European Journal of Operational Research 310 (2023) 415–431

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

Loss function-based change point detection in risk measures

Emese Lazar[a,*], Shixuan Wang[b], Xiaohan Xue[c]

[a] ICMA Centre, Henley Business School, University of Reading, Reading RG6 6DL, United Kingdom
[b] Department of Economics, University of Reading, Reading RG6 6EL, United Kingdom
[c] Finance Group, Norwich Business School, University of East Anglia, Norwich NR4 7TJ, United Kingdom

ARTICLE INFO

Article history:
Received 20 January 2022
Accepted 25 March 2023
Available online 29 March 2023

Keywords:
Risk analysis
Change point detection
Loss function
Risk measures
Stationary bootstrap

ABSTRACT

We propose a new test to detect change points in financial risk measures, based on the cumulative sum (CUSUM) procedure applied to the Wilcoxon statistic within a popular class of loss functions for risk measures. The proposed test efficiently captures change points jointly in two risk measure series: Value-at-Risk (VaR) and Expected Shortfall (ES), estimated by (semi)parametric models. We derive the asymptotic distribution of the proposed statistic and adopt a stationary bootstrapping technique to obtain the *p*-values of the test statistic. Monte Carlo simulation results show that our proposed test has better size control and higher power than the alternative tests under various change point scenarios. An empirical study of risk measures based on the S&P 500 index illustrates that our proposed test is able to detect change points that are consistent with well-known market events.

1. Introduction

Measuring market risk plays a central role not only in the area of risk management but also in the broader context of financial markets. Value-at-Risk (VaR) and Expected Shortfall (ES) are two prevalent risk measures dominating in contemporary financial regulation (Leung et al., 2021). VaR provides financial institutions with a loss level that occurs in the worst situations at a given confidence level; ES, as an alternative to VaR, is the expectation of losses, conditional on their realization lying below VaR. As for the estimation of these two measures, Engle & Manganelli (2004) distinguish three main categories: nonparametric, parametric, and semiparametric approaches. In a univariate framework, some of the models for financial risk measures include Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) family models (Bali & Theodossiou, 2007), score-driven models (Patton et al., 2019), and CAViaR-ES models (Taylor, 2019).[1]

It is worth mentioning that the presence of change points in time series may cause misleading statistical inference under the assumption of stationarity (Clements & Hendry, 1996; Diebold & Inoue, 2001; Loschi et al., 2007; Mikosch & Stărică, 2004; Stock & Watson, 1996). Related empirical evidence has been extensively documented, especially in stock returns (Pástor & Stambaugh, 2001), volatility (Inclan & Tiao, 1994), correlation dynamics (Barassi et al., 2020), and macroeconomic time series (Pesaran & Timmermann, 2007). There is a vast literature on tests for change points in time-series; some of these detect changes in a historical dataset (Aue et al., 2009; Csörgő & Horváth, 1997), whereas others monitor changes in a sequential manner (Berkes et al., 2004; Dette & Gösmann, 2020; Horváth et al., 2020a). Also, these tests can differ in terms of their objective function given by, e.g., the likelihood for volatility models (Chen & Hong, 2016) and copula models (Ye et al., 2012) or the loss function for quantile regressions (Qu, 2008). We refer the readers to Aue & Horváth (2013) for a detailed literature review.

In addition to the theoretical contribution, a large section of the literature concerns the practical application of change points detection, particularly in economics and finance. Regarding applications in macroeconomics, Andersson et al. (2006) discuss a large number of studies on the detection of turning points in business cycles. Horváth et al. (2020a) propose a sequential monitoring scheme and apply it to detect the changes in real US GDP series. A noteworthy application in the housing markets is provided by Horváth et al. (2022). In the finance literature, several studies investigate the financial contagion between markets during the subprime crisis by implementing a change point analysis in tail risk, e.g., Ye et al. (2012). A sequential monitoring procedure is employed by Ji et al. (2020) to detect left-quantile changes in asset returns in order to

* Corresponding author.
E-mail addresses: e.lazar@icmacentre.ac.uk (E. Lazar), shixuan.wang@reading.ac.uk (S. Wang), x.xue@pgr.reading.ac.uk (X. Xue).

[1] This class of semiparametric models has been extended to incorporate the intraday or high-frequency information (Gerlach & Wang, 2020; Lazar & Xue, 2020; Meng & Taylor, 2020) and combine with networks (Bonaccolto et al., 2022).

https://doi.org/10.1016/j.ejor.2023.03.033

search for safe-haven assets during the COVID-19 pandemic. By detecting change points in the distributions of the daily returns of the constituent stocks of the S&P 500 index, Horváth et al. (2021) associate the detected change points with well known financial events, e.g., the subprime mortgage crisis and COVID-19 pandemic.

In applications of risk management, the existence of change points can cause estimation errors for VaR and ES, as argued in Hoga (2017, hereafter Hoga) and Fan et al. (2018, hereafter FGP). These papers use an innovative self-normalized estimator *à la* Zhang & Lavitas (2018) to detect change points when the risk measures are estimated in a nonparametric way. Specifically, Hoga investigates change points in the VaR process, and FGP consider changes in the ES process. Since regulatory capital requirements in Basel Committee on Banking Supervision (2019) are linked to ES estimates, it would be prudent to detect change points in this process. Also, if change points are detected in the ES series alone, then the effect of VaR on ES is ignored. Since ES is elicitable only jointly with VaR,[2] it is meaningful to detect change points in the (VaR, ES) tuple.

To fill this gap, our study extends the current literature by proposing a test to detect change points in the VaR and ES series simultaneously, which are estimated by (semi)parametric models. We construct this test using the FZ loss functions proposed by Fissler & Ziegel (2016). Since the FZ loss functions are minimized for the true values of VaR and ES, changes in the parameter values of the model cause breaks in the process of the VaR and ES estimates, which will result in changes in the loss series. Our framework of detecting change points in the VaR and ES series based on loss values is general and can accommodate for any type of (semi)parametric estimation models.

Our first contribution is to propose a test to detect change points in both VaR and ES risk measures simultaneously, based on the FZ loss functions. The general framework is closely related to the likelihood ratio test to detect changes in volatility, and the test for structural changes in quantile regressions proposed by Qu (2008). Due to the dominance of the indicator term in the FZ loss functions, the presence of extreme values (spikes), when returns exceed VaR, is one of the main characteristics of the loss series. However, the commonly used cumulative sum (CUSUM) test can be affected by the presence of outliers (Fearnhead & Rigaill, 2019). To address this problem, we adopt a more suitable alternative, namely the Wilcoxon test (Dehling et al., 2013) to detect change points in the loss process.[3] We call this procedure the loss-based Wilcoxon test, and we shed light on its advantages in detecting joint change points in time series of VaR and ES simultaneously.

Secondly, our paper contributes to the current literature by deriving the asymptotic behavior of our test statistic under weak dependence. Also, to improve the finite sample performance of the proposed test, we adopt a stationary bootstrap method based on Politis & Romano (1994), which follows the strand of literature in the area (Hušková & Kirch, 2008; Quaedvlieg, 2021). Furthermore, we prove that the stationary bootstrap is valid for the loss-based Wilcoxon test.

