E-backtesting

Qiuqi Wang* Ruodu Wang† Johanna Ziegel‡

26th February 2025

Abstract

In the recent Basel Accords, the Expected Shortfall (ES) replaces the Value-at-Risk (VaR) as the standard risk measure for market risk in the banking sector, making it the most important risk measure in financial regulation. One of the most challenging tasks in risk modeling practice is to backtest ES forecasts provided by financial institutions. To design a model-free backtesting procedure for ES, we make use of the recently developed techniques of e-values and e-processes. Backtest e-statistics are introduced to formulate e-processes for risk measure forecasts, and unique forms of backtest e-statistics for VaR and ES are characterized using recent results on identification functions. For a given backtest e-statistic, a few criteria for optimally constructing the e-processes are studied. The proposed method can be naturally applied to many other risk measures and statistical quantities. We conduct extensive simulation studies and data analysis to illustrate the advantages of the model-free backtesting method, and compare it with the ones in the literature.

Keywords: E-values, e-processes, Expected Shortfall, Value-at-Risk, identification function.

1 Introduction

Forecasting risk measures is important for financial institutions to calculate capital reserves for risk management purposes. Regulators are responsible to monitor whether risk forecasts are correctly reported by conducting hypothesis tests known as backtests (see e.g., Christoffersen, 2011; McNeil et al., 2015, for general treatments). Regulatory backtests have several features distinct from traditional testing problems; see Acerbi and Székely (2014) and Nolde and Ziegel (2017). First, risk forecasts and realized losses arrive sequentially over time. Second, due to frequently

*Maurice R. Greenberg School of Risk Science, Georgia State University, U.S.A. ✉ qwang30@gsu.edu

†Department of Statistics and Actuarial Science, University of Waterloo, Canada. ✉ wang@uwaterloo.ca

‡Seminar of Statistics, ETH Zurich, Switzerland. ✉ ziegel@stat.math.ethz.ch

changing portfolio positions and the complicated temporal nature of financial data, the losses and risk predictions are neither independent nor identically distributed, and they do not follow any standard time-series models. Third, the tester (e.g., a regulator) is concerned about risk measure underestimation, which means high insolvency risk, whereas overestimation (i.e., being conservative) is secondary or acceptable. Fourth, the tester does not necessarily accurately know the underlying model used by a financial institution to produce risk predictions.

In financial practice, a well-adopted simple approach exists for backtesting the Value-at-Risk (VaR), which is the so-called three-zone approach based on binomial tests described in BCBS (2013); this approach is *model-free* in the sense that one directly tests the risk forecast without testing any specific family of models. In other statistical context, model-free methods are often called "nonparametric", but we use the term "model-free" to emphasize our specific setting of requiring no distributional forecast. More recently, the Basel Committee on Banking Supervision (BCBS, 2016, 2019) replaced VaR by the Expected Shortfall (ES) as the standard regulatory measure for market risk, mostly due to the convenient properties of ES, in particular, being able to capture tail risk.[1] However, as discussed by Gneiting (2011), ES is not elicitable, and backtesting ES is substantially more challenging than VaR. Table 1 summarizes the main features of existing methods backtesting ES. To the best of our knowledge, there is no model-free non-asymptotic backtesting method for ES. Moreover, except for Hoga and Demetrescu (2023), most of the backtesting methods in the existing literature only work for a fixed data size, and are thus not valid under optional stopping, or equivalently, not anytime valid (see e.g., Ramdas et al., 2023). This creates limitations to financial regulation practice where early rejections are highly desirable.

In this paper, we develop a model-free backtesting method for risk measures, including ES, using the concepts of e-values and e-tests (Shafer, 2021; Vovk and Wang, 2021; Grünwald et al., 2024). E-tests have important advantages over classical statistical tests (p-tests) based on p-values. Wang and Ramdas (2022, Section 2) collect many reasons for using e-values and e-tests, regarding high-dimensional asymptotics, composite models, sequential (any-time valid) inference, information accumulation, and robustness to model misspecification and dependence; other advantages of e-values have been illustrated by Vovk and Wang (2021), Vovk et al. (2022) and Grünwald et al. (2024). As a particularly relevant feature to our context, our proposed e-tests allow regulators to get alerted early as the e-process accumulates to a reasonably large value. This is different from scientific discoveries (such as genome studies) where a scientist may not be entitled to reject a

[1] Quoting BCBS (2016, p.1): *Use of ES will help to ensure a more prudent capture of "tail risk" and capital adequacy during periods of significant financial market stress.* See also Wang and Zitikis (2021) for an axiomatic justification of ES in financial regulation.

Table 1: Comparison of existing backtesting methods for ES

Literature	Parametric or dependence assumptions	Forecast structural assumptions	Fixed sample size	Asymptotic test	Reliance on VaR or distributional forecasts
MF00	yes	yes	yes	yes	yes
AS14	yes	yes	yes	yes	yes
DE17	yes	yes	yes	yes	yes
NZ17	yes	yes	yes	yes	yes
SQPQ21	yes	yes	yes	yes	yes
BD22	yes	yes	yes	yes	no
HD22	yes	yes	no	no	yes
This paper	no	no	no	no	yes

Notes: We use shortcuts MF00 for McNeil and Frey (2000), AS14 for Acerbi and Székely (2014), DE17 for Du and Escanciano (2017), NZ17 for Nolde and Ziegel (2017), SQPQ21 for Su et al. (2021), BD22 for Bayer and Dimitriadis (2022), and HD22 for Hoga and Demetrescu (2023). Parametric or dependence assumptions refer to those on loss distributions, time series models, stationarity, or strong mixing. Forecast structural assumptions refer to requirements on the forms and properties of risk forecasts. Acerbi and Székely (2014) proposed three methods of backtesting ES; The first two methods do not require specific forms of ES forecasts, but the third method requires ES to be estimated as realized ranks.

hypothesis based on merely "substantial" evidence. Noticing this, a multi-zone approach similar to the three-zone approach can be developed by setting different e-value thresholds in financial regulation. A simple illustrative example of our methodology is presented in Section 2.

The main contribution of this paper is four-fold: First, we define backtest e-statistics and propose new backtesting methods. In particular, we obtain backtest e-statistics for ES in Section 3 (Theorem 1), allowing us to construct e-processes to backtest ES, as well as other risk measures in Section 4 (Theorem 2). The backtesting method for VaR and ES is discussed in Section 4.2. Second, with backtest e-statistics chosen, the next important step to construct an e-process is choosing a suitable betting process, which we address in Section 5. We propose three methods to calculate the betting processes based on data. It turns out that these methods are asymptotically optimal (equivalent to an oracle betting process) in different situations (Theorem 3). Third, we characterize backtest e-statistics for the mean, the variance, VaR (Theorem 4), and ES (Theorem 5) in Section 6. All backtest e-statistics for these functionals take similar forms as mixtures between the constant function equal to 1 and a simple backtest e-statistic. Finally, through the simulation study and data analysis in Sections 7 and 8, we demonstrate detailed procedures of backtesting VaR and ES using e-values for practical operations of financial regulations.

Five sections are put in the e-companion. Section A shows the betting processes calculated via

Taylor approximation for VaR and ES. Section B discusses the link between backtest e-statistics and identification functions, useful for the characterization results in Section 6. Except for a short proof of Theorem 1, proofs of all results are relegated to Section C. Section D contains additional simulation and data analyses and some necessary details of the ones in the paper. Section E discusses a deliberate over-forecast strategy by the financial institution and how our methods can address this issue.

1.1 Related literature

Besides financial regulation, evaluating forecasting models and methods for major economic variables is also essential in the decision-making processes of government institutions and regulatory authorities. Earlier work on predictive ability tests and forecast selection includes Diebold and Mariano (1995), whose method was extended by West (1996), Clark and McCracken (2001), and Giacomini and White (2006). Unconditional backtests of VaR were considered by Kupiec (1995) on testing Bernoulli distributions, which were extended by Christoffersen (1998) to include testing independence of the VaR-violations. Engle and Manganelli (2004) tested conditional autoregressive VaR; Escanciano and Olmo (2010) further studied backtesting Value-at-Risk by considering the estimation risk; Berkowitz et al. (2011) unified existing evaluation methods of VaR; and Ziggel et al. (2014) proposed a Monte Carlo simulation-based backtesting method for VaR.

Due to its increasing importance and challenging nature, there are ample studies in the more recent literature on backtesting ES with different approaches. McNeil and Frey (2000) proposed bootstrap tests with iid innovations; Acerbi and Székely (2014, 2017) studied three backtesting methods under independent losses; Du and Escanciano (2017) designed parametric tests using cumulative violations; their approach was extended by Su et al. (2021) who adopted an efficient empirical likelihood method backtesting ES. Nolde and Ziegel (2017) studied comparative backtests among forecasting methods, and considered conditional and unconditional calibration tests for risk measures. The backtests using e-processes that we propose in this paper are testing conditional calibration. Indeed, they can be seen as natural anytime-valid analogues of the conditional calibration tests in Nolde and Ziegel (2017). Bayer and Dimitriadis (2022) introduced backtesting through a linear regression model; and Hoga and Demetrescu (2023) proposed sequential monitoring based on parametric distributions. The main features of these approaches are summarized in Table 1. See also Moldenhauer and Pitera (2017) for a method of backtesting empirical ES forecasts.

The literature on e-values has also been growing fast recently. E-values were used in the early literature in different disguises, although the term "e-value" was proposed by Vovk and Wang (2021).

For instance, e-values and e-tests were essentially used in the work of Wald (1945) and Darling and Robbins (1967), and they are central to the ideas of testing by betting and martingales (Shafer et al., 2011; Shafer and Vovk, 2019) and universal inference (Wasserman et al., 2020). E-values are shown to be useful in multiple hypothesis testing with dependence (Vovk and Wang, 2021), Bayesian models with optional sampling (Grünwald et al., 2024), false discovery rate control (Wang and Ramdas, 2022), high-dimensional regression (Ren and Barber, 2024), and many other statistical applications. Confidence sequences for tail risk measures including ES are studied by Agrawal et al. (2021), see also Casgrain et al. (2024). The recent survey paper of Ramdas et al. (2023) contains a review of this topic. Compared to p-values, e-values have their own interpretations, such as wealth levels by betting against the null hypothesis, generalized likelihood ratios, and post-hoc valid tests. We refer to the recent monograph Ramdas and Wang (2024) for a general treatment of e-values, including their statistical interpretations.

2 The e-backtesting procedure: An illustrative example

To prepare for the general methodology of e-backtesting, we first present a simple example of testing the mean and the variance. Denote by $[T] = \{1, \dots, T\}$ for any positive integer T. Suppose that we observe a sequence of iid data $(L_t)_{t\in[T]}$ from the random variable L. For $r > 0$ and $z \in \mathbb{R}$, we would like to test the nonparametric null hypothesis

$$H_0 : \mathrm{var}(L) \leqslant r \text{ and } \mathbb{E}[L] = z$$

against the alternative hypothesis H_1 that $\mathrm{var}(L) > r$, where $\mathbb{E}[L]$ is the expectation of L and $\mathrm{var}(L)$ is the variance of L. We are concerned about the variance being large, but not the mean. We consider sequential testing; that is, for each time point t we decide whether we reject H_0 based on $(L_s)_{s\in[t]}$ or continue, until all T data points are observed. Our approach based on an e-process can be described in the following steps.

(i) Pick a value $\alpha \in (0, 1)$, which is a significance level.

(ii) Choose a function $e : \mathbb{R}^3 \to \mathbb{R}$ such that $\mathbb{E}[e(L, r, z)] \leqslant 1$ under the null hypothesis, and $\mathbb{E}[e(L, r, z)] > 1$ under the alternative hypothesis. Define $X_t = e(L_t, r, z)$ for each $t \in [T]$. In this particular setting, the function e can be chosen as $e(x, r, z) = (x - z)^2/r$.

(iii) For each $t \in [T]$, choose a value $\lambda_t \in [0, 1]$ that can depend on $(L_s)_{s\in[t-1]}$.

(iv) Construct a stochastic process: $M_0 = 1$ and

$$M_t = (1 - \lambda_t + \lambda_t X_t) M_{t-1} = \prod_{s=1}^{t} (1 - \lambda_s + \lambda_s X_s), \quad t \in [T].$$

(v) Reject H_0 at time t if we observe $M_t \geqslant 1/\alpha$.

Under the null hypothesis, regardless of how $(\lambda_t)_{t\in[T]}$ is chosen, $(M_t)_{t\in[T]}$ is a non-negative supermartingale with respect to the σ-algebra $\sigma(L_1, \dots, L_t)$, which can be easily checked by

$$\mathbb{E}[M_t | \sigma(L_1, \dots, L_{t-1})] = (1 - \lambda_t + \lambda_t \mathbb{E}[X_t]) M_{t-1} \leqslant M_{t-1}.$$

By Ville's inequality, we have $\mathbb{P}(\sup_{t\in[T]} M_t \geqslant 1/\alpha) \leqslant \alpha \mathbb{E}[M_0] = \alpha$ (Ville, 1939); see Theorem 2. Therefore, we have type-I error control at level α. Moreover, under the alternative hypothesis H_1, because $\mathbb{E}[X_t] > 1$, $(M_t)_{t\in[T]}$ has a growing mean (it is in fact a submartingale), and we may hope that it grows with high probability as t increases, and thus it is able to reject the null hypothesis. Indeed, with properly chosen $(\lambda_t)_{t\in[T]}$, this process grows exponentially in the iid setting under H_1.

The rest of the paper formalizes the above idea via the theory of e-values. We will prove the claims above in a much more general framework of risk measures and sequential forecasts. Some key ingredients are explained below.

1. Variance and mean are replaced by general functionals ρ (risk measure to backtest, e.g., capital requirement for the portfolio) and ϕ (auxiliary information, sometimes not needed), respectively.

2. The random variable $X_t = e(L_t, r, z)$ in (ii) is called an e-variable, and its realization an e-value. The functions e for general risk measures will be formally defined in Section 3 with examples for several important cases. The functions e do not depend on the specific distributions of the loss L, thus they are model-free. Their characterizations are studied in Section 6.

3. In our general setting, the data are not iid, and a bank provides sequential risk forecasts (r_t, z_t) each day for their portfolio loss L_t in the next day. As formally studied in Section 4, the null hypothesis is

$$H_0: \quad r_t \geqslant \rho(L_t | \mathcal{F}_{t-1}) \text{ and } z_t = \phi(L_t | \mathcal{F}_{t-1}) \quad \text{for } t \in [T],$$

and the alternative is $r_t < \rho(L_t | \mathcal{F}_{t-1})$ for some t (i.e., the bank is under-forecasting their

risk), where $\mathcal{F}_t$ is the σ-algebra of available information at time t.

4. The choice of $(\lambda_t)_{t\in[T]}$ in (iii) is essential for the power of the above procedure. We offer several ways to compute $(\lambda_t)_{t\in[T]}$ and study their optimality in Section 5.

5. We allow T to be infinite or data-dependent, which is practically useful as we usually do not know a priori how many data points we will sample in sequential testing. The test can stop at any time. The flexibility of T is a feature of e-tests, which allows us to address situations more general than the ones considered in the literature with fixed T.

6. The level α can be chosen differently from traditional tests, because tests using e-values have a different type of guarantee in addition to a type-I error control. Values of α like $1/5$ and $1/10$ are already useful in the e-value literature (although their type-I error guarantee is not very small). This is discussed in Section 4.

3 E-values and backtest e-statistics

The approach in Section 2 builds on the theory of e-values. This section first reviews e-values and risk measures, and then introduces backtest e-statistics, which are further illustrated with examples, including the important case of ES.

3.1 E-values and risk measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be any probability space. A (composite) hypothesis H is a set of probability measures on $(\Omega, \mathcal{F})$. A hypothesis H is simple if it is a singleton. Following the terminology of Vovk and Wang (2021), an *e-variable* for H is a random variable $E : \Omega \to [0, \infty]$ such that $\mathbb{E}^P[E] \leqslant 1$ for each $P \in H$. We denote by $\mathcal{E}_H$ the set of e-variables for a hypothesis H and by $\mathcal{E}_P$ the set of e-variables for the simple hypothesis $\{P\}$. An *e-test* rejects the hypothesis H if a realized e-variable, called an *e-value*, is larger than a given threshold. A common rule of thumb is that an e-value of 10 represents strong evidence, see Remark 3 for more discussion on thresholds for e-values. Markov's inequality guarantees that the reciprocal of an e-value is a conservative p-value. A non-negative stochastic process $(E_t)_{t\in K}$, $K \subseteq \mathbb{N}$, adapted to a given filtration, is an *e-process* for H if $\mathbb{E}^P[E_\tau] \leqslant 1$ for all stopping times τ taking values in K and each $P \in H$.

Let d be a positive integer. The *model space* $\mathcal{M}$ is a set of distributions on $\mathbb{R}$. The value of the functional $\psi = (\rho, \phi_1, \dots, \phi_{d-1}) : \mathcal{M} \to \mathbb{R}^d$ represents the collection of available statistical information, where ρ is the risk prediction to be tested, and $\phi = (\phi_1, \dots, \phi_{d-1})$ contains auxiliary

information. A simple example to keep in mind for now is that ρ is the variance (which we aim to backtest) and ϕ is the mean (which helps to backtest the variance) as in Section 2. In this case, $d = 2$. If $d = 1$, then the only available information is the predicted value of ρ. We also write $\psi(X) = \psi(F)$ for a random variable $X \overset{\mathrm{d}}{\sim} F \in \mathcal{M}$, that is, when F is the distribution of X.

Let $\mathcal{M}_q$, $\mathcal{M}_\infty$, and $\mathcal{M}_0$ represent the set of distributions on $\mathbb{R}$ with finite q-th moment for $q \in (0, \infty)$, that of compactly supported distributions on $\mathbb{R}$, and that of all distributions on $\mathbb{R}$, respectively. For level $p \in (0, 1)$, the *Value-at-Risk (VaR)* is defined as the lower p-quantile:

$$\mathrm{VaR}_p(F) = Q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geqslant p\}, \quad F \in \mathcal{M}_0,$$

and the *Expected Shortfall (ES)* is defined as

$$\mathrm{ES}_p(F) = \frac{1}{1-p} \int_p^1 \mathrm{VaR}_\alpha(F)\, \mathrm{d}\alpha, \quad F \in \mathcal{M}_1.$$

VaR and ES belong to the class of dual utilities in Yaari (1987) and Schmeidler (1989). The pair $(\rho, \phi) = (\mathrm{ES}_p, \mathrm{VaR}_p)$ is a main object of interest in our paper. For $\psi = (\rho, \phi_1, \dots, \phi_{d-1})$, the natural domains of $\rho, \phi_1, \dots, \phi_{d-1}$ are not necessarily identical, and the domain of ψ is their intersection.

3.2 Backtest e-statistics

Next, we introduce the key tool we use for our e-tests.

Definition 1 (Point and one-sided e-statistics)**.** Let $\mathcal{P} \subseteq \mathcal{M}$. A *$\mathcal{P}$-point e-statistic for $\psi : \mathcal{M} \to \mathbb{R}^d$* is a measurable function $e : \mathbb{R}^{d+1} \to [0, \infty]$ satisfying $\int_{\mathbb{R}} e(x, \psi(F))\, \mathrm{d}F(x) \leqslant 1$ for each $F \in \mathcal{P}$. A measurable function $e : \mathbb{R}^{d+1} \to [0, \infty]$ is a *$\mathcal{P}$-one-sided e-statistic for $\psi = (\rho, \phi)$* if

$$\int_{\mathbb{R}} e(x, r, \phi(F))\, \mathrm{d}F(x) \leqslant 1 \text{ for each } F \in \mathcal{P} \text{ and } r \geqslant \rho(F).$$

If $\mathcal{P} = \mathcal{M}$, we speak of point or one-sided e-statistics for ψ, respectively, without mentioning $\mathcal{P}$.

We have seen e-statistics used in the example in Section 2. Using the language of e-values, a $\mathcal{P}$-point e-statistic is a function e such that $e(X, r, z)$ is an e-variable for the hypothesis

$$H_0 : X \text{ follows any } F \in \mathcal{P} \text{ with } \rho(F) = r \text{ and } \phi(F) = z$$

for each (r, z) in the range $\psi(\mathcal{M})$ of $\psi = (\rho, \phi)$. Here, the conditions in the null hypothesis are

equalities. In case $d = 1$, $e(X, r)$ is an e-variable for the hypothesis

$$H_0 : X \text{ follows any } F \in \mathcal{P} \text{ with } \rho(F) = r.$$

A $\mathcal{P}$-one-sided e-statistic for ψ is a function e such that $e(X, r, z)$ is an e-variable for the hypothesis

$$H_0 : X \text{ follows any } F \in \mathcal{P} \text{ with } \rho(F) \leqslant r \text{ and } \phi(F) = z$$

for each (r, z) in the range of ψ (see Section 2). It is clear that one-sided e-statistics are always point e-statistics. The next definition states desirable properties of e-statistics under the alternative.

Definition 2 (Backtest e-statistics)**.** Let $\mathcal{P} \subseteq \mathcal{M}$. A $\mathcal{P}$-one-sided e-statistic $e : \mathbb{R}^{d+1} \to [0, \infty]$ for $\psi = (\rho, \phi)$ is a *$\mathcal{P}$-backtest e-statistic for ψ* if $\int_{\mathbb{R}} e(x, r, z)\, \mathrm{d}F(x) > 1$ for all $(r, z) \in \psi(\mathcal{P})$ and $F \in \mathcal{P}$ with $\rho(F) > r$. The backtest e-statistic e is called *monotone* if, in addition, $r \mapsto e(x, r, z)$ is decreasing for each (x, z).

The property of having a mean larger than 1 under the alternative is crucial for the e-statistic. If the risk ρ is underestimated, then a backtest e-statistic will have a mean that is larger than 1, regardless of whether the prediction of the auxiliary functional ϕ is truthful. The function e in Section 2 is a backtest e-statistic for $\psi = (\mathrm{var}, \mathbb{E})$, with details explained in Example 2.

In our backtesting procedure, detailed in Section 4, we will use e-variables of the form $(1 - \lambda) + \lambda e(X, r, z)$ for some $\lambda \in [0, 1)$ and loss random variable X, such that $e(X, r, z)$ is an e-variable for the null hypothesis that (r, z) is correctly specified. The following lemma justifies that using a backtest e-statistic for ψ yields positive e-power for some λ. The e-power of an e-variable E for a probability measure Q is defined as $\mathbb{E}^Q[\log E]$ (Vovk and Wang, 2024b). As we see in Section 5, the e-power is closely related to the Kelly criterion and growth-optimal portfolios, a central concept for e-values (e.g., Shafer and Vovk, 2019; Grünwald et al., 2024).

Lemma 1. *For any random variable $E \geqslant 0$, we have*

$$\mathbb{E}[E] > 1 \iff \mathbb{E}[\log((1 - \lambda) + \lambda E)] > 0 \text{ for some } \lambda \in [0, 1].$$

By Lemma 1, if e is a backtest e-statistic and $r < \rho(F)$, then for any z there exists $\lambda \in (0, 1)$ such that $(1 - \lambda) + \lambda e(X, r, z)$ has positive e-power. This condition is sufficient for establishing consistency of our backtesting procedures in some settings; see Theorem 3 in Section 5.

We are only interested in forecast values (r, z) in the set $\psi(\mathcal{P})$. Any forecast values outside $\psi(\mathcal{P})$ can be rejected automatically. Our idea of backtest e-statistics specifically addresses the un-

derestimation of ρ, which is consistent with the motivation in regulatory backtesting. If a backtest e-statistic is monotone, then an overestimation of the risk is rewarded: An institution being scrutinized by the regulator can deliberately report a higher risk value (which typically means higher capital reserve) to pass to the regulatory test, thus rewarding prudence.

3.3 Examples

Below, we give a few examples of backtest e-statistics for some common risk measures. Throughout, we use the convention that $0/0 = 1$ and $1/0 = \infty$, and let $\mathbb{R}_+ = [0, \infty)$.

Example 1 (Backtest e-statistic for the mean)**.** Let $\mathcal{P}$ be the set of distributions on $\mathbb{R}_+$ in $\mathcal{M}_1$. Define the function $e(x, r) = x/r$ for $x, r \geqslant 0$. In this case, we have $\mathbb{E}[e(X, r)] \leqslant 1$ for all random variables X with distribution in $\mathcal{P}$ and $r \geqslant \mathbb{E}[X]$. Moreover, for any such X, $\mathbb{E}[X] > r \geqslant 0$ implies $\mathbb{E}[e(X, r)] > 1$. Therefore, e is a monotone $\mathcal{P}$-backtest e-statistic for the mean.

