagnostic tests, whose technical details are summarized in Appendix B.2-B.6 for the reader's convenience.

To assess the in-sample VaR estimation performance, we consider the in-sample Dynamic Quantile (DQ) test by Engle and Manganelli (2004). In particular, we consider the test in its conditional coverage and independence versions as described in Dumitrescu et al. (2012). The asymptotic theory for these tests was originally derived by Engle and Manganelli (2004) for the pure CAViaR models. Therefore, the results presented hereafter refer to the ES_no case only, involving CAViaR models estimated by minimizing the aggregated quantile loss.

While the in-sample DQ test assesses the goodness-of-fit of CAViaR models, its out-of-sample counterpart can be seen as a general test for evaluating the statistical properties of a set of VaR

---

[4]Fewer models are considered in the out-of-sample analysis, excluding "ES=SkKu" due to computational complexity and "Loss=FZ0" due to its similar behavior to "Loss=ALS".

[5]It should be noted that performing a complete estimation for all rolling windows in the out-of-sample exercise has been highly time-consuming. As a result, we perform a full estimation for the initial window and subsequently at 500-day time intervals, and, instead of repeating the complete estimation process for each subsequent window, we update the parameters at regular intervals. To accomplish this, we perform parameter optimization every 50 observations, starting from the results obtained in the previous step. This parameter updating allows us to refine the estimation without repeating the entire process. In all other rolling windows, we maintain the parameters obtained from the previous window.

11

forecasts, regardless of the model. This includes testing for unbiasedness, independent hits, and the independence of quantile estimates, as outlined by Engle and Manganelli (2004). In our case, the out-of-sample DQ (OOS-DQ) test can effectively evaluate the properties of the VaR forecasts generated by joint dynamic VaR-ES models.

Further, we jointly assess the statistical accuracy of VaR and ES estimates, both in and out-of-sample, employing two regression-based testing procedures, i.e., the regression-based calibration tests by Patton et al. (2019), henceforth PZC, and the ESR test by Bayer and Dimitriadis (2022). The former includes separate calibration tests for VaR and ES while the latter test is specific for ES diagnostics. Moreover, the PZC tests are based on OLS auxiliary regression equations where the standardized generalized residuals (as in Patton et al., 2019) are regressed on their past values as well as on VaR and ES forecasts, respectively.

The ESR approach, instead, is based on three separate test statistics: the Auxiliary, the Strict and the Strict Intercept Backtest, which can be seen as an extension of Mincer-Zarnowitz regression to a semi-parametric setting, relying on the minimization of the consistent loss functions proposed by Fissler and Ziegel (2015). The Auxiliary and Strict test statistics are computed regressing returns on the ES forecasts and test the ES coefficients for joint (0, 1) values. Specifically, the Auxiliary test requires an auxiliary VaR forecast, while the Strict Intercept tests whether the expected shortfall of the forecast error ($r_t - ES_t$) is zero.[6]

Finally, the out-of-sample forecasting accuracy of each of the models is assessed by comparing the average values of the FZ0 loss achieved over the forecasting period.

4.4. Testing the properties of the risk estimates

In this section, we analyze the properties of the in-sample risk estimates over the full-sample for the set of 823 assets, postponing to a later subsection the generation of out-of-sample risk forecasts. Also, for the sake of brevity, we only discuss results for $\alpha = 0.01$, while the results for $\alpha = 0.025$ and $\alpha = 0.05$ are contained in the Web Appendix *Tables W.1 and W.2*.

The *In-sample DQ* section of Table C.1 reports the results of the in-sample DQ test in its conditional coverage ($CC - DQ_{IS}$) and independence ($ID - DQ_{IS}$) versions, respectively. For each model, the table provides the non-rejection frequency at the 5% significance level. Higher values in the table indicate better performance, as they correspond to models less frequently rejected by the tests.

Although the $CC - DQ_{IS}$ provides a comprehensive evaluation of the risk estimation performance, it has a *portmanteau* nature, which overlooks the clustering features of the hit series and neglects their coverage properties. By contrast, the $ID - DQ_{IS}$ test offers a complementary perspective to the previous test, since it focuses explicitly on clustering. Combining the information from both tests makes it possible to gain deeper insight into the reasons behind any model underperformance. The test findings for $\alpha = 0.01$ (similar results hold for the other risk levels) can be summarized as follows:

---

[6]Bayer and Dimitriadis (2022) also consider a one-sided version of the Strict Intercept that is particularly useful for regulatory evaluations. Since our main interest is simply in the assessment of forecasting accuracy, in this paper we only consider the two-sided version of the test.

$CC - DQ_{IS}$: multiplicative models outperform their additive counterparts with the `mlt_lev` resulting the best model at all risk levels. This model is not rejected at the 5% level in approximately 70% of cases, closely followed by the `mlt_skk`. The `mlt_sim` yields slightly lower rates than models incorporating information on realized skewness and kurtosis. The non-rejection frequency of additive models is much lower, being on average close to 20%, with the highest rate being recorded for the `add_sim` model.

$ID - DQ_{IS}$: multiplicative and additive models are characterized by similar performances, suggesting that the high rejection rate of the latter class is mostly due to lack of coverage rather than to hit clustering.

Second, to appreciate the contribution of the additional information in the form of realized skewness and kurtosis in the `skk` models, in the *Wald test* section of Table C.1 we assess the significance of the skewness and kurtosis coefficients involved in the VaR and ES dynamics, respectively. Specifically, we test the null $a_1 = a_2 = a_3 = 0$, for VaR, and $b_1 = b_2 = 0$, for ES, against a two-sided alternative.

Again, the test results in terms of empirical non-rejection frequencies are summarized over the panel of assets considered. For the plain CAViaR models, the null $a_1 = a_2 = a_3 = 0$ is almost always rejected at the usual 5% significance level, for all risk levels considered. Differently, for joint VaR-ES models, the non-rejection frequency increases with the risk level α. Namely, the percentage of non-rejections is close to 0 for $\alpha = 0.01$ but it increases to values up to $\approx 40\%$ for $\alpha = 0.05$. The discrepancy between non-rejection frequencies for pure VaR and joint VaR-ES models is likely to be due to the fact that, for each class of models, testing is based on a different asymptotic distribution: we rely on the theory derived by Engle and Manganelli (2004), for the EM loss, and on Patton et al. (2019), for ALS and FZ0. The test results are only marginally affected by the choice of the joint loss, AL or FZ0, used for estimation.

Moving to the analysis of ES dynamics, we find that the non-rejection frequencies of the null $b_1 = b_2 = 0$ are substantially higher than the values observed for VaR parameters and are clearly affected by the risk level. Namely, they approximately lie in the range 49%-61%, for $\alpha = 0.01$, 55%-72%, for $\alpha = 0.025$, and 62%-79%, for $\alpha = 0.05$. Results are very close for models based on ALS and FZ0 losses. Overall, we conclude that the inclusion of realized skewness and kurtosis measures in the ES equation is less strongly supported than for the VaR.

Next, we focus on the in-sample PZC and ESR calibration tests. First, the *Calibration test* section of Table C.1 for $\alpha = 0.01$ (see Web Appendix *Tables W.1 and W.2* for $\alpha = 0.025$ and $\alpha = 0.05$) reports the results of the former tests for VaR and ES. The main findings arising from the table can be summarized as follows:

- for $\alpha \geq 0.025$ (Tables W.1 and W.2), all models yield remarkably good non-rejection frequencies, with values ranging from 72% to 94%.

- For both VaR and ES, we record a decay of the non-rejection frequency at the 0.01 risk level (Table C.1). This is particularly relevant for VaR since $\alpha = 0.01$ is the mandatory level indicated by the Basel Committee.

- Models based on ALS and FZ0 losses return very close performances.

13

- Comparing simpler models (*_sim) with more complicated specifications (*_skk and *_lev), we find that there is no clear winner but the ranking depends on the functional form and risk level.