Thirdly, we use Monte Carlo (MC) simulations to show that our proposed test has good size control and higher power in finite samples than the alternatives. Additionally, our study on risk measures of the S&P 500 index returns provides an empirical application to demonstrate the practical usage of our proposed test. We present evidence that the loss-based Wilcoxon test can detect change points that are consistent with well-known market events.

[2] There is no (strictly) consistent loss function for ES that does not also contain VaR (Fissler & Ziegel, 2016).

[3] In order to construct the Wilcoxon test statistic, we initially obtain the ranks of loss values and then feed the ranks into the CUSUM procedure. More details can be found in Section 2.2.

The paper is structured as follows: Section 2 briefly discusses the established methods for change point detection in risk measures, introduces the FZ family of loss functions and the Wilcoxon test statistic, and presents some theoretical results related to the asymptotic distribution and the validity of bootstrapping; Section 3 shows the simulation results; Section 4 contains an empirical application based on the S&P 500 index; Section 5 concludes the paper.

2. Test statistic for change point detection

This section briefly presents the literature on existing tests for change point detection in risk measures. To enrich the literature, we propose a test to detect change points in (VaR, ES) risk measures simultaneously based on the FZ loss functions introduced by Fissler & Ziegel (2016). In this section, we formulate the test problem and derive the asymptotic theorem for our test statistic. For finite samples, we apply the stationary bootstrap method to obtain the *p*-values of our test. Also, we verify the validity of the bootstrap method.

2.1. Established methods for change point detection in risk measures

Detecting change points in risk measures has attracted recent attention from academia. Here we summarize some of the most related studies. Hoga (2017) proposes a test to retrospectively detect change points in extreme quantiles of time series. The main intuition behind the proposed test is to compare the extreme quantiles, estimated by a nonparametric method, namely Weissman's estimator, over different time periods. It is a difficult task to choose a bandwidth parameter for the long-run variance estimator, and sometimes a data-dependent bandwidth can lead to non-monotonic power of the test (Shao & Zhang, 2010; Vogelsang, 1999; Zhang & Lavitas, 2018). In order to confront the challenge of the estimation of asymptotic long-run variance, Hoga's test is built on the self-normalization framework proposed by Shao & Zhang (2010). In a subsequent study, Fan et al. (2018) develop a change point test for nonparametric ES estimates for weakly dependent time series. Their proposed test is based on monitoring changes in the tail structure and uses self-normalization.

The tests discussed above are designed for detecting changes in unconditional risk measures (which are estimated over different time periods). On the other hand, our proposed tests are for conditional risk estimates obtained *via* (semi)parametric risk models. To run the test, we first calculate the loss series associated with the estimated (semi)parametric risk measures. Our test is inspired by the Wilcoxon test proposed by Dehling et al. (2013), which now is based on the loss series. This test is general and can accommodate for any type of (semi)parametric joint estimation methods for VaR and ES. Another advantage of our proposed test is that it can identify changes in VaR and ES jointly, instead of detecting changes in single risk measure, which is in line with the requirements of the Basel Committee on Banking Supervision (2019).

An alternative test that we can compare our proposed test with is based on Rényi-type statistics. Traditional CUSUM-based tests can be ineffective when changes occur near the start or the end of a sequence of observations; this has been addressed by Horváth et al. (2020b), who propose change point detection methods which rely on weighting or trimming schemes based on Rényi-type statistics. We consider a variant of our test that uses Rényi-type statistics, which we call the "Rényi-type test". This is a joint test for changes in VaR and ES risk estimates, and is based on the FZ loss function. In this test, instead of using the Wilcoxon statistic, we calculate the test statistic relying on a trimmed sample, i.e., after removing subsamples at the start and at the end of the entire sample, leading to high power for the change point detection near the

Table 1
Loss functions in the FZ family with different degrees of positive homogeneity b.

b	FZ loss function
0	$L^{FZ0}(r,v,e,\boldsymbol{\theta};\alpha) = -\frac{1}{\alpha e(\boldsymbol{\theta})}\mathbf{1}[r \le v(\boldsymbol{\theta})]\cdot[v(\boldsymbol{\theta}) - r] + \frac{v(\boldsymbol{\theta})}{e(\boldsymbol{\theta})} + \log(-e(\boldsymbol{\theta})) - 1$
-1	$L^{FZ1}(r,v,e,\boldsymbol{\theta};\alpha) = \frac{1}{e(\boldsymbol{\theta})^2}\left\{\frac{1}{\alpha}\mathbf{1}[r \le v(\boldsymbol{\theta})]\cdot[v(\boldsymbol{\theta}) - r] - [v(\boldsymbol{\theta}) - e(\boldsymbol{\theta})]\right\} + \frac{1}{e(\boldsymbol{\theta})}$
0.5	$L^{FZ2}(r,v,e,\boldsymbol{\theta};\alpha) = \frac{1}{2\sqrt{-e(\boldsymbol{\theta})}}\left\{\frac{1}{\alpha}\mathbf{1}[r \le v(\boldsymbol{\theta})]\cdot[v(\boldsymbol{\theta}) - r] - [v(\boldsymbol{\theta}) - e(\boldsymbol{\theta})]\right\} + \sqrt{-e(\boldsymbol{\theta})}$

beginning and end of the sample. More details can be found in Section 3.2.

As benchmarks, we consider two further tests, specifically the self-normalized CUSUM test for VaR and the self-normalized CUSUM test for ES. These are based on the self-normalized change point test proposed by Shao & Zhang (2010) and use conditional risk measures estimated by (semi)parametric risk models. These tests are described in more detail in Section 3.2.

2.2. Loss functions

Let $\{r_t\}_{t=1}^T$ be a series of observed returns measured over an arbitrary frequency, such as daily. (Semi)parametric models can be used to estimate the corresponding conditional risk measures, VaR and ES, denoted by $\{v_t(\boldsymbol{\theta})\}_{t=1}^T$ and $\{e_t(\boldsymbol{\theta})\}_{t=1}^T$, at a specified significance level α, where $\boldsymbol{\theta}$ denotes the parameter vector of the model. Some major results about the consistency of model parameter estimators for selected (semi)parametric models can be found in Francq & Zakoïan (2015) and Patton et al. (2019). We summarize these results in the Supplemental Appendix.

Fissler & Ziegel (2016) introduce the FZ family of loss functions stated below, used to evaluate the (VaR, ES) tuple of risk measures:

$$\begin{aligned} &L^{FZ}(r_t, v_t, e_t, \boldsymbol{\theta}; \alpha) \\ &= \left\{\mathbf{1}[r_t \le v_t(\boldsymbol{\theta})] - \alpha\right\}\left[G_1(v_t(\boldsymbol{\theta})) - G_1(r_t) + \frac{1}{\alpha}v_t(\boldsymbol{\theta})G_2(e_t(\boldsymbol{\theta}))\right] \\ &-G_2(e_t(\boldsymbol{\theta}))\left\{\frac{1}{\alpha}\mathbf{1}[r_t \le v_t(\boldsymbol{\theta})]r_t - e_t(\boldsymbol{\theta})\right\} - \mathcal{G}_2(e_t(\boldsymbol{\theta})), \end{aligned} \tag{1}$$

where G_1 is weakly increasing, G_2 is strictly increasing and strictly positive, and $\mathcal{G}'_2 = G_2$.[4]

For the specification function G_1 in (1), we use $G_1(z) = 0$, which follows the reasoning of Nolde & Ziegel (2017). We consider the second specification function $\mathcal{G}_2$ with different degrees of positive homogeneity[5] $b = -1$, $b = 0$, and $b = 0.5$, which follow the choices of Dimitriadis & Bayer (2019), specified as: $\mathcal{G}_2(z) = -\frac{1}{z}$, $\mathcal{G}_2(z) = -\log(-z)$, and $\mathcal{G}_2(z) = -\sqrt{-z}$, respectively, where z must be negative.