Example 2 (Backtest e-statistic for $(\mathrm{var}, \mathbb{E})$)**.** Consider $(\mathrm{var}, \mathbb{E}) : \mathcal{M}_2 \to \mathbb{R}^2$ as in Section 2. The function $e(x, r, z) = (x - z)^2/r$ for $x, z \in \mathbb{R}$ and $r \geqslant 0$ is a monotone backtest e-statistic for $(\mathrm{var}, \mathbb{E})$. To see this, for all random variables X with distribution in $\mathcal{M}_2$, we have

$$\mathbb{E}[e(X, r, \mathbb{E}[X])] = \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{r} \leqslant 1$$

for $r \geqslant \mathrm{var}(X)$. Moreover, since $z = \mathbb{E}[X]$ minimizes $\mathbb{E}[(X - z)^2]$ over $z \in \mathbb{R}$ and $\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$, $\mathrm{var}(X) > r \geqslant 0$ implies

$$\mathbb{E}[e(X, r, z)] = \frac{\mathbb{E}[(X - z)^2]}{r} \geqslant \frac{\mathrm{var}(X)}{r} > 1.$$

Example 3 (Backtest e-statistic for a quantile)**.** Take $p \in (0, 1)$. Define the function

$$e_p^Q(x, r) = \frac{1}{1 - p} \mathbb{1}_{\{x > r\}}, \quad x, r \in \mathbb{R}. \tag{1}$$

We have $\mathbb{E}[e_p^Q(X, r)] \leqslant 1$ for any random variable X with distribution F and $r \geqslant Q_p(F)$. Moreover, $Q_p(F) > r$ implies $\mathbb{P}(X > r) > 1 - p$, and hence $\mathbb{E}[e_p^Q(X, r)] > 1$. Therefore, e_p^Q is a monotone backtest e-statistic forthe p-quantile.

Example 4 (Backtest e-statistic for an expected loss)**.** For some $a \in \mathbb{R}$, let $\ell : \mathbb{R} \to [a, \infty)$ be a function that is interpreted as a loss. Define the function $e(x, r) = (\ell(x) - a)/(r - a)$ for $x \in \mathbb{R}$ and $r \geqslant a$. Analogously to Example 1, e is a monotone backtest e-statistic for the expected loss $F \mapsto \int \ell \, \mathrm{d}F$ on its natural domain.

The choice of a backtest e-statistic e for ψ is not necessarily unique. For instance, a linear combination of e with 1 with the weight between 0 and 1 is also a backtest e-statistic for ψ. Depending on the specific situation, either e-statistic may be useful in practice.

Remark 1. The functional $\psi = (\rho, \phi) = (\mathrm{var}, \mathbb{E})$ in Example 2 is an example of a Bayes pair; that is, there exists a measurable function $L : \mathbb{R}^{d+1} \to \mathbb{R}$, called the *loss function*, such that

$$\phi(F) \in \operatorname*{arg\,min}_{z\in\mathbb{R}^d} \int L(z,x)\,\mathrm{d}F(x) \text{ and } \rho(F) = \min_{z\in\mathbb{R}^d} \int L(z,x)\,\mathrm{d}F(x), \quad F \in \mathcal{M}, \tag{2}$$

where $\int L(z,x)\,\mathrm{d}F(x)$ is assumed to be well-defined for each $z \in \mathbb{R}^d$, $F \in \mathcal{M}$ (Fissler and Ziegel, 2016; Frongillo and Kash, 2021; Embrechts et al., 2021). The function L is the square loss function in the case of $(\mathrm{var}, \mathbb{E})$. Bayes pairs often admit backtest e-statistics. A typical example commonly used in risk management practice is $(\mathrm{ES}, \mathrm{VaR})$ treated below,[2] which is our main focus.

We will see that, for $p \in (0,1)$, the function

$$e_p^{\mathrm{ES}}(x, r, z) = \frac{(x-z)_+}{(1-p)(r-z)}, \quad x \in \mathbb{R},\ z \leqslant r \tag{3}$$

defines a backtest e-statistic for $(\mathrm{ES}_p, \mathrm{VaR}_p)$, where y_+ denotes $\max\{y, 0\}$ for any real number y. Recall the convention that $0/0 = 1$ and $1/0 = \infty$, and set $e_p^{\mathrm{ES}}(x, r, z) = \infty$ if $r < z$, which is a case of no relevance since $\mathrm{ES}_p(F) \geqslant \mathrm{VaR}_p(F)$ for any $F \in \mathcal{M}_1$.

Theorem 1. *The function e_p^{ES} is a monotone backtest e-statistic for $(\mathrm{ES}_p, \mathrm{VaR}_p)$.*

Proof. By the VaR-ES relation of Rockafellar and Uryasev (2002), for any random variable X with finite mean,

$$\mathrm{VaR}_p(X) \in \operatorname*{arg\,min}_{z\in\mathbb{R}} \left\{ z + \frac{1}{1-p}\mathbb{E}[(X-z)_+] \right\}, \tag{4}$$

and

$$\mathrm{ES}_p(X) = \min_{z\in\mathbb{R}} \left\{ z + \frac{1}{1-p}\mathbb{E}[(X-z)_+] \right\}. \tag{5}$$

This indicates that $(\mathrm{ES}_p, \mathrm{VaR}_p)$ is a Bayes pair by (2) with loss function $L : (z,x) \mapsto z + (x-z)_+/(1-p)$. Since $L(z,x) \geqslant z$, for $r \geqslant z$, we have that $e_p^{\mathrm{ES}}(x,r,z) = (L(z,x)-z)/(r-z) \geqslant 0$, and it is decreasing in r. We have for $r \geqslant \mathrm{ES}_p(X)$ that

$$\mathbb{E}\left[\frac{L(\mathrm{VaR}_p(X), X) - \mathrm{VaR}_p(X)}{r - \mathrm{VaR}_p(X)}\right] = \frac{\mathrm{ES}_p(X) - \mathrm{VaR}_p(X)}{r - \mathrm{VaR}_p(X)} \leqslant 1.$$

[2] We sometimes omit the probability level p in VaR_p, ES_p and $(\mathrm{VaR}_p, \mathrm{ES}_p)$ in the text (but never in equations).

Furthermore, for $z < r \leqslant \mathrm{ES}_p(X)$,

$$\mathbb{E}\left[\frac{L(z,X)-z}{r-z}\right] \geqslant \frac{\mathrm{ES}_p(X)-z}{r-z} \geqslant 1$$

with equality if and only if $r = \mathrm{ES}_p(X)$. □

While Examples 1-4 and Theorem 1 show that interesting backtest e-statistics exist, much more can be said about their general structure; see Section 6.

4 E-backtesting risk measures

We next present general methodology for backtesting risk measures via e-statistics in a sequential setting. We will use one-sided e-statistics in Definition 1. Backtest e-statistics in Definition 2 are relevant to have powerful backtests (see Section 5) but we first consider validity in this section.

4.1 General risk measures

Let T be any time horizon, which may be fixed, infinite, or adaptive, i.e., depending on the data observed. Let the σ-algebra $\mathcal{F}_t$ represent all available information up to time $t \in [T]$, such that $\mathcal{F}_m \subseteq \mathcal{F}_n$ for all $m \leqslant n$. Let $(L_t)_{t\in[T]}$ be a sequence of realized losses that are adapted to the filtration $(\mathcal{F}_t)_{t\in[T]}$. Denote by $\rho(L_t|\mathcal{F}_{t-1})$ and $\phi(L_t|\mathcal{F}_{t-1})$, respectively, the values of ρ and ϕ applied to the conditional distribution of L_t given $\mathcal{F}_{t-1}$. Let r_t and z_t be forecasts for $\rho(L_t|\mathcal{F}_{t-1})$ and $\phi(L_t|\mathcal{F}_{t-1})$ made at time $t-1$, respectively. Note that $\rho(L_t|\mathcal{F}_{t-1})$ and $\phi(L_t|\mathcal{F}_{t-1})$ are random variables and $\mathcal{F}_{t-1}$-measurable for all functionals of interest (see e.g., Fissler and Holzmann, 2022).

We assume that the risk forecasts r_t and z_t are obtained based on past market information and all other possible factors that may affect the decisions of risk predictors in financial institutions. For instance, the information may even include throwing a die or random events such as coffee spilling; all these events up to time $t-1$ are included in $\mathcal{F}_{t-1}$.

We test the following null hypothesis:

$$H_0: \quad r_t \geqslant \rho(L_t|\mathcal{F}_{t-1}) \text{ and } z_t = \phi(L_t|\mathcal{F}_{t-1}) \quad \text{for } t \in [T]. \tag{6}$$

Rejecting (6) implies, in particular, rejecting $r_t = \rho(L_t|\mathcal{F}_{t-1})$ and $z_t = \phi(L_t|\mathcal{F}_{t-1})$ for all $t \in [T]$. In the special case that $d = 1$ (i.e. we do not need auxiliary information from ϕ), (6) becomes

$$H_0: \quad r_t \geqslant \rho(L_t|\mathcal{F}_{t-1}) \quad \text{for } t \in [T].$$

Remark 2. Since ρ is the regulatory risk measure of interest, over-predicting ρ is conservative. On the other hand, ϕ represents some additional statistical information and it may not relate to measuring financial risk. Hence, over-predicting ϕ is not necessarily conservative. See Example 5 below for a sanity check. Therefore, it is more natural to test an equality of the auxiliary information z_t in (6) instead of an inequality; note also that this hypothesis is still more lenient than testing a specified loss distribution. For the case where a financial institution is conservative for both the risk measures ρ and ϕ, see Section 4.2.

For a nonnegative function $e : \mathbb{R}^{d+1} \to [0, \infty]$, let $X_t = e(L_t, r_t, z_t)$ for each t. We construct the following stochastic process: $M_0 = 1$ and

$$M_t(\boldsymbol{\lambda}) = (1 - \lambda_t + \lambda_t X_t) M_{t-1}(\boldsymbol{\lambda}) = \prod_{s=1}^{t} (1 - \lambda_s + \lambda_s X_s), \quad t \in [T], \tag{7}$$

where the process $\boldsymbol{\lambda} = (\lambda_t)_{t \in [T]}$ is chosen such that λ_t is a function of $(L_{s-1}, r_s, z_s)_{s \in [t]}$ and takes values in $[0, 1]$ for $t \in [T]$.[3] Suppose that e is a one-sided e-statistic for $(\rho, \phi) : \mathcal{M} \to \mathbb{R} \times \mathbb{R}^{d-1}$. In the construction of (7), the user-chosen inputs are the process $\boldsymbol{\lambda}$ and the e-statistic e.

We have by definition that X_t is an e-variable conditional on $\mathcal{F}_{t-1}$ under H_0 for all $t \in [T]$. As suggested by Vovk and Wang (2024a), the only admissible (or unwasteful) way to combine these e-variables is through the martingale function (7); see also the universal representation result in Proposition 3 of Waudby-Smith and Ramdas (2024). The e-process in (7) may be interpreted as the payoffs of a betting strategy against the null hypothesis H_0 as in Shafer and Vovk (2019). In this betting game, the initial capital is $M_0 = 1$ and all the money is invested at each step, split between a random payoff X_t and a fixed payoff 1. The payoff per capital at each step is $1 - \lambda_t + \lambda_t X_t$ for $t \in [T]$. As a result, the player earns money at step t if $X_t > 1$, meaning that there is some evidence against the null hypothesis, which is the interpretation of an e-value larger than 1. In this sense, we call the process $\boldsymbol{\lambda}$ in (7) a *betting process* (see step (iii) in Section 2). In particular, $\lambda_t < 1$ can be interpreted as keeping part of one's money in the pocket, and $\lambda_t = 0$ means one does not bet at all in this step. The following theorem follows from Ville's well-known inequality (Ville, 1939), indicating that $(M_t(\boldsymbol{\lambda}))_{t \in \{0, \dots, T\}}$ in (7) is a non-negative supermartingale, so in particular an e-process under the null hypothesis in (6).

Theorem 2. *Suppose that e is a one-sided e-statistic for $(\rho, \phi) : \mathcal{M} \to \mathbb{R} \times \mathbb{R}^{d-1}$. Under H_0 in (6),*

[3] More generally, we may allow λ_t to be $\mathcal{F}_{t-1}$-measurable instead of $\sigma((L_{s-1}, r_s, z_s)_{s \in [t]})$-measurable, but this adds no further methodological value.

$(M_t(\boldsymbol{\lambda}))_{t\in\{0,\dots,T\}}$ *in* (7) *is a non-negative supermartingale with* $M_0 = 1$*, and for each* $\alpha \in (0,1)$*,*

$$\mathbb{P}\left(\sup_{t\in\{0,\dots,T\}} M_t(\boldsymbol{\lambda}) \geqslant \frac{1}{\alpha}\right) \leqslant \alpha.$$

Based on Theorem 2, we will use the e-test that arises from the e-variable $M_\tau(\boldsymbol{\lambda})$ where $M(\boldsymbol{\lambda})$ is the e-process given by (7) and τ is the stopping time $\min\{T, \inf\{t \geqslant 0 : M_t(\boldsymbol{\lambda}) \geqslant 1/\alpha\}\}$ with $\inf \varnothing = \infty$. This is common practice in testing with e-values.

Remark 3. The choice of α in the threshold $1/\alpha$ for e-processes has been extensively discussed in the literature; see Shafer (2021); Vovk and Wang (2021); Grünwald et al. (2024) and Wang and Ramdas (2022). Theorem 2 only gives a minimal guarantee of type-I error α, but the guarantee is actually much stronger, although not in terms of a provable smaller type-I error. In particular, it gives an *anytime-valid* type-I error control, that is, validity under arbitrary stopping times and optional continuation. The empirical type-I error is usually much smaller than α for a moderate sample size (see Section 7), and it can be close to α for a very large sample size (see Section D.3). Choosing the threshold also depends on the objective of the backtest. E-value and p-value tests require different types of guarantee in general. If the goal is for early warning and not a decisive rejection, as in financial regulation, then a threshold 2 or 5 is useful. If the goal is to reject the null with a provable type-I error α, as in the traditional setting, then a threshold $1/\alpha$ should be chosen. In our simulation and empirical studies, we will use the thresholds 2, 5 and 10, which (although only with type-I error guarantee of 50%, 20% and 10% by Theorem 2) roughly correspond to minor, substantial and strong evidence for e-values according to Vovk and Wang (2021).

4.2 E-backtesting Value-at-Risk and Expected Shortfall

To put our general ideas in the context of financial regulation, we focus on backtesting VaR and ES in this section. Let L_t be the random loss at time t. For the case of backtesting VaR, $(r_t)_{t\in[T]}$ are the forecasts for $\mathrm{VaR}_p(L_t|\mathcal{F}_{t-1})$, $p \in (0,1)$. As we see in Example 3, the function $e_p^Q(L_t, r_t)$ defined in (1) is an e-variable under the following null hypothesis that we are testing:

$$H_0: \quad r_t \geqslant \mathrm{VaR}_p(L_t|\mathcal{F}_{t-1}), \qquad t \in [T]. \tag{8}$$

For backtesting ES, $(r_t)_{t\in[T]}$ and $(z_t)_{t\in[T]}$ are the forecasts for $\mathrm{ES}_p(L_t|\mathcal{F}_{t-1})$ and $\mathrm{VaR}_p(L_t|\mathcal{F}_{t-1})$, $p \in (0,1)$, respectively. By Theorem 1, $e_p^{\mathrm{ES}}(L_t, r_t, z_t)$ is an e-variable under the null hypothesis

$$H_0: \quad r_t \geqslant \mathrm{ES}_p(L_t|\mathcal{F}_{t-1}) \text{ and } z_t = \mathrm{VaR}_p(L_t|\mathcal{F}_{t-1}), \qquad t \in [T]. \tag{9}$$

In practice, a financial institution may use a conservative model for risk management purposes, which leads to underestimation of both VaR and ES. In the following proposition, we illustrate that $e_p^{\mathrm{ES}}(L_t, r_t, z_t)$ is a valid e-variable in case both $\mathrm{VaR}_p(L_t|\mathcal{F}_{t-1})$ and $\mathrm{ES}_p(L_t|\mathcal{F}_{t-1})$ are over-predicted, together with their difference.

Proposition 1. *For $p \in (0,1)$, $(e_p^{\mathrm{ES}}(L_t, r_t, z_t))_{t\in[T]}$ are e-variables for*

$$H_0 : \ z_t \geqslant \mathrm{VaR}_p(L_t|\mathcal{F}_{t-1}) \quad \textit{and} \quad r_t - z_t \geqslant \mathrm{ES}_p(L_t|\mathcal{F}_{t-1}) - \mathrm{VaR}_p(L_t|\mathcal{F}_{t-1}), \ t \in [T]. \tag{10}$$

In practice, the equality $z_t = \mathrm{VaR}_p(L_t|\mathcal{F}_{t-1})$ in (9) is unlikely to hold exactly. For example, even with correctly specified forecasting models, there will still be estimation error in parameters. The obtained e-values should be seen as a quantification of how strong the evidence is against H_0; this is a useful feature of e-values, as discussed by e.g., Shafer (2021) and Grünwald et al. (2024). On the other hand, H_0 in (10) can hold in practice as it is formulated using inequalities.

The hypothesis H_0 in (10) is stronger than

$$H_0 : z_t \geqslant \mathrm{VaR}_p(L_t|\mathcal{F}_{t-1}) \ \text{ and } \ r_t \geqslant \mathrm{ES}_p(L_t|\mathcal{F}_{t-1}). \tag{11}$$

In contrast to (10), $e_p^{\mathrm{ES}}(L_t, r_t, z_t)$ is not necessarily an e-variable for (11). For instance, $\mathbb{E}[e_p^{\mathrm{ES}}(L_t, r_t, z_t)] = \infty$ if $r_t = z_t = \mathrm{ES}_p(L_t|\mathcal{F}_{t-1})$ and $\mathbb{P}(L_t > z_t) > 0$. It implies that over-predicting $\mathrm{VaR}_p(L_t|\mathcal{F}_{t-1})$ does not always lead to a smaller e-value $e_p^{\mathrm{ES}}(L_t, r_t, z_t)$; in contrast, over-predicting $\mathrm{ES}_p(L_t|\mathcal{F}_{t-1})$ always reduces the resulting e-value. The following example shows that a poor VaR forecast could result in large e-values although it is obtained by over-prediction and satisfies (11).

Example 5. For $p \in (0,1)$, a continuously distributed random variable X with $a = \mathrm{VaR}_p(X) < \mathrm{ES}_p(X) = 1$ (this implies $\mathbb{P}(X \leqslant 1) < 1$) has a heavier tail than Y with $\mathrm{VaR}_p(Y) = \mathrm{ES}_p(Y) = 1$ (this could happen if Y has a point-mass at 1, and it implies $\mathbb{P}(Y \leqslant 1) = 1$). Thus, intuitively, a forecaster producing the random loss X is more conservative than that producing Y. This shows that over-predicting both VaR_p and ES_p does not always mean that the forecaster is more conservative about the risk. Our backtest e-statistic e_p^{ES} can detect this, because $\mathbb{E}[e_p^{\mathrm{ES}}(X, 1, a)] \leqslant 1$ while $\mathbb{E}[e_p^{\mathrm{ES}}(X, 1, 1)] = \infty$, thus correctly rejecting the forecast $(1, 1)$ of $(\mathrm{ES}_p(X), \mathrm{VaR}_p(X))$ but not rejecting the truthful forecast $(1, a)$, although $a < 1$.

The following example collects some practical situations of conservative forecasts. In each case, $e_p^{\mathrm{ES}}(L_t, r_t, z_t)$ yields a valid e-variable.

Example 6. (i) $r_t = \mathrm{ES}_q(L_t|\mathcal{F}_{t-1})$ and $z_t = \mathrm{VaR}_q(L_t|\mathcal{F}_{t-1})$ for $q > p$:

$$\mathbb{E}[e_p^{\mathrm{ES}}(L_t, r_t, z_t)] = \frac{1-q}{1-p}\mathbb{E}[e_q^{\mathrm{ES}}(L_t, \mathrm{ES}_q(L_t|\mathcal{F}_{t-1}), \mathrm{VaR}_q(L_t|\mathcal{F}_{t-1}))] = \frac{1-q}{1-p} < 1.$$

In this situation, VaR_p and ES_p are over-predicted by lifting the confidence level p to q. For instance, this may represent the output of a stress-testing scenario which amplifies the probability of extreme losses.

(ii) $r_t = c_1\mathrm{ES}_p(L_t|\mathcal{F}_{t-1})$ and $z_t = c_2\mathrm{VaR}_p(L_t|\mathcal{F}_{t-1}) \geqslant 0$ for $c_1 \geqslant c_2 \geqslant 1$, as justified by Proposition 1. In particular, VaR_p and ES_p can be over-predicted by the same multiplicative factor.

(iii) $r_t = \mathrm{ES}_p(L_t|\mathcal{F}_{t-1}) + b_1$ and $z_t = \mathrm{VaR}_p(L_t|\mathcal{F}_{t-1}) + b_2$ for $b_1 \geqslant b_2 \geqslant 0$, as justified by Proposition 1. In particular, VaR_p and ES_p can be over-predicted by the same absolute amount.

Remark 4. Our e-backtesting method of ES and the method based on cumulative violations introduced in Du and Escanciano (2017) (we call it the cumulative violation method) have several different features. First, the cumulative violation method requires distributional forecasts $\hat{u}_t(\hat{\theta})$ as input based on some parametric distribution; our e-backtesting method needs ES and VaR forecasts that can be arbitrarily reported. This provides more flexibility in practice in the sense that our method does not require special treatment of estimation effects as in Du and Escanciano (2017) and Hoga and Demetrescu (2023). Second, the cumulative violation method is a two-sided test and focuses on detecting model misspecification; our method is a one-sided test focusing only on the underestimation of ES. This means that we do not reject the null as long as ES is not underestimated even though the forecasts are obtained based on a wrong model or no specific model is assumed. Third, the cumulative violation method relies on a fixed sample size T and an asymptotic model, which means its statistical validity requires it to be only evaluated at the end of the sampling period T that is large enough; our method is sequential and is valid at any stopping time, where detections can be achieved much earlier. This is desirable in risk management applications in timely detecting insufficient risk predictions. Most other classical backtesting methodologies become invalid when evaluated before the end of the pre-specified time period set for testing; see Table 1.

Remark 5. As explained in Section 3, our e-processes are designed to test underestimation of risk measures. If one is interested in testing overestimation, for the case of VaR, one can simply use $-L_t$ instead of L_t, and build an e-process. This leads to a two-sided test by averaging two e-processes, one for each side. On the other hand, testing overestimation of ES seems to be very challenging, and our current method does not apply. This asymmetry between the two sides is due to the nature of ES as a one-sided tail risk measure.

5 Choosing the betting process

One of the essential steps in the testing procedure is choosing a betting process $\boldsymbol{\lambda} = (\lambda_t)_{t\in[T]}$ in (7). Throughout this section, e is a $\mathcal{P}$-one-sided e-statistic for $\psi = (\rho, \phi) : \mathcal{M} \to \mathbb{R} \times \mathbb{R}^{d-1}$ and $\mathcal{P} \subseteq \mathcal{M}$. We omit $\mathcal{P}$ when $\mathcal{P}$ is the domain of ψ. Any predictable process $\boldsymbol{\lambda}$ with values in $[0, 1]$ yields a supermartingale in (7) under H_0, and thus the testing procedure is valid at all stopping times by Theorem 2. However, the statistical power of the tests, and the growth of the process $(M_t(\boldsymbol{\lambda}))_{t\in\{0,\dots,T\}}$ if the null hypothesis H_0 is false, heavily depends on a good choice of $\boldsymbol{\lambda}$. The betting process is chosen by the tester, e.g., a regulator or an internal model risk examiner.

5.1 GRO, GREE, GREL and GREM methods

Our methods are related to maximizing the expected log-capital originally proposed by Kelly (1956), adopted by Grünwald et al. (2024) in their GRO (growth-rate optimal) criterion, and studied by Shafer (2021) and Waudby-Smith and Ramdas (2024) for testing by betting. For an e-variable E and a probability measure Q representing an alternative hypothesis, the key quantity to consider is $\mathbb{E}^Q[\log E]$, which is called the e-power of E under Q by Vovk and Wang (2024b).

Let T be the time horizon of interest, which can be a finite integer or ∞. Let Q_t, $t \in [T]$, be the distribution of L_t given the information contained in $\mathcal{F}_{t-1}$. When choosing the betting process $(\lambda_t)_{t\in[T]}$, we fix an upper bound $\gamma \in (0, 1)$ and restrict $\lambda_t \in [0, \gamma]$ for all $t \in [T]$. The upper bound γ is not restrictive and only prevents some ill-behaving cases. We can set $\gamma = 1/2$ in our context (see Remark 7 below). Below we formally introduce a few methods to determine the betting process.

1. **GRO** (growth-rate optimal): We compute $\boldsymbol{\lambda}^{\mathrm{GRO}} = (\lambda_t^{\mathrm{GRO}})_{t\in[T]}$ by

$$\lambda_t^{\mathrm{GRO}} = \lambda_t^{\mathrm{GRO}}(r, z) = \operatorname*{arg\,max}_{\lambda\in[0,\gamma]} \mathbb{E}^{Q_t}[\log(1 - \lambda + \lambda e(L_t, r, z))], \tag{12}$$

and we plug in $(r, z) = (r_t, z_t)$. The optimal λ_t^{GRO} in (12) can be calculated through a convex program as the function $\lambda \mapsto \log(1 - \lambda + \lambda e(L_t, r, z))$ is concave. This requires the knowledge of the conditional distribution Q_t of L_t given $\mathcal{F}_{t-1}$. In practice, Q_t is unknown to the tester, and one may need to choose a model to approximate Q_t. If the probability measure Q_t is unknown but from a certain family, one can use the method of mixtures or mixture martingales (see e.g., de la Peña et al., 2004, 2009) of alternative scenarios to obtain an e-process close to that based on the unknown true model. The optimizer λ_t^{GRO} may not be unique in some special cases, e.g., $e(L_t, r, z)$ is the constant 1 or the expectation in (12) is ∞, but it is unique in most practical cases.

Without specifying a particular model for the alternative hypothesis, we propose the following

three methods to choose $\boldsymbol{\lambda}$, all trying to approximate the distribution Q_t under the alternative hypothesis using the sample until time point $t-1$.