Finally, to assess the "calibration" of ES forecasts, the *ES calibration test* section in Table C.1 for $\alpha = 0.01$ (see Web Appendix *Tables W.1 and W.2* for $\alpha = 0.025$ and $\alpha = 0.05$) reports the non-rejection frequencies of the three ESR tests proposed by Bayer and Dimitriadis (2022)[7]. It is worth noting that due to numerical problems in the computation of the test statistic, this could not be computed for some of the assets in our panel, in addition to those that had been previously excluded due to convergence issues in the estimation of the reference risk models: the number of valid assets for each configuration, determined by a combination of available models and risk levels, ranges from a minimum of 644 to a maximum of 796 assets out of 823.

Compared to the calibration test by Patton et al. (2019), the ESR reveals a much lower discriminatory power returning non-rejection frequencies very close to unity for all models and risk levels. Again, we do not report any apparent differences in model performances based on the ALS and FZ0 losses.

4.5. *Out-of-sample forecasting comparison*

This section presents the results of the out-of-sample forecasting analysis. First, the performance of the models under analysis is assessed by computing the following test statistics and diagnostics over the out-of-sample period

- DQ tests for independence and conditional coverage

- VaR and ES calibration tests by Patton et al. (2019)

- ESR tests for ES calibration by Bayer and Dimitriadis (2022).

As in the previous section, test results across the whole panel of assets are summarized in terms of empirical non-rejection frequencies. Also, we only discuss results for $\alpha = 0.01$ in Table C.2 while results for $\alpha = 0.025$ and $\alpha = 0.05$ have been reported in the Web Appendix *Tables W.3 and W.4*.

Similarly to what was observed in the full sample analysis, in a limited number of cases it has not been possible to calculate the *p*-values of ESR tests due to failures in the estimation of the auxiliary regression model underlying the test. Overall, depending on risk level, specific test of interest, and sample size, the available number of stocks has been found to range between 371 and 406 out of 406 potentially available stocks.

The findings of the analysis can be succinctly summarized as:

- DQ tests: the non-rejection frequencies are very low for the shortest estimation window $T = 500$ but they tend to increase with the sample size although, even for $T = 2000$, they barely exceed 40%, for independence tests, only in a few isolated cases. Overall some stylized facts arise. Plain VaR models on average perform better than joint VaR-ES models while the inclusion of information on skewness and kurtosis does not bring any evident advantages.

---

[7]The tests were implemented using the esback R-library provided by the same authors, freely available from CRAN at the URL: https://cran.r-project.org/web/packages/esback/index.html

- VaR calibration tests: the performances are very poor for the shortest sample size $T = 500$ but tend to improve as T increases. The model performances also depend on the value of the risk level α with the best results obtained for $\alpha = 0.025$. In terms of model specifications, `add_sim` and `mlt_sim` yield the highest non-rejection frequencies that exceed 70% for T=2000 and $\alpha = 0.025$ when the EM loss is used. When comparing plain VaR and joint VaR-ES models, there are no clear performance gaps.

- ES calibration tests: the results are qualitatively not different from what was observed for the VaR tests. Hence, similar considerations hold.

- ESR tests: the performance of the "strict" and "auxiliary" tests improves as the sample size increases although the performance gap across different sample sizes is less evident than for the other regression-based VaR and ES calibration tests. As above, even in this case, we record the best performances for the `add_sim` and `mlt_sim` models reaching, in some cases, non-rejection frequencies close to 80%. As far as the "strict intercept test" is concerned, the differences across different models and sample sizes are much less evident and the non-rejection frequency is > 80% in all instances. Again, the information on realized skewness and kurtosis does not appear to lead to improvements in terms of forecasting performances.

Finally, we assess and compare the forecasting accuracy of the different models based on the out-of-sample values of the following strictly consistent scoring functions: quantile loss (E) for VaR and AL-score for joint VaR and ES forecasting (ALS). In terms of median loss (Table C.3), the multiplicative model without skewness and kurtosis information (`mlt_sim`) achieves the minimum loss value in most cases for both quantile and ALS scoring functions. It is only slightly outperformed by its additive counterpart (`add_sim`) in one instance for pure VaR models and in two instances for joint forecasts of VaR and ES. A similar trend is observed when considering average ranks (Table C.4). The `mlt_sim` model consistently delivers the minimum average rank, except in the case of joint VaR and ES forecasts at the 0.05 level and for $T = 1000$, where it ranks second, closely following the `add_sim` model that also does not use skewness and kurtosis information.

In conclusion, the key insights from the assessment of forecasting performance can be summarized as follows:

- Incorporating information on realized skewness and kurtosis does not enhance forecasting accuracy;

- Simpler models are preferable to more complex ones, as the latter are more vulnerable to computational issues;

- The *multiplicative* specification is generally preferable to the more popular *additive* approach.

5. Concluding remarks

In this paper, we have presented a forecasting comparison of several semi-parametric risk forecasting models. Our work presents some important elements of novelty and potential interest for

practitioners and researchers alike. First, the comparison is based on an unusually large set of 823 stocks: to the best of our knowledge, there are no other contributions relying on such a large dataset in the tail-risk forecasting literature. Also, the availability of such a rich data environment has a positive impact on the reliability of the regularities that emerge from the empirical analysis, giving them a good degree of external validity. Second, we assess the potential contribution coming from considering information on some recently proposed realized skewness and kurtosis measures. Third, we provide deeper insight into the selection of the functional form of the semi-parametric model used to generate forecasts.

The results of our analysis clearly indicate that, at the forecasting stage, simple models should be preferred to more complicated ones with a preference for multiplicative GARCH-type specifications. Realized skewness and kurtosis measures do not apparently provide valuable information for improving the accuracy of tail risk forecasts even if in most cases, their coefficients turn out to be significant in the full-sample analysis. By the same token, they may prove useful in generating improved density forecasts, a task that we leave for future research.

When we shift the focus to the functional form of the dynamic risk model, an interesting and original finding from our extensive empirical investigation is that the standard *CaViaR-like* additive model specification outperformed by the less commonly used (in a semi-parametric framework) *GARCH-like* multiplicative parameterization.

References

Amaya, D., Christoffersen, P., Jacobs, K., and Vasquez, A. (2015). Does realized skewness predict the cross-section of equity returns? *Journal of Financial Economics*, 118(1):135 – 167.

Andersen, T. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39:885–905.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2010). Parametric and nonparametric volatility measurement. In *Handbook of financial econometrics: Tools and techniques*, pages 67–137. Elsevier.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453):42–55.

Andersen, T. G., Dobrev, D., and Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics*, 169(1):75–93.

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

Bae, K. and Lee, S. (2021). Realized higher-order comoments. *Quantitative Finance*, 21(3):421–429.

Barndorff-Nielsen, O. E., Kinnebrock, S., and Shephard, N. (2008). Measuring downside risk - realised semivariance. Research paper no. 2008-42, CREATES.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Measuring the impact of jumps in multivariate price processes using bipower covariation.

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *The Review of Financial Studies*, 9(1):69–107.

Bayer, S. and Dimitriadis, T. (2020). *esback: Expected Shortfall Backtesting*. R package version 0.3.0.

Bayer, S. and Dimitriadis, T. (2022). Regression-based expected shortfall backtesting. *Journal of Financial Econometrics*, 20(3):437–471.

Bollerslev, T., Li, J., Patton, A. J., and Quaedvlieg, R. (2020). Realized semicovariances. *Econometrica*, 88(4):1515–1551.

Choe, G. H. and Lee, K. (2014). High moment variations and their application. *Journal of Futures Markets*, 34(11):1040–1061.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Dumitrescu, E., Hurlin, C., and Pham, V. (2012). Value-at-risk: From dynamic quantile to dynamic binary tests. *Finance*, 33:79–112.

Engle, R. F. and Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381.

Fissler, T. and Ziegel, J. F. (2015). Higher order elicitability and Osband's principle. *ArXiv e-prints*.

Fissler, T., Ziegel, J. F., and Gneiting, T. (2015). Expected shortfall is jointly elicitable with value at risk-implications for backtesting. *arXiv preprint arXiv:1507.00244*.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.

Hill, G. W. and Davis, A. W. (1968). Generalized Asymptotic Expansions of Cornish-Fisher Type. *The Annals of Mathematical Statistics*, 39(4):1264 – 1273.