In our study, we use the three loss functions corresponding to the above specifications, detailed in Table 1, to compute the time series of loss values.

To provide some intuition, L^{FZ0} can be reformulated as:

$$\begin{aligned} &L^{FZ0}(r, v, e, \boldsymbol{\theta}; \alpha) \\ &= \begin{cases} -\frac{1}{\alpha e(\boldsymbol{\theta})}[v(\boldsymbol{\theta}) - r] + \frac{v(\boldsymbol{\theta})}{e(\boldsymbol{\theta})} + \log(-e(\boldsymbol{\theta})) - 1, & \text{if } r \le v(\boldsymbol{\theta}), \\ \frac{v(\boldsymbol{\theta})}{e(\boldsymbol{\theta})} + \log(-e(\boldsymbol{\theta})) - 1, & \text{if } r > v(\boldsymbol{\theta}). \end{cases} \end{aligned}$$

The probability of the first outcome is α, and the probability of the second one is $1-\alpha$. Thus, the distribution of the loss value can be generally considered as a mixing distribution with mixing parameter α.

To get a better understanding of the time series properties of the risk measures and loss series, we test, using simulations based on a GARCH(1,1)-skewed t data generating process (DGP), for the presence of (1) autocorrelation, (2) conditional heteroskedasticity, (3) unit root, and (4) outliers against the normal distribution in these series.[6] The results show that the loss series possibly has weak autocorrelation, but we found no evidence of conditional heteroskedasticity. Also, in our setup we found that the loss series is stationary and it is affected by outliers (rejecting normality) which can be linked to VaR exceptions (causing spikes in the loss series).

2.3. Hypotheses and test statistic

The distribution of $\{r_t\}_{t=1}^T$ and the values of $\{v_t(\boldsymbol{\theta})\}_{t=1}^T$ and $\{e_t(\boldsymbol{\theta})\}_{t=1}^T$ all depend on the model parameter vector which can be time varying, hence it will be denoted by $\boldsymbol{\theta}_t$. Thus, in this case, a procedure for detecting a change point can be conducted by testing the null hypothesis: $\boldsymbol{\theta}_1 = \ldots = \boldsymbol{\theta}_T$, against the alternative hypothesis that there is one unknown change point k^*, that is: $\boldsymbol{\theta}_1 = \ldots = \boldsymbol{\theta}_{k^*} \neq \boldsymbol{\theta}_{k^*+1} = \ldots = \boldsymbol{\theta}_T$. The true values of VaR and ES will lead to the minimal loss values for the given returns. If there is a change point, the parameter values estimated using the time period between 1 and k^* will be different from the parameter values estimated from the whole sample, so the VaR and ES estimates based on the parameters obtained from the whole sample will deviate from the true values, leading to an increase in their loss values.

We can formulate a test for the hypotheses above using the loss series. In this framework, the loss values can be expressed as:

$$L_t = \begin{cases} \mu_0 + \varepsilon_t, & \text{if } 1 \le t \le k^* \\ \mu_A + \varepsilon_t, & \text{if } k^* + 1 \le t \le T, \end{cases}$$

where μ_0 and μ_A are unknown parameters and $\mathbb{E}[\varepsilon_t] = 0$ for $1 \le t \le T$. The null hypothesis of no change point in the loss series can be written as:

$$H_0 : \mu_0 = \mu_A,$$

versus the alternative hypothesis[7] of one change point in the loss series:

$$H_1 : \mu_0 \neq \mu_A.$$

The CUSUM test is commonly used to detect change points of a process. However, this test has the limitation that it can be disturbed by the presence of outliers or extremely heavy-tailed noise (Fearnhead & Rigaill, 2019; Gerstenberger, 2018). As shown in Section 2.1, outliers (against normality) commonly exist in the loss series, due to the VaR exceptions, and thus making the CUSUM test less suitable to be applied directly on the "raw" loss series. As highlighted by Gerstenberger (2018), the Wilcoxon test statistic is a rank-type statistic and has the inherent advantage that it is not

[4] Using the FZ loss function for estimation and forecast evaluation requires choosing G_1 and G_2. The selection of a proper set of (G_1, G_2) remains an open question. For a more elaborate discussion on this, see Patton et al. (2019).

[5] A loss function L is called positively homogeneous of degree b if for all r, v and e, $L(cr, cv, ce) = c^b L(r, v, e)$, for all $c > 0$.

[6] The simulation setup and results are reported in Table S.1 of the Supplemental Appendix. We use the Ljung–Box test, Engle's ARCH test, the Augmented Dickey-Fuller (ADF) test and Grubb's test.

[7] For simplicity, in this study we consider the alternative hypothesis that there exists only one change point k^* occurring in the series.

affected by outliers. Therefore, we employ the Wilcoxon test to detect change points in the rank of the loss series. The general form of the Wilcoxon test statistic is defined as:

$$W_T := \max_{1\le k\le T} |W_{k,T}|, \text{ where } W_{k,T} := \sum_{i=1}^{k} R_i - \frac{k}{T}\sum_{i=1}^{T} R_i, \tag{2}$$

where $R_i = \text{rank}(L_i) = \sum_{j=1}^{T} \mathbf{1}\{L_j \le L_i\}$, for $i = 1, \dots, T$. Inspired by Betken (2016), our test statistic based on ranks is given below:

$$W_T = \max_{1\le k\le T} \left| \sum_{i=1}^{k} R_i - \frac{k}{T}\sum_{i=1}^{T} R_i \right| = \max_{1\le k\le T} \left| \sum_{i=1}^{k}\sum_{j=k+1}^{T} \left\{ \mathbf{1}[L_i \le L_j] - \frac{1}{2} \right\} \right|. \tag{3}$$