2. **GREE** (growth-rate for empirical e-statistics): Let E follow the empirical distribution of the sample $e(L_s, r_s, z_s)_{s\leqslant t-1}$. We compute $\boldsymbol{\lambda}^{\mathrm{GREE}} = (\lambda_t^{\mathrm{GREE}})_{t\in[T]}$ by

$$\lambda_t^{\mathrm{GREE}} = \operatorname*{arg\,max}_{\lambda\in[0,\gamma]} \mathbb{E}[\log(1-\lambda+\lambda E)] = \operatorname*{arg\,max}_{\lambda\in[0,\gamma]} \frac{1}{t-1}\sum_{s=1}^{t-1} \log(1-\lambda+\lambda e(L_s, r_s, z_s)). \tag{13}$$

3. **GREL** (growth-rate for empirical losses): Let L follow the empirical distribution of the sample $(L_s)_{s\leqslant t-1}$. We compute $\boldsymbol{\lambda}^{\mathrm{GREL}} = (\lambda_t^{\mathrm{GREL}})_{t\in[T]}$ by

$$\begin{aligned} \lambda_t^{\mathrm{GREL}} = \lambda_t^{\mathrm{GREL}}(r,z) &= \operatorname*{arg\,max}_{\lambda\in[0,\gamma]} \mathbb{E}[\log(1-\lambda+\lambda e(L,r,z))] \\ &= \operatorname*{arg\,max}_{\lambda\in[0,\gamma]} \frac{1}{t-1}\sum_{s=1}^{t-1} \log(1-\lambda+\lambda e(L_s,r,z)), \end{aligned} \tag{14}$$

and we plug in $(r,z) = (r_t, z_t)$.

Both (13) and (14) are plug-in methods for the measure Q_t, one using sample e-statistics and the other using sample losses. The problems (13) and (14) can be solved directly via convex programming. The GREE and GREL methods are equivalent when the risk forecasts r_t and z_t are constant across $t\in[T]$. To explain the main difference between GREE and GREL, GREE decides λ_t without using the current forecast (r_t, z_t) as input but uses (r_s, z_s) for $s<t$, whereas GREL uses (r_t, z_t) to compute λ_t, but discards (r_s, z_s) for $s<t$. The forecast (r_t, z_t) supplied by the bank could be informative, noisy, or even misleading, and there is no uniformly superior choice. The GREE and GREL methods are asymptotically optimal in different practical situations, which we study in Section 5.2. For both methods, one may use a moving window to compute the empirical distributions instead of using all previous data points; see Section 8 for financial data analysis.

To take advantage of both methods, we propose an e-process by taking the average.

4. **GREM** (growth-rate for empirical mixture): We choose $\boldsymbol{\lambda}^{\mathrm{GREM}} = (\lambda_t^{\mathrm{GREM}})_{t\in[T]}$ such that

$$M_t(\boldsymbol{\lambda}^{\mathrm{GREM}}) = \frac{M_t(\boldsymbol{\lambda}^{\mathrm{GREE}})}{2} + \frac{M_t(\boldsymbol{\lambda}^{\mathrm{GREL}})}{2},$$

where M is defined in (7). By Lemma 1 of Vovk and Wang (2024a), there exists a betting process for the GREM method. More precisely,

$$\lambda_t^{\mathrm{GREM}} = \frac{M_{t-1}(\boldsymbol{\lambda}^{\mathrm{GREE}})\lambda_t^{\mathrm{GREE}} + M_{t-1}(\boldsymbol{\lambda}^{\mathrm{GREL}})\lambda_t^{\mathrm{GREL}}}{M_{t-1}(\boldsymbol{\lambda}^{\mathrm{GREE}}) + M_{t-1}(\boldsymbol{\lambda}^{\mathrm{GREL}})}.$$

The GREM method is asymptotically optimal for the practical cases when either the GREE or the GREL method is optimal (see Theorem 3). Other ways of averaging than using equal weights is possible, if the tester has side information on which method likely works better.

Our construction of the e-processes is based on the same idea that underlines the GRAPA method introduced in Waudby-Smith and Ramdas (2024, Section B.2). Two main differences are that we build e-variables from backtest e-statistics and we use risk forecasts as input variables for the betting strategy. Constructing a betting process depending on predictions has previously been explored by Henzi and Ziegel (2022).

In all simulation and empirical results, we use the monotone backtest e-statistics $e = e_p^Q$ for VaR_p and $e = e_p^{\mathrm{ES}}$ for $(\mathrm{ES}_p, \mathrm{VaR}_p)$. These choices are justified by results in Section 6, which we briefly explain below.

Remark 6. Usually, growth-rate optimality (GRO) refers to the optimality among the set of *all* e-variables for the null $r \geqslant \rho(L_t|\mathcal{F}_{t-1})$, $z = \phi(L_t|\mathcal{F}_{t-1})$, conditional on $\mathcal{F}_{t-1}$. *A priori*, GRO as defined at (12) is just a restricted GRO, that is, GRO amongst the subset of e-variables of the form $1-\lambda+\lambda e\,(L_t, r, z)$. However, as we will see from results in Section 6, for the important examples of the mean, the variance, and VaR, the seemingly restricted GRO is actually the usual unrestricted GRO, which implies essentially that the presented approach cannot be improved without imposing specific model assumptions. The same holds true for the case of ES under an additional assumption.

A simple way to get an approximation of (13) and (14) is to use a Taylor expansion $\log(1+y) \approx y - y^2/2$ at $y = 0$ and the first-order condition. This leads to

$$\lambda_t^{\mathrm{GREL}} \approx 0 \vee \frac{\sum_{s=1}^{t-1} e(L_s, r, z) - t + 1}{\sum_{s=1}^{t-1} (e(L_s, r, z) - 1)^2} \wedge \gamma \tag{15}$$

for the GREL method (14), where $a \vee b$ represents the maximum of a, b and $a \wedge b$ represents the minimum. We replace (r, z) in (15) by (r_s, z_s) for the GREE method (13). The special cases of (15) for VaR and ES are given in Section A.

Remark 7. We restrict the betting process by the upper bound $\gamma < 1$ to avoid the e-process collapsing to 0. For illustration, suppose that for each t, $X_t = e(L_t, r, z)$ given $\mathcal{F}_{t-1}$ takes value 0 with small probability and value 2 with large probability, so its expected value is larger than 1 and the null hypothesis is not true. As long as we do not observe 0 up to time t, the empirical distribution is concentrated at 2, leading to an optimal strategy $\lambda_t^{\mathrm{GREE}} = 1$. This betting process yields an e-process that becomes 0 as soon as we observe a 0 from X_t and therefore should be avoided. In all our numerical and data experiments, the optimal λ_t from each method is typically

quite small (< 0.1) for tail risk measures like VaR and ES. Hence, it is harmless to set $\gamma = 1/2$ by default in our context. A different approach to obtain the betting process is to put a distribution on λ and estimate by a "posterior"; see Agrawal et al. (2021) and Jang et al. (2023).

Remark 8. In our setting, the financial institution can arbitrarily report risk forecasts. This allows the possibility for the bank to "game" the regulator by intentionally over-reporting the risk forecasts for a period of time before it starts under-reporting in another period of time. Such a gaming problem also happens for most traditional backtesting methods. Our method provides an efficient solution to such an issue by using a rolling window (e.g., 250 or 500 days) to calculate the betting process instead of using all past data. During the over-prediction period, the e-values $e(L_t, r_t, z_t)$ tend to be small on average with mean smaller than 1. By (15) and Proposition 2 below, the resulting betting process with the GREE or GREL method will be equal or close to 0, yielding an e-process being almost a constant 1. After the bank starts to under-predict, the past data will exit the rolling window gradually. The GREE or GREL method will pick up the large e-values and start to choose a positive betting process. Moreover, how the betting process will be chosen is not known to the bank.[4] This also increases the cost of gaming by the bank since the regulator can actively choose a positive betting process to detect evidence against the under-reports if she has sufficient reason to realize the bank under-predicts for a certain period. More details and simulation studies can be found in Section E.

5.2 Optimality of betting processes

Next, we discuss the optimality of the betting process. We first define an intuitive notion of asymptotic optimality. For asymptotic results discussed in this section, we will assume an infinite time horizon; that is, we consider $t \in \mathbb{N}$. All statements on probability and convergence are with respect to the true probability generating the data.

Definition 3. For $(L_{t-1}, r_t, z_t)_{t\in\mathbb{N}}$ adapted to $(\mathcal{F}_{t-1})_{t\in\mathbb{N}}$ and a given function $e : \mathbb{R}^{d+1} \to [0, \infty]$,

(i) two betting processes $\boldsymbol{\lambda} = (\lambda_t)_{t\in\mathbb{N}}$ and $\boldsymbol{\lambda}' = (\lambda'_t)_{t\in\mathbb{N}}$ are *asymptotically equivalent*, denoted by $\boldsymbol{\lambda} \simeq \boldsymbol{\lambda}'$, if

$$\frac{1}{T}(\log M_T(\boldsymbol{\lambda}) - \log M_T(\boldsymbol{\lambda}')) \xrightarrow{L^1} 0 \quad \text{as } T \to \infty,$$

where M is defined in (7);[5]

[4] Under the null hypothesis, the e-process $M(\boldsymbol{\lambda})$ is a supermartingale, no matter how $\boldsymbol{\lambda}$ is chosen. Therefore, the regulator can announce λ_t after seeing the forecast (r_t, z_t) and before seeing the next loss data point L_t.

[5] The L^1-convergence $Z_T \xrightarrow{L^1} Z$ means $\mathbb{E}[|Z_T - Z|] \to 0$ as $T \to \infty$.

(ii) a betting process $\boldsymbol{\lambda}$ is *asymptotically optimal* if $\boldsymbol{\lambda} \simeq (\lambda_t^{\mathrm{GRO}}(r_t, z_t))_{t\in\mathbb{N}}$.

Intuitively, the asymptotic equivalence between two betting processes means that the long-term growth rates of the two resulting e-processes are the same. The asymptotic optimality of a betting process is defined by the asymptotic equivalence using the GRO method as a benchmark because GRO is the best-performing method if we know the full distributional information of the losses.

The following proposition characterizes the situations where the betting processes in the GRO method do not reach 0 and 1. In our formulation, $\boldsymbol{\lambda}$ is not allowed to reach 1 due to the upper bound $\gamma < 1$, but we nevertheless give a theoretical condition that the unconstrained optimizer is less than 1.

Proposition 2. *For $(r, z) \in \mathbb{R} \times \mathbb{R}^{d-1}$ and $t \in \mathbb{N}$ and any optimizer λ_t^{GRO} of (12), the following statements hold.*

(i) $\lambda_t^{\mathrm{GRO}}(r, z) > 0$ if and only if $\mathbb{E}^{Q_t}[e(L_t, r, z)|\mathcal{F}_{t-1}] > 1$.

(ii) With $\gamma = 1$ in (12), $\lambda_t^{\mathrm{GRO}}(r, z) < 1$ if and only if $\mathbb{E}^{Q_t}[1/e(L_t, r, z)|\mathcal{F}_{t-1}] > 1$.

The observation in Proposition 2 also holds true for Q_t replaced by an empirical measure as in (13) and (14). Hence, GREE or GREL will choose $\lambda_t > 0$ only when the empirical mean of the e-statistic (specified differently in the two methods) is larger than 1.

Below we present an assumption for the asymptotic analysis. The condition is very weak because the interesting case in backtesting is when $\mathbb{E}^{Q_t}[\log(e(L_t, r, z))]$ is small. Denote by $\psi^*(\mathcal{P}) \subseteq \psi(\mathcal{P})$ as the set of all values $(r, z) \in \psi(\mathcal{P})$ such that $e(x, r, z) < \infty$ for all $x \in \mathbb{R}$.

Assumption 1. For all $(r, z) \in \psi^*(\mathcal{P})$, $\sup_{t\in\mathbb{N}} \mathbb{E}^{Q_t}[|\log(e(L_t, r, z))|] < \infty$.

The following theorem addresses the asymptotic optimality of the GREE, GREL and GREM methods in different situations, as well as their consistency. We say that a method with e-process M has asymptotic full power if $Q(\sup_{t\in\{0,\dots,T\}} M_t \geqslant 1/\alpha) \to 1$ as $T \to \infty$ for all $\alpha \in (0, 1)$, where Q is the data generating probability.

Theorem 3. *For $(L_{t-1}, r_t, z_t)_{t\in\mathbb{N}}$ adapted to $(\mathcal{F}_{t-1})_{t\in\mathbb{N}}$ such that (r_t, z_t) takes values in $\psi^*(\mathcal{P})$ and $e : \mathbb{R}^{d+1} \to [0, \infty]$, under Assumption 1 and using $\gamma = 1$ in the betting strategies, the following statements hold.*

(i) $(\lambda_t^{\mathrm{GREE}})_{t\in\mathbb{N}}$ is asymptotically optimal if $(e(L_t, r_t, z_t))_{t\in\mathbb{N}}$ is iid and $(r_t, z_t)_{t\in\mathbb{N}}$ is deterministic.

(ii) $(\lambda_t^{\mathrm{GREL}}(r_t, z_t))_{t\in\mathbb{N}}$ is asymptotically optimal if $(L_t)_{t\in\mathbb{N}}$ is iid and either:

(a) $(r_t, z_t)_{t\in\mathbb{N}}$ *takes finitely many possible values in* $\mathbb{R}^d$.

(b) (r_t, z_t), $t \in \mathbb{N}$, *are in a common compact set,* $e(x, r, z)$ *is continuous in* (r, z), *and* $(r_t, z_t) \xrightarrow{\mathrm{p}} (r_0, z_0)$ *as* $t \to \infty$ *for some* $(r_0, z_0) \in \mathbb{R}^d$.

(iii) $(\lambda_t^{\mathrm{GREM}})_{t\in\mathbb{N}}$ *is asymptotically optimal if either* $(\lambda_t^{\mathrm{GREE}})_{t\in\mathbb{N}}$ *or* $(\lambda_t^{\mathrm{GREL}}(r_t, z_t))_{t\in\mathbb{N}}$ *is asymptotically optimal.*

Moreover, if $r_t < \rho(L_t|\mathcal{F}_{t-1}) - \varepsilon$ *for some* $\varepsilon > 0$ *and each* $t \in \mathbb{N}$, *then GREE in (i), GREL in (ii), and GREM in (i) and (ii) have asymptotic power one, using any backtest e-statistic* e *for* ψ.

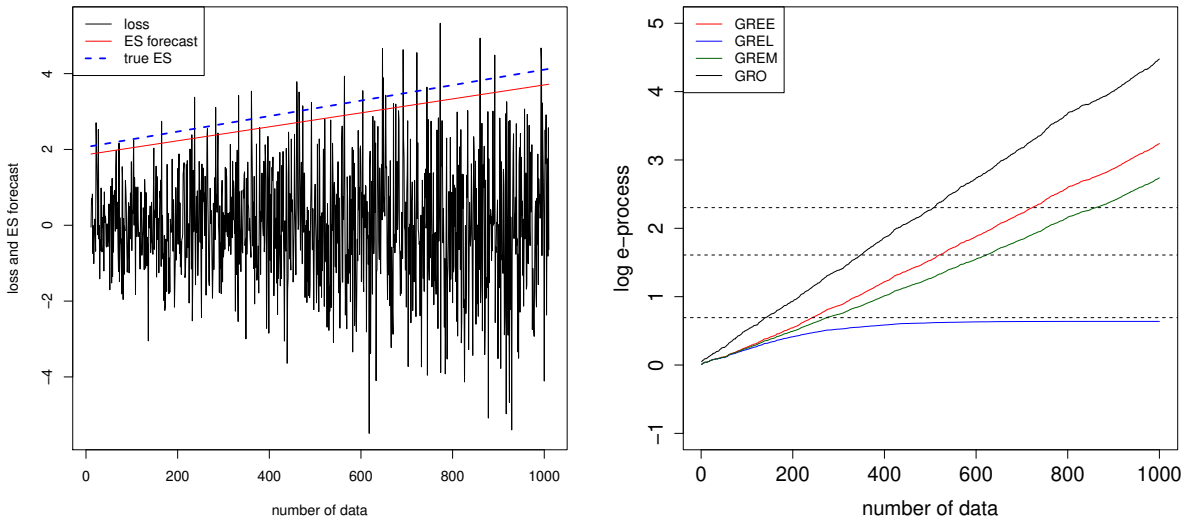
Any choice of the betting strategy $\boldsymbol{\lambda}$ yields a valid and non-asymptotic test, and here $T \to \infty$ is taken only to discuss optimality of the power. The asymptotic optimality results in Theorem 3 are based on strong, and perhaps unrealistic, assumptions; they are imposed for technical reasons. Nevertheless, we obtain some useful insight on the comparison between GREE and GREL. Intuitively, the GREE method should outperform the GREL method when the backtest e-statistics $e(L_t, r_t, z_t)$, $t \in \mathbb{N}$, are iid and (r_t, z_t) is not informative about how to choose λ_t (i.e., they are noise), while the GREL method should outperform the GREE method when the losses L_t, $t \in \mathbb{N}$, are iid and (r_t, z_t) is informative on how to choose λ_t;[6] recall that GREL uses (r_t, z_t) whereas GREE does not. We illustrate the insights for the comparison between the GREE and GREL methods through Example 7 below. In most practical cases, we do not know clear patterns of the losses and forecasts as they arrive sequentially over time. In this sense, the GREM method is recommended.

Example 7. Let the size of training data be $l = 10$, the sample size for testing be $n = 1,000$, and $Z_1, \dots, Z_{n+l}$ be iid samples simulated from the standard normal distribution. We report the average performance of backtesting methods over $1,000$ simulations.

(a) The iid condition of the whole backtest e-statistics implies that the GREE method works better than the GREL method when losses and risk forecasts exhibit co-movements over time. Such situations are common in the financial market; for instance, risk forecasts will increase over time when a company is extending its business. Assume that $L_t = (1+t/(n+l))Z_t$ for $t \in [n+l]$. This model represents the case where the financial institution's investment generates iid cash flow but the institution increases the investment amount over time. Following the increasing trend of the investment, the risk forecaster announces the under-estimated forecasts of $\mathrm{VaR}_{0.95}(L_t|\mathcal{F}_{t-1})$ and $\mathrm{ES}_{0.95}(L_t|\mathcal{F}_{t-1})$ as $z_t = 1.48(1 + t/(n + l))$ and $r_t = 1.86(1 + t/(n + l))$, respectively, for

[6] In this case, we can statistically infer the value of $\rho(L_t)$ from past data, and thus the values of (r_t, z_t) are informative about whether it is likely an over-prediction (which means we should choose a large λ_t) or an under-prediction (which means we should choose a small or zero λ_t).

Figure 1: Realized losses and ES forecasts with a linear extending business (left panel); average log-transformed e-processes obtained by different methods over $1,000$ simulations (right panel)



$t \in [n+l]$. Figure 1 plots the realized losses L_t, ES forecasts r_t, and the e-processes obtained by the GRO, GREE, GREL and GREM methods for $t = l+1, \dots, n+l$. We observe from Figure 1 that the GREE e-process dominates the GREL e-process. This is consistent with the result of Theorem 3 by noting the co-movements of the losses and the VaR and ES forecasts which makes the backtest e-statistics $(e_p(L_t, r_t, z_t))_{t\in[n+l]}$ iid.

(b) We consider a non-linear business cycle. Take the random losses to be $L_t = Z_t(1+\sin(\theta t))$ for $t \in [n+l]$, where $\theta = 0.01$. The risk forecasts of $\mathrm{VaR}_{0.95}(L_t|\mathcal{F}_{t-1})$ and $\mathrm{ES}_{0.95}(L_t|\mathcal{F}_{t-1})$ also have a similar trend but are under-estimated. Namely, we have $z_t = 1.48(1+\sin(\theta t))$ for VaR and $r_t = 1.86(1+\sin(\theta t))$ for ES. The losses and forecasts, and the average log e-processes for different methods are plotted in Figure 2. Similarly to (a), we also observe better performance of the GREE method than GREL because of the overall iid pattern of the whole e-statistics.

(c) It is expected from Theorem 3 that the GREL method will dominate the GREE method when the losses exhibit an iid pattern and there is no clear evidence of co-movements between losses and risk forecasts. Let the random losses be $Z_1, \dots, Z_{n+l}$; thus, they are iid. Suppose that the risk forecaster announces the forecasts of $\mathrm{VaR}_{0.95}(Z_t|\mathcal{F}_{t-1})$ and $\mathrm{ES}_{0.95}(Z_t|\mathcal{F}_{t-1})$ to be $z_t = 1.64 + \varepsilon_t$ and $r_t = 2.06 + \varepsilon_t$, respectively, for $t \in [n+l]$, where $\varepsilon_1, \dots, \varepsilon_{n+l}$ are iid samples uniformly distributed on the support $\{\pm i/10 : i = 0, \dots, 5\}$. In this case, the forecaster is able to obtain risk forecasts close to the true values but is subject to a forecasting error $(\varepsilon_t)_{t\in[n+l]}$.

Figure 2: Realized losses and ES forecasts with a non-linear business cycle (left panel); average log-transformed e-processes obtained by different methods over $1,000$ simulations (right panel)

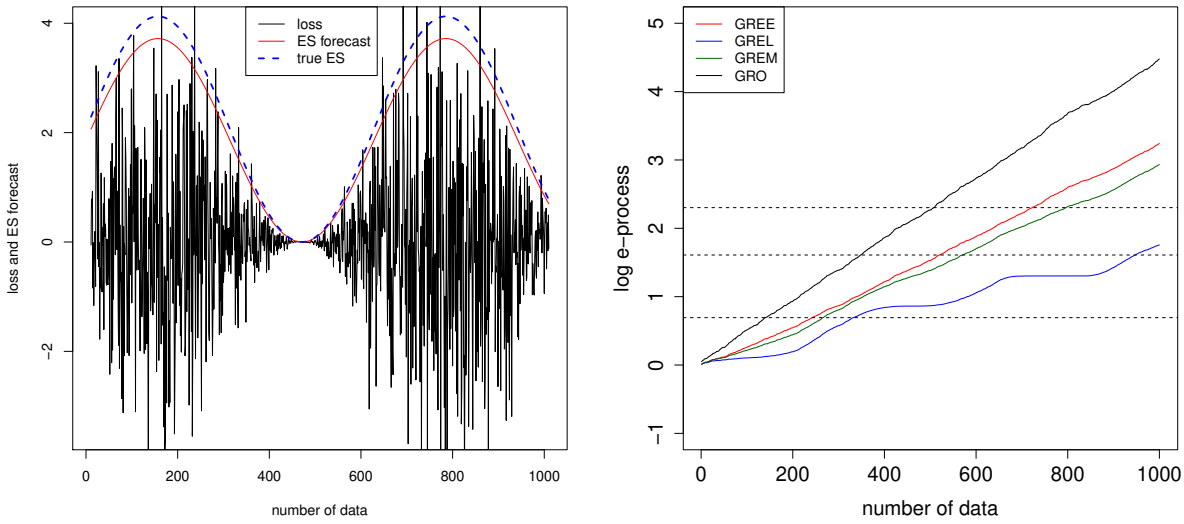


Figure 3 plots the realized losses Z_t, ES forecasts r_t, and the corresponding e-processes obtained by the GRO, GREE, GREL and GREM methods for $t = l+1, \dots, n+l$. We observe from Figure 3 that the GREL method outperforms the GREE method. This example shows that the GREL method is able to detect evidence against risk forecasts due to downward fluctuations of the forecasts, while GREE does not perform well in this case because it only uses historical forecasts whose average is close to the true value.

6 Characterizing backtest e-statistics

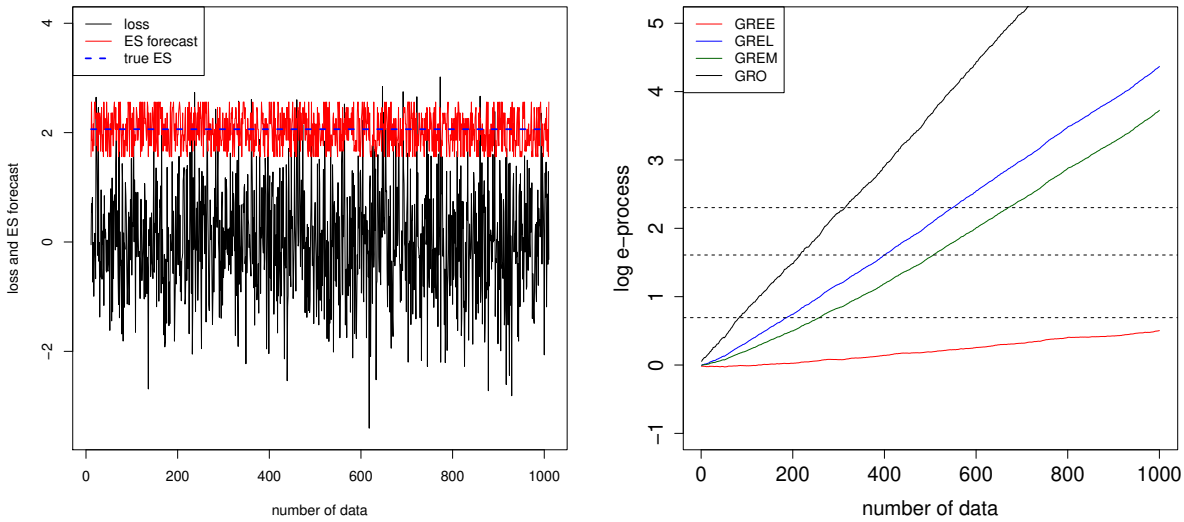
In this section, we present several results on the characterization of backtest e-statistics. These results justify the unique roles of the e-statistics we introduced in Section 3, and their growth-rate optimality among all e-variables as discussed in Section 5. A main practical message is that e_p^Q and e_p^{ES} are essentially the only useful choices for VaR and (ES,VaR), respectively, in building up e-processes in Section 4.[7]

6.1 Necessary conditions for the existence of backtest e-statistics

Not all functionals ρ on $\mathcal{M}$ admit backtest e-statistics that are solely based on the information of ρ. Below we give a necessary condition for a backtest e-statistic to exist. A functional $\rho : \mathcal{M} \to \mathbb{R}$

[7] These results are not needed for the statistical validity of our methodology.