Jorion, P. (1997). *Value at Risk: The New Benchmark for Controlling Market Risk*. McGraw-Hill.

Liu, Z., Wang, K., and Liu, J. (2014). Realized skewness at high frequency and the link to a conditional market premium. *Asian Finance Association (AsFA) 2013 Conference*.

Mei, D., Liu, J., Ma, F., and Chen, W. (2017). Forecasting stock market volatility: Do realized skewness and kurtosis help? *Physica A: Statistical Mechanics and its Applications*, 481:153–159.

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

Neuberger, A. (2012). Realized Skewness. *The Review of Financial Studies*, 25(11):3423–3455.

Neuberger, A. and Payne, R. (2021). The skewness of the stock market over long horizons. *The Review of Financial Studies*, 34(3):1572–1616.

Patton, A. J., Ziegel, J. F., and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, 211(2):388 – 413.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics*, 37(1):121–133.

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, 36(2):428–441.

Wang, C., Gerlach, R., and Chen, Q. (2023). A semi-parametric conditional autoregressive joint value-at-risk and expected shortfall modeling framework incorporating realized measures. *Quantitative Finance*, 23(2):309–334.

Xiao, Z. and Koenker, R. (2009). Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models. *Journal of the American Statistical Association*, 104(488):1696–1712.

17

Appendix A: Asymptotic distribution of the estimators

A.1. *Standard errors estimation for pure* VaR *models*

The theoretical results presented in this section are based on Engle and Manganelli (2004) and assumptions therein. It is worth noting that, although Engle and Manganelli (2004) focuses on additive CAViaR models, their framework readily applies to the multiplicative VaR models class. In the following, we assume that the return process r_t has conditional α-quantile given by $v_t(\boldsymbol{\theta}_0)$. The estimated VaR, $v_t(\hat{\boldsymbol{\theta}}^{(v)})$, is obtained by replacing $\boldsymbol{\theta}_0$ with the minimizer of the aggregated quantile loss function. Applying the results in Section 4 of Engle and Manganelli (2004), $\hat{\boldsymbol{\theta}}^{(v)}$ can be shown to be consistent and asymptotically normal. In particular, we have

$$\sqrt{T} A_T^{-1/2} D_T(\hat{\boldsymbol{\theta}}^{(v)} - \boldsymbol{\theta}_0) \xrightarrow{d} MVN(\mathbf{0}, I) \quad \text{as } T \to \infty,$$

where $A_T = \mathrm{E}\left[T^{-1}\alpha(1-\alpha)\sum_{t=1}^{T}\nabla' v_t(\boldsymbol{\theta}_0^{(v)})\nabla v_t(\boldsymbol{\theta}_0^{(v)})\right]$, $D_T = \mathrm{E}\left[T^{-1}\sum_{t=1}^{T} h_t(0|\mathcal{I}_{t-1})\nabla' v_t(\boldsymbol{\theta}_0^{(v)})\nabla v_t(\boldsymbol{\theta}_0^{(v)})\right]$, $h_t(0|\mathcal{I}_{t-1})$ is the conditional density of $\eta_t = r_t - v_t$ evaluated at 0 and $\nabla v_t(\boldsymbol{\theta}^{(v)}) = \partial v_t(\boldsymbol{\theta}^{(v)})/\partial \boldsymbol{\theta}^{(v)}$. Consistent estimates of A_T and D_T can be then obtained as follows

$$\widehat{A_T} = T^{-1}\alpha(1-\alpha)\nabla^{\top} v(\boldsymbol{\theta}_0^{(v)})\nabla(v\boldsymbol{\theta}_0^{(v)}),$$

where $\nabla v(\boldsymbol{\theta_0})$ is the $(T \times p)$ matrix whose t-th row is $\nabla^{\top} v_t(\boldsymbol{\theta}_0^{(v)})$, and

$$\widehat{D_T} = (2\,T\,\widehat{c_T})^{-1}\sum_{t=1}^{T} I(|r_t - v_t(\hat{\boldsymbol{\theta}}^{(v)})| < \widehat{c_T})\nabla' v_t(\hat{\boldsymbol{\theta}}^{(v)})\nabla v_t(\hat{\boldsymbol{\theta}}^{(v)}).$$

Analytical expressions for the elements of $\nabla v_t(\hat{\boldsymbol{\theta}}^{(v)})$ have been derived and reported in Web Appendix *Derivatives*. Following Engle and Manganelli (2004), the bandwidth $\widehat{c_T}$ is set as: $\widehat{c_T} = 40$, for $\alpha = 0.01$, $\widehat{c_T} = 60$, for $\alpha = 0.05$. For the case $0.01 < \alpha < 0.05$, which is not considered in the paper by Engle and Manganelli (2004), we estimate the bandwidth by linear interpolation. So the final estimated asymptotic variance and covariance matrix of $\hat{\boldsymbol{\theta}}$ is computed as $\widehat{\Sigma}_{\hat{\boldsymbol{\theta}}} = \frac{1}{T}(\alpha)(1-\alpha)\widehat{D_T}^{-1}\widehat{A_T}\widehat{D_T}^{-1}$ and estimated standard errors are computed as $\widehat{se}(\hat{\boldsymbol{\theta}}) = \sqrt{\mathrm{diag}\left(\widehat{\Sigma}_{\hat{\boldsymbol{\theta}}}^{-1}\right)}$.

Letting $\mathbf{a} \subset \boldsymbol{\theta}$, the above results can be used to test $\mathbf{a} = \mathbf{0}$. Note that we can write $R\boldsymbol{\theta} = \mathbf{a}$ where

$$R = \left(\begin{array}{c|c} \multicolumn{2}{c}{\mathbf{0}_{n-2,n}} \\ \hline \mathbf{0}_{2,n} & I_2 \end{array}\right)$$

$\mathbf{0}_{m,n}$ indicate a $(m \times n)$ matrix of zeroes and I_n be a $(n \times n)$ identity matrix. Relying on the asymptotic normality of $\widehat{\boldsymbol{\theta}}$, it can then be shown that, under the null $R\boldsymbol{\theta} = \mathbf{0}$, $(R\widehat{\boldsymbol{\theta}})^{\top}(R\widehat{V}_T R^{\top})^{-1}(R\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi_2^2$, as $T \to \infty$.

Standard errors for joint VaR-ES *models*

Patton et al. (2019) prove consistency and asymptotic normality for the estimator $\hat{\boldsymbol{\theta}}^{(j)}$. The theoretical results presented in this section rely on the theory developed in Patton et al. (2019) and

assumptions therein. In the presentation of the following results, we assume that the return process r_t has theoretical α-level VaR and ES given by $v_t(\boldsymbol{\theta}_0^{(j)})$ and $e_t(\boldsymbol{\theta}_0^{(j)})$, respectively. The estimated VaR and ES, $v_t(\hat{\boldsymbol{\theta}}^{(j)})$ and $e_t(\hat{\boldsymbol{\theta}}^{(j)})$, are obtained by replacing $\boldsymbol{\theta}^{(j)}$ with the minimizer of the strictly consistent loss function used for estimation. It is worth noting that, although the results in Patton et al. (2019) are derived for estimators based on the FZ0 loss function, they can be immediately extended to estimators obtained by minimizing different loss functions, such as the ALS.