Definition 2.1. The estimator for the time of a change point $\hat{k}_W$ is defined as the value that maximizes the loss-based Wilcoxon test statistic,

$$\hat{k}_W = \hat{k}_W(T) := \min\{k : |W_{k,T}| = W_T\}. \tag{4}$$

2.4. Theoretical results and stationary bootstrap

In this section, we investigate the asymptotic distribution of our proposed Wilcoxon-type statistic in (3). This can be treated as a U-statistic (Csörgő & Horváth, 1988; Dehling et al., 2017) with the kernel:

$$h_W(X, Y) = \mathbf{1}[X \le Y] - \frac{1}{2}. \tag{5}$$

We can define the U-process as below:

$$U_T(\tau) = \sum_{i=1}^{\lfloor \tau T \rfloor} \sum_{j=\lfloor \tau T \rfloor + 1}^{T} h_W(L_i, L_j) = \sum_{i=1}^{\lfloor \tau T \rfloor} \sum_{j=\lfloor \tau T \rfloor + 1}^{T} \left\{ \mathbf{1}[L_i \le L_j] - \frac{1}{2} \right\}, \tag{6}$$

where $0 \le \tau \le 1$, and $\lfloor \cdot \rfloor$ denotes the integer part of a number. Thus the Wilcoxon change point test statistic in (3) can be written as:

$$W_T = \max_{0\le\tau\le 1} \left| U_T(\tau) \right|. \tag{7}$$

The kernel $h_W(X, Y)$ is antisymmetric, so it satisfies:

$$h_W(X, Y) = -h_W(Y, X). \tag{8}$$

In this case, $\mathbb{E}[h_W(L_i, L_j)] = 0$ and similarly to the symmetric case we let $\tilde{h}_W(X) = \mathbb{E}[h_W(X, L_i)]$. Following Csörgő & Horváth (1988), it is reasonable to assume that:

$$0 < \mathbb{E}[h_W^2(L_i, L_j)] < \infty \quad \text{and} \quad 0 < \sigma_W^2 = \mathbb{E}[\tilde{h}_W^2(L_i)] < \infty. \tag{9}$$

To derive the asymptotic distribution of the process $U_T(\tau)$, we consider the following assumptions.

Assumption 2.1.

(A) The process $\{r_t\}_{t=1}^T$ is strictly stationary and satisfies $\mathbb{E}[r_t] = 0$, and $\mathbb{E}[|r_t|^s] < \infty$, for some $s > 2$;

(B) The loss series $\{L_t\}_{t=1}^T$ is strictly stationary and ergodic, and it satisfies $\mathbb{E}[|L_t|^\gamma] < \infty$, for some $\gamma > 0$;

(C) For any integer $1 \le t \le T$, the cumulative distribution function F of L_t is continuous on the real line with a density f that is bounded;

(D) $h_W(L_1, L_2)$ given in (5) is an antisymmetric kernel, such that for a $\delta > 0$, $M > 0$:

$$\int\int |h_W(L_1, L_2)|^{2+\delta} dF(L_1) dF(L_2) \le M,$$

$$\forall k \in \mathbb{N}_0 : \int |h_W(L_1, L_{1+k})|^{2+\delta} dP(L_1, L_{1+k}) \le M;$$

(E) $\{r_t, v_t(\boldsymbol{\theta}), e_t(\boldsymbol{\theta})\}_{t=1}^T$ is strong mixing (α-mixing) with $\alpha(T) = O\left(T^{-(q-2)/q}\right)$ for some $q > 2$; $\{L_t(r_t, v_t(\boldsymbol{\theta}), e_t(\boldsymbol{\theta}))\}_{t=1}^T$ is strong mixing with the coefficient $\alpha(T) = O(T^{-\rho})$ for a $\rho > \frac{3\gamma\delta+\delta+5\gamma+2}{2\gamma\delta}$.

Assumption 2.1 (A) is a standard moment and stationarity condition for the loss series. Assumption 2.1 (B) is the condition on the continuous and bounded density of the loss series. Assumption 2.1 (C) requires the moment bound for the given kernel $h_W(L_1, L_2)$, which is consistent with Borovkova et al. (2001) and Dehling & Wendler (2010). Patton et al. (2019) provide the same dependence condition as Assumption 2.1 (D) for $\{r_t, v_t(\boldsymbol{\theta}), e_t(\boldsymbol{\theta})\}_{t=1}^T$ to support the central limit theorem for the loss series; if the first half of Assumption 2.1 (D) holds, the sequence of loss $L_t(r_t, v_t(\boldsymbol{\theta}), e_t(\boldsymbol{\theta}))$ is α-mixing with a decay rate at least as fast as that of $\left\{r_t, v_t(\boldsymbol{\theta}), e_t(\boldsymbol{\theta})\right\}_{t=1}^T$ (Patton et al., 2019). Thus, it is reasonable to assume the mixing condition for the loss series with the coefficient provided by Dehling & Wendler (2010).

Theorem 2.1. *Under the null hypothesis, assume that (8), (9), and Assumption 2.1 hold. Then as $T \to \infty$, we have:*

$$\sup_{0\le\tau\le 1} \left| \frac{1}{T^{3/2}} U_T(\tau) - \sigma_W B_T(\tau) \right| = o_P(1),$$

where $B_T(\tau), 0 \le \tau \le 1$ is a sequence of Brownian bridges, and:

$$\sigma_W^2 = Var(F(L_1)) + 2\sum_{j=2}^{\infty} Cov\left(F(L_1), F(L_j)\right).$$

The proof of Theorem 2.1 is provided in Appendix A. One way to implement such a test is by estimating the long-run variance and using the asymptotic limit to obtain the p-values. The detailed procedure is described in the Supplemental Appendix. However, as often found in the literature, the empirical size obtained when relying on the asymptotic limit in finite samples may differ significantly from the prespecified significance level. The Supplemental Appendix shows that the loss-based Wilcoxon test based on the asymptotic distribution with two long-run variance estimators is generally oversized, especially for small samples. As such, instead of estimating the long-run variance σ_W^2 above, we are going to use bootstrapping to obtain the p-values. In the following, we will elaborate the bootstrapping algorithm.

It is well known that bootstrap methods have been widely used to avoid the finite sample size distortions (see Barendse & Patton, 2022; Chen & Hong, 2016; Chen & Fang, 2019, for more examples). Thus, we propose to obtain the p-value of the test statistic W_T by using stationary bootstrapping in the following way. For a given return series $\{r_t\}_{t=1}^T$, we calculate the test statistic W_T using (7). Then, we adopt the stationary bootstrap method of Politis & Romano (1994) to generate N_B bootstrapped return series $\{r_t^*\}_{t=1}^T$ using the expected block length ℓ.[8] For each bootstrapped series, we estimate the bootstrapped VaR and ES denoted by $\{v_t^*(\hat{\boldsymbol{\theta}}_T^*)\}_{t=1}^T$ and $\{e_t^*(\hat{\boldsymbol{\theta}}_T^*)\}_{t=1}^T$, where $\hat{\boldsymbol{\theta}}_T^*$ is the parameter vector estimated from the bootstrapped returns $\{r_t^*\}_{t=1}^T$. Then we compute the loss series denoted by $\{L_t^*\}_{t=1}^T$. Applying (6) and (7) for each bootstrapped series j, we compute the bootstrapped U-process, $U_T^{*(j)}$ and the bootstrapped statistic $W_T^{*(j)}$. Then, we define the set of the bootstrapped statistics $\mathcal{W}_T^* = \{W_T^{*(1)}, \dots, W_T^{*(N_B)}\}$. After that, we calculate the frequency that the statistic W_T is below $W_T^{*(j)}$, and this is

[8] In this study, we follow Hoga to set the expected block length as $0.08T$, which can consistently produce satisfactory results in various settings. It is possible to select the optimal block length for stationary bootstrapping, please see Politis & White (2004) and Patton et al. (2009) for more details.

the bootstrapped p-value. The detailed procedure can be found in Algorithm 1.

Algorithm 1: Bootstrap procedure to obtain p-value, $\texttt{bootstrap}(\{r_t\}_{t=1}^T, W_T, N_B)$.