Figure 3: Realized losses and ES forecasts with iid losses (left panel); average log-transformed e-processes obtained by different methods over $1,000$ simulations (right panel)



is *monotone* if $\rho(F) \leqslant \rho(G)$ for all $F \leqslant_1 G$, where $\leqslant_1$ is the usual stochastic order; namely, $F \leqslant_1 G$ if and only if $F \geqslant G$ pointwise on $\mathbb{R}$. We also say that ρ is *uncapped* if for each $F \in \mathcal{M}$ and $r > \rho(F)$, there exists $\bar{F} \in \mathcal{M}$ such that $\bar{F} \geqslant_1 F$ and $\rho(\bar{F}) = r$. All monetary risk measures (Föllmer and Schied, 2016) are monotone and uncapped. A functional $\rho : \mathcal{M} \to \mathbb{R}$ is *quasi-convex* if $\rho(\lambda F + (1-\lambda)G) \leqslant \max\{\rho(F), \rho(G)\}$ for all $\lambda \in [0,1]$ and $F, G \in \mathcal{M}$. Similarly, ρ is *quasi-concave* if $-\rho$ is quasi-convex, and ρ is *quasi-linear* if it is both quasi-convex and quasi-concave.

Proposition 3. *Suppose that $\rho : \mathcal{M} \to \mathbb{R}$ is monotone and uncapped. If there exists a backtest e-statistic, then ρ is quasi-convex.*

When $\mathcal{M}$ is convex, quasi-convexity of ρ is equivalent to the condition that the set $\{F \in \mathcal{M} : \rho(F) \leqslant r\}$ is convex for each $r \in \mathbb{R}$. The requirement in Proposition 3 rules out a large class of coherent risk measures including ES.[8] As is shown in the following proposition, if the backtest e-statistic for ρ is monotone, then ρ is necessarily quasi-linear, which is stronger than the quasi-convexity in Proposition 3, and this result does not require that ρ is monotone and uncapped.

Proposition 4. *If there exists a monotone backtest e-statistic $e : \mathbb{R}^2 \to [0, \infty]$ for $\rho : \mathcal{M} \to \mathbb{R}$, then ρ is quasi-linear.*

[8]In particular, all comonotonic-additive coherent risk measures except for the mean are monotone and uncapped but not quasi-convex (see e.g., Wang et al. 2020, Theorem 3). A risk measure is coherent (Artzner et al., 1999) if it is subadditive, cash additive, monotone, and positively homogeneous as a mapping from a set of random variables to real numbers.

We say a functional $\rho : \mathcal{M} \to \mathbb{R}$ has *convex level sets (CxLS)* if the set $\{F \in \mathcal{M} : \rho(F) = r\}$ is convex for each $r \in \mathbb{R}$. Quasi-linearity of ρ is stronger than the condition that ρ has CxLS, and they are equivalent when $\mathcal{M}$ is convex and ρ is monotone. Functionals with CxLS have been studied extensively in the recent literature due to their connection to elicitability and backtesting (Gneiting, 2011; Ziegel, 2016). For a recent summary of related results, see Wang and Wei (2020).

6.2 Characterizing backtest e-statistics for common risk measures

First, we characterize all backtest e-statistics for the mean and for $(\mathrm{var}, \mathbb{E})$; see also Examples 1–2.

Proposition 5 (Backtest e-statistics for the mean)**.** *Let $a \in \mathbb{R}$ and $\mathcal{P} \subseteq \mathcal{M}_1$ be the set of distributions with support in $[a, \infty)$. Any $\mathcal{P}$-one-sided e-statistic for the mean is bounded above by*

$$e'(x, r) = 1 + h(r)\frac{x - r}{r - a}, \quad x \geqslant a,\ r \geqslant a, \tag{16}$$

for some function h on $[a, \infty)$ with $0 \leqslant h \leqslant 1$. The upper bound e' is a one-sided e-statistic for the mean. Moreover, e' is a backtest e-statistic for the mean, if and only if $h > 0$. The functions h and $r \mapsto (r - a)/h(r)$ are increasing if and only if e' is a monotone backtest e-statistic for the mean.

Proposition 5 shows in particular that any e-variable for the null hypothesis $\mathbb{E}[L] \leqslant r$, $L \geqslant a$ that is a function of L must be dominated by an e-variable $e'(L, r)$ as given in (16).

Proposition 6 (Backtest e-statistics for the variance)**.** *Any one-sided e-statistic on $\mathcal{M}_2$ for $\psi = (\mathrm{var}, \mathbb{E})$ is bounded above by*

$$e'(x, r, z) = 1 + h(r, z)\frac{(z - x)^2 - r}{r}, \quad x, z \in \mathbb{R},\ r \geqslant 0, \tag{17}$$

for some function h on $[0, \infty) \times \mathbb{R}$ with $0 \leqslant h \leqslant 1$. The upper bound e' is a one-sided e-statistic for ψ. Moreover, e' is a backtest e-statistic for ψ, if and only if $h > 0$. The functions $r \mapsto h(r, z)$ and $r \mapsto r/h(r, z)$ are increasing for all $z \in \mathbb{R}$ if and only if e' is a monotone backtest e-statistic for ψ.

Proposition 6 shows in particular that any e-variable for the null hypothesis $\mathrm{var}[L] \leqslant r$ and $\mathbb{E}[L] = z$ that is a function of L must be dominated by an e-variable $e'(L, r, z)$ as given in (17).

The simplest backtest e-statistic for VaR is given in (1) in Example 3. The following theorem shows that all backtest e-statistics for VaR are essentially dominated by this simple choice.

Theorem 4 (Backtest e-statistics for VaR)**.** *Let $p \in (0,1)$. Any one-sided e-statistic for VaR_p is bounded above by*

$$e'(x,r) = 1 + h(r)\frac{p - \mathbb{1}_{\{x \leqslant r\}}}{1-p}, \tag{18}$$

for some function h on $\mathbb{R}$ with $0 \leqslant h \leqslant 1$. The upper bound e' is a one-sided e-statisic for VaR_p. Moreover, e' is a backtest e-statistic for VaR_p, if and only if $h > 0$. The function h is constant if and only if e' is a monotone backtest e-statistic for VaR_p.

Theorem 4 shows in particular that any e-variable for the null hypothesis $\mathrm{VaR}_p(L) \leqslant r$ that is a function of L must be dominated by an e-variable $e'(L,r)$ as given in (18).

Next, we consider backtest e-statistics for $(\mathrm{ES}, \mathrm{VaR})$. It is straightforward that ES_p is monotone, uncapped, and $\{F \in \mathcal{M} : \mathrm{ES}_p(F) \leqslant r\}$ is not convex.[9] Hence, Proposition 3 implies that there is no backtest e-statistic for ES_p using solely the information of ES_p. A similar point was made in Acerbi and Székely (2017) that ES_p is not backtestable in some specific sense. By Theorem 1, there exists a backtest e-statistic e_p^{ES} for $(\mathrm{ES}_p, \mathrm{VaR}_p)$. In particular, $e_p^{\mathrm{ES}}(L,r,z)$ is an e-variable for the hypothesis $H_0 : \mathrm{ES}_p(L) \leqslant r$ and $\mathrm{VaR}_p(L) = z$. The following theorem characterizes one-sided and backtest e-statistics for $(\mathrm{ES}_p, \mathrm{VaR}_p)$ under the assumption that they are increasing in the loss L.

Theorem 5 (Backtest e-statistics for ES)**.** *Let $p \in (0,1)$. Any one-sided e-statistic on $\mathcal{M}_1$ for $\psi = (\mathrm{ES}_p, \mathrm{VaR}_p)$ that is increasing in its first argument (i.e., realized loss) is bounded above by*

$$e'(x,r,z) = 1 + h(r,z)\left(\frac{(x-z)_+}{(1-p)(r-z)} - 1\right) + k(r,z)\frac{p - \mathbb{1}_{\{x \leqslant z\}}}{1-p}, \quad x \in \mathbb{R},\ z < r,$$

for some functions h, k with $0 \leqslant h, k \leqslant 1$ and $h+k \leqslant 1$. The upper bound e' is a one-sided e-statistic for ψ. Moreover, e' is a backtest e-statistic for ψ, if and only if $h > 0$ and $k = 0$. The functions $r \mapsto h(r,z)$ and $r \mapsto (r-z)/h(r,z)$ are increasing for all $z < r$ if and only if e' is a monotone backtest e-statistic for ψ.

It is clear that in Theoroem 5, if $e'(x,r,z)$ is continuous in x, then $k = 0$ and $e' = 1 - h + he_p^{\mathrm{ES}}$.

Theorems 4 and 5 illustrate the essential roles of e_p^Q and e_p^{ES} among all possible choices of e-statistics for VaR and the pair (ES,VaR). All choices of one-sided e-statistics for VaR_p are dominated by an e-statistic of the form $1 - \lambda + \lambda e_p^Q(x,r)$, and all those for $(\mathrm{ES}_p, \mathrm{VaR}_p)$ that are relevant for risk management are upper bounded by $1 - \lambda + \lambda e_p^{\mathrm{ES}}(x,r,z)$, where λ is a function taking values in $[0,1]$. Therefore, in view of (7), the e-statistics e can be without loss of generality chosen as e_p^Q for VaR_p and e_p^{ES} for $(\mathrm{ES}_p, \mathrm{VaR}_p)$, and λ can be chosen separately depending on the risk forecasts.

[9] It might be interesting to note that ES_p is concave on $\mathcal{M}$, implying that the set $\{F \in \mathcal{M} : \mathrm{ES}_p(F) \geqslant r\}$ is convex for each r; see Theorem 3 of Wang et al. (2020).

Characterization results of backtest e-statistics for risk measures can be obtained by using the link between backtest e-statistics and identification functions that is presented in Section B, or by direct arguments. The first approach has the advantage of being more general but in the special cases we considered here, we would only obtain the results under slightly stronger conditions.

Results in this section have implications on the GRO e-variables discussed in Remark 6. A GRO e-variable for the considered risk measures must have the form $1-\lambda+\lambda e\left(L_t, r, z\right)$ for suitable e in Section 3; in case of ES, the optimality is restricted to e-variables increasing in the loss L_t.

7 Simulation studies

In this section, we provide simulation studies on backtesting VaR and ES. This illustrates the details of our backtesting methodology numerically. Furthermore, we examine how different factors affect the quality of the backtesting procedure, especially the impact of the choice of the betting process in (7). We evaluate the backtesting performance when the risk measures are under-reported, over-reported, or reported exactly by the risk forecaster.

For all e-tests, we report evidence against the forecasts when the e-process exceeds thresholds 2, 5, or 10. We call such evidence a detection. From the practical viewpoint, the three thresholds we choose form four zones for levels of alerts to financial institutions. This is in a similar sense to the standard three-zone approach for backtesting VaR in the financial industry.

Remark 9. In classical statistical terminology, what we call a *detection* is a *rejection* of the null hypothesis based on our e-test with thresholds 2, 5, and 10, respectively. Our choice of using "detection" is to emphasize that having detected evidence of moderate size such as 2 with the e-test is a useful early warning that risk predictions might not be prudent enough. Recall that Jeffrey's threshold of e-values for "substantial" evidence is 3.2 and for "decisive" evidence is 10; see Shafer (2021) and Vovk and Wang (2021) for more discussions on observing moderately large e-values.

The simulation and data analysis in Sections 7 and 8, together with those in Section D, illustrate our main methodology.

7.1 Backtests via stationary time series

We apply our e-backtesting procedure to a setting with time series to test the hypotheses (8) and (9). We simulate data from an AR(1)–GARCH(1, 1) process:

$$L_t = \mu_t + \sigma_t Z_t, \quad \mu_t = -0.05 + 0.3 L_{t-1}, \quad \sigma_t^2 = 0.01 + 0.1\sigma_{t-1}^2 Z_{t-1}^2 + 0.85\sigma_{t-1}^2, \quad t \in \mathbb{N}$$

Table 2: Percentage of detections (in %) for $\text{VaR}_{0.99}$ forecasts based on $1,000$ simulations of time series, 500 trading days, and the GREM method: we use an AR(1)–GARCH(1, 1) model with four different innovations for the prediction; -10% and $+10\%$ mean that the prediction has been manually under-reported or over-reported by 10%; a detection means that the e-process reaches the corresponding threshold

	normal			t			skewed-t			true		
threshold	2	5	10	2	5	10	2	5	10	2	5	10
-10%	99.6	98.0	94.3	97.8	88.8	76.1	46.0	14.4	5.9	38.3	10.7	4.5
exact	97.0	87.9	75.0	86.8	60.9	40.2	17.8	2.5	0.5	15.0	1.7	0.2
$+10\%$	86.1	62.4	41.4	62.5	26.3	11.9	6.3	0.4	0	3.9	0.3	0

where $\{Z_t\}_{t\in\mathbb{N}}$ is a sequence of iid innovations following a skewed-t distribution with shape parameter $\nu = 5$ and skewness parameter $\gamma = 1.5$. These parameter values are from Nolde and Ziegel (2017) and we use them for a direct comparison. In total, $1,000$ independent simulations are produced, each of which includes a sample of size 500 used for backtesting. A rolling window of size 500 is applied for risk estimation at each time spot t.

For forecasting, we assume that the data follow an AR(1)–GARCH(1, 1) process $\{L_t\}_{t\in\mathbb{N}}$ with $L_t = \mu_t + \sigma_t Z_t$, where $\{Z_t\}_{t\in\mathbb{N}}$ is assumed to be a sequence of iid innovations with mean 0 and variance 1, following a normal, t-, or skewed-t distribution. Thus, the forecaster has a correct time-series structure with possibly incorrect innovation. Here, $\{\mu_t\}_{t\in\mathbb{N}}$ and $\{\sigma_t\}_{t\in\mathbb{N}}$ are adapted to $(\mathcal{F}_{t-1})_{t\in\mathbb{N}}$. In addition, as a benchmark (column "true" in Tables 2 and 3), we produce a series of forecasts with skewed-t innovations and the true $\{\mu_t\}_{t\in\mathbb{N}}$ and $\{\sigma_t\}_{t\in\mathbb{N}}$ as the data generating process. The details of the forecasting procedure are described in Section D.1. The risk forecaster deliberately under-reports, over-reports, or reports the exact point forecasts of VaR_p or $(\text{ES}_p, \text{VaR}_p)$ she obtains. For backtesting, the e-processes in (7) are calculated with the betting process $(\lambda_t)_{t\in[T]}$ chosen by the GREM method using Taylor approximation via (15). The results for the GREE and GREL methods and their comparison are demonstrated in Section D.2. We detect evidence against the forecasts when the e-processes exceed thresholds 2, 5, or 10. We first present results for backtesting $\text{VaR}_{0.99}$. The percentage of detections, the average number of days taken to detect evidence against the forecasts (conditional on detection occurring), and the average final log-transformed e-values are shown in Tables 2 and 3.

As expected, the results show that evidence against normal and t-innovations is more likely to be detected than against skewed-t innovations which is the true model. The percentage of detections for exact true forecasts, or the type-I error, is 0.2% for threshold 10. Under-reporting VaR leads

Table 3: The average number of days taken to detect evidence against $\text{VaR}_{0.99}$ forecasts conditional on detection, based on $1,000$ simulations of time series, 500 trading days, and the GREM method: the models are the same as in Table 2; numbers in brackets are average final log-transformed e-values

	normal				t				skewed-t				true			
threshold	2	5	10		2	5	10		2	5	10		2	5	10	
−10%	116	185	228	(5.489)	156	238	285	(3.390)	230	284	332	(0.3707)	220	312	347	(0.2200)
exact	158	239	287	(3.311)	200	284	322	(1.759)	219	244	217	(−0.07467)	234	310	441	(−0.1466)
+10%	196	277	316	(1.858)	224	305	351	(0.7341)	183	227	–	(−0.2135)	209	315	–	(−0.2428)

Table 4: Percentage of detections (%) for $\text{ES}_{0.975}$ forecasts based on simulations of time series, 500 trading days, and the GREM method: we use an AR(1)–GARCH(1, 1) model with four different innovations for the prediction; −10% ES (both) and +10% ES (both) mean that the prediction of ES (both VaR and ES) has been manually under-reported or over-reported by 10%; a detection means that the e-process reaches the corresponding threshold

	normal			t			skewed-t			true		
threshold	2	5	10	2	5	10	2	5	10	2	5	10
−10% ES	99.8	99.5	98.5	98.4	88.8	77.1	47.6	16.1	6.2	35.5	9.2	3.6
−10% both	99.8	99.5	98.1	98.5	91.4	82.0	48.0	15.7	6.5	36.1	10.1	4.2
exact	99.3	95.7	88.3	88.1	63.9	43.1	18.8	4.0	0.8	11.9	1.7	0.5
+10% both	95.2	80.4	61.9	64.9	27.6	9.9	7.1	1.0	0	4.2	0.1	0.1
+10% ES	94.8	79.8	62.1	70.0	34.9	15.6	7.9	1.1	0.1	4.6	0.2	0.1

to earlier detections than reporting the exact VaR forecasts and the converse holds true for over-reporting. For the exact and over-reporting true forecasts, rejections are essentially type-I errors. If such an error is made, it tends to be early, and this is confirmed by the simulation results.

Results on backtests of ES are reported in Tables 4 and 5. The results in Table 4 confirm our intuition that under-reporting or using a wrong innovation can be detected with a large probability, whereas forecasts from the true model and their more conservative versions appear the opposite. Moreover, under-reporting (resp. over-reporting) both of ES and VaR and under-reporting (resp. over-reporting) only ES have similar performance in terms of probability of detection and time of detection. The average time to detection (Table 5) is useful for risk management since early warnings (threshold 2) are often issued after about a fourth of the sampling time, and decisive warnings (threshold 10) after about half of the considered trading days.

A comparison with the traditional backtest method of Nolde and Ziegel (2017) is presented in Section D.3.

Table 5: The average number of days taken to detect evidence against $\mathrm{ES}_{0.975}$ forecasts conditional on detection; simulations of time series, 500 trading days, and the GREM method: the models are the same as in Table 4; "–" represents no detection; numbers in brackets are average final log-transformed e-values

	normal				t				skewed-t				true			
threshold	2	5	10		2	5	10		2	5	10		2	5	10	
−10% ES	89	137	176	(6.671)	151	223	277	(3.428)	224	271	264	(0.5072)	219	303	346	(0.2544)
−10% both	94	146	189	(6.347)	141	215	271	(3.720)	218	256	251	(0.4679)	218	282	317	(0.2292)
exact	129	201	247	(4.311)	185	267	311	(1.953)	198	195	251	(−0.04676)	194	252	300	(−0.1549)
+10% both	171	250	292	(2.737)	217	283	289	(0.8275)	141	206	–	(−0.2072)	148	25	57	(−0.2522)
+10% ES	168	247	297	(2.702)	207	286	298	(0.9865)	147	158	165	(−0.2323)	145	186	57	(−0.2819)

7.2 Monitoring structural change of time series

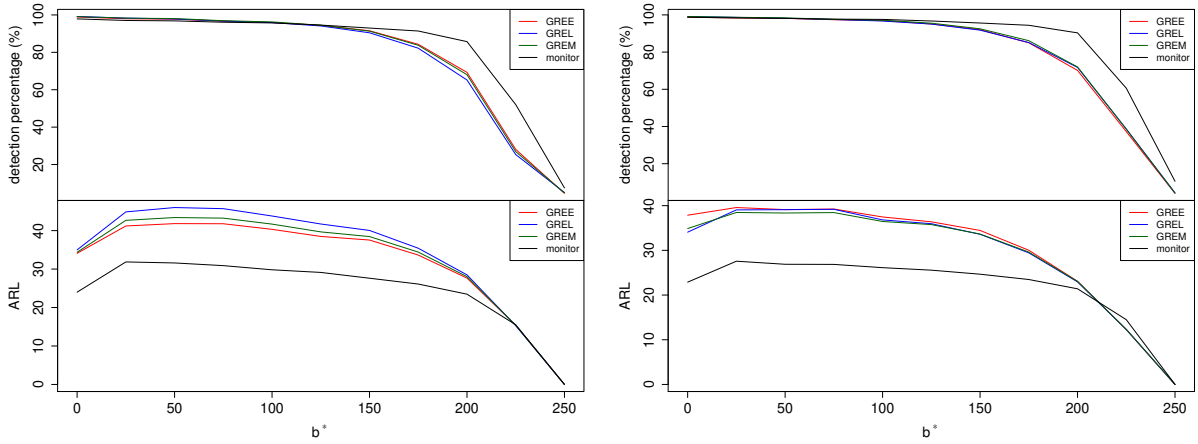
We examine the power of our e-backtesting method to monitor the structural change of simulated time series data. We refer to Chu et al. (1996) and Berkes et al. (2004) for earlier work on monitoring the structural change of datasets. For a comparison with the results in Hoga and Demetrescu (2023), we use the same setup as described in their Section 6, and call their method the sequential monitoring method. We simulate the losses $\{L_t\}_{t\in\mathbb{N}}$ following the GARCH$(1,1)$ process:

$$L_t = -\sigma_t Z_t, \quad \sigma_t^2 = 0.00001 + 0.04L_{t-1}^2 + \beta_t \sigma_{t-1}^2,$$

where $\{Z_t\}_{t\in\mathbb{N}}$ is a sequence of iid innovations following a skewed-t distribution with shape parameter $\nu = 5$ and skewness parameter $\gamma = 0.95$, $\beta_t = 0.7 + 0.25\mathbb{1}_{\{t>b^*\}}$ and $b^* \in [0, 250]$ represents the time after which the model is subject to a structural change. We simulate 250 presampled data for forecasting risk measures and another 250 data for backtesting.

We choose the probability level for VaR_p and ES_p to be $p = 0.95$. Via the presampled data, the forecaster obtains the forecasts of $\mathrm{VaR}_{0.95}(L_t|\mathcal{F}_{t-1})$ and $\mathrm{ES}_{0.95}(L_t|\mathcal{F}_{t-1})$ using empirical VaR and ES of the residuals and the estimated model parameters $\hat{\boldsymbol{\theta}} = (\hat{\omega}, \hat{\alpha}, \hat{\beta})$. See Section D.4 for details of the forecasting procedure. Due to the model-free nature, we only use the losses and forecasts (L_t, r_t, z_t) for our e-backtesting method, while the sequential monitoring method also uses the estimated volatility $\sigma_t(\hat{\boldsymbol{\theta}})$ by assuming the GARCH model of the losses. As suggested by Hoga and Demetrescu (2023), the Monte Carlo simulations detector with a rolling window performs the best among others for both VaR and ES monitoring. Therefore, we take this method for comparison with ours. We choose the size $m = 50$ of the rolling window. The significance level of the sequential monitoring method is set to be 5%, while we choose the rejection threshold of our e-backtesting

Figure 4: Top panels: percentage of detections (%) of $\text{VaR}_{0.95}$ (left panels) and $\text{ES}_{0.95}$ (right panels) forecasts over $10,000$ simulations of time series and 250 trading days with structural changes at b^*; Bottom panels: average run lengths (ARLs) of backtesting procedures; black lines ("monitor") represent results of the sequential monitoring method



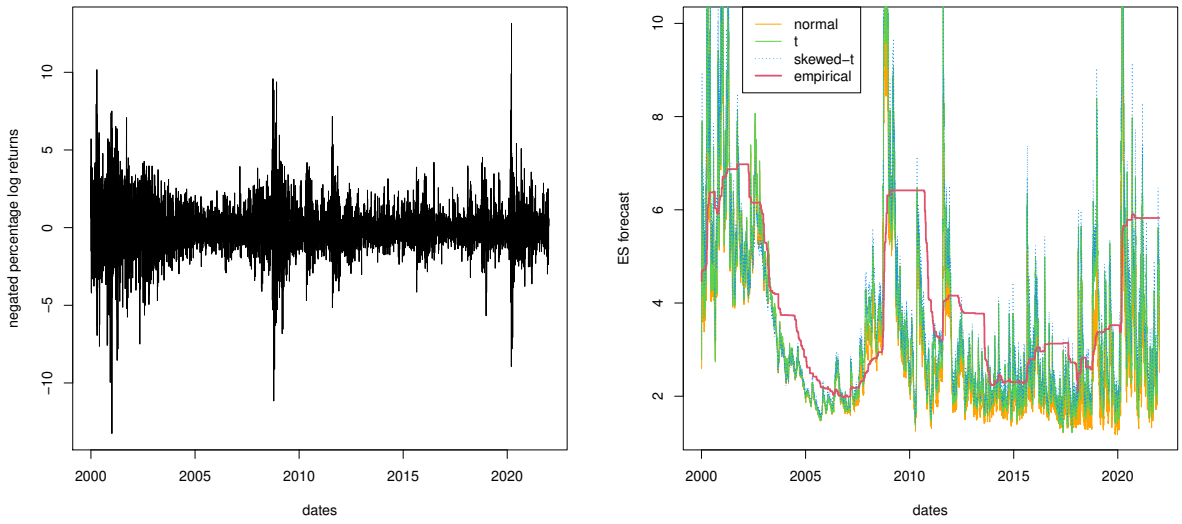
method to be $1/5\% = 20$ (see Theorem 2).