In particular, the asymptotic distribution of $\hat{\boldsymbol{\theta}}^{(j)}$ is given by

$$\sqrt{T} A_0^{-1/2} D_0 (\hat{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}_0^{(j)}) \xrightarrow{d} MVN(\mathbf{0}, I) \quad \text{as } T \to \infty.$$

Consistent estimates of A_0 and D_0 can be obtained as follows

$$\widehat{A_T} = T^{-1} \sum_{t=1}^{T} \lambda_t(\hat{\boldsymbol{\theta}}^{(j)}) \lambda_t^{\top}(\hat{\boldsymbol{\theta}}^{(j)}) \tag{11}$$

$$\widehat{D_T} = T^{-1} \sum_{t=1}^{T} \left\{ \frac{1}{2c_T} I(|r_t - v_t| < c_T) \frac{\nabla^{\top} e_t(\hat{\boldsymbol{\theta}}^{(j)}) \nabla v_t(\hat{\boldsymbol{\theta}}^{(j)})}{-\alpha e_t(\hat{\boldsymbol{\theta}}^{(j)})} + \frac{\nabla^{\top} e_t(\hat{\boldsymbol{\theta}}^{(j)}) \nabla e_t(\hat{\boldsymbol{\theta}}^{(j)})}{e_t^2(\hat{\boldsymbol{\theta}}^{(j)})} \right\} \tag{12}$$

where the bandwidth c_T is set equal to $T^{-1/3}$, as suggested by Patton et al. (2019) and $\lambda_t(\boldsymbol{\theta}^{(j)}) = \partial L_t^{(\alpha)}/\partial \boldsymbol{\theta}^{(j)}$, with $L_t^{(\alpha)}$ denoting the strictly consistent loss function used for estimation. When $L_t^{(\alpha)} \equiv FZ0_t^{(\alpha)}$, $\lambda_t(\boldsymbol{\theta}^{(j)})$ is given by

$$\begin{aligned}
\lambda_t(\boldsymbol{\theta}^{(j)}) &= \frac{\partial FZ0_t^{(\alpha)}}{\partial \boldsymbol{\theta}^{(j)}} = \nabla^{\top} v_t(\boldsymbol{\theta}^{(j)}) \frac{1}{-e_t(\boldsymbol{\theta}^{(j)})} \left[\frac{1}{\alpha} I\{v_t(\boldsymbol{\theta}^{(j)})\} - 1 \right] \\
&+ \nabla^{\top} e_t(\boldsymbol{\theta}^{(j)}) \frac{1}{e_t(\boldsymbol{\theta}^{(j)})^2} \left[\frac{1}{\alpha} I\{v_t(\boldsymbol{\theta}^{(j)})\} \{v_t(\boldsymbol{\theta}^{(j)}) - r_t\} - v_t(\boldsymbol{\theta}^{(j)}) + e_t \right].
\end{aligned}$$

If the ALS loss is used, the formula above becomes

$$\begin{aligned}
\lambda_t(\boldsymbol{\theta}^{(j)}) &= \frac{\partial ALS_t^{(\alpha)}}{\partial \boldsymbol{\theta}^{(j)}} = \nabla^{\top} v_t(\boldsymbol{\theta}^{(j)}) \frac{1}{-e_t(\boldsymbol{\theta}^{(j)})} \left[\frac{1}{\alpha} I\{v_t(\boldsymbol{\theta}^{(j)})\} - 1 \right] \\
&+ \nabla^{\top} e_t(\boldsymbol{\theta}^{(j)}) \frac{1}{e_t(\boldsymbol{\theta}^{(j)})^2} \left[\frac{1}{\alpha} I\{v_t(\boldsymbol{\theta}^{(j)})\} \{v_t(\boldsymbol{\theta}^{(j)}) - r_t\} - v_t(\boldsymbol{\theta}^{(j)}) + e_t(\boldsymbol{\theta}^{(j)}) + r_t \right]
\end{aligned}$$

that differs from (13) for the return r_t in the last term on the RHS. Analytical expressions for the elements of $\nabla e_t(\hat{\boldsymbol{\theta}}^{(j)})$ have been derived and reported in Web Appendix *Derivatives*. Finally, an estimate of the asymptotic variance and covariance matrix of $\boldsymbol{\theta}^{(j)}$ is then given by $\widehat{\Sigma}_{\hat{\boldsymbol{\theta}}} = T^{-1} \widehat{D_T^{-1}} \widehat{A_T} \widehat{D_T}$ and standard errors are computed as $\widehat{se}(\hat{\boldsymbol{\theta}}) = \sqrt{\text{diag}\left(\widehat{\Sigma}_{\hat{\boldsymbol{\theta}}}^{-1}\right)}$.

Appendix B: Diagnostic tests for VaR and ES

B.2. In-sample hit test

The in-sample hit test is proposed by Engle and Manganelli (2004) as a model-based diagnostic test for detecting misspecified CAViaR models. The test relies on the *Hit* variables defined as

$Hit_t = I(v_t) - \alpha$. Replacing v_t by its estimated counterpart $\hat{v}_t$, the estimated hits are obtained as $\widehat{Hit}_t = I(\hat{v}_t) - \alpha$ and stacked together into $\widehat{\mathbf{Hit}} = (\widehat{Hit}_{q+1}, \ldots, \widehat{Hit}_T)^\top$. Then, letting $\theta^{(v)}$ denote the vector of CAViaR coefficients, define $\mathbf{X}_t(\theta_0^{(v)})$ as $\mathbf{X}_t(\theta_0^{(v)}) = (Hit_{t-1}, \ldots, Hit_{t-q}, \mathbf{z}_{t-1}^\top)$, where $\mathbf{z}_t$ is a $(m \times 1)$ vector of $\mathcal{I}_t$ measurable instruments, for $t = q+1, \ldots, T$. For example, $\mathbf{z}_{t-1}^\top$ could include estimated past VaR values or realized measures of skewness and kurtosis. Let then $\mathbf{X}(\theta_0)$ be the matrix whose generic row is given by $\mathbf{X}_t(\theta_0^{(v)})$, so that $\mathbf{X}(\theta^{(v)})$ is of dimension $(T - q) \times (q + m)$. Then, define $(q + m) \times (T - q)$-dimensional matrix

$$\mathbf{M}_T = \mathbf{X}^\top(\theta_0^{(v)}) - \mathrm{E}\left\{ T^{-1}\mathbf{X}^\top(\theta_0^{(v)})\mathbf{H}\nabla v_t(\theta_0^{(v)}) \right\} D_T^{-1},$$

where $\mathbf{H}$ is a diagonal matrix with diagonal entries given by $h_t(0|\mathcal{I}_{t-1})$ as defined in A.1. Under the assumptions from Engle and Manganelli (2004), we have

$$\left\{ \alpha(1 - \alpha)\,\mathrm{E}\left(T^{-1}\mathbf{M}_T\mathbf{M}_T^\top \right) \right\}^{-1/2} T^{-1/2}\mathbf{X}^\top(\hat{\theta}^{(v)})\,\widehat{\mathbf{Hit}} \xrightarrow{d} N(\mathbf{0}_{q+m}, \mathbf{I}_{q+m}),$$

where $\mathbf{0}_{q+m}$ is the $q + m$ vector of zeros, and $\mathbf{I}_{q+m}$ is the $q + m$ dimensional identity matrix. In the above result, the matrix $\mathbf{M}_T$ can be estimated as

$$\widehat{\mathbf{M}}_T = \mathbf{X}^\top(\hat{\theta}^{(v)}) - \left\{ (2T\widehat{c}_T)^{-1} \sum_{t=1}^{T} I(|r_t - v_t(\hat{\theta}^{(v)})| < \widehat{c}_T)\,\mathbf{X}_t^\top(\hat{\theta}^{(v)})\nabla v_t(\hat{\theta}^{(v)}) \right\} \widehat{D}_T^{-1}\mathbf{g}^\top(\hat{\theta}^{(v)}),$$

with $\mathbf{g}(\theta^{(v)}) = \{\nabla v_{q+1}(\theta^{(v)}), \ldots, \nabla v_T(\theta^{(v)})\}^\top$. It can then be proved that

$$DQ_{IS} = \frac{\widehat{\mathbf{Hit}}^\top \mathbf{X}(\hat{\theta}^{(v)})(\widehat{\mathbf{M}}_T\widehat{\mathbf{M}}_T^\top)^{-1}\mathbf{X}^\top(\hat{\theta}^{(v)})\widehat{\mathbf{Hit}}}{\alpha(1 - \alpha)} \xrightarrow{d} \chi_q^2,$$

that in the reminder will be denoted as the *In-Sample Dynamic Quantile* test statistic.