Input: $\{r_t\}_{t=1}^T$, W_T, N_B
Output: p-value (p)
Initialization: $j = 0$
repeat//Bootstrap j//
  $j = j + 1$
  Generate the bootstrapped returns ~$\{r_t^*\}_{t=1}^T$ using the stationary bootstrap
  Estimate the bootstrapped risk measure series $\{v_t^*(\hat{\boldsymbol{\theta}}_T^*)\}_{t=1}^T$ and $\{e_t^*(\hat{\boldsymbol{\theta}}_T^*)\}_{t=1}^T$
  Compute the bootstrapped loss series $\{L_t^*\}_{t=1}^T$
  Compute the bootstrapped statistic $W_T^{*(j)}$
until $j = N_B$;
Using $\{W_T^{*(1)},\ldots,W_T^{*(N_B)}\}$ compute $p = \frac{1}{N_B}\sum_{j=1}^{N_B} \mathbf{1}[W_T^{*(j)} > W_T]$
return p.

To verify the validity of the bootstrap method, we obtain the asymptotic distribution of the bootstrapped statistic W_T^*, which is computed based on (7) using the bootstrapped data. Then we show that it asymptotically converges to the limit distribution of the statistic W_T under the null hypothesis. To conduct the verification, we consider the following proposition, which is needed for the proof of our results.

Assumption 2.2. $\{r_t^*\}_{t=1}^T$ is generated by the stationary bootstrap with geometric block lengths with success probability $p_T = cT^{-a}$, where $a, c \in (0, 1)$.

Proposition 2.1 (Politis & Romano, 1994)**.** *If Assumption 2.1 (A) holds, and additionally Assumption 2.2 holds, then the pseudo time series $\{r_t^*\}_{t=1}^T$ is stationary.*

This proposition implies that the stationary bootstrapping ensures the stationarity of the process. In this study, we resample the return series $\{r_t\}_{t=1}^T$ instead of resampling the loss series $\{L_t\}_{t=1}^T$ directly.[9] The following theorem states the asymptotic behavior of the statistics of the bootstrapped loss series.

Theorem 2.2. *Under the null hypothesis, assume that (8), (9), Assumptions 2.1 and 2.2 hold. Let ℓ be the expected block length with $\ell \to \infty$ and also $T/\ell \to \infty$ as $T \to \infty$. Then we have the following convergence result for the bootstrapped process U_T^* obtained with expected block length ℓ:*

$$|Var^*(T^{-3/2}U_T^*(\tau)) - Var(T^{-3/2}U_T(\tau))| \xrightarrow{P} 0, \tag{10}$$

$$\sup_{x\in\mathbb{R}} |P^*(T^{-3/2}U_T^*(\tau) \le x) - P(T^{-3/2}U_T(\tau) \le x)| \xrightarrow{P} 0, \tag{11}$$

where Var^* and P^* denote the variance and probability with respect to the probability measure induced by the stationary bootstrap.

The proof of this theorem can be found in Appendix B.

Recall that $W_T^{*(j)}$, $1 \le j \le N_B$, denotes the bootstrapped statistic calculated similarly to W_T defined in (7). Next, we show that the asymptotic distribution of the bootstrapped statistic W_T^* coincides with the asymptotic distribution of W_T under the null hypothesis. The empirical distribution function of $W_T^{*(j)}$ is calculated as:

$$Q_{T,N_B}(w) = \frac{1}{N_B}\sum_{1\le j\le N_B} \mathbf{1}[W_T^{*(j)} \le w], \quad w \in \mathbb{R}. \tag{12}$$

[9] We found that resampling the loss series $\{L_t(r_t^*)\}_{t=1}^T$ directly would lead to a higher empirical size, especially for small sample sizes (the related simulation results can be found in Table S.3 of the Supplemental Appendix).

Based on Eqs. (7) and (12), as well as Theorems 2.1 and 2.2, we obtain the following result:

Corollary 2.1. *If the assumptions of Theorem 2.2 hold, then under H_0 we have:*

$$\sup_{w\in\mathbb{R}} |P(W_T \le w) - Q_{T,N_B}(w)| \xrightarrow{P} 0, \quad \text{where } N_B \to \infty \text{ and } T \to \infty. \tag{13}$$

This corollary demonstrates that the proposed bootstrap methodology is appropriate to be used to obtain the p-value of the loss-based Wilcoxon test statistic. In the next section, we implement a simulation study to show that the bootstrap methodology has the correct size under the null hypothesis and has high power under the alternative hypothesis.

3. Simulation analysis

Based on the test framework proposed above, we implement a comprehensive simulation study to evaluate the performance of the loss-based Wilcoxon test under the null and alternative hypotheses. In the following, we present the simulation design and analyse the simulation results.

3.1. Simulation design

We perform a simulation study to investigate the size and power of the proposed test in finite samples. Under the null hypothesis, the DGP of the return series is a univariate GARCH process as given below:

$$\begin{aligned} &r_t = \sigma_t u_t, \quad u_t \sim i.i.d.\ skewed\ t\ (\nu_1, \lambda_1),\\ &\sigma_t^2 = \omega_1 + \beta_1\sigma_{t-1}^2 + \gamma_1 r_{t-1}^2, \quad t = 1,\ldots,T, \end{aligned} \tag{14}$$

where r_t is the simulated return process generated by the product of u_t, which follows the standardized skewed t distribution of Hansen (1994), with Degrees of Freedom (DoF) ν_1 and skewness λ_1, with the density function given by:[10]

$$g(u|\nu,\lambda) = \begin{cases} bc\left[1 + \frac{1}{\nu-2}\left(\frac{bu+a}{1-\lambda}\right)^2\right]^{-(\nu+1)/2}, & \text{if } u < -a/b,\\ bc\left[1 + \frac{1}{\nu-2}\left(\frac{bu+a}{1+\lambda}\right)^2\right]^{-(\nu+1)/2}, & \text{if } u \ge -a/b, \end{cases}$$

and conditional volatility σ_t given by a GARCH(1,1) specification. For the simulations, we choose the sample sizes of $T \in \{1000, 3000\}$ to study the finite sample properties and convergence of the test.[11]

Under the alternative hypothesis, the DGP of the returns is the process $r_t = \sigma_t u_t$ with:

$$\begin{cases} \sigma_t^2 = \omega_1 + \beta_1\sigma_{t-1}^2 + \gamma_1 r_{t-1}^2, & u_t \sim i.i.d.\ skewed\ t\ (\nu_1,\lambda_1), \quad \text{if } 1 < t \le \lfloor \pi T \rfloor,\\ \sigma_t^2 = \omega_2 + \beta_2\sigma_{t-1}^2 + \gamma_2 r_{t-1}^2, & u_t \sim i.i.d.\ skewed\ t\ (\nu_2,\lambda_2), \quad \text{if } \lfloor \pi T \rfloor < t \le T, \end{cases} \tag{15}$$

where one of the parameters changes its value after $\lfloor \pi T \rfloor$ which is the location of the change point in the process. In this study, we consider $\pi \in \{0.5, 0.75\}$.[12] This change in the return series will eventually cause a change point in the VaR and ES as well, and our main purpose is to investigate the detection of change points in the VaR and ES processes at $\alpha = 1\%$.[13]

[10] In the density function, $2 < \nu < \infty$, and $-1 < \lambda < 1$. The constants a, b and c are given by $a = 4\lambda c\left(\frac{\nu-2}{\nu-1}\right), b^2 = 1 + 3\lambda^2 - a^2$, and $c = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma(\nu/2)}$.