Figure 4 plots the average results we get based on $10,000$ simulations, where the betting processes of e-backtesting are chosen by the GREE, GREL or GREM method. The top panels plot the percentage of detections over the total $10,000$ simulations, including those before and after the structural changes at $t = b^* + 1$, while the bottom panels show the average number of trading days from the structural changes at $t = b^* + 1$ to detections through backtesting, given that detections occur after $t = b^* + 1$. We call this quantity the average run length (ARL) as in Hoga and Demetrescu (2023). We observe that the sequential monitoring method outperforms the GREE, GREL, and GREM methods in terms of the detection percentage and the detection speed (reflected by ARL). This is not surprising, as our backtesting method is nonparametric and requires less model assumptions.[10] On the other hand, our test is one-sided while the sequential monitoring method is a two-sided test. The GREE, GREL, and GREM methods exhibit reasonable performance for all values of b^*. From the ARL plots, the GREE, GREL and GREM methods detect evidence against the forecasts around 0 to 30 days later than the sequential monitoring method.

8 Financial data analysis

[10]The backtesting methods based on cumulative violations in Du and Escanciano (2017) and Hoga and Demetrescu (2023) require calculating the realized uniform variables known as probability integral transforms. For this, one needs to assume a parametric model (e.g., t-distribution) for the losses.

Figure 5: Negated percentage log-returns of the NASDAQ Composite index (left panel); $\mathrm{ES}_{0.875}$ and $\mathrm{ES}_{0.975}$ forecasts fitted by normal, t-, skewed-t distributions, and empirical risk forecasts (right panel) from Jan 3, 2000 to Dec 31, 2021



8.1 The NASDAQ index

We calculate the negated percentage log-returns using the data of the NASDAQ Composite index from Jan 16, 1996 to Dec 31, 2021. An AR(1)-GARCH(1, 1) model is fitted to the data with a moving estimation window of 500 data points. The e-processes in (7) are calculated with the betting process $(\lambda_t)_{t\in[T]}$ chosen by the GREE, GREL or GREM method. Different from the backtesting methods used in Section 7.1, for each $t \in [T]$, the empirical mean in (15) is calculated using a moving window of data in the past 500 days. This choice is made to reflect the practice of risk modeling where more recent data represent the current market and economic conditions better. Therefore, the first 500 forecasts use 500 data points each, and we start the backtesting procedure after the first 500 forecasts are available, thus after the first $1,000$ data points. The sample size for backtesting is $5,536$, corresponding to forecasts of risk measures from Jan 3, 2000 to Dec 31, 2021. We plot the negated log-returns and the forecasts of $\mathrm{ES}_{0.975}$ fitted by normal, t-, and skewed-t distributions for the innovations over time in Figure 5. In addition to the parametric methods, we also plot the empirical risk forecasts with a nonparametric rolling window approach in Figure 5.

We present backtesting results using data from Jan 3, 2005 to Dec 31, 2021 to examine the impact of the 2007 – 2008 financial crisis. Figure 6 shows the e-processes over time. Table 6 demonstrates the average $\mathrm{ES}_{0.975}$ forecasts and the number of days taken to detect evidence against the forecasts, where the second last row contains the results for ES forecasts deliberately over-

Figure 6: Log-transformed e-processes testing $\mathrm{ES}_{0.975}$ with respect to the number of days for the NASDAQ index from Jan 3, 2005 to Dec 31, 2021; left panel: GREE method, middle panel: GREL method, right panel: GREM method

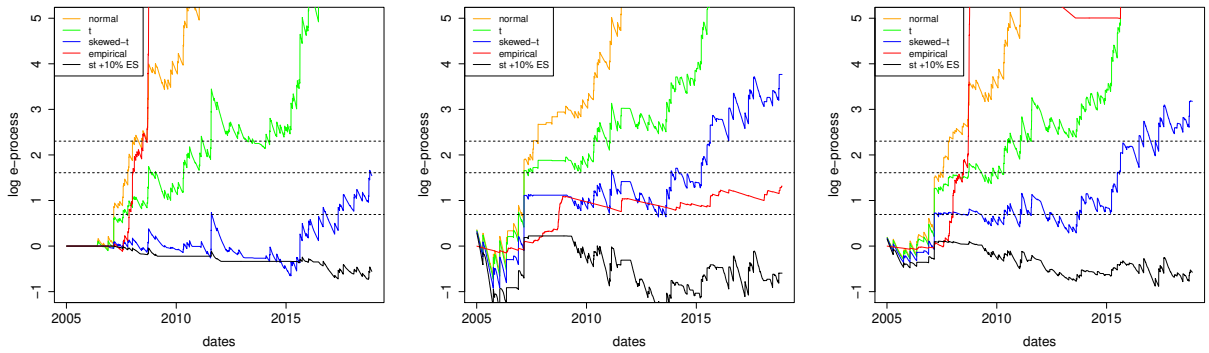


Table 6: Average $\mathrm{ES}_{0.975}$ forecasts (boldface in brackets) and the number of days taken to detect evidence against the forecasts for the NASDAQ index from Jan 3, 2005 to Dec 31, 2021; "–" means no detection is detected till Dec 31, 2021

		GREE			GREL			GREM		
threshold		2	5	10	2	5	10	2	5	10
normal	(**2.676**)	540	704	756	479	540	650	540	610	713
t	(**2.997**)	650	941	1545	479	540	1344	540	933	1381
skewed-t	(**3.202**)	1661	3477	–	540	1545	2676	540	2639	2889
st +10% ES	(**3.522**)	–	–	–	–	–	–	–	–	–
empirical	(**3.656**)	719	758	876	941	3823	–	756	862	931

reported by 10% assuming skewed-t innovations as a forecasting model that is prudent.

Most of the detections in Table 6 happen around 500 – 700 trading days after Jan 3, 2005, where significant losses occurred during the financial crisis. Correspondingly, there are sharp jumps of the e-processes in Figure 6 at around 500 – 700 trading days. In general, we observe that detections for lower thresholds 2 and 5 are significantly earlier than those for the final threshold 10. This allows regulators to get alerted much earlier than using the traditional p-tests when e-processes exceed the first threshold 2 or further exceed 5. The backtesting procedure may be stopped when an e-process exceeds 10, which indicates a "decisive" failure of the underlying model used by the financial institution.

The GREL method generally yields larger e-processes than the GREE method except for the empirical forecasts. This may be because the sharp increase of losses upon the occurrence of the financial crisis violates the growth trend and co-movements of the losses and the risk forecasts,

making the GREE method not favorable compared with the GREL method as discussed in Example 7. It seems from the result that the GREL method is more likely to detect evidence against the risk forecasts for extreme events (e.g., financial crisis) causing an abnormally sharp increase in losses. The GREL method does not perform well in detecting evidence against the empirical forecasts for both VaR and ES. This is expected because the empirical forecasts and the betting process of the GREL method are both obtained only by the information of the empirical distribution of losses, making GREL lack additional information to reject the empirical forecasts. Compared with the GREE and GREL methods, the performance of the GREM method is more stable in different cases where underestimation is detected. Therefore, the GREM method is recommended as a default choice for implementation.

Although the two papers use different time frames for the real data analysis, our method is able to detect evidence against $\mathrm{ES}_{0.975}$ forecasts as efficiently as Nolde and Ziegel (2017) (who tested forecasts with normal and skewed-t innovations). However, due to the nature of sequential and non-asymptotic tests, our method can tell the time when underestimation happens by observing the fast growth of e-processes. This feature is useful for practical financial regulation and cannot be achieved by traditional asymptotic tests based on p-values. A simple real data result for comparison to a p-value test is demonstrated in Section D.5.

8.2 Optimized portfolios

Apart from the NASDAQ index, we perform the e-backtesting procedure on data of a portfolio of $n = 22$ stocks from Jan 5, 2001 to Dec 31, 2021. Suppose that a bank invests in the above portfolio. After each trading day at time $t \in [T]$, the weights

$$\mathbf{w}_t = (w_t^1, \dots, w_t^n) \in \Delta_n = \left\{ (w_1, \dots, w_n) \in [0,1]^n : \sum_{i=1}^n w_i = 1 \right\}$$

are determined by a mean-variance criterion. Specifically, the bank solves the following optimization problem:[11]

$$\max_{\mathbf{w}_t \in \Delta_n} \mathbb{E}[-\mathbf{w}_t^\top \mathbf{L}_t] - \frac{\gamma}{2} \mathrm{var}(-\mathbf{w}_t^\top \mathbf{L}_t),$$

[11] We use the mean-variance strategy to illustrate our method for its simplicity, despite its performance may not be empirically satisfactory; see e.g., DeMiguel et al. (2009). Recall that our backtesting method does not require knowledge of the trading strategy or the statistical model, and can be applied to any trading strategy. There are many other portfolio strategies that can be considered; see e.g., Basak and Shapiro (2001) and Mencía and Sentana (2009).

Table 7: Average $\mathrm{ES}_{0.975}$ forecasts (boldface in brackets) and the number of days taken to detect evidence against the forecasts for portfolio data from Jan 3, 2005 to Dec 31, 2021; "–" means no detection is detected till Dec 31, 2021

		GREE			GREL			GREM		
threshold		2	5	10	2	5	10	2	5	10
normal	(**2.817**)	547	730	767	438	541	541	461	541	714
t	(**3.191**)	767	934	3036	1009	2207	2411	778	2207	2502
skewed-t	(**3.304**)	767	–	–	469	1009	2502	541	2411	2972
st +10% ES	(**3.635**)	–	–	–	–	–	–	–	–	–

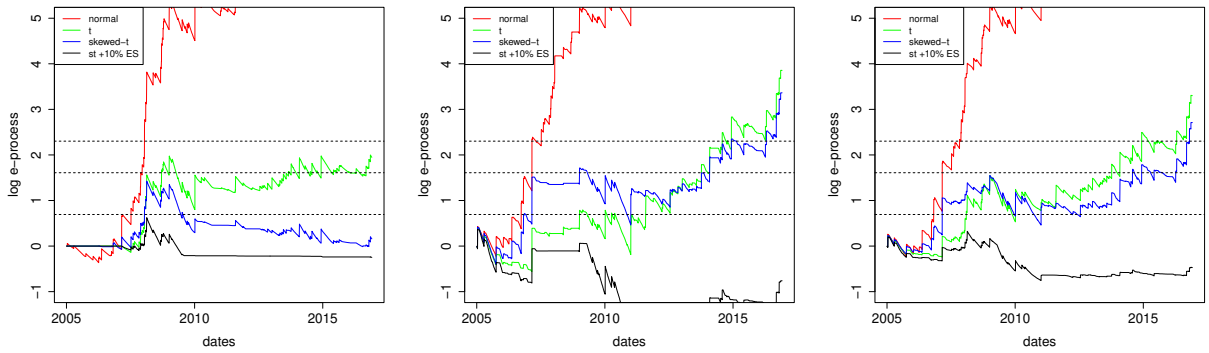
where $\mathbf{L}_t = (L_t^1, \dots, L_t^n)$ is the vector of negated percentage log-returns for all stocks in the portfolio with each $\{L_t^i\}_{t\in[T]}$ being modeled by an AR(1)–GARCH(1, 1) process for $i \in [n]$. The bank reports VaR and ES of the weighted portfolio by assuming $\mathbf{w}_t^\top \mathbf{L}_t$ to be normal, t-, or skewed-t distributed. Some of the assumptions in the estimation procedure are simplistic, and hence we do not expect to obtain precise risk forecasts. Suppose a financial institution reports its risk forecasts based on the naive approach described above. We are more likely to get detections if the simplistic assumptions lead to underestimation. The detailed setup and the list of stocks can be seen in Section D.6.

Table 7 shows the average forecasts of $\mathrm{ES}_{0.975}$ and backtesting results with different innovation distributions. The e-processes are plotted in Figure 7. The portfolio data differ from the simulated time series in the sense that the random losses and risk predictions exhibit much more complicated temporal dependence. Detections are obtained in most of the cases for thresholds 2 and 5 before large losses come in during the financial crisis in 2008. Due to the model-free nature, our e-backtesting method is able to detect evidence against risk forecasts when losses and risk forecasts exhibit complicated temporal dependence. This enables regulations for most real portfolio investments in financial markets.

9 Concluding remarks

The e-backtesting method proposed in this paper is the first model-free and non-asymptotic backtest for ES, the most important risk measure in financial regulation implemented by BCBS (2016). Our methodology contributes to the backtesting issues of ES, which have been a central point of discussions in the risk management literature. Our methods are constructed using the recently developed notions of e-values and e-processes, which are shown to be promising in many application domains of statistics other than risk management. Some topics on which e-values become

Figure 7: Log-transformed e-processes testing $\mathrm{ES}_{0.975}$ with respect to the number of days for portfolio data from Jan 3, 2005 to Dec 31, 2021; left panel: GREE method, middle panel: GREL method, right panel: GREM method



useful include sequential testing (Shafer, 2021; Grünwald et al., 2024), multiple testing and false discovery control (Vovk and Wang, 2021; Wang and Ramdas, 2022), probability forecast evaluation (Henzi and Ziegel, 2022), meta-analysis in biomedical sciences (ter Schure and Grünwald, 2021), and composite hypotheses (Waudby-Smith and Ramdas, 2024). Our paper connects two active areas of research through theoretical results and methodologies, and we expect more techniques from either world to be applicable to solve problems from the other.

Our e-test procedures feature advantages of e-values, including validity for all stopping times and feasibility for no assumptions on the underlying models. Central to our proposed backtesting method, we use backtest e-statistics, which are useful also for traditional testing problems, although the main focus of the paper is backtesting. The characterization results in Section 6 give guidelines for choosing backtest e-statistics. Remarkably, for VaR and ES, essentially unique optimal choices of backtest e-statistics are identified, leaving little doubt on how to choose them in applications.

If the sample size of a test is fixed, and accurate forecasts for the risk model are available and to be tested together with forecasts for the risk measure, then traditional model-based methods may be recommended to use in practice, as they often have better power than our e-backtests. In the more realistic situations where the sample size is not fixed, or no models are to be tested along with the risk measure forecasts, our e-backtests are useful, and their multifaceted attractive features are illustrated by our study.

As for any other new statistical methodology, e-backtests have their own limitations, challenges, and undeveloped extensions. As the main limitation, since e-backtests require very little information on the underlying model, they could be less powerful than traditional model-based or p-value-based approaches. Therefore, there is a trade-off between flexibility and power that a risk practitioner has

to keep in mind. For future directions, an important task is to obtain theoretically optimal betting processes using some data-driven procedures under practical assumptions. The methodology can be extended to more general risk measures and economic indices useful in different contexts, each demanding its own backtest e-statistics and backtesting procedure.

Acknowledgments

We thank Zaichao Du, Chao Gao, Glenn Harrison, Marcin Pitera, Aaditya Ramdas and Neil Xu for helpful comments on the paper. RW acknowledges financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada (Grant Nos. RGPIN-2024-03728, CRC-2022-00141). JZ acknowledge financial support from the Swiss National Science Foundation.

References

C. Acerbi and B. Székely. Backtesting Expected Shortfall. *Risk Magazine*, 27(11):76–81, 2014.

C. Acerbi and B. Székely. General properties of backtestable statistics. *Preprint, SSRN: 2905109*, 2017.

S. Agrawal, W. Koolen, and S. Juneja. Optimal best-arm identification methods for tail-risk measures. *Advances in Neural Information Processing Systems*, 34:25578–25590, 2021.

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

Suleyman Basak and Alexander Shapiro. Value-at-risk-based risk management: optimal policies and asset prices. *The review of financial studies*, 14(2):371–405, 2001.

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

BCBS. *Consultative Document: Fundamental Review of the Trading Book (Oct 2013)*. Bank for International Settlements, 2013. https://www.bis.org/publ/bcbs219.pdf.

BCBS. *Minimum Capital Requirements for Market Risk (Feb 2016)*. Bank for International Settlements, 2016. https://www.bis.org/bcbs/publ/d352.pdf.

BCBS. *Minimum Capital Requirements for Market Risk (Jan 2019)*. Bank for International Settlements, 2019. https://www.bis.org/bcbs/publ/d457.pdf.

I. Berkes, E. Gombay, L. Horváth, and P. Kokoszka. Sequential change-point detection in GARCH(p, q) models. *Econometric Theory*, 20(6):1140–1167, 2004.

J. Berkowitz, P. Christoffersen, and D. Pelletier. Evaluating Value-at-Risk models with desk-level data. *Management Science*, 57(12):2213–2227, 2011.

P. Casgrain, M. Larsson, and J. Ziegel. Anytime-valid sequential testing for elicitable functionals via supermartingales. *Bernoulli*, 30:1347–1373, 2024.

P. F. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39(4):841–862, 1998.

P. F. Christoffersen. *Elements of Financial Risk Management*. Academic Press, second edition, 2011.

J. C. S. Chu, M. Stinchcombe, and H. White. Monitoring structural change. *Econometrica*, 64(5): 1045–1065, 1996.

T. E. Clark and M. W. McCracken. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110, 2001.

D. A. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

V. H. de la Peña, M. J. Klass, and T. L. Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3):1902–1933, 2004.

V. H. de la Peña, T. L. Lai, and Q.-M. Shao. *Self-Normalized Processes: Limit Theory and Statistical Applications. Probability and Its Applications*. Springer, Berlin, first edition, 2009.

V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13:253–263, 1995.

T. Dimitriadis, T. Fissler, and J. F. Ziegel. Osband's principle for identification functions. *Statistical Papers*, 2023. doi: https://doi.org/10.1007/s00362-023-01428-x.

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

P. Embrechts, T. Mao, Q. Wang, and R. Wang. Bayes risk, elicitability, and the Expected Shortfall. *Mathematical Finance*, 31(4):1190–1217, 2021.

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

J. C. Escanciano and J. Olmo. Backtesting parametric value-at-risk with estimation risk. *Journal of Business & Economic Statistics*, 28(1):36–51, 2010.

T. Fissler. *On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space*. PhD Thesis, University of Bern, 2017.

T. Fissler and H. Holzmann. Measurability of functionals and of ideal point forecasts. *Electronic*

Journal of Statistics, 16(2):5019–5034, 2022.

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

H. Föllmer and A. Schied. *Stochastic Finance*. De Gruyter, Berlin, fourth edition, 2016.

R. Frongillo and I. A. Kash. Elicitation complexity of statistical properties. *Biometrika*, 108(4): 857–879, 2021.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578, 2006.

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

P. Grünwald, R. de Heide, and W. M. Koolen. Safe testing. *Journal of the Royal Statistical Society, Series B*, 2024. To appear.

A. Henzi and J. F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, 109(3):647–663, 2022.

Y. Hoga and M. Demetrescu. Monitoring Value-at-Risk and Expected Shortfall forecasts. *Management Science*, 69(5):2954–2971, 2023.

K. Jang, K.-S. Jun, I. Kuzborskij, and F. Orabona. Tighter pac-bayes bounds through coin-betting. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2240–2264. PMLR, 2023.

J. L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35(4):917–926, 1956.

P. Kupiec. Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2):73–84, 1995.

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

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ, revised edition, 2015.

Javier Mencía and Enrique Sentana. Multivariate location–scale mixtures of normals and mean–variance–skewness portfolio allocation. *Journal of Econometrics*, 153(2):105–121, 2009.

F. Moldenhauer and M. Pitera. Backtesting expected shortfall: a simple recipe? *Journal of Risk*, 22(1):17—-42, 2017.

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

A. Patton, J. F. Ziegel, and R. Chen. Dynamic semiparametric models for Expected Shortfall (and

Value-at-Risk). *Journal of Econometrics*, 211(2):388–413, 2019.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *arXiv preprint arXiv:2410.23614*, 2024.

Z. Ren and R. F. Barber. Derandomized knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B*, 86(1):122–154, 2024.

S. Resnick. *A Probability Path*. Springer, Birkhäuser Boston, MA, first edition, 2019.

R. T. Rockafellar and S. Uryasev. Conditional Value-at-Risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

D. Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57 (3):571–587, 1989.

G. Shafer. The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184(2):407–431, 2021.

G. Shafer and V. Vovk. *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons, 2019.

G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test martingales, bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

I. Steinwart, C. Pasin, R. Williamson, and S. Zhang. Elicitation and identification of properties. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 482–526, Barcelona, Spain, 2014. PMLR.

Q. Su, Z. Qin, L. Peng, and G. Qin. Efficiently backtesting conditional Value-at-Risk and conditional Expected Shortfall. *Journal of the American Statistical Association*, 116(536):2041–2052, 2021.

J. ter Schure and P. Grünwald. ALL-IN meta-analysis: breathing life into living systematic reviews. *Preprint, arXiv: 2109. 12141*, 2021.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

J. Ville. Étude critique de la notion de collectif. *Thèses de l'entre-deux-guerres*, 1939.

V. Vovk and R. Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49(3):1736–1754, 2021.

V. Vovk and R. Wang. Merging sequential e-values via martingales. *Electronic Journal of Statistics*, 18(1):1185–1205, 2024a.

V. Vovk and R. Wang. Nonparametric e-tests of symmetry. *The New England Journal of Statistics in Data Science*, 2:261–270, 2024b.

V. Vovk, B. Wang, and R. Wang. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375, 2022.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society, Series B*, 84(3):822–852, 2022.

R. Wang and Y. Wei. Risk functionals with convex level sets. *Mathematical Finance*, 30(4):1337–1367, 2020.

R. Wang and R. Zitikis. An axiomatic foundation for the Expected Shortfall. *Management Science*, 67(3):1413–1429, 2021.

R. Wang, Y. Wei, and G. E. Willmot. Characterization, robustness and aggregation of signed Choquet integrals. *Mathematics of Operations Research*, 45(3):993–1015, 2020.

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of Royal Statistical Society, Series B*, 86(1):1–27, 2024.

K. D. West. Asymptotic inference about predictive ability. *Econometrica*, 64:1067–1084, 1996.

M. E. Yaari. The dual theory of choice under risk. *Econometrica*, 55(1):95–115, 1987.

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

D. Ziggel, T. Berens, G. N. F. Weiß, and D. Wied. A new set of improved Value-at-Risk backtests. *Journal of Banking & Finance*, 48:29–41, 2014.

E-companion to "E-backtesting"

This e-companion contains four appendices. Section A contains the betting processes calculated via Taylor approximation for VaR and ES. Section B discusses the link between backtest e-statistics and identification functions. Section C contains proofs of all results. Section D contains some details of the simulation and data analysis.

A Taylor approximation formulas for GREE and GREL

We give formulas for the betting processes of the GREE and GREL methods for VaR and ES via Taylor approximation. For the GREL method, the special case of VaR, that is, taking $e = e_p^Q$ in (15), yields

$$\lambda_t^{\mathrm{GREL}} \approx 0 \vee \frac{(1-p)\left((t-1)p - \sum_{s=1}^{t-1} \mathbb{1}_{\{L_s \leqslant r\}}\right)}{(t-1)p^2 + (1-2p)\sum_{s=1}^{t-1} \mathbb{1}_{\{L_s \leqslant r\}}} \wedge \gamma.$$

For the special case of ES, taking $e = e_p^{\mathrm{ES}}$ in (15), the approximation is

$$\lambda_t^{\mathrm{GREL}} \approx 0 \vee \frac{(1-p)(r-z)\left(\sum_{s=1}^{t-1}(L_s - z)_+ - (t-1)(1-p)(r-z)\right)}{\sum_{s=1}^{t-1}((L_s - z)_+ - (1-p)(r-z))^2} \wedge \gamma.$$

The corresponding formulas for the GREE method are obtained by replacing r and (r, z) by r_s and (r_s, z_s) in the s-th summand in above formulas, respectively.

B Link between backtest e-statistics and identification functions

The link between backtest e-statistics and identification functions is useful for deriving the characterization results of backtest e-statistics. An integrable function $V : \mathbb{R}^{d+1} \to \mathbb{R}^d$ is said to be an $\mathcal{M}$-*identification function* for a functional $\psi : \mathcal{M} \to \mathbb{R}^d$ if $\int_{\mathbb{R}} V(x, \psi(F))\,\mathrm{d}F(x) = \mathbf{0}$ for all $F \in \mathcal{M}$. Furthermore, V is said to be *strict* if

$$\int_{\mathbb{R}} V(x, y)\,\mathrm{d}F(x) = \mathbf{0} \iff y = \psi(F)$$

for all $F \in \mathcal{M}$ and $y \in \mathbb{R}^d$ (Fissler and Ziegel, 2016). We say that ψ is *identifiable* if there exists a strict $\mathcal{M}$-identification function for ψ.

There is a connection between backtest e-statistics and identification functions. Let $e : \mathbb{R}^2 \to \mathbb{R}$

be a backtest e-statistic for $\rho : \mathcal{M} \to \mathbb{R}$. For $F \in \mathcal{M}$ and $r \geqslant \rho(F) > r'$ it holds that

$$\int e(x, r) \, \mathrm{d}F(x) \leqslant 1 < \int e(x, r') \, \mathrm{d}F(x),$$

and hence, $1 - e(x, r)$ is often a strict identification function for ρ. Since identifiability of a functional coincides with eliciability under some assumptions detailed in Steinwart et al. (2014), Proposition 4 is not surprising since elicitable functionals are known to have CxLS.