B.3. *Out-of-sample diagnostic tests for* VaR *forecasts: the out-of-sample Dynamic Quantile test (Engle and Manganelli, 2004)*

Formally, let $(\hat{v}_{T+1}, \ldots, \hat{v}_{T+H})$ be a sequence of (1-step ahead) out-of-sample VaR forecasts. The OOS-DQ test statistic is computed as

$$DQ_{OOS} = \frac{\widetilde{\mathbf{Hit}}(\hat{\theta}^{(v)})^\top \widetilde{\mathbf{X}}(\hat{\theta}^{(v)})\{\widetilde{\mathbf{X}}(\hat{\theta}^{(v)})^\top \widetilde{\mathbf{X}}(\hat{\theta}^{(v)})\}^{-1}\widetilde{\mathbf{X}}(\hat{\theta}^{(v)})^\top \widetilde{\mathbf{Hit}}(\hat{\theta}^{(v)})^\top}{H\alpha(1 - \alpha)}, \tag{13}$$

where $\theta^{(v)} \subseteq \theta$ indicates the subvector of model parameters involved in the VaR model and

$$\widetilde{\mathbf{Hit}}(\hat{\theta}^{(v)}) = (\widetilde{Hit}_{T+q+1}, \ldots, \widetilde{Hit}_{T+H}),$$

is the series of out-of-sample hits based on parameters estimated relying on information up to time T; $\widetilde{\mathbf{X}}(\hat{\theta}^{(v)})$ is a $(H - q) \times (q + 2)$ matrix such that its $(t - q)$-th row is given by

$$\widetilde{\mathbf{X}}_t(\hat{\theta}^{(v)}) = (1, \widetilde{Hit}_{t-1}, \ldots, \widetilde{Hit}_{t-q}, v_t(\hat{\theta}^{(v)})),$$

20

for $t = q+1, \ldots, H$, and where $\hat{\boldsymbol{\theta}}^{(v)}$ is the estimate of $\boldsymbol{\theta}^{(v)}$. It can be shown that, under the assumption in Engle and Manganelli (2004), $DQ_{OOS} \xrightarrow{d} \chi^2_{q+2}$.

The out-of-sample DQ test could also be implemented by augmenting the $\widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}_T)$ with other instruments such as past volatility measures, such as realized variances, squared or absolute returns. In this case, the degrees of freedom of the χ^2 distribution should be changed accordingly.

The out-of-sample DQ test, in the above-presented configuration, can be seen as a portmanteau test for the correct specification of the VaR forecasting model and, in this sense, we will refer to this test as the correct conditional coverage DQ_{OOS} test, abbreviated $CC - DQ_{OOS}$. Differently, removing the constant term from the specification of $\widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}_T)$ would yield a different version of the out-of-sample DQ test, that we will call the independence DQ_{OOS} or, abbreviated, the $ID - DQ_{OOS}$. The name derives from the fact that $ID - DQ_{OOS}$ can detect serial correlation in the sequence of hits but, due to the missing constant term, cannot be used to assess correct coverage of VaR forecasts. The asymptotic distribution for the $ID - DQ_{OOS}$ will be given by a χ^2_{q+1} random variable. Similarly, it is possible to define analogous conditional coverage and independence versions of the in-sample DQ test. In the remainder, these will be labeled as $CC - DQ_{IS}$ and $ID - DQ_{IS}$, respectively.

B.4. Diagnostic tests for ES forecasts

B.5. The ESR backtests (Bayer and Dimitriadis, 2022)

In Bayer and Dimitriadis (2022), authors propose a set of backtesting procedures for ES regression (ESR), which build upon the testing approach introduced by Mincer and Zarnowitz (1969). Specifically, the authors propose three ESR backtests: the "auxiliary", "strict" and "strict intercept" ESR.

The "auxiliary" ESR test is based on the bivariate regression model

$$r_t = \beta_0 + \beta_1 \hat{v}_t + u_t^v, \tag{14}$$
$$r_t = \gamma_0 + \gamma_1 \hat{e}_t + u_t^e, \tag{15}$$

for $t = 1, \ldots, T$, where $\mathrm{E}(u_t^e | \mathcal{I}_{t-1}, r_t < \hat{v}_t) = 0$ and $Q_\alpha(u_t^v | \mathcal{I}_{t-1}) = 0$. The idea is that a series of ES forecasts from a correctly specified ES model should satisfy the following relation

$$\mathrm{E}(r_t | \mathcal{I}_{t-1}, r_t < \hat{v}_t) = \gamma_0 + \gamma_1 \hat{e}_t,$$

with $(\gamma_0, \gamma_1)^\top = (0, 1)^\top$. This hypothesis can be tested by fitting the regression model in (14-15) through the minimization of a strictly consistent joint (VaR, ES) loss function. This leads to the following Wald-type test statistic

$$T_{A-ESR} = T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\widehat{\Omega}_{\boldsymbol{\gamma}}^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\top \xrightarrow{d} \chi^2_2,$$

where $\boldsymbol{\gamma}_0 = (0, 1)^\top$, $\hat{\boldsymbol{\gamma}}$ is a consistent estimator of $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ and $\widehat{\Omega}_{\boldsymbol{\gamma}}$ is a consistent estimator of the covariance of $\hat{\boldsymbol{\gamma}}$ (see Bayer and Dimitriadis, 2022). The "strict" ESR test is based on a similar framework but $\hat{v}_t$ in equation (14) is replaced by $\hat{e}_t$. Finally, the "strict intercept" test replaces equation (15) by the following

$$r_t - \hat{e}_t = \gamma_1 + u_t^e.$$

21

The null is now given by $\gamma_1 = 0$ against a one-sided or a two-sided alternative [8]. The test is performed by computing a standard t-type statistic based on the estimated asymptotic variance of $\hat{\gamma}_1$. The one-sided alternative is of interest for regulatory and, in general, risk management purposes. In this paper, our interest is mainly in detecting deviations from the ideal situation of correct specification of the risk forecasting models. Hence, we will focus only on the situation where a two-sided alternative is considered. To implement the ESR backtests, we use the `Esback` R package provided by the same authors (Bayer and Dimitriadis, 2020) for our empirical analysis.

B.6. Regression based calibration tests for VaR and ES (*Patton et al., 2019*)

Patton et al. (2019) propose OLS regression-based calibration tests for assessing the quality of VaR and ES forecasts. In the auxiliary regression equations used for implementing the tests, the dependent variables are given by the standardized generalized residuals

$$\lambda^s_{v,t} = I(r_t \le \hat{v}_t) - \alpha \qquad \lambda^s_{e,t} = \frac{1}{\alpha} I(r_t \le \hat{v}_t)\frac{r_t}{\hat{e}_t} - 1$$

for VaR and ES, respectively. Both $\lambda^s_{v,t}$ and $\lambda^s_{e,t}$ are conditionally zero mean under correct specification of the VaR and ES models $E(\lambda^s_{v,t}|\mathcal{I}_{t-1}) = 0$ and $E(\lambda^s_{e,t}|\mathcal{I}_{t-1}) = 0$, for all t. It is also worth noting that $\lambda^s_{v,t} = Hit_t$, the hit variable already defined for DQ tests.

The test procedures, henceforth denoted as the PZC tests, are based on fitting by OLS the following regression models

$$\lambda^s_{v,t} = a_{0,v} + a_{1,v}\lambda^s_{v,t-1} + a_{2,v}\hat{v}_t + u^v_t \tag{16}$$

$$\lambda^s_{e,t} = a_{0,e} + a_{1,e}\lambda^s_{v,t-1} + a_{2,e}\hat{e}_t + u^v_t, \tag{17}$$

where, under correct specifications, we have $\mathbf{a}_v = (a_{0,v}, a_{1,v}, a_{2,v})^\top = \mathbf{0}$ and $\mathbf{a}_e = (a_{0,e}, a_{1,e}, a_{2,e})^\top = \mathbf{0}$. The statistics for testing these hypotheses are computed as

$$PZC_v = \hat{\mathbf{a}}_v^\top \widehat{\Omega}_v^{-1} \hat{\mathbf{a}}_v \xrightarrow{d} \chi^2_2, \qquad PZC_e = \hat{\mathbf{a}}_e^\top \widehat{\Omega}_e^{-1} \hat{\mathbf{a}}_e \xrightarrow{d} \chi^2_2,$$

where $\widehat{\Omega}_v$ ($\widehat{\Omega}_e$) is a consistent estimator of the asymptotic covariance matrix of $\hat{\mathbf{a}}_v$ ($\hat{\mathbf{a}}_e$). In our empirical analysis, following Patton et al. (2019), to estimate Ω_v and Ω_e a Newey-West estimator with 20 lags is used.