[11] These sample sizes are in line with the sample sizes used in the literature on risk measurement (see Patton et al., 2019).

[12] We follow Hoga (2017) in selecting these two locations for the change point.

[13] Results for $\alpha = 5\%$ are consistent with the results reported here, and are available upon request.

Regarding parameter values, we set $(\omega_1, \beta_1, \gamma_1, \nu_1, \lambda_1) = (0.05, 0.9, 0.05, 16.5, -0.5)$. Under the null hypothesis, $(\omega_2, \beta_2, \gamma_2, \nu_2, \lambda_2) = (\omega_1, \beta_1, \gamma_1, \nu_1, \lambda_1)$ in (15), meaning no change points in the process. For the alternative hypothesis, we consider six different scenarios of change points to evaluate the empirical power of the proposed test. Each break consists of a change in the value of one parameter as follows:

- H_1^{A1}: an increase of 0.04 in the volatility persistence parameter, i.e. $\beta_2 = 0.94$;
- H_1^{A2}: a decrease of 0.04 in the volatility persistence parameter, i.e. $\beta_2 = 0.86$;
- H_1^{B1}: an increase of 0.04 in the volatility reaction parameter, i.e. $\gamma_2 = 0.09$;
- H_1^{B2}: a decrease of 0.04 in the volatility reaction parameter, i.e. $\gamma_2 = 0.01$;
- H_1^{C1}: a decrease of 13.5 in the DoF parameter, i.e. $\nu_2 = 3$;
- H_1^{C2}: a decrease of 14 in the DoF parameter,[14] i.e. $\nu_2 = 2.5$.

In addition to the above alternatives,[15] we follow Andreou & Ghysels (2002) to examine whether the presence of outliers affects our test results under the null hypothesis. We conjecture that the existence of outliers should not lead to the rejection of the test, i.e. an effective test would not mistakenly consider outliers as change points:

- H_0^D: $(\omega_2, \beta_2, \gamma_2, \nu_2, \lambda_2) = (\omega_1, \beta_1, \gamma_1, \nu_1, \lambda_1)$, when 12 randomly selected returns in the simulated process are multiplied by 5.

In the simulation, we consider the eight DGPs detailed above. For the estimation of VaR and ES, we use the following three (semi)parametric models: GARCH(1,1)-skewed t (G-Skt), GARCH(1,1)-Gaussian (G-G) and the Generalized Autoregressive Score (GAS) model in a hybrid framework (Hybrid).[16] In terms of the loss function, we choose loss functions with three different degrees of positive homogeneity: L^{FZ0}, L^{FZ1}, and L^{FZ2}, given in Table 1.

For each combination of (DGP, estimation method, loss function), we compute the rejection rates of the proposed test according to the procedure explained below. For each simulation i, we simulate return series of length T, denoted by $\{r_t\}_{t=1}^T$.[17] We then estimate the VaR and ES series using the given model, and we denote the estimated risk series as $\{v_t(\hat{\boldsymbol{\theta}}_T)\}_{t=1}^T$ and $\{e_t(\hat{\boldsymbol{\theta}}_T)\}_{t=1}^T$. Following this, we calculate the loss series $\{L_t\}_{t=1}^T$ for the given loss function. Then, based on (7) we compute the loss-based Wilcoxon statistic W_T for the loss series. By calling the bootstrap procedure in Algorithm 1 with $N_B = 1000$, we obtain the p-value of simulation i, denoted by $p(i)$. If $p(i)$ is below the significance level a, then the null hypothesis is rejected for simulation i.[18] By repeating this simulation $N_S = 1000$ times, we obtain the rejection rate ζ as the frequency of $p(i)$ being lower than a in the total number of simulations. The detailed procedure can be found in Algorithm 2.

In terms of the simulation results, we expect that the empirical size converges to a, the significance level under the null hypothesis, as the number of observations increases. Under the alternative hypothesis, the expectation is that the empirical power is high and converges to 1 with the sample size. When adding outliers to the process without change points, the empirical rejection rate should be close to a if the change point test is not sensitive to outliers. Our setup allows us to explore the sensitivity of the test to the choice of risk estimation model, loss function, type and location of change point and sample size.[19]

Algorithm 2: Monte Carlo simulation procedure for loss-based Wilcoxon test.

Input: N_S, N_B, T, a
Output: rejection rate (ζ)
Initialization: $i = 0$
repeat//Simulation i//
  $i = i + 1$
  Simulate $\{r_t\}_{t=1}^T$ using the specified DGP with sample size T
  Estimate the risk measure series $\{v_t(\hat{\boldsymbol{\theta}}_T)\}_{t=1}^T$ and $\{e_t(\hat{\boldsymbol{\theta}}_T)\}_{t=1}^T$
  Calculate the loss values $\{L_t\}_{t=1}^T$
  Calculate:
$$W_T = \max_k \left| \sum_{c=1}^{k} \sum_{d=k+1}^{T} \left\{ \mathbf{1}[L_c \le L_d] - \frac{1}{2} \right\} \right|.$$
  Obtain p-value by calling Algorithm 1:$p(i)$ = `Bootstrap`$(\{r_t\}_{t=1}^T, W_T, N_B)$.
until $i = N_S$;
Using the p-values: $\{p(1),\ldots,p(N_S)\}$ compute the rejection rate $\zeta = \frac{1}{N_S} \sum_{i=1}^{N_S} \mathbf{1}[p(i) < a]$
return ζ.

3.2. Simulation results

The simulation results commence with the evaluation of the proposed loss-based Wilcoxon test in identifying change points in risk measures when the underlying process is generated from the DGP in (14) and (15) with the parameter values given in Section 4.1. Table 2 shows the size and power of the test based on the bootstrapping procedure at 5% significance level. In the table, the panel for H_0 shows the empirical sizes under the null hypothesis. As expected, all of the empirical sizes for the Wilcoxon test are close to the significance level. As the sample size increases, the empirical size gets closer to 5% in general.

For the alternative hypotheses, we consider the change points detailed in Section 4.1. The results in Table 2 reveal that our test has a strong power in detecting change points in the volatility parameters (H_1^{A1}, H_1^{A2}, H_1^{B1}, H_1^{B2}) and reasonable power in detecting change points in the DoF (H_1^{C1}, H_1^{C2}). The power of the test improves when T increases from 1000 to 3000 for all DGPs and loss functions. The table also shows that the power of the test is sensitive to the location of change point. The rejection rate modestly falls when the location of change point moves to $\lfloor 0.75T \rfloor$. However, as the sample size increases, the test can successfully detect the change point that occurs even at $\lfloor 0.75T \rfloor$. Also, the results show that our test is not sensitive to the presence of outliers (H_0^D).