A backtest e-statistic is called *non-conservative* if $\int_{\mathbb{R}} e(x, \psi(F)) \, \mathrm{d}F(x) = 1$ for all $P \in \mathcal{P}$.

Proposition EC.1. *Let $e : \mathbb{R}^{d+1} \to [0, \infty]$ be a non-conservative backtest e-statistic for $\psi = (\rho, \phi) : \mathcal{M} \to \mathbb{R}^d$, and assume that ϕ has a $\mathcal{P}$-identification function v. We have $V(x, r, z) = (v(x, z), 1 - e(x, r, z))^\top$ is a $\mathcal{P}$-identification function for ψ.*

Proof. Let $F \in \mathcal{P}$. By assumption,

$$\int V(x, \rho(F), \phi(F)) \, \mathrm{d}F(x) = \left(\int v(x, \phi(F)) \, \mathrm{d}F(x), 1 - \int e(x, \rho(F), \phi(F)) \, \mathrm{d}F(x) \right)^\top = \mathbf{0}. \quad \square$$

The connection of backtest e-statistics to identification functions is useful because under some regularity conditions there are characterization results for all possible identification functions for a functional (Fissler, 2017; Dimitriadis et al., 2020). Roughly speaking, given a monotone backtest e-statistic e for ψ, then all other possible monotone backtest e-statistics e' must be of the form

$$e'(x, r) = 1 + h(r)(e(x, r) - 1)$$

for some non-negative function h. Clearly, h must fulfill further criteria to ensure that e' is a monotone backtest e-statistic for ψ.

A further consequence of these considerations is that for a functional ρ with monotone backtest e-statistic, there must be an identification function $V(x, r)$ that is bounded below by -1. This rules out a number of functionals including the expectation without further conditions on $\mathcal{M}$.

C Omitted proofs of all results

Proof. Proof of Lemma 1. The backward direction is a simple application of Jensen's inequality, which yields

$$0 < \mathbb{E}[\log(1 - \lambda + \lambda E)] \leqslant \log \mathbb{E}[1 - \lambda + \lambda E] \implies \mathbb{E}[E] > 1.$$

To show the forward direction, it suffices to verify $\mathbb{E}[\log(1 - \lambda + \lambda E)] > 1$ for $\lambda > 0$ small enough.

Note that $\mathbb{E}[E] > 1$ implies $\mathbb{E}[E \wedge K] > 1$ for some $K \geqslant 1$. We denote by $Y = E \wedge K$ and let $x_- = (-x)_+$ for $x \in \mathbb{R}$. Since $\mathbb{E}[(Y-1)_+] - \mathbb{E}[(Y-1)_-] = \mathbb{E}[Y-1] > 0$, there exists some $\varepsilon \in (0,1)$ such that

$$\frac{1}{1+\varepsilon}\mathbb{E}[(Y-1)_+] - \frac{1}{1-\varepsilon}\mathbb{E}[(Y-1)_-] > 0.$$

Note that $\log(1+x) \geqslant x/(1+\varepsilon)$ for $x \in [0,\varepsilon)$ and $\log(1+x) \geqslant x/(1-\varepsilon)$ for $x \in (-\varepsilon, 0)$, that is,

$$\log(1+x) \geqslant \frac{x_+}{1+\varepsilon} - \frac{x_-}{1-\varepsilon} \quad \text{for } x \in (-\varepsilon, \varepsilon).$$

Hence, for $\lambda \in (0, \varepsilon/K)$, implying $\lambda(Y-1) \in (-\varepsilon, \varepsilon)$, we have

$$\begin{aligned}\mathbb{E}[\log(1-\lambda+\lambda E)] &\geqslant \mathbb{E}[\log(1+\lambda(Y-1))] \\ &\geqslant \frac{1}{1+\varepsilon}\mathbb{E}[\lambda(Y-1)_+] - \frac{1}{1-\varepsilon}\mathbb{E}[\lambda(Y-1)_-] > 0,\end{aligned}$$

thus showing the desired inequality. □

Proof. Proof of Proposition 1. Suppose that H_0 in (10) holds. By the VaR-ES relation in (4) and (5),

$$\mathbb{E}[e_p^{\mathrm{ES}}(L,r,z)] = \frac{\mathbb{E}\left[(L-z)_+\right]}{(1-p)(r-z)} \leqslant \frac{\mathbb{E}\left[(L-\mathrm{VaR}_p(L))_+\right]}{(1-p)(\mathrm{ES}_p(L)-\mathrm{VaR}_p(L))} = 1.$$

Hence, $e_p^{\mathrm{ES}}(L,r,z)$ is an e-variable for (10). □

Proof. Proof of Proposition 2. For all $(r,z) \in \mathbb{R} \times \mathbb{R}^{d-1}$ and $t \in \mathbb{N}$, write $X_t = e(L_t, r, z)$.

(i) Note that $\lambda_t^{\mathrm{GRO}}(r,z) > 0$ is equivalent to $\mathbb{E}^{Q_t}[\log(1-\lambda+\lambda X_t)|\mathcal{F}_{t-1}] > 0$ for some λ that is $\mathcal{F}_{t-1}$-measurable. Therefore, the equivalence statement follows from Lemma 1.

(ii) For the "⇐" direction, suppose that $\lambda_t^{\mathrm{GRO}}(r,z) = 1$. It is clear that $Q_t(X_t = 0) = 0$. It follows by continuity of $\lambda \mapsto \mathbb{E}^{Q_t}[\log(1-\lambda+\lambda X_t)|\mathcal{F}_{t-1}]$ and $\lambda \mapsto \mathbb{E}^{Q_t}[(X_t-1)/(1-\lambda+\lambda X_t)]$ that

$$0 \leqslant \frac{\mathrm{d}}{\mathrm{d}\lambda}\mathbb{E}^{Q_t}[\log(1-\lambda+\lambda X_t)|\mathcal{F}_{t-1}]\Big|_{\lambda=1} = \mathbb{E}^{Q_t}\left[\frac{X_t-1}{X_t}\middle|\mathcal{F}_{t-1}\right].$$

Hence, $\mathbb{E}^{Q_t}[1/X_t|\mathcal{F}_{t-1}] \leqslant 1$. For the "⇒" direction, suppose that $\mathbb{E}^{Q_t}[1/X_t|\mathcal{F}_{t-1}] \leqslant 1$. It follows that

$$\begin{aligned}\mathbb{E}^{Q_t}[\log(1-\lambda+\lambda X_t) - \log(X_t)|\mathcal{F}_{t-1}] &= \mathbb{E}^{Q_t}\left[\log\left(\frac{1-\lambda}{X_t}+\lambda\right)\middle|\mathcal{F}_{t-1}\right] \\ &\leqslant \mathbb{E}^{Q_t}\left[(1-\lambda)\left(\frac{1}{X_t}-1\right)\middle|\mathcal{F}_{t-1}\right] \leqslant 0.\end{aligned}$$

By strict concavity of $\lambda \mapsto \mathbb{E}^{Q_t}[\log(1-\lambda+\lambda X_t)|\mathcal{F}_{t-1}]$, we have $\lambda_t^{\mathrm{GRO}}(r,z) = 1$, where the upper

bound $\log(X_t)$ is obtained. $\square$

Next we show Theorem 3, which relies on Lemma EC.1 and Proposition EC.2 below.

Lemma EC.1. *If $M, M_t : [0,1] \to L^0$ are convex for all $t \in \mathbb{N}$, M is continuous and $M_t(\lambda) \xrightarrow{\text{a.s.}} M(\lambda)$ as $t \to \infty$ for all $\lambda \in [0,1]$, then $\sup_{\lambda\in[0,1]} |M_t(\lambda) - M(\lambda)| \xrightarrow{\text{p}} 0$ as $t \to \infty$.*

Proof. For all $t \in \mathbb{N}$, define an affine function $\psi_t : [0,1] \to L^0$ such that $\psi_t(0) = M_t(0)$ and $\psi_t(1) = M_t(1)$. This is clear that ψ_t converges uniformly to the affine function $\psi : [0,1] \to L^0$ such that $\psi(0) = M(0)$ and $\psi(1) = M(1)$. Therefore, replacing M_t by $M_t - \psi_t$ and M by $M - \psi$, we assume without loss of generality that $M_t(0) = M(0) = M_t(1) = M(1) = 0$ for all $t \in \mathbb{N}$.

For all $\eta > 0$, take $\varepsilon = \eta/4$. By continuity of M, there exists $\delta_0 > 0$, such that $|M(\lambda) - M(\lambda')| < \varepsilon$ for all $|\lambda - \lambda'| < \delta$. By convexity of M, there exists $\tilde{\lambda} \in (0,1)$, such that M is decreasing on $[0, \tilde{\lambda}]$ and increasing on $[\tilde{\lambda}, 1]$. Define $K = (M(1) - M(\tilde{\lambda}))/(1 - \tilde{\lambda})$ and $\delta = \min\{\varepsilon/K, \delta_0\}$. There exist $\lambda_1 < \cdots < \lambda_{I-1}$ for $I \in \mathbb{N} \setminus \{1\}$, such that $\lambda_{i+1} - \lambda_i < \delta$ for all $i \in \{0, \dots, I-1\}$, where we write $\lambda_0 = 0$ and $\lambda_I = 1$. It follows that $\delta \leqslant 1 - \lambda_i$ for all $i \in \{0, \dots, I-1\}$.

Because $M_t(\lambda) \xrightarrow{\text{a.s.}} M(\lambda)$ as $t \to \infty$ for all $\lambda \in [0,1]$, there exists an event A with $\mathbb{P}(A) = 1$ as follows: There exists $T \in \mathbb{N}$, such that for all $t > T$, we have $|M_t(\lambda_i) - M(\lambda_i)| < \varepsilon$ for all $i = 0, \dots, I$. For all $\lambda \in [0,1]$, there exists $i \in \{0, \dots, I\}$, such that $\lambda \in [\lambda_i, \lambda_{i+1}]$. Without loss of generality, we assume $0 \leqslant \lambda_i \leqslant \lambda \leqslant \lambda_{i+1} \leqslant \tilde{\lambda}$. The case of $\tilde{\lambda} \leqslant \lambda_i \leqslant \lambda \leqslant \lambda_{i+1} \leqslant 1$ can be shown analogously by symmetry. If A holds, then

$$\begin{aligned}
|M_t(\lambda) - M(\lambda)| &\leqslant |M_t(\lambda) - M(\lambda_i)| \\
&< \varepsilon + \frac{M(1) - M(\lambda_{i+1}) + \varepsilon}{1 - \lambda_{i+1}}\delta + |M(\lambda_i) - M(\lambda_{i+1})| \\
&< 3\varepsilon + K \min\left\{\frac{\varepsilon}{K}, \delta_0\right\} \leqslant 4\varepsilon = \eta.
\end{aligned}$$

Therefore, we have $\mathbb{P}(\sup_{\lambda\in[0,1]} |M_t(\lambda) - M(\lambda)| \geqslant \eta) = 0$ for all $t > T$. Hence, $\sup_{\lambda\in[0,1]} |M_t(\lambda) - M(\lambda)| \xrightarrow{\text{p}} 0$ as $t \to \infty$. $\square$

The next proposition says that under the iid assumption, the betting process computed from empirical distributions is asymptotically equivalent to that computed from the true distribution.

Proposition EC.2. *Let $X_1, X_2, \dots$ be nonnegative iid random variables with $\mathbb{E}[\log(X_1)] < \infty$. Let*

$$\lambda_t = \operatorname*{arg\,max}_{\lambda\in[0,\gamma]} \frac{1}{t-1}\sum_{s=1}^{t-1} \log(1 - \lambda + \lambda X_s); \quad \lambda^* = \operatorname*{arg\,max}_{\lambda\in[0,\gamma]} \mathbb{E}[\log(1 - \lambda + \lambda X_t)], \quad t \in \mathbb{N}.$$

We have $T^{-1}\sum_{t=1}^{T}(\log(1 - \lambda_t + \lambda_t X_t)) - \log(1 - \lambda^* + \lambda^* X_t)) \xrightarrow{L^1} 0$ as $T \to \infty$.

Proof. Write $M(\lambda) = \mathbb{E}[\log(1-\lambda+\lambda X_t)]$ and

$$M_t(\lambda) = \frac{1}{t-1}\sum_{s=1}^{t-1}\log(1-\lambda+\lambda X_s) \quad \text{for } \lambda \in [0,\gamma],\ t \in \mathbb{N}.$$

By the strong law of large numbers, we have $M_t(\lambda) \xrightarrow{\text{a.s.}} M(\lambda)$ as $t \to \infty$ for all $\lambda \in [0,\gamma]$. Since the functions M and M_t are concave for all $t \in \mathbb{N}$, by Lemma EC.1, we have $\sup_{\lambda\in[0,\gamma]} |M_t(\lambda) - M(\lambda)| \xrightarrow{\text{a.s.}} 0$. For all $\varepsilon > 0$, we have

$$\sup_{\lambda:|\lambda-\lambda^*|\geqslant\varepsilon} M(\lambda) \leqslant M(\lambda^*)$$

by the definition of λ^* and the concavity of M. For all $t \in \mathbb{N}$, we have $M_t(\lambda_t) \geqslant M_t(\lambda^*)$ by the definition of λ_t. Therefore, we have by Theorem 5.7 of van der Vaart (1998) that $\lambda_t \xrightarrow{\text{p}} \lambda^*$ as $t \to \infty$. Because λ_t is bounded for all $t \in \mathbb{N}$, $\{\lambda_t\}_{t\in\mathbb{N}}$ is uniformly integrable. It follows that $\lambda_t \to \lambda^*$ with respect to the L^1-norm as $t \to \infty$, denoted by $\lambda_t \xrightarrow{L^1} \lambda^*$; see e.g., Resnick (2019, Theorem 6.6.1).

Next, we show that

$$\frac{1}{T}\sum_{t=1}^{T}(\log(1-\lambda_t+\lambda_t X_t) - \log(1-\lambda^*+\lambda^* X_t)) \xrightarrow{L^1} 0 \tag{EC.1}$$

as $T \to \infty$. To simplify notation, write $Y_t(\lambda) = 1-\lambda+\lambda X_t = 1+(X_t-1)\lambda$, where λ is a variable.

For all $\varepsilon_1 > 0$ and $t \in \mathbb{N}$, by continuity and the monotone convergence theorem, there exists $\delta > 0$, such that $\log(1-\delta) > -\varepsilon_1$ and $\mathbb{E}[\log(1+(X_t-1)\delta)|X_t > 1] < \varepsilon_1$. If $\lambda^* = 0$, we have $\lambda_t \xrightarrow{\text{p}} 0$ as $t \to \infty$. Hence, there exists $N > 0$, such that for all $t > N$, $\mathbb{P}(\lambda_t > \delta) < \varepsilon_1$. We write $x_- = \min\{x, 0\}$ and $x_+ = \max\{x, 0\}$ for $x \in \mathbb{R}$. It follows that

$$\begin{aligned}\mathbb{E}[(\log Y_t(\lambda_t))_-] &\geqslant \mathbb{E}[\log(1-\lambda_t)] \geqslant \log(1-\gamma)\mathbb{P}(\delta < \lambda_t \leqslant \gamma) + \log(1-\delta)\mathbb{P}(0 \leqslant \lambda_t \leqslant \delta)\\ &\geqslant (\log(1-\gamma)-1)\varepsilon_1,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[(\log Y_t(\lambda_t))_+] &= \mathbb{E}[\log Y_t(\lambda_t)(\mathbb{1}_{\{X_t>1, 0\leqslant\lambda_t\leqslant\delta\}} + \mathbb{1}_{\{X_t>1,\delta<\lambda_t\leqslant\gamma\}})]\\ &\leqslant \mathbb{E}[\log Y_t(\delta)|X_t > 1]\mathbb{P}(0 \leqslant \lambda_t \leqslant \delta) + \mathbb{E}[\log Y_t(\gamma)|X_t > 1]\mathbb{P}(\delta < \lambda_t \leqslant \gamma)\\ &\leqslant (1 + \mathbb{E}[\log Y_t(\gamma)|X_t > 1])\varepsilon_1.\end{aligned}$$

It is clear that $\sup_{t\in\mathbb{N}} \mathbb{E}[\log Y_t(\gamma)|X_t > 1]$ is bounded because $\sup_{t\in\mathbb{N}} \mathbb{E}[\log(X_t)] < \infty$. Hence, $\mathbb{E}[|\log Y_t(\lambda_t)|] < M_1\varepsilon_1$ for some $M_1 > 0$ and for all $t > N$. Therefore, there exists $N_1 > 0$, such

that for all $T > N_1$,

$$\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T}\log Y_t(\lambda_t)\right|\right] \leqslant \frac{1}{T}\sum_{t=1}^{N}\mathbb{E}[|\log Y_t(\lambda_t)|] + \frac{1}{T}\sum_{t=N+1}^{T}\mathbb{E}[|\log Y_t(\lambda_t)|] \leqslant \varepsilon_1 + \frac{(T-N)M_1}{T}\varepsilon_1.$$

Hence (EC.1) holds for $\lambda^* = 0$.

If $\lambda^* \in (0, \gamma]$, we write $K = \max\{1/\lambda^*, 1/(1-\lambda^*)\} < \infty$. Thus

$$\frac{X_t - 1}{1 + (X_t - 1)\lambda^*} \in [-K, K] \quad \text{for all } t \in \mathbb{N}.$$

For all $\varepsilon_2 > 0$, it is clear by continuity that there exists $\delta > 0$, such that $\log(1-\delta/K) > -\varepsilon_2$. Because $\lambda_t \xrightarrow{L^1} \lambda^*$ as $t \to \infty$, there exists $N > 0$, such that $\mathbb{P}(|\lambda_t - \lambda^*| > \delta) < \varepsilon_2$ and $\mathbb{E}[|\lambda_t - \lambda^*|] < \varepsilon_2$ for all $t > N$. It follows that, for all $t > N$,

$$\begin{aligned}
&\mathbb{E}\left[(\log Y_t(\lambda_t) - \log Y_t(\lambda^*))_-\right] \\
&= \mathbb{E}\left[\log\left(1 + \frac{X_t - 1}{Y_t(\lambda^*)}(\lambda_t - \lambda^*)\right)_-\right] \\
&= \mathbb{E}\left[\log\left(1 + \frac{X_t - 1}{Y_t(\lambda^*)}(\lambda_t - \lambda^*)\right)\left(\mathbb{1}_{\{X_t \geqslant 1, 0 \leqslant \lambda_t \leqslant \lambda^*\}} + \mathbb{1}_{\{X_t < 1, \lambda^* < \lambda_t \geqslant \gamma\}}\right)\right] \\
&\geqslant \mathbb{P}(|\lambda_t - \lambda^*| \leqslant \delta)\log(1-\delta/K) + \mathbb{P}(|\lambda_t - \lambda^*| > \delta)\Big(\mathbb{E}[\log Y_t(\lambda_t)|X_t \geqslant 1, 0 \leqslant \lambda_t < \lambda^* - \delta] \\
&\qquad + \mathbb{E}[\log Y_t(\lambda_t)|X_t < 1, \lambda^* + \delta \leqslant \lambda_t < \gamma] - \mathbb{E}[\log Y_t(\lambda^*)]\Big) \\
&\geqslant (\log(1-\gamma) - \mathbb{E}[\log Y_t(\lambda^*)] - 1)\varepsilon_2,
\end{aligned}$$

and

$$\mathbb{E}\left[(\log Y_t(\lambda_t) - \log Y_t(\lambda^*))_+\right] = \mathbb{E}\left[\log\left(1 + \frac{X_t - 1}{Y_t(\lambda^*)}(\lambda_t - \lambda^*)\right)_+\right] \leqslant K\mathbb{E}[|\lambda_t - \lambda^*|] < K\varepsilon_2.$$

Because $\mathbb{E}[\log Y_t(\lambda^*)]$ is bounded, we have $E\left[|\log Y_t(\lambda_t) - \log Y_t(\lambda^*)|\right] < M_2\varepsilon_2$ for some $M_2 > 0$ and for all $t > N$. Similar argument as the case of $\lambda^* = 0$ leads to (EC.1). □

Proposition EC.2 gives a simplified illustration of the asymptotic optimality of the GREE method, which uses historical one-sided e-statistics as iid input. This gives rise to a proof of Theorem 3.

Proof. Proof of Theorem 3. For (i), because $(r_t, z_t)_{t\in\mathbb{N}}$ is deterministic and $(e(L_t, r_t, z_t))_{t\in\mathbb{N}}$ is iid,

we have for all $t \in \mathbb{N}$,

$$\lambda_t^{\mathrm{GRO}}(r_t, z_t) = \arg\max_{\lambda\in[0,\gamma]} \mathbb{E}^{Q_t}[\log(1-\lambda+\lambda e(L_t, r_t, z_t))|\mathcal{F}_{t-1}] = \arg\max_{\lambda\in[0,\gamma]} \mathbb{E}[\log(1-\lambda+\lambda X_t)|\mathcal{F}_{t-1}] =: \lambda^*,$$

where $X_t = e(L_t, r_t, z_t)$. The statement thus follows directly from Proposition EC.2.

For (ii), we first show the result for any fixed $(r, z) \in \psi^*(\mathcal{P})$. This follows directly from Proposition EC.2 by taking $X_t = e(L_t, r, z)$, $\lambda_t = \lambda_t^{\mathrm{GREL}}(r, z)$, and $\lambda^* = \lambda_t^{\mathrm{GRO}}(r, z)$ for $t \in \mathbb{N}$.

(a) Suppose that $(r_t, z_t)_{t\in\mathbb{N}}$ takes finitely many possible values in $\mathbb{R}^{\mathbb{N}}$. Let $(M_t)_{t\in\mathbb{N}}$ be defined in (7), $\boldsymbol{\lambda}^{\mathrm{GREL}} = (\lambda_t^{\mathrm{GREL}}(r_t, z_t))_{t\in\mathbb{N}}$ and $\boldsymbol{\lambda}^{\mathrm{GRO}} = (\lambda^{\mathrm{GRO}}(r_t, z_t))_{t\in\mathbb{N}}$. We have

$$\frac{1}{T}(\log(M_T(\boldsymbol{\lambda}^{\mathrm{GREL}})) - \log(M_T(\boldsymbol{\lambda}^{\mathrm{GRO}}))) \xrightarrow{L^1} 0$$

by taking mixtures of all possible values of $(r_t, z_t)_{t\in\mathbb{N}}$ that are finitely many.

(b) It suffices to show the result for $d = 1$ and the general case holds similarly. Since $e(x, r)$ is continuous in r and r_t, $t \in \mathbb{N}$, are in a common compact set, $e(x, r)$ is uniformly continuous with respect to r. Define $M(r, \lambda) = \mathbb{E}^{Q_t}[\log(1 - \lambda + \lambda e(L_t, r))|\mathcal{F}_{t-1}]$ with $r \in \mathbb{R}$, $\lambda \in [0, \gamma]$ and $L_t \sim Q_t$ for $t \in \mathbb{N}$.

Let Q_t be the empirical probability measure $\widehat{Q}_{t-1}$ for all $t \in \mathbb{N}$. We now prove that for all $t \in \mathbb{N}$ and $\varepsilon > 0$, there exists $\delta_1 > 0$, such that for all $|r - r'| < \delta_1$, $|\lambda_t^{\mathrm{GREL}}(r) - \lambda_t^{\mathrm{GREL}}(r')| \leqslant \varepsilon$. Suppose that the negated statement is true. Hence there exists $t \in \mathbb{N}$ and $\varepsilon_0 > 0$, such that for all $\delta > 0$, there exist $|r_\delta - r'_\delta| < \delta$, $|\lambda_t^{\mathrm{GREL}}(r_\delta) - \lambda_t^{\mathrm{GREL}}(r'_\delta)| > \varepsilon_0$. Because $\lambda_t^{\mathrm{GREL}}(r) = \arg\max_{\lambda\in[0,1]} M(r, \lambda)$ and $M(r, \lambda)$ is strictly concave in λ, we have

$$\min\left\{M(r_\delta, \lambda_t^{\mathrm{GREL}}(r_\delta)) - M(r_\delta, \lambda_t^{\mathrm{GREL}}(r'_\delta)),\ M(r'_\delta, \lambda_t^{\mathrm{GREL}}(r'_\delta)) - M(r'_\delta, \lambda_t^{\mathrm{GREL}}(r_\delta))\right\} > l$$

for some $l > 0$. By uniform continuity of $M(r, \lambda)$ with respect to r, there exists $\delta_0 > 0$, such that for all $|r - r'| < \delta_0$,

$$\max\left\{\left|M(r, \lambda_t^{\mathrm{GREL}}(r)) - M(r', \lambda_t^{\mathrm{GREL}}(r))\right|, \left|M(r, \lambda_t^{\mathrm{GREL}}(r')) - M(r', \lambda_t^{\mathrm{GREL}}(r'))\right|\right\} < l.$$

Therefore,

$$\begin{aligned} 2l &< M(r_{\delta_0}, \lambda_t^{\mathrm{GREL}}(r_{\delta_0})) - M(r_{\delta_0}, \lambda_t^{\mathrm{GREL}}(r'_{\delta_0})) + M(r'_{\delta_0}, \lambda_t^{\mathrm{GREL}}(r'_{\delta_0})) - M(r'_{\delta_0}, \lambda_t^{\mathrm{GREL}}(r_{\delta_0})) \\ &\leqslant \left|M(r_{\delta_0}, \lambda_t^{\mathrm{GREL}}(r_{\delta_0})) - M(r'_{\delta_0}, \lambda_t^{\mathrm{GREL}}(r_{\delta_0}))\right| + \left|M(r_{\delta_0}, \lambda_t^{\mathrm{GREL}}(r'_{\delta_0})) - M(r'_{\delta_0}, \lambda_t^{\mathrm{GREL}}(r'_{\delta_0}))\right| < 2l. \end{aligned}$$

This leads to a contradiction.