---

[8]Differently from the "strict" and "auxiliary" tests for which only a two-sided alternative was allowed.

Appendix C: Tables

Table C.1: Calibration tests for VaR and ES models Patton et al. (2019); in-sample DQ conditional coverage (CC_{IS}) and independence test (ID_{IS}); Wald tests for skewness and kurtosis coefficients in VaR and ES models; "Str.", "Aux.", "Str.I." ES Regression (ESR) calibration test (Bayer and Dimitriadis (2022)): non-rejection frequencies at the 0.05 significance level (full sample).

VaR	ES	Loss	In-sample DQ CC_{IS}	ID_{IS}	Wald test $a_i = 0$	$b_i = 0$	Calibration test VaR	ES	ESR calibration test "Str."	"Aux."	"Str.I."
mlt_lev	—	EM	0.712	0.727	0.005	—	0.617	—	—	—	—
mlt_skk	—	EM	0.701	0.690	0.002	—	0.661	—	—	—	—
mlt_sim	—	EM	0.684	0.589	—	—	0.684	—	—	—	—
add_lev	—	EM	0.198	0.746	0.010	—	0.594	—	—	—	—
add_skk	—	EM	0.193	0.679	0.004	—	0.625	—	—	—	—
add_sim	—	EM	0.235	0.562	—	—	0.655	—	—	—	—
mlt_lev	sim	ALS	—	—	0.016	—	0.548	0.695	0.988	0.987	0.995
mlt_skk	sim	ALS	—	—	0.017	—	0.584	0.734	0.984	0.983	0.995
mlt_sim	sim	ALS	—	—	—	—	0.612	0.783	0.980	0.982	0.996
add_lev	sim	ALS	—	—	0.016	—	0.563	0.741	0.984	0.984	0.994
add_skk	sim	ALS	—	—	0.010	—	0.601	0.778	0.980	0.978	0.991
add_sim	sim	ALS	—	—	—	—	0.588	0.682	0.983	0.982	0.993
mlt_lev	skk	ALS	—	—	0.017	0.532	0.546	0.670	0.969	0.961	0.999
mlt_skk	skk	ALS	—	—	0.010	0.539	0.597	0.688	0.962	0.947	0.999
mlt_sim	skk	ALS	—	—	—	0.494	0.638	0.718	0.978	0.931	0.999
add_lev	skk	ALS	—	—	0.012	0.610	0.598	0.722	0.975	0.946	0.997
add_skk	skk	ALS	—	—	0.015	0.589	0.604	0.730	0.975	0.946	0.997
add_sim	skk	ALS	—	—	—	0.527	0.618	0.651	0.973	0.927	0.997
mlt_lev	sim	FZ0	—	—	0.021	—	0.554	0.716	0.982	0.985	0.995
mlt_skk	sim	FZ0	—	—	0.021	—	0.590	0.727	0.985	0.985	0.994
mlt_sim	sim	FZ0	—	—	—	—	0.614	0.783	0.982	0.980	0.994
add_lev	sim	FZ0	—	—	0.011	—	0.560	0.737	0.984	0.983	0.994
add_skk	sim	FZ0	—	—	0.019	—	0.605	0.761	0.974	0.979	0.992
add_sim	sim	FZ0	—	—	—	—	0.597	0.686	0.983	0.983	0.993
mlt_lev	skk	FZ0	—	—	0.018	0.521	0.551	0.651	0.965	0.953	0.999
mlt_skk	skk	FZ0	—	—	0.007	0.527	0.590	0.699	0.970	0.952	0.997
mlt_sim	skk	FZ0	—	—	—	0.487	0.623	0.711	0.967	0.934	0.999
add_lev	skk	FZ0	—	—	0.019	0.597	0.593	0.732	0.970	0.950	0.999
add_skk	skk	FZ0	—	—	0.010	0.592	0.618	0.738	0.968	0.942	0.997
add_sim	skk	FZ0	—	—	—	0.539	0.609	0.651	0.972	0.929	0.997

The top of the table shows $\alpha = 0.01$.

Table C.2: Out-of-sample DQ independence test (ID$_{OOS}$) and conditional coverage test (CC$_{OOS}$); calibration tests for VaR and ES models Patton et al. (2019); "Strict", "Auxiliary" and "Strict Intercept" ES Regression (ESR) calibration test (Bayer and Dimitriadis (2022)): non-rejection frequencies at the 0.05 significance level and number of valid assets (out-of-sample data T_{in}).

| | $\alpha = 0.01$ | | Out-of-sample DQ | | Calibration test | | ESR calibration test | | |
	VaR	ES	ID$_{OOS}$	CC$_{OOS}$	VaR	ES	"Str."	"Aux."	"Str.I."
$T_{in} = 500$	mlt_lev	—	0.030	0.002	0.002	—	—	—	—
	mlt_skk	—	0.010	0.000	0.005	—	—	—	—
	mlt_sim	—	0.039	0.017	0.084	—	—	—	—
	add_lev	—	0.025	0.000	0.002	—	—	—	—
	add_skk	—	0.015	0.000	0.000	—	—	—	—
	add_sim	—	0.052	0.022	0.074	—	—	—	—
	mlt_lev	sim	0.052	0.022	0.000	0.025	0.169	0.176	0.817
	mlt_skk	sim	0.020	0.007	0.007	0.057	0.186	0.211	0.838
	mlt_sim	sim	0.042	0.022	0.067	0.099	0.370	0.366	0.867
	add_lev	sim	0.022	0.005	0.000	0.074	0.266	0.261	0.884
	add_skk	sim	0.049	0.015	0.007	0.074	0.343	0.340	0.935
	add_sim	sim	0.015	0.007	0.049	0.092	0.317	0.348	0.898
$T_{in} = 1000$	mlt_lev	—	0.180	0.086	0.165	—	—	—	—
	mlt_skk	—	0.133	0.076	0.207	—	—	—	—
	mlt_sim	—	0.224	0.185	0.405	—	—	—	—
	add_lev	—	0.126	0.067	0.163	—	—	—	—
	add_skk	—	0.126	0.047	0.175	—	—	—	—
	add_sim	—	0.175	0.131	0.430	—	—	—	—
	mlt_lev	sim	0.143	0.089	0.064	0.114	0.409	0.415	0.948
	mlt_skk	sim	0.143	0.101	0.106	0.146	0.457	0.463	0.944
	mlt_sim	sim	0.190	0.170	0.326	0.304	0.641	0.688	0.964
	add_lev	sim	0.094	0.057	0.059	0.126	0.491	0.499	0.944
	add_skk	sim	0.126	0.069	0.099	0.168	0.487	0.487	0.939
	add_sim	sim	0.145	0.123	0.277	0.296	0.551	0.580	0.960
$T_{in} = 2000$	mlt_lev	—	0.330	0.241	0.449	—	—	—	—
	mlt_skk	—	0.310	0.227	0.491	—	—	—	—
	mlt_sim	—	0.419	0.362	0.531	—	—	—	—
	add_lev	—	0.281	0.200	0.459	—	—	—	—
	add_skk	—	0.268	0.192	0.447	—	—	—	—
	add_sim	—	0.377	0.335	0.578	—	—	—	—
	mlt_lev	sim	0.291	0.232	0.316	0.326	0.593	0.587	0.885
	mlt_skk	sim	0.234	0.192	0.358	0.356	0.629	0.643	0.886
	mlt_sim	sim	0.335	0.291	0.499	0.488	0.740	0.751	0.866
	add_lev	sim	0.303	0.207	0.294	0.311	0.656	0.644	0.880
	add_skk	sim	0.222	0.150	0.259	0.323	0.630	0.618	0.896
	add_sim	sim	0.300	0.241	0.486	0.472	0.735	0.735	0.872

Table C.3: Median loss function values across all assets (out-of-sample data, T_{in} = 500, 1000, 2000). For VaR models, the average quantile loss is reported (×1000). For ES models, the ALS scoring function is considered. The best model within each class is reported in boldface.