In the following, we compare our proposed loss-based Wilcoxon test with five alternative tests in terms of size and power, under the same simulation settings and hypotheses as detailed before. For the first two alternative tests, we consider (i) the self-normalized CUSUM (SN-CUSUM) test for VaR and (ii) the SN-CUSUM for ES, which detect change points in the VaR and ES processes individually. Following Shao & Zhang (2010), the two test statistics are defined as:

$$V_T^v = \sup_k \frac{\left[T^{-\frac{1}{2}} \sum_{t=1}^{k} v_t(\hat{\boldsymbol{\theta}}_T) - \frac{k}{T} \sum_{t=1}^{T} v_t(\hat{\boldsymbol{\theta}}_T) \right]^2}{T^{-2} \left[\sum_{t=1}^{k} S_{v,t}^2(1,k) + \sum_{t=k+1}^{T} S_{v,t}^2(k+1,T) \right]^{\frac{1}{2}}}, \tag{16}$$

[14] We are aware that these values of ν_2 mean that the fourth moment of the simulated returns does not exist. Nevertheless, these values of ν_2 are useful for illustrative purposes. The literature considers DGPs with less than four finite moments, such as in Berkes et al. (2003).

[15] These values are chosen so that the first two moments of the simulated returns still exist.

[16] More details about the models can be found in Table S.4 of the Supplemental Appendix.

[17] For simplicity, we disregard the dependence on i in the notation for $\{r_t\}_{t=1}^T$, $\{v_t(\hat{\boldsymbol{\theta}}_T)\}_{t=1}^T$, $\{e_t(\hat{\boldsymbol{\theta}}_T)\}_{t=1}^T$ and $\{L_t\}_{t=1}^T$.

[18] Here, we only consider the case of $a = 5\%$; the results for other values of a are available on request.

[19] If model misspecification risk is present, then the ordering of models is affected by the choice of loss function (Patton, 2020); in this case, the size and power properties of our proposed test might be affected by the choice of loss function.

Table 2
Empirical size and power of the loss-based Wilcoxon test for a change point.

	$\pi = 0.5$						$\pi = 0.75$					
	$T = 1000$			$T = 3000$			$T = 1000$			$T = 3000$		
	G-Skt	G-G	Hybrid	G-Skt	G-G	Hybrid	G-Skt	G-G	Hybrid	G-Skt	G-G	Hybrid
H_0: Univariate GARCH(1,1)-skewed t, with $(\omega_1, \gamma_1, \beta_1, \nu_1, \lambda_1)$=(0.05, 0.05, 0.9, 16.5, −0.5)												
L^{FZ0}	0.045	0.044	0.030	0.047	0.063	0.039	0.045	0.044	0.030	0.047	0.063	0.039
L^{FZ1}	0.045	0.044	0.030	0.047	0.064	0.039	0.045	0.044	0.030	0.047	0.064	0.039
L^{FZ2}	0.045	0.044	0.030	0.047	0.064	0.038	0.045	0.044	0.030	0.047	0.064	0.038
H_1^{A1}: An increase of 0.04 in the volatility persistence parameter, i.e. $\beta_2 = 0.94$												
L^{FZ0}	0.992	0.992	0.923	1.000	1.000	1.000	0.788	0.769	0.326	0.996	0.975	0.919
L^{FZ1}	0.992	0.992	0.923	1.000	1.000	1.000	0.788	0.770	0.326	0.996	0.975	0.918
L^{FZ2}	0.992	0.992	0.923	1.000	1.000	1.000	0.788	0.773	0.328	0.996	0.975	0.918
H_1^{A2}: A decrease of 0.04 in the volatility persistence parameter, i.e. $\beta_2 = 0.86$												
L^{FZ0}	0.627	0.623	0.373	0.988	0.963	0.770	0.232	0.220	0.118	0.712	0.695	0.355
L^{FZ1}	0.627	0.623	0.373	0.988	0.963	0.770	0.232	0.220	0.117	0.712	0.695	0.354
L^{FZ2}	0.627	0.623	0.373	0.988	0.963	0.770	0.232	0.221	0.117	0.712	0.695	0.355
H_1^{B1}: An increase of 0.04 in the volatility reaction parameter, i.e. $\gamma_2 = 0.09$												
L^{FZ0}	0.912	0.911	0.715	1.000	1.000	0.985	0.608	0.605	0.225	0.939	0.898	0.747
L^{FZ1}	0.912	0.911	0.715	1.000	1.000	0.985	0.608	0.604	0.225	0.939	0.898	0.747
L^{FZ2}	0.912	0.911	0.715	1.000	1.000	0.985	0.608	0.604	0.225	0.939	0.898	0.747
H_1^{B2}: A decrease of 0.04 in the volatility reaction parameter, i.e. $\gamma_2 = 0.01$												
L^{FZ0}	0.529	0.524	0.350	0.987	0.951	0.731	0.196	0.194	0.114	0.694	0.644	0.363
L^{FZ1}	0.529	0.524	0.349	0.987	0.951	0.731	0.196	0.194	0.114	0.694	0.643	0.363
L^{FZ2}	0.528	0.524	0.349	0.987	0.951	0.730	0.196	0.194	0.113	0.694	0.643	0.363
H_1^{C1}: A decrease of 13.5 in the DoF parameter, i.e. $\nu_2 = 3$												
L^{FZ0}	0.293	0.290	0.176	0.777	0.758	0.283	0.164	0.159	0.092	0.393	0.354	0.169
L^{FZ1}	0.293	0.290	0.176	0.777	0.758	0.281	0.165	0.159	0.093	0.393	0.354	0.169
L^{FZ2}	0.293	0.290	0.176	0.776	0.757	0.282	0.166	0.159	0.093	0.393	0.354	0.169
H_1^{C2}: A decrease of 14 in the DoF parameter, i.e. $\nu_2 = 2.5$												
L^{FZ0}	0.636	0.627	0.358	0.996	0.988	0.449	0.330	0.326	0.164	0.593	0.552	0.389
L^{FZ1}	0.636	0.627	0.357	0.996	0.987	0.448	0.331	0.324	0.165	0.593	0.552	0.389
L^{FZ2}	0.636	0.627	0.358	0.996	0.987	0.448	0.331	0.324	0.164	0.593	0.552	0.389
H_0^D: 12 randomly selected returns in the simulated process multiplied by 5												
L^{FZ0}	0.041	0.042	0.041	0.035	0.052	0.039	0.041	0.042	0.041	0.035	0.052	0.039
L^{FZ1}	0.041	0.042	0.040	0.035	0.052	0.039	0.041	0.042	0.040	0.035	0.052	0.039
L^{FZ2}	0.041	0.042	0.041	0.035	0.052	0.039	0.041	0.042	0.041	0.035	0.052	0.039

Note: Empirical size and power, for $a = 5\%$, of the loss-based Wilcoxon test under various hypotheses *via* 1000 simulations, for three types of risk measures (GARCH(1,1)-skewed t, GARCH(1,1)-Gaussian and GAS-Hybrid) and three FZ loss functions with different degrees of positive homogeneity. VaR and ES are jointly estimated at 1% level. We consider two sample sizes: 1000 and 3000, and different locations of the change point at $\lfloor \pi T \rfloor$ with $\pi = 0.5$ and 0.75.