Similarly, we can show that there exists $\delta_2 > 0$, such that for all $|r - r'| < \delta_2$, $|\lambda^{\mathrm{GRO}}(r) - \lambda^{\mathrm{GRO}}(r')| \leqslant \varepsilon$ by taking Q_t to the probability measure Q for the iid random variables L_t, $t \in \mathbb{N}$. Take $\hat{\delta} = \min\{\delta_1, \delta_2\}$. Because $r_t \xrightarrow{\mathrm{p}} r_0$, for all $\eta > 0$, there exists $N \in \mathbb{N}$, such that $Q(|r_t - r_0| < \hat{\delta}) > 1 - \eta$ for all $t > N$. It follows that

$$\max\left\{Q(|\lambda_t^{\mathrm{GREL}}(r_t) - \lambda_t^{\mathrm{GREL}}(r_0)| > \varepsilon),\ Q(|\lambda^{\mathrm{GRO}}(r_t) - \lambda^{\mathrm{GRO}}(r_0)| > \varepsilon)\right\} \leqslant Q(|r_t - r_0| \geqslant \hat{\delta}) < \eta.$$

Since we also have $\lambda_t^{\mathrm{GREL}}(r_0) \xrightarrow{\mathrm{p}} \lambda^{\mathrm{GRO}}(r_0)$ as $t \to \infty$, it is clear that $\lambda_t^{\mathrm{GREL}}(r_t) \xrightarrow{\mathrm{p}} \lambda^{\mathrm{GRO}}(r_0)$ and $\lambda^{\mathrm{GRO}}(r_t) \xrightarrow{\mathrm{p}} \lambda^{\mathrm{GRO}}(r_0)$ as $t \to \infty$. By boundedness of the betting processes, we have $\lambda_t^{\mathrm{GREL}}(r_t) \xrightarrow{L^1} \lambda^{\mathrm{GRO}}(r_0)$ and $\lambda^{\mathrm{GRO}}(r_t) \xrightarrow{L^1} \lambda^{\mathrm{GRO}}(r_0)$ as $t \to \infty$. The result thus holds by (EC.1).

For (iii), write $\boldsymbol{\lambda}^{\mathrm{GREE}} = (\lambda_t^{\mathrm{GREE}})_{t\in\mathbb{N}}$ and $\boldsymbol{\lambda}^{\mathrm{GREM}} = (\lambda_t^{\mathrm{GREM}})_{t\in\mathbb{N}}$. It suffices to notice that $M_T(\boldsymbol{\lambda}^{\mathrm{GREM}}) \geqslant \max\{M_T(\boldsymbol{\lambda}^{\mathrm{GREL}}), M_T(\boldsymbol{\lambda}^{\mathrm{GREE}})\}/2$, and by taking a limit as $T \to \infty$ we obtain the asymptotic optimality of $\boldsymbol{\lambda}^{\mathrm{GREM}}$ from that of $\boldsymbol{\lambda}^{\mathrm{GREE}}$ or $\boldsymbol{\lambda}^{\mathrm{GREL}}$.

To show the last statement on consistency, note that under the stated conditions, we can verify via Proposition 2 that in each case,

$$\lim_{T\to\infty} \frac{1}{T} \log(M_T(\boldsymbol{\lambda}^{\mathrm{GRO}})) > 0.$$

This implies that the GRO method has asymptotic full power, and hence the same holds for the other methods, because they were shown above to be asymptotically optimal under the corresponding conditions. □

Proof. Proof of Proposition 3. Let e be a backtest e-statistic for ρ. Write $\mathcal{M}_r(\rho) = \{F \in \mathcal{M} : \rho(F) \leqslant r\}$. Take $F, G \in \mathcal{M}_r(\rho)$ satisfying $\rho(F) = \rho(G) = r$. We have $\int e(x,r)\,\mathrm{d}F(x) \leqslant 1$ and $\int e(x,r)\,\mathrm{d}G(x) \leqslant 1$, and hence $\int e(x,r)\,\mathrm{d}(\lambda F + (1-\lambda)G)(x) \leqslant 1$. Thus, $\rho(\lambda F + (1-\lambda)G) \leqslant r$. Next, for any $F, G \in \mathcal{M}_r(\rho)$, without loss of generality we assume $q := \rho(F) \geqslant \rho(G)$. Take $\bar{G} \in \mathcal{M}$ such that $\bar{G} \geqslant_1 G$ and $\rho(\bar{G}) = q$. From the above analysis, we know that $\lambda F + (1-\lambda)\bar{G} \in \mathcal{M}_q(\rho) \subseteq \mathcal{M}_r(\rho)$. Since $\lambda F + (1-\lambda)G \leqslant_1 \lambda F + (1-\lambda)\bar{G}$, we have $\lambda F + (1-\lambda)G \in \mathcal{M}_r(\rho)$. □

Proof. Proof of Proposition 4. Take $F, G \in \mathcal{M}$, $r \in \mathbb{R}$, $\lambda \in [0,1]$, and write $H_\lambda = \lambda F + (1-\lambda)G$. First, suppose that $\rho(F), \rho(G) \leqslant r$. Since $(x,r) \mapsto e(x,r)$ is decreasing in r, we have $\int e(x,r)\,\mathrm{d}F(x) \leqslant 1$ and $\int e(x,r)\,\mathrm{d}G(x) \leqslant 1$, and hence $\int e(x,r)\,\mathrm{d}H_\lambda(x) \leqslant 1$ for all $\lambda \in [0,1]$. This implies $\rho(H_\lambda) \leqslant r$. Further, suppose that $\rho(F), \rho(G) \geqslant r$. Assume that $\rho(H_\lambda) < r$. Write $q = \rho(H_\lambda)$. There exists $\varepsilon > 0$, such that $q + \varepsilon < r$. Since $(x,r) \mapsto e(x,r)$ is decreasing in r,

$\int e(x, q+\varepsilon)\,\mathrm{d}H_\lambda(x) \leqslant 1$, $\int e(x, q+\varepsilon)\,\mathrm{d}F(x) > 1$, and $\int e(x, q+\varepsilon)\,\mathrm{d}G(x) > 1$. This leads to a contradiction. Therefore, $\rho(H_\lambda) \geqslant r$. Summarizing the above arguments, $\rho(F), \rho(G) \leqslant r$ implies $\rho(H_\lambda) \leqslant r$, and $\rho(F), \rho(G) \geqslant r$ implies $\rho(H_\lambda) \geqslant r$. This gives the quasi-linearity of ρ. □

The proof of Proposition 5 relies on the following lemma.

Lemma EC.2. *If a function $g : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies $\mathbb{E}[g(L)] \leqslant 1$ for all $L \geqslant 0$ with $\mathbb{E}[L] \leqslant r$, then there exists $h \in [0,1]$ such that $g(x) \leqslant (1-h) + hx/r$ for $x \in \mathbb{R}$.*

Proof. Without loss of generality we can assume $r = 1$. First, it is easy to note that $g(y) \leqslant y$ for $y > 1$; indeed, if $g(y) > y$ then taking a random variable X with $\mathbb{P}(X = y) = 1/y$ and 0 otherwise gives $\mathbb{E}[g(X)] > 1$ and breaks the assumption. Moreover, $g(y) \leqslant 1$ for $y \leqslant 1$ is also clear, which in particular implies $g(1) \leqslant 1$.

Suppose for the purpose of contradiction that the statement in the lemma does not hold. This means that for each $\lambda \in [0,1]$, either (a) $g(x) > (1-\lambda)+\lambda x$ for some $x < 1$ or (b) $g(y) > (1-\lambda)+\lambda y$ for some $y > 1$. Since $g(y) \leqslant y$ for $y > 1$ and $g(x) \leqslant 1$ for $x < 1$, we know that $\lambda = 1$ implies (a) and $\lambda = 0$ implies (b).

We claim that there exists $\lambda_0 \in (0,1)$ for which both (a) and (b) happen. To show this claim, let

$$\Lambda_0 = \{\lambda \in [0,1] : g(y) > (1-\lambda) + \lambda y \text{ for some } y > 1\};$$
$$\Lambda_1 = \{\lambda \in [0,1] : g(x) > (1-\lambda) + \lambda x \text{ for some } x < 1\}.$$

Clearly, the above arguments show $\Lambda_0 \cup \Lambda_1 = [0,1]$, $0 \in \Lambda_0$, and $1 \in \Lambda_1$. Moreover, since the function $(1-\lambda)+\lambda x$ is monotone in λ for either $x < 1$ or $x > 1$, we know that both Λ_0 and Λ_1 are intervals. Let $\lambda_* = \sup \Lambda_0$ and $\lambda^* = \inf \Lambda_1$. We will argue $\lambda_* \notin \Lambda_0$ and $\lambda^* \notin \Lambda_1$. If $\lambda_* \in \Lambda_0$, then there exists $y > 1$ such that $g(y) > (1-\lambda_*) + \lambda_* y$. By continuity, there exists $\hat{\lambda}_* > \lambda_*$ such that $g(y) > (1-\hat{\lambda}_*) + \hat{\lambda}_* y$, showing that $\hat{\lambda}_* \in \Lambda_0$, contradicting the definition of λ_*. Therefore, $\lambda_* \notin \Lambda_0$. Similarly, $\lambda^* \notin \Lambda_1$, following the same argument. If $\lambda_* = \lambda^*$, then this point is not contained in $\Lambda_0 \cup \Lambda_1$, a contradiction to $\Lambda_0 \cup \Lambda_1 = [0,1]$. Hence, it must be $\lambda_* > \lambda^*$, which implies that $\Lambda_0 \cap \Lambda_1$ is not empty.

Let $x_0 < 1$ and $y_0 > 1$ be such that

$$g(x_0) > 1 - \lambda_0 + \lambda_0 x_0 \quad \text{and} \quad g(y_0) > 1 - \lambda_0 + \lambda_0 y_0.$$

Let X be distributed as $\mathbb{P}(X = y_0) = (1-x_0)/(y_0 - x_0)$ and $\mathbb{P}(X = x_0) = (y_0 - 1)/(y_0 - x_0)$, which

clearly satisfies $\mathbb{E}[X] = 1$. Moreover,

$$\begin{aligned}\mathbb{E}[g(X)] &= \frac{1-x_0}{y_0-x_0} g(y_0) + \frac{y_0-1}{y_0-x_0} g(x_0) \\ &> \frac{1-x_0}{y_0-x_0} (1-\lambda_0+\lambda_0 y_0) + \frac{y_0-1}{y_0-x_0} (1-\lambda_0+\lambda_0 x_0) = 1.\end{aligned}$$

This yields a contradiction. □

Proof. Proof of Proposition 5. The first part of the proposition is immediate from Lemma EC.2. Example 1 implies that e' is a one-sided e-statistic for the mean, and a backtest e-statistic if $h > 0$. Assume now that e' is a backtest e-statistic for the mean. Considering any L with $\mathbb{E}L = r' > r$, we find that $h > 0$. The condition that e' is monotone implies the remaining conditions. □

The proof of Proposition 6 relies on the following lemma.

Lemma EC.3. *If a function $g : \mathbb{R} \to \mathbb{R}_+$ satisfies $\mathbb{E}[g(L)] \leqslant 1$ for all L with $\mathbb{E}[L] = 0$, $\mathbb{E}[L^2] \leqslant r$, then there exists $h \in [0,1]$ such that $g(x) \leqslant (1-h) + hx^2/r$ for $x \in \mathbb{R}$.*

Proof. The proof is similar to that of Lemma EC.2. Without loss of generality we assume $r = 1$. First, it is easy to note that $g(y) \leqslant y^2$ for $y > 1$; indeed, if $g(y) > y^2$ then taking a random variable X with $\mathbb{P}(X = y) = \mathbb{P}(X = -y) = y^{-2}/2$ and 0 otherwise gives $\mathbb{E}[g(X)] > 1$ and breaks the assumption. Moreover, $g(y) \leqslant 1$ for $y \leqslant 1$ is also clear, which in particular implies $g(1) \leqslant 1$.

Suppose for the purpose of contradiction that the statement in the lemma does not hold. This means that for each $\lambda \in [0,1]$, either (a) $g(x) > (1-\lambda)+\lambda x^2$ for some $x < 1$ or (b) $g(y) > (1-\lambda)+\lambda y^2$ for some $y > 1$. Since $g(y) \leqslant y^2$ for $y > 1$ and $g(x) \leqslant 1$ for $x < 1$, we know that $\lambda = 1$ implies (a) and $\lambda = 0$ implies (b).

We see that there exists $\lambda_0 \in (0,1)$ for which both (a) and (b) happen; this was shown in the proof of Lemma EC.2.

Let $x_0 < 1$ and $y_0 > 1$ be such that

$$g(x_0) > 1 - \lambda_0 + \lambda_0 x_0^2 \quad \text{and} \quad g(y_0) > 1 - \lambda_0 + \lambda_0 y_0^2.$$

Let B be a Bernoulli random variable with mean $1/2$. Let Y be independent of B and such that $\mathbb{P}(Y = y_0) = (1-x_0^2)/(y_0^2 - x_0^2)$ and $\mathbb{P}(Y = x_0) = (y_0^2 - 1)/(y_0^2 - x_0^2)$. It is clear that $\mathbb{E}[Y^2] = 1$.

Let $X = BY$, which clearly satisfies $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$. Moreover,

$$\begin{aligned}\mathbb{E}[g(X)] &= \frac{1-x_0^2}{y_0^2-x_0^2}g(y_0) + \frac{y_0^2-1}{y_0^2-x_0^2}g(x_0)\\ &> \frac{1-x_0^2}{y_0^2-x_0^2}\left(1-\lambda_0+\lambda_0 y_0^2\right) + \frac{y_0^2-1}{y_0^2-x_0^2}\left(1-\lambda_0+\lambda_0 x_0^2\right) = 1.\end{aligned}$$

This yields a contradiction. □

Proof. Proof of Proposition 6. The first part of the proposition follows from Lemma EC.3. Example 2 implies that e' is a one-sided e-statistic for $(\mathrm{var}, \mathbb{E})$, and it is a backtest e-statistic if $h > 0$. Suppose now that e' is a backtest e-statistic. Taking a random variable L with finite variance $\mathrm{var}(L) > r$, it follows that $h > 0$. The final statement is directly obtained with the same argument as in the proof of Proposition 5.

□

For the proof of Theorem 4, we show the following lemma.

Lemma EC.4. *Fix $r \in \mathbb{R}$ and $p \in (0,1)$. If a function $g : \mathbb{R} \to \mathbb{R}_+$ satisfies $\mathbb{E}[g(L)] \leqslant 1$ for all L with $\mathrm{VaR}_p(L) \leqslant r$, then there exists $h \in [0,1]$ such that*

$$g(x) \leqslant 1 + h\,\frac{p - \mathbb{1}_{\{x \leqslant r\}}}{1-p}, \qquad \text{for } x \in \mathbb{R}.$$

Proof. Fix $r \in \mathbb{R}$ and consider the hypothesis

$$H_0 : \mathrm{VaR}_p(L) \leqslant r.$$

The distribution δ_x for $x \leqslant r$ is in H_0. Therefore, $g(x) \leqslant 1$ for $x \leqslant r$.

For any $q \geqslant p$ and $\varepsilon > 0$, $\varepsilon' \geqslant 0$, the distribution $q\delta_{r-\varepsilon'} + (1-q)\delta_{r+\varepsilon}$ is in H_0. Therefore, $qY(r-\varepsilon') + (1-q)Y(r+\varepsilon) \leqslant 1$. Solving for $g(r+\varepsilon)$ and taking the minimum over q, we find that

$$g(r+\varepsilon) \leqslant \frac{1 - g(r-\varepsilon')p}{1-p},$$

which implies

$$g(r+\varepsilon) \leqslant \frac{1 - \sup_{x \leqslant r} g(x)p}{1-p}.$$

Rearranging this inequality, we obtain $p \sup_{x \leqslant r} g(x) \leqslant 1 - (1-p)g(r+\varepsilon)$, hence

$$p \sup_{x \leqslant r} g(x) + (1-p) \sup_{x > r} g(x) \leqslant 1.$$

Denote by $h = 1 - \sup_{x \leqslant r} g(x) \in [0,1]$ and $b = \sup_{x>r} g(x)$. Clearly, $g \leqslant f$, where

$$f(x) = (1-h)\mathbb{1}_{\{x \leqslant r\}} + b\mathbb{1}_{\{x>r\}}, \quad x \in \mathbb{R}.$$

Taking a supremum over two-point distributions, it is easy to see that $\mathbb{E}[f(L)] \leqslant 1$ from $\mathbb{E}[g(L)] \leqslant 1$. Take any random variable L with $\mathbb{P}(L > r) = 1 - p$. Clearly L satisfies H_0. It follows that

$$(1-h)p + b(1-p) = \mathbb{E}[f(L)] \leqslant 1,$$

leading to

$$b \leqslant \frac{1-p+hp}{1-p} = 1 - h + \frac{h}{1-p}.$$

Therefore,

$$g(x) \leqslant f(x) \leqslant 1 - h + \frac{h}{1-p}\mathbb{1}_{\{x>r\}}, \quad x \in \mathbb{R},$$

as desired. □

Proof. Proof of Theorem 4. The first part of the theorem follows from Lemma EC.4. Example 4 shows that e' is a one-sided e-statistic for VaR_p. In order to obtain that e' is a backtest e-statistic, it is necessary and sufficient that $h > 0$. Fix $x \in \mathbb{R}$. For $r \geqslant x$, the function $r \mapsto e'(x,r)$ is decreasing if and only if h is increasing; for $r < x$, the same function is decreasing if and only if h is decreasing. Since these considerations hold for any $x \in \mathbb{R}$, h has to be constant if e' is monotone.

□

We use the following lemma for the proof of Theorem 5.

Lemma EC.5. *Fix (r,z) with $r \geqslant z$ and $p \in (0,1)$. If an increasing function $g : \mathbb{R} \to \mathbb{R}_+$ satisfies $\mathbb{E}[g(L)] \leqslant 1$ for all L with $\mathrm{ES}_p(L) \leqslant r$ and $\mathrm{VaR}_p(L) = z$, then there exist $h, k \in [0,1]$ with $h + k \leqslant 1$ such that*

$$g(x) \leqslant (1-h-k) + h\frac{(x-z)_+}{(1-p)(r-z)} + k\frac{\mathbb{1}_{\{x>z\}}}{1-p}, \quad \text{for } x \in \mathbb{R}. \tag{EC.2}$$

Proof. Without loss of generality, assume $z = 0$. First note that the condition on g implies $g(0) \leqslant 1$. Write $\theta = g(0) \in [0,1]$. Consider

$$H_0 : \mathrm{ES}_p(L) \leqslant r \text{ and } \mathrm{VaR}_p(L) = 0.$$

Let X be any e-variable independent of a Bernoulli random variable B with mean p. Let $L = rX(1-B)$. It is clear that $\mathrm{VaR}_p(L) = 0$ and $\mathrm{ES}_p(L) = r\mathbb{E}[X] \leqslant r$. Hence, L satisfies H_0. It

follows that $\mathbb{E}[g(L)] \leqslant 1$, leading to

$$p\theta + (1-p)\mathbb{E}[g(rX)] \leqslant 1,$$

and thus

$$\mathbb{E}\left[\frac{1-p}{1-p\theta}g(rX)\right] \leqslant 1.$$

Since X is an arbitrary e-variable, by Lemma EC.2, there exists $\lambda \in [0,1]$ such that

$$\frac{1-p}{1-p\theta}g(rx) \leqslant 1-\lambda+\lambda x \quad \text{for all } x \geqslant 0.$$

Equivalently, we can write

$$g(x) \leqslant \frac{1-p\theta}{1-p}\left(1-\lambda+\lambda\frac{x}{r}\right) \quad \text{for all } x \geqslant 0.$$

This implies

$$\theta = g(0) \leqslant \frac{1-p\theta}{1-p}(1-\lambda),$$

and thus

$$(1-\lambda)(1-p\theta) - (1-p)\theta = 1-\lambda-\theta+p\theta\lambda \geqslant 0.$$

Also, monotonicity of g implies $g(x) \leqslant g(0) = \theta$ for $x \leqslant 0$. Putting these together, we get

$$\begin{aligned} g(x) &\leqslant \theta \mathbb{1}_{\{x\leqslant 0\}} + \frac{1-p\theta}{1-p}\left(1-\lambda+\lambda\frac{x}{r}\right)\mathbb{1}_{\{x>0\}} \\ &= \theta + \left(\frac{1-p\theta}{1-p}(1-\lambda)-\theta\right)\mathbb{1}_{\{x>0\}} + \frac{1-p\theta}{1-p}\lambda\frac{x_+}{r} \\ &= \theta + (1-\lambda-\theta+p\theta\lambda)\frac{\mathbb{1}_{\{x>0\}}}{1-p} + (1-p\theta)\lambda\frac{x_+}{(1-p)r}. \end{aligned}$$

Noting that $\theta + (1-\lambda-\theta+p\theta\lambda) + (1-p\theta)\lambda = 1$, and each of the three terms is in $[0,1]$, we obtain that (EC.2) holds. □

Proof. Proof of Theorem 5. The first part of the theorem follows directly from Lemma EC.5. Suppose that e' is a backtest e-statistic for $(\mathrm{ES}_p, \mathrm{VaR}_p)$. For all $r \in \mathbb{R}$, $z \leqslant r$, $\varepsilon > 0$, and for some $q \in (p,1]$, consider a random variable

$$X = ((r-z)(1-p)/(1-q)+\varepsilon)\,\mathbb{1}_A + z,$$

where $\mathbb{P}(A) = 1-q$. It follows that $\mathbb{E}[(X-z)_+] = (1-p)(r-z)+\varepsilon(1-q)$ and $\mathrm{ES}_p(X) = r+\varepsilon(1-q)/(1-p) > r$, since $\mathrm{VaR}_p(X) = z$. Then,

$$\begin{aligned} 1 < \mathbb{E}[e'(x,r,z)] &= 1 + h(r,z)\left(\frac{\mathbb{E}[(X-z)_+]}{(1-p)(r-z)} - 1\right) + k(r,z)\frac{p-\mathbb{P}(X \leqslant z)}{1-p} \\ &= 1 + h(r,z)\frac{(1-q)\varepsilon}{(1-p)(r-z)} + k(r,z)\frac{p-q}{1-p}. \end{aligned}$$

Arbitrariness of ε implies that $k = 0$ and $h > 0$. The final statement is obtained by the same argument as in the proof of Proposition 5. □

D Supplementary simulation and data analysis

D.1 Forecasting procedure for stationary time series data

This section describes the details of the forecasting procedure for VaR and ES in Section 7.1. We assume that the data generated above follow an AR(1)–GARCH(1, 1) process $\{L_t\}_{t\in\mathbb{N}}$ with $L_t = \mu_t + \sigma_t Z_t$, where $\{Z_t\}_{t\in\mathbb{N}}$ is assumed to be a sequence of iid innovations with mean 0 and variance 1, and $\{\mu_t\}_{t\in\mathbb{N}}$ and $\{\sigma_t\}_{t\in\mathbb{N}}$ are $\mathcal{F}_{t-1}$-measurable. Specifically, we have

$$\mu_t = c + \psi L_{t-1} \quad \text{and} \quad \sigma_t^2 = \alpha_0 + \alpha_1\sigma_{t-1}^2 Z_{t-1}^2 + \beta\sigma_{t-1}^2, \quad t \in \mathbb{N}. \tag{EC.3}$$

We first assume the innovations Z_t to follow a normal, t-, or skewed-t distribution, and estimate $\{\hat{\mu}_t\}_{t\in\mathbb{N}}$ and $\{\hat{\sigma}_t^2\}_{t\in\mathbb{N}}$ through obtaining the maximum likelihood estimators of $(c, \psi, \alpha_0, \alpha_1, \beta)$ under the assumption on the distribution of Z_t. For t- and skewed-t distributions, parameters are estimated by the maximum likelihood method via the standardized residuals $\{(L_t - \hat{\mu}_t)/\hat{\sigma}_t\}_{t\in\mathbb{N}}$. For a risk measure ρ ($\rho = \mathrm{VaR}_p$ or $\rho = \mathrm{ES}_p$), the value of $\rho(Z_t)$ can be calculated explicitly for the assumed parametric models (see e.g., McNeil et al. (2015), Nolde and Ziegel (2017) and Patton et al. (2019)). For estimation, the estimated parameters are plugged into these formulas resulting in estimates $\widehat{\rho(Z_t)}$. The final risk predictions are then $\hat{\mu}_t + \hat{\sigma}_t\widehat{\rho(Z_t)}$, where $\hat{\mu}_t$ and $\hat{\sigma}_t$ are computed from (EC.3) with the estimated parameters. Table EC.1 shows the average of the forecasts of VaR and ES at different levels over all 1,000 trials and all trading days, where the last line shows the average forecasts of VaR and ES using the true information of the data generating process.