VaR	ES	T_{in} = 500			T_{in} = 1000			T_{in} = 2000		
		0.01	0.025	0.05	0.01	0.025	0.05	0.01	0.025	0.05
mlt_lev	—	0.999	1.667	2.538	0.852	1.490	2.322	0.818	1.464	2.319
mlt_skk	—	0.982	1.643	2.521	0.842	1.486	2.309	0.801	1.458	2.320
mlt_sim	—	**0.940**	**1.602**	**2.454**	0.822	**1.469**	**2.291**	**0.790**	**1.443**	**2.308**
add_lev	—	0.990	1.663	2.524	0.855	1.494	2.330	0.813	1.465	2.322
add_skk	—	0.992	1.648	2.537	0.859	1.490	2.320	0.816	1.467	2.325
add_sim	—	0.958	1.639	2.569	**0.820**	1.472	2.306	0.798	1.448	2.315
mlt_lev	sim	-0.411	-1.312	-1.739	-1.031	-1.654	-1.977	-1.395	-1.789	-2.042
mlt_skk	sim	-0.484	-1.365	-1.765	-1.082	-1.645	-1.984	-1.399	-1.797	-2.043
mlt_sim	sim	**-1.002**	**-1.592**	-1.888	**-1.319**	**-1.784**	**-2.045**	**-1.506**	**-1.844**	-2.055
add_lev	sim	-0.354	-1.316	-1.737	-1.014	-1.651	-1.965	-1.357	-1.762	-2.026
add_skk	sim	-0.442	-1.331	-1.695	-0.968	-1.604	-1.954	-1.297	-1.756	-2.023
add_sim	sim	-0.665	-1.568	**-1.900**	-1.302	-1.782	-2.043	-1.491	-1.836	**-2.064**

Table C.4: Average ranks across all assets based on quantile loss and ALS scoring functions (out-of-sample data, T_{in} = 500, 1000, 2000).The best model within each class is reported in boldface.

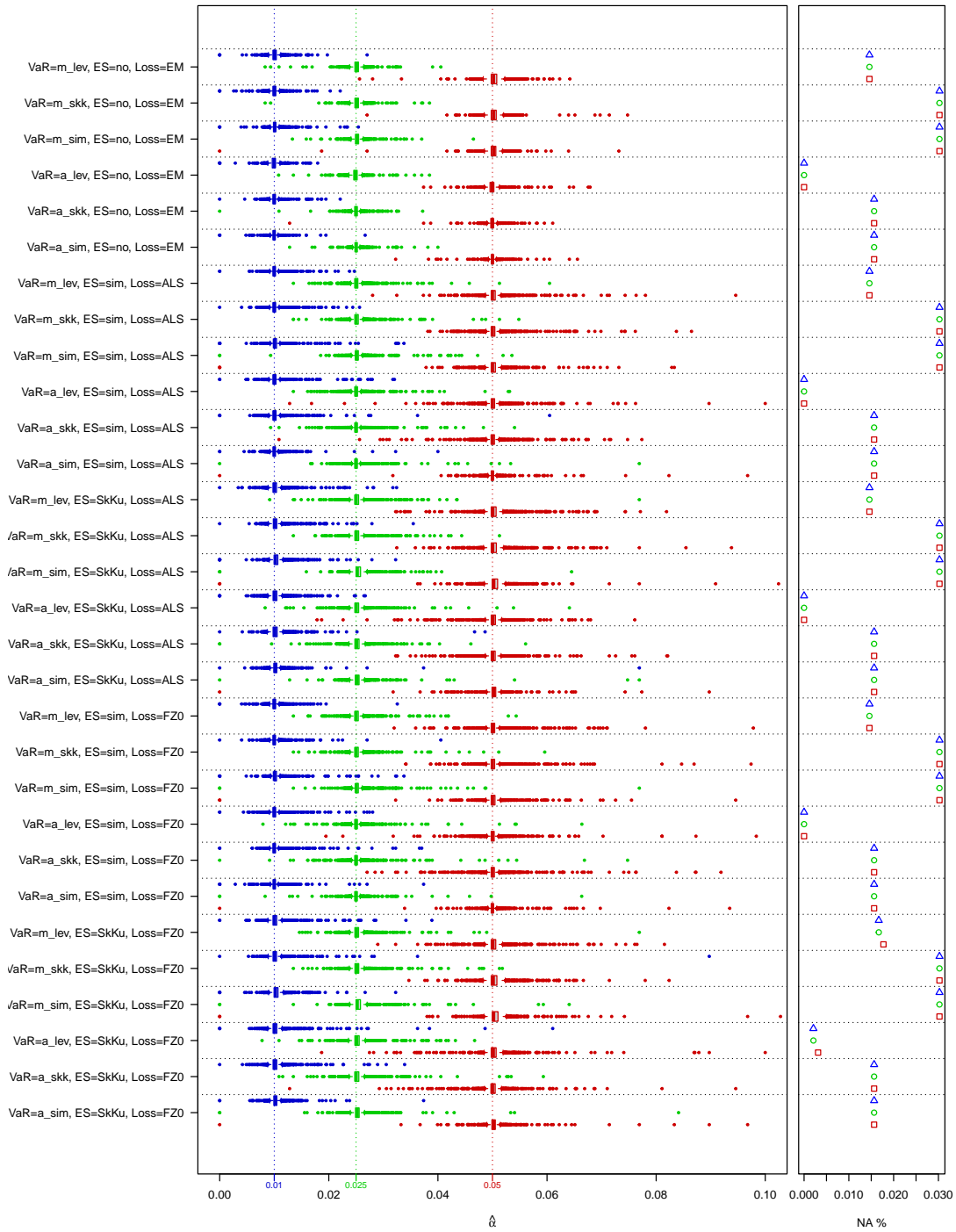| VaR	ES	T_{in} = 500			T_{in} = 1000			T_{in} = 2000		
		0.01	0.025	0.05	0.01	0.025	0.05	0.01	0.025	0.05
mlt_lev	—	4.441	4.374	4.153	4.303	4.241	4.155	4.204	4.034	3.975
mlt_skk	—	3.436	3.303	3.192	3.554	3.594	3.483	3.643	3.466	3.581
mlt_sim	—	**2.002**	**1.709**	**1.820**	**2.020**	**2.027**	**2.081**	**2.261**	**2.241**	**2.264**
add_lev	—	4.209	4.562	4.569	4.475	4.781	4.751	4.387	4.562	4.369
add_skk	—	3.990	4.143	4.081	4.143	4.113	4.217	4.165	4.236	4.200
add_sim	—	2.921	2.909	3.185	2.505	2.244	2.313	2.340	2.461	2.611
mlt_lev	sim	3.956	4.160	4.025	4.076	4.030	4.084	4.002	4.002	3.803
mlt_skk	sim	3.739	3.685	3.749	3.766	3.938	3.778	3.778	3.682	3.734
mlt_sim	sim	**2.067**	**2.229**	**2.377**	**2.057**	**2.167**	2.246	**2.094**	**2.131**	**2.091**
add_lev	sim	4.101	4.283	4.165	4.234	4.458	4.507	4.446	4.549	4.623
add_skk	sim	3.823	4.126	4.227	4.251	4.350	4.406	4.515	4.505	4.552
add_sim	sim	3.315	2.517	2.458	2.616	2.057	**1.978**	2.165	**2.131**	2.197