Table EC.1: Average point forecasts of VaR and ES at different levels over 1,000 simulations of time series and 500 trading days; values in boldface are underestimated by at least 10% compared with values in the last line

	$\overline{\text{VaR}}_{0.95}$	$\overline{\text{VaR}}_{0.99}$	$\overline{\text{VaR}}_{0.875}$	$\overline{\text{ES}}_{0.875}$	$\overline{\text{VaR}}_{0.975}$	$\overline{\text{ES}}_{0.975}$
normal	0.619	**0.906**	0.411	**0.620**	**0.752**	**0.910**
t	**0.534**	**0.999**	**0.300**	**0.576**	**0.722**	**1.065**
skewed-t	0.676	1.281	0.369	0.727	0.922	1.358
true	0.674	1.271	0.368	0.723	0.918	1.343

D.2 Comparing GREE and GREL methods for stationary time series

This section serves as a supplement to Section 7.1 by demonstrating the results of backtesting VaR and ES using the GREE and GREL methods through Taylor approximation in (15). Meanwhile, we compare the performance of the GREE and GREL methods. The results of VaR are shown in Tables EC.2 and EC.3 and those for ES are shown in Tables EC.4 and EC.5. The GREL method is better than the GREE method in terms of percentage of detections in all cases of VaR and ES. This is consistent with the result in Theorem 3 because for the time series data, the losses used by the GREL method are relatively closer to an iid pattern compared to the whole e-statistics used by the GREE method. This is also not a contradiction to the slightly longer expected time to detection conditional on the detection of the GREL method, noting that the GREL method detects more often.

Table EC.2: Percentage of detections (%) for $\text{VaR}_{0.99}$ forecasts over 1,000 simulations of time series and 500 trading days

	GREE								
	normal			t			skewed-t		
threshold	2	5	10	2	5	10	2	5	10
−10%	99.0	96.5	92.1	95.5	84.2	72.4	38.3	13.7	5.6
exact	95.0	84.4	72.0	79.5	55.4	37.2	14.0	2.8	0.8
+10%	80.5	56.8	38.2	52.3	22.0	9.8	5.3	0.6	0
	GREL								
	normal			t			skewed-t		
threshold	2	5	10	2	5	10	2	5	10
−10%	99.7	98.3	94.5	98.2	88.7	76.7	51.0	15.1	6.6
exact	97.8	88.6	75.9	87.4	62.4	39.8	24.3	3.2	0.6
+10%	87.7	65.0	43.9	67.5	28.9	13.6	10.8	0.4	0.1

Table EC.3: The average number of days taken to detect evidence against $\text{VaR}_{0.99}$ forecasts conditional on detection over $1,000$ simulations of time series and 500 trading days; numbers in brackets are average final log-transformed e-values

	GREE											
	normal				t				skewed-t			
threshold	2	5	10		2	5	10		2	5	10	
−10%	123	186	228	(5.475)	159	236	278	(3.327)	206	260	300	(0.2856)
exact	164	239	283	(3.236)	197	272	311	(1.638)	189	229	265	(−0.1012)
+10%	197	268	300	(1.734)	217	280	318	(0.5933)	158	224	–	(−0.1706)
	GREL											
	normal				t				skewed-t			
threshold	2	5	10		2	5	10		2	5	10	
−10%	116	185	233	(5.338)	158	241	293	(3.290)	239	281	330	(0.3492)
exact	160	241	295	(3.240)	196	286	332	(1.736)	233	238	289	(−0.1463)
+10%	189	284	330	(1.849)	226	304	358	(0.7599)	230	211	377	(−0.3472)

Table EC.4: Percentage of detections (%) for $\text{ES}_{0.975}$ forecasts over $1,000$ simulations of time series and 500 trading days

	GREE								
	normal			t			skewed-t		
threshold	2	5	10	2	5	10	2	5	10
−10% ES	99.7	98.9	97.9	96.7	85.6	73.8	39.9	14.9	6.1
−10% both	99.8	98.9	97.3	97.9	89.2	79.3	41.1	13.6	6.4
exact	98.6	93.2	86.1	83.0	59.8	41.1	13.5	3.2	1.1
+10% both	91.1	75.7	59.8	54.2	22.3	10.2	5.1	0.8	0
+10% ES	91.0	76.1	60.2	60.6	28.3	13.8	6.0	1.1	0.1
	GREL								
	normal			t			skewed-t		
threshold	2	5	10	2	5	10	2	5	10
−10% ES	99.9	99.2	98.2	97.5	86.2	74.1	49.3	16.6	6.5
−10% both	99.9	99.0	97.6	98.1	89.2	78.8	49.9	17.4	6.6
exact	99.2	94.6	85.7	87.6	62.4	41.9	25.3	5.6	0.9
+10% both	94.1	77.3	56.2	66.0	31.6	12.5	11.5	1.5	0.1
+10% ES	94.4	76.7	57.4	72.7	36.1	16.9	13.1	1.7	0.2

Table EC.5: The average number of days taken to detect evidence against $\mathrm{ES}_{0.975}$ forecasts conditional on detection over 1,000 simulations of time series and 500 trading days; "–" represents no detection; numbers in brackets are average final log-transformed e-values

	GREE											
	normal				t				skewed-t			
threshold	2	5	10		2	5	10		2	5	10	
−10% ES	95	141	181	(6.596)	158	226	274	(3.365)	213	250	234	(0.3991)
−10% both	95	152	194	(6.341)	149	220	269	(3.674)	207	242	224	(0.3594)
exact	139	201	250	(4.278)	193	265	307	(1.822)	208	233	220	(−0.1011)
+10% both	177	249	292	(2.669)	219	267	286	(0.6485)	106	109	–	(−0.1843)
+10% ES	174	248	295	(2.625)	210	266	288	(0.8098)	90	96	156	(−0.2005)
	GREL											
	normal				t				skewed-t			
threshold	2	5	10		2	5	10		2	5	10	
−10% ES	97	147	189	(6.344)	155	231	282	(3.238)	235	264	277	(0.4990)
−10% both	99	154	201	(5.963)	146	221	271	(3.511)	223	263	278	(0.4565)
exact	134	209	258	(4.027)	191	266	318	(1.892)	208	233	220	(−0.09266)
+10% both	174	257	291	(2.577)	217	289	298	(0.8661)	186	207	70	(−0.3171)
+10% ES	173	254	296	(2.557)	215	282	301	(1.007)	189	185	271	(−0.3653)

D.3 Type-I errors for backtesting stationary time series

As shown by Remark 3 and the results in Section 7, the practical type-I errors are usually much smaller than the theoretical bound α in Theorem 2 when the detection threshold is $1/\alpha$ and the betting process is chosen properly. To see the type-I error specified by Theorem 2, we simulate the same time series data as that in Section 7.1 but with different sample sizes up to 10^5 for backtesting. Under the true model, we report risk forecasts with skewed-t innovations and the true $\{\mu_t\}_{t\in\mathbb{N}}$ and $\{\sigma_t\}_{t\in\mathbb{N}}$ as the data generating process. Table EC.6 shows the type-I errors for different detection thresholds and methods of choosing the betting processes. As a comparison, the bottom of Table EC.6 also shows the type-I errors for the traditional backtest in McNeil and Frey (2000) and Nolde and Ziegel (2017) with significance levels 5%, 2%, and 1%, where the hypothesis is rejected if it is rejected at any sample size (this is of course invalid). We can see that the type-I errors increase as the sample size increases. For sample size 10^5 (usually much smaller in practice), the type-I errors for the e-backtesting approaches are smaller than but close to the theoretical bounds indicated by Theorem 2. The type-I error of the traditional backtest explodes for the large sample size 10^5 due to its lack of optional validity, as expected.

Table EC.6: Percentage of detections (%) for true $\mathrm{VaR}_{0.99}$ forecasts over 1,000 simulations of time series and different sample sizes

sample size	10^5			10^4			10^3			500		
threshold $(1/\alpha)$	2	5	10	2	5	10	2	5	10	2	5	10
constant $\lambda_t = 0.01$	38.9	16.7	8.6	38.9	16.7	8.6	37.4	15.1	7.1	36.3	10.9	4.3
GREE	32.5	11.0	5.7	24.1	7.7	3.8	14.9	3.0	0.9	10.2	1.1	0.2
GREL	41.8	16.4	8.7	41.7	16.0	8.2	26.8	5.8	2.0	20.1	3.0	0.4
GREM	38.8	14.1	7.1	34.9	12.0	5.9	20.4	4.0	1.2	15.0	1.7	0.2
Significance level	5%	2%	1%	5%	2%	1%	5%	2%	1%	5%	2%	1%
Traditional	89.7	82.4	73.8	14.6	7.8	4.5	2.2	0.6	0.2	1.7	0.5	0.3

D.4 Forecasting procedure for time series with structural change

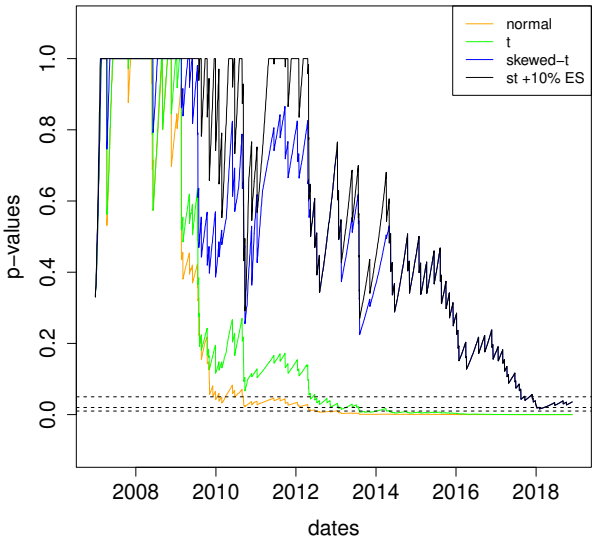
This section provides details for the forecasting procedure for time series data with structural change in Section 7.2. After a burn-in period of length 1,000, 500 data points are simulated, within which 250 presampled data $L_1, \dots, L_{250}$ are for forecasting risk measures and the rest 250 data $L_{251}, \dots, L_{500}$ are for backtesting. The forecaster obtains the estimates $\hat{\boldsymbol{\theta}}$ of the model parameters $\boldsymbol{\theta} = (\omega, \alpha, \beta_t)$ once using the standard Gaussian QML for the presampled 250 losses. For $t \in \{251, \dots, 500\}$, the forecasts of $\mathrm{VaR}_{0.95}(L_t|\mathcal{F}_{t-1})$ and $\mathrm{ES}_{0.95}(L_t|\mathcal{F}_{t-1})$ are obtained by $z_t = \sigma_t(\hat{\boldsymbol{\theta}})\widehat{\mathrm{VaR}}_{0.95}$ and $r_t = \sigma_t(\hat{\boldsymbol{\theta}})\widehat{\mathrm{ES}}_{0.95}$, respectively, where $\widehat{\mathrm{VaR}}_{0.95}$ and $\widehat{\mathrm{ES}}_{0.95}$ are empirical forecasts of VaR and ES using presampled residuals $L_1/\sigma_1(\hat{\boldsymbol{\theta}}), \dots, L_{250}/\sigma_{250}(\hat{\boldsymbol{\theta}})$. We choose the size $m = 50$ of the rolling window and a leg $d = 5$ of autocorrelations.[12]

D.5 P-value trajectories for NASDAQ data analysis

We perform the traditional backtest in McNeil and Frey (2000) and Nolde and Ziegel (2017) for the same NASDAQ index used in Section 8.1 for $(\mathrm{VaR}_{0.975}, \mathrm{ES}_{0.975})$. For an ad hoc comparison to the e-processes in Figure 6, we do the traditional backtest "sequentially" and plot a process of p-values over the trading days in Figure 8. During the 2007-2008 financial crisis, we observe similar downward jumps for all p-value trajectories. However, all forecasts, no matter whether they are conservative or not, are eventually rejected in 2021. The result clearly shows that "sequentially" performing a traditional p-value test violates validity.

[12] The leg $d = 5$ is not necessary for the Monte Carlo simulations detector. We choose it to be consistent with the simulation setting of Hoga and Demetrescu (2023) for comparison.

Figure 8: Left panel: p-value trajectories testing $(\mathrm{VaR}_{0.975}, \mathrm{ES}_{0.975})$ with respect to number of days for the NASDAQ index from Jan 3, 2005 to Dec 31, 2021; right panel: the number of days taken for the traditional p-value test to detect evidence against the $(\mathrm{VaR}_{0.975}, \mathrm{ES}_{0.975})$ forecasts



threshold	0.05	0.02	0.01
normal	756	1337	1381
t	1354	1529	1659
skewed-t	2677	2776	3477
st +10% ES	2677	2776	3477

D.6 Detailed setup of data analysis for optimized portfolios

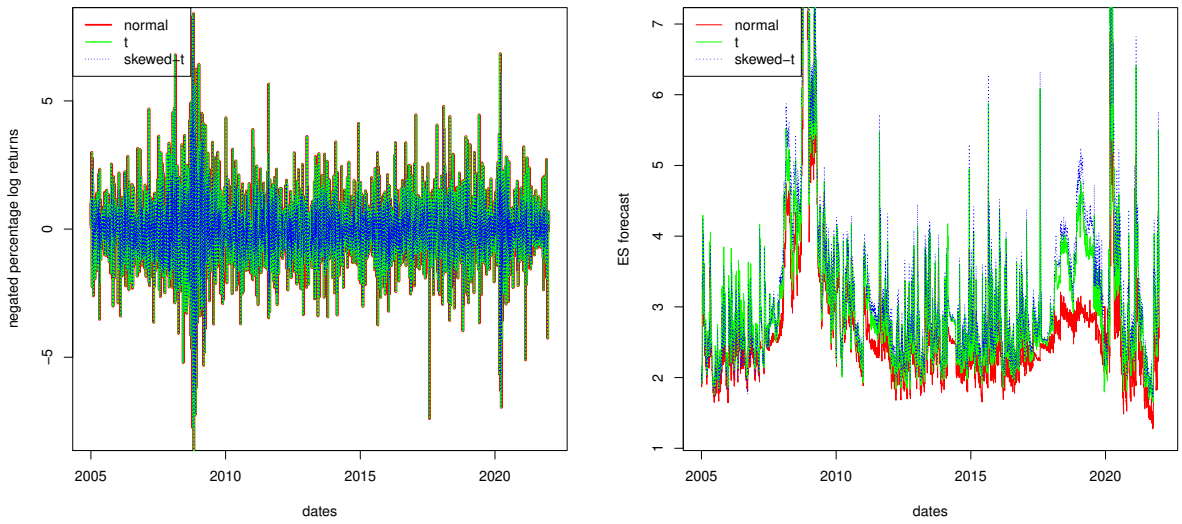
For the data of optimized portfolios we use in Section 8.2, the first 500 data points are used for the initial forecast and another 500 data points are for computing the first value of the betting process of the backtesting procedure. The final sample for backtesting contains 4280 negated percentage log-returns from Jan 3, 2005 to Dec 31, 2021. The selected stocks are those with the largest market caps in the 11 S&P 500 sectors divided by the GICS level 1 index as of Jan 3, 2005. The list of selected stocks is shown in Table EC.7.

Table EC.7: 22 selected stocks in S&P 500 sectors divided by GICS level 1 as of Jan 3, 2005 for the portfolio

Communication Services	**Customer Discretionary**	**Consumer Staples**
Verizon Communications Inc.	Time Warner Inc.	The Procter & Gamble Co.
AT&T Inc.	The Home Depot, Inc.	Walmart Inc.
Energy	**Financials**	**Health Care**
Exxon Mobil Corp.	Citigroup Inc.	Johnson & Johnson
Chevron Corp.	Bank of America Corp.	Pfizer Inc.
Industrials	**Information Technology**	**Materials**
United Parcel Service Inc.	International Business Machines Corp.	EI du Pont de Nemours and Co.
General Electric Co.	Microsoft Corp.	The Dow Chemical Co.
Real Estate	**Utilities**	
Weyerhaeuser Co.	Exelon Corp.	
Simon Property Group Inc.	The Southern Co.	

For forecasting, we assume $L_t^i = \mu_t^i + \sigma_t^i Z_t^i$ with μ_t^i and σ_t^i defined to be the same as (EC.3) for all $i \in [n]$. The innovations $\{Z_t^i\}_{t\in\mathbb{N}}$ are iid with respect to time with mean 0 and variance 1 for $i \in [n]$, assumed to be normal, t-, or skewed-t distributed. If a stock delists from the S&P 500 during the period from Jan 3, 2005 to Dec 31, 2021, it is removed from the portfolio as soon as it delists with all of its weight redistributed to the other stocks in the portfolio. The bank reports the VaR and the ES of the weighted portfolio by assuming $\mathbf{w}_t^\top \mathbf{L}_t$ to be normal, t-, or skewed-t distributed, respectively, with the mean $\sum_{i=1}^n w_t^i \mu_t^i$ and the variance $(w_t^1\sigma_t^1, \dots, w_t^n\sigma_t^n)^\top \Sigma_t (w_t^1\sigma_t^1, \dots, w_t^n\sigma_t^n)$, where Σ_t is the covariance matrix of $(Z_t^1, \dots, Z_t^n)$. The assumption is true when the innovations follow the normal distribution or follow the t-distribution with the same degree of freedom. We use this assumption to approximate the true distribution of the portfolio for t-distributions of different degrees of freedom and skewed-t distributions. The parameters of the t- and skewed-t distributions of the weighted portfolio are estimated by the maximum likelihood method assuming the negated percentage log-return of the portfolio to be the AR(1)-GARCH(1, 1) process with innovations belonging to the same class of distribution. Figure 9 shows the negated log-returns of the portfolio and the forecasts of $\mathrm{ES}_{0.975}$ over time assuming different innovation distributions.

Figure 9: Portfolio data fitted by different distribution from Jan 3, 2005 to Dec 31, 2021; left panel: negated percentage log-returns, right panel: $\mathrm{ES}_{0.975}$ forecasts



E Deliberate over-forecast strategy

E.1 Deliberate over-forecast strategy and a solution to this

Due to the flexibility of our setting that allows for arbitrary reports of the risk forecasts, it is possible that the financial institution tries to"game" the regulator to escape the penalty from a detection. Specifically, the financial institution may first deliberately produce conservative risk forecasts to keep the e-process at a low level due to its multiplicative structure, and then start under-reporting risks to avoid the e-process exceeding the detection threshold quickly. Below we show that this issue can be addressed by our e-backtesting method with a properly calculated betting process.

By Proposition 2 and (15), a positive betting process can only be chosen if the empirical mean of e-values is larger than 1. Thus, when the realized e-values $e(L_t, r_t, z_t)$ are very small (due to over-forecasting), the GREE or GREL method will choose a small betting process (equal or close to 0) during the over-reporting period. This keeps the e-process being almost a constant 1. However, if we calculate the betting process using the empirical distribution of all past data, it may still take many data points for the GREE or GREL method to realize the under-reports after a long period of over-prediction. This problem can be solved by a slight modification of our method. Instead of using all past data points, we use a rolling window (e.g., 250 or 500 days) to determine the betting process in the GREE or GREL method. This guarantees the past over-report data to exit

the window gradually, and thus prevents the GREE or GREL method from being dragged by the over-prediction for too long. Moreover, since the bank is not supposed to know the betting process $\boldsymbol{\lambda}$, the gaming cost for the bank increases to cover later under-prediction using intentional over-prediction. The regulator can actively choose a positive betting process to detect evidence against the null faster if she suspects that the bank under-predicts in a certain period.

Most of the existing backtesting approaches (such as the standard test for VaR by counting the number of breaches) are also faced with similar problems of being gamed by the bank through an intentionally long over-prediction period before the under-reports (leading to fewer breaches than theoretical). Our method provides a better solution to the issue than other methods mostly due to its nature of sequential testing and its strategic choices of the betting processes.

E.2 A simulation of stationary time series with gamed forecasts

We perform a simulation to demonstrate that our method can address this problem efficiently. We simulate the same AR(1)–GARCH(1, 1) process as that in Section 7.1, with a sample size 2,000 and rolling window 500. Assume that the financial institution first produces 1,000 conservative risk forecasts, under AR(1)–GARCH(1, 1), by assuming skewed-t innovations and deliberately over-reporting 10% of the exact forecasts; while it then produces 1,000 aggressive risk forecasts by assuming normal innovations. When calculating the betting process for our e-backtesting method, different from Section 7.1, we use a moving window of size 500 instead of using all historical data.[13]

Table EC.8: Percentage of detections (%), the average number of days (after 1,000 days) taken to detect evidence against risk forecasts, and average final log-transformed e-values for $\mathrm{VaR}_{0.99}$ and $\mathrm{ES}_{0.975}$ over 1,000 simulations of time series and 2000 trading days using the GREM method

	$\mathrm{VaR}_{0.99}$			$\mathrm{ES}_{0.975}$		
threshold	2	5	10	2	5	10
percentage of detections	99.7	98.1	94.8	100	100	99.9
days taken for detections	302	467	541	249	409	488
final log e-value		6.118			7.332	

[13] For the first 500 data points, the moving window sizes are less than 500.

Figure 10: Average log-transformed e-processes testing $\mathrm{VaR}_{0.99}$ (left) and $\mathrm{ES}_{0.975}$ (right) with respect to the number of data points over $1,000$ simulations of time series and 2000 trading days

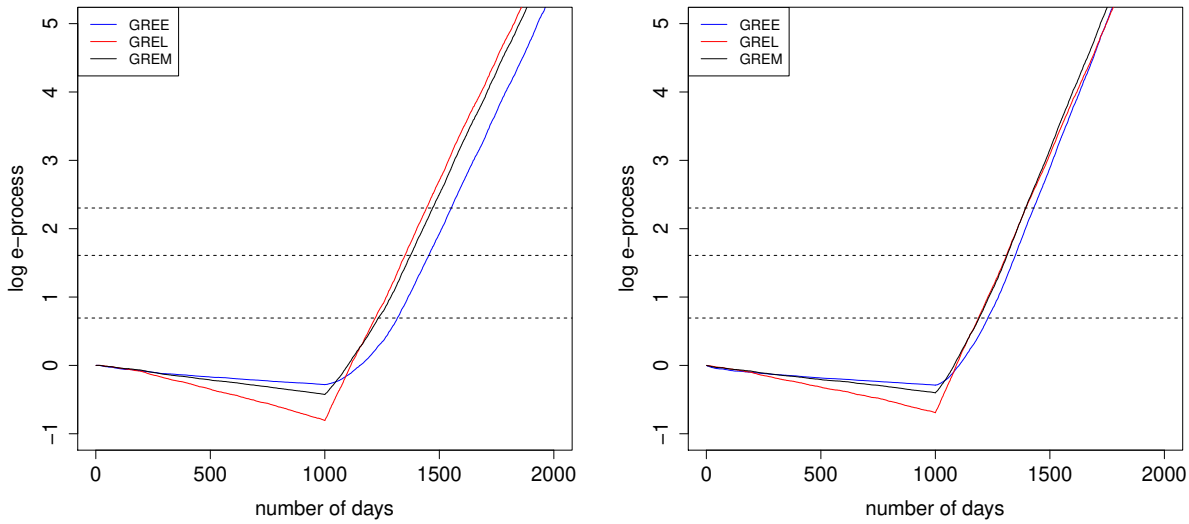


Table EC.8 shows the average result over $1,000$ simulations for the GREM method. After receiving $1,000$ conservative forecasts from the financial institution, the regulator detects evidence against $\mathrm{VaR}_{0.99}$ and $\mathrm{ES}_{0.975}$ forecasts at threshold 2 after 250 to 300 days, which are 100 to 150 days later than those in Tables 3 and 5 but are still timely for the regulator to get alerted. Part of the delay is also caused by the fact that the moving window approach gives less accurate estimates than using all historical data, assuming that the data are stationary. This suggests that the e-process is not seriously undermined by the over-reporting strategy of the financial institution. Figure 10 shows that the log e-process for the GREE method is always close to 0 during the over-reporting period and grows steadily after the under-reports start to come in. On the other hand, the log e-process for the GREL method first drops a little lower to a point above -1, and then increases sharply as soon as the financial institution starts to under-report. The drop in the e-process for the GREL method does happen during the over-report period. However, the GREM method shares the advantages of both GREE and GREL methods, which guarantees that the e-process stays near 1 (the GREE method will automatically choose the betting process near 0) during the over-reporting period and grows quickly during the under-reporting period (the GREL method quickly starts to choose a positive betting process). In practice, we suggest the moving window approach and the GREM method to resolve similar gaming issues, although there may still be some inevitable delays.

References

T. Dimitriadis, T. Fissler, and J. F. Ziegel. Osband's principle for identification functions. *Statistical Papers*, 2023. doi: https://doi.org/10.1007/s00362-023-01428-x.

T. Fissler. *On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space*. PhD thesis, University of Bern, 2017.

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

Y. Hoga and M. Demetrescu. Monitoring Value-at-Risk and Expected Shortfall Forecasts. *Management Science*, 69(5):2954–2971, 2023.

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

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ, revised edition, 2015.

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

A. Patton, J. F. Ziegel and R. Chen. Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk). *Journal of Econometrics*, 211(2):388–413, 2019.

S. Resnick. *A Probability Path*. Springer, Birkhäuser Boston, MA, first edition, 2019.

I. Steinwart, C. Pasin, R. Williamson, and S. Zhang. Elicitation and identification of properties. *JMLR: Workshop and Conference Proceedings*, 35:1–45, 2014.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.