Appendix D: Figures

Figure D.1: In-sample coverage for the models listed on the Y-axis. The left panel displays the estimated coverage probabilities ($\hat{\alpha}$), while the right panel shows the percentage of cases where the model estimation failed. Different colors represent different true coverage levels: blue for $\alpha = 0.01$, green for $\alpha = 0.025$, and red for $\alpha = 0.05$. Each point corresponds to one stock.
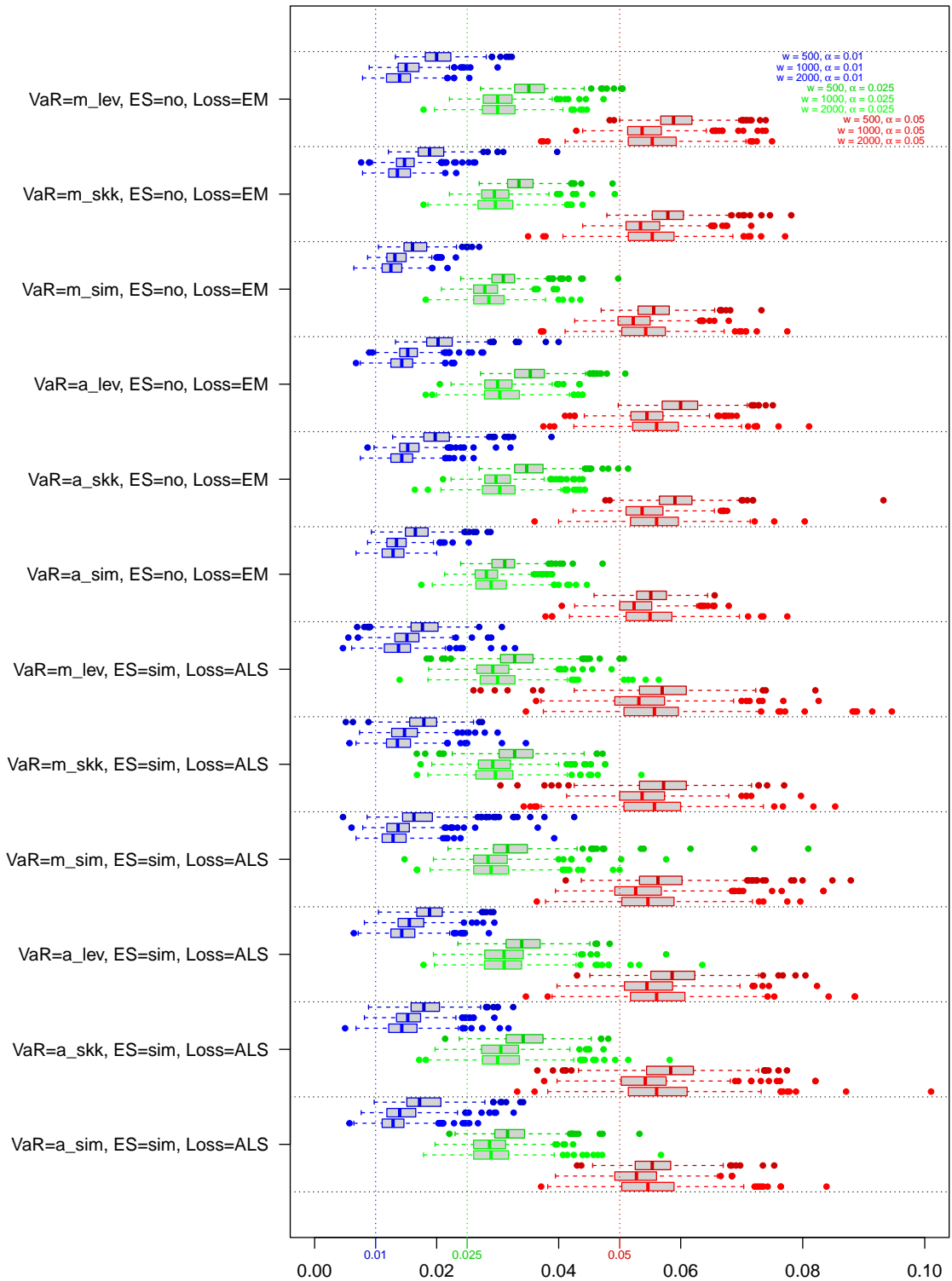
Figure D.2: Out-sample-coverage estimated probabilities ($\hat{\alpha}$) for the models listed on the Y-axis. Different colors represent different true coverage levels: blue for $\alpha = 0.01$, green for $\alpha = 0.025$, and red for $\alpha = 0.05$. Different intensities of the colours represent different widths of the rolling window, as depicted in top-right corner. Each point corresponds to one stock.
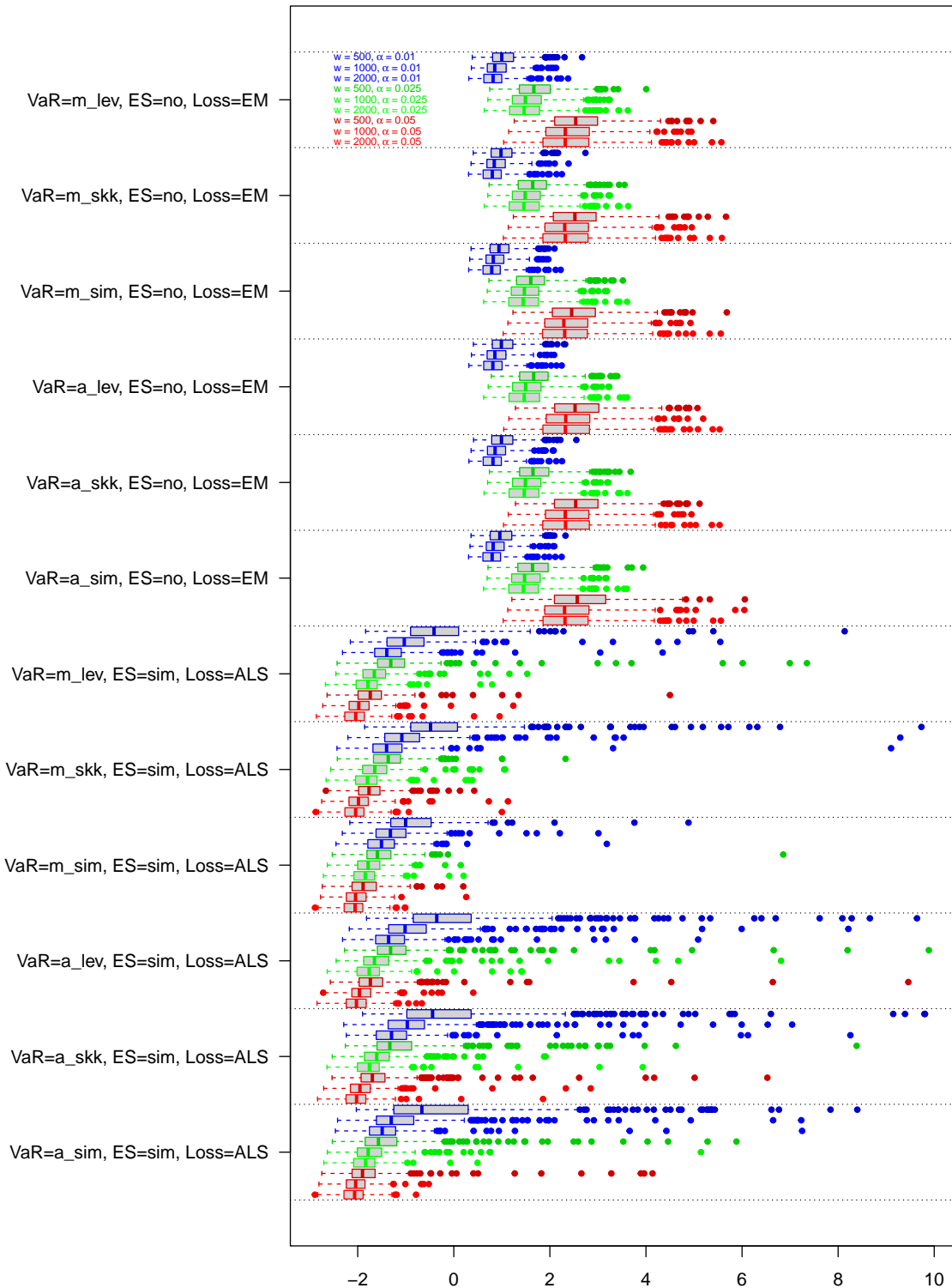
Figure D.3: Value of the loss function in out-sample-performance for the models listed on the Y-axis. Different colors represent different true coverage levels: blue for $\alpha = 0.01$, green for $\alpha = 0.025$, and red for $\alpha = 0.05$. Different intensities of the colours represent different widths of the rolling window, as depicted in top-right corner. Each point corresponds to one stock.

29