$$V_T^e = \sup_k \frac{\left[T^{-\frac{1}{2}} \sum_{t=1}^{k} e_t(\hat{\boldsymbol{\theta}}_T) - \frac{k}{T} \sum_{t=1}^{T} e_t(\hat{\boldsymbol{\theta}}_T)\right]^2}{T^{-2}\left[\sum_{t=1}^{k} S_{e,t}^2(1,k) + \sum_{t=k+1}^{T} S_{e,t}^2(k+1,T)\right]^{\frac{1}{2}}}, \tag{17}$$

where $v_t(\hat{\boldsymbol{\theta}}_T)$ and $e_t(\hat{\boldsymbol{\theta}}_T)$ are the estimated VaR and ES, and $S_{v,t}(j,k) = \sum_{h=j}^{t}[v_h(\hat{\boldsymbol{\theta}}_T) - \bar{v}_{j,k}]$, $\bar{v}_{j,k} = \frac{1}{k-j+1}\sum_{t=j}^{k} v_t(\hat{\boldsymbol{\theta}}_T)$, as well as $S_{e,t}(j,k) = \sum_{h=j}^{t}[e_h(\hat{\boldsymbol{\theta}}_T) - \bar{e}_{j,k}]$, $\bar{e}_{j,k} = \frac{1}{k-j+1}\sum_{t=j}^{k} e_t(\hat{\boldsymbol{\theta}}_T)$. Table 3 presents the empirical size and power simulation results of the SN-CUSUM tests for VaR and ES. The sizes of the SN-CUSUM tests are close to the significance level, but their powers are generally less than the power of our test for all loss functions considered.

One disadvantage of the standard CUSUM test is the low power in detecting change points occurring in relatively early or late segments of the sample period. As an alternative, Horváth et al. (2020b) propose a Rényi-type statistic for change point detection to mitigate this problem. However, when the change point happens around the middle of the sample period, the detecting power of the Rényi-type test is relatively low. The Rényi-type test works under the assumption that there is no change point occurring within the two trimmed domains, at the beginning and at the end of the sample defined by the trimming parameter τ_0. Thus, we consider the alternative test (*iii*) a Rényi-type test based on the rank of loss values. Specifically, the test statistic is a Rényi-type formulation of the loss-based Wilcoxon test statistic:

$$D_T := \max_{\lfloor \tau_0 T \rfloor \le k \le T - \lfloor \tau_0 T \rfloor} \left| \frac{1}{k}\sum_{i=1}^{k} R_i - \frac{1}{T-k}\sum_{i=k+1}^{T} R_i \right| \tag{18}$$

with trimming parameter τ_0.[20]

In addition to these, we consider the following two recently developed tests: (*iv*) the change point test for VaR of Hoga, and (*v*) the change point test for ES of FGP.[21] These two tests are based on the self-normalized variance estimator of Shao & Zhang (2010).

Table 4 presents the simulations results for alternative tests (*iii*) to (*v*) (in columns Rényi, Hoga, and FGP, respectively). The results

[20] We use the FZ0 loss function to compute the loss values for GARCH(1,1)-skewed t risk estimates. We implement the stationary bootstrapping based MC simulation for the Rényi-type loss-based Wilcoxon test, instead of applying the asymptotic distribution that assumes normally distributed loss values.

[21] We choose the historical quantile to estimate VaR and ES, in line with FGP, instead of applying the Weissman estimator for VaR used by Hoga. Based on our simulations, the critical value at 5% significance level is 80.21 for $\tau_0 = 0.2$, which is very close to the one given by Hoga.

Table 3
Empirical size and power of the SN-CUSUM test for a change point in VaR and ES.

	$\pi = 0.5$						$\pi = 0.75$					
	$T = 1000$			$T = 3000$			$T = 1000$			$T = 3000$		
	G-Skt	G-G	Hybrid	G-Skt	G-G	Hybrid	G-Skt	G-G	Hybrid	G-Skt	G-G	Hybrid
H_0: Univariate GARCH(1,1)-skewed t, with $(\omega_1, \gamma_1, \beta_1, \nu_1, \lambda_1)$=(0.05, 0.05, 0.9, 16.5, −0.5)												
VaR	0.034	0.034	0.037	0.066	0.066	0.044	0.034	0.034	0.037	0.066	0.066	0.044
ES	0.034	0.034	0.043	0.066	0.066	0.043	0.034	0.034	0.043	0.066	0.066	0.043
H_1^{A1}: An increase of 0.04 in the volatility persistence parameter, i.e. $\beta_2 = 0.94$												
VaR	0.629	0.629	0.548	0.772	0.772	0.765	0.724	0.724	0.595	0.924	0.924	0.884
ES	0.629	0.629	0.546	0.772	0.772	0.765	0.724	0.724	0.600	0.924	0.924	0.884
H_1^{A2}: A decrease of 0.04 in the volatility persistence parameter, i.e. $\beta_2 = 0.86$												
VaR	0.361	0.361	0.307	0.783	0.783	0.676	0.100	0.100	0.104	0.353	0.353	0.317
ES	0.361	0.361	0.305	0.783	0.783	0.675	0.100	0.100	0.100	0.353	0.353	0.313
H_1^{B1}: An increase of 0.04 in the volatility reaction parameter, i.e. $\gamma_2 = 0.09$												
VaR	0.372	0.372	0.362	0.558	0.558	0.847	0.518	0.518	0.421	0.776	0.776	0.744
ES	0.372	0.372	0.361	0.558	0.558	0.847	0.518	0.518	0.421	0.776	0.776	0.743
H_1^{B2}: A decrease of 0.04 in the volatility reaction parameter, i.e. $\gamma_2 = 0.01$												
VaR	0.295	0.295	0.303	0.769	0.769	0.683	0.091	0.091	0.092	0.338	0.338	0.296
ES	0.295	0.295	0.296	0.769	0.769	0.687	0.091	0.091	0.093	0.338	0.338	0.297
H_1^{C1}: A decrease of 13.5 in the DoF parameter, i.e. $\nu_2 = 3$												
VaR	0.161	0.161	0.208	0.268	0.268	0.225	0.076	0.076	0.155	0.171	0.171	0.226
ES	0.161	0.161	0.210	0.268	0.268	0.223	0.076	0.076	0.157	0.171	0.171	0.225
H_1^{C2}: A decrease of 14 in the DoF parameter, i.e. $\nu_2 = 2.5$												
VaR	0.392	0.392	0.352	0.606	0.606	0.322	0.180	0.180	0.225	0.337	0.340	0.303
ES	0.392	0.392	0.341	0.606	0.606	0.319	0.180	0.180	0.223	0.337	0.340	0.302
H_0^D: 12 randomly selected returns in the simulated process multiplied by 5												
VaR	0.033	0.033	0.047	0.046	0.046	0.048	0.033	0.033	0.047	0.046	0.046	0.048
ES	0.033	0.033	0.047	0.046	0.046	0.047	0.033	0.033	0.047	0.046	0.046	0.047

Note: Empirical size and power, for $a = 5\%$, of the SN-CUSUM test for VaR and ES, considered individually, under various hypotheses *via* 1000 simulations, for three types of risk measures (GARCH(1,1)-skewed t, GARCH(1,1)-Gaussian and GAS-Hybrid). VaR and ES are jointly estimated at 1% level. We consider two sample sizes: 1000 and 3000, and different locations of the change point at $\lfloor \pi T \rfloor$ with $\pi = 0.5$ and 0.75.

Table 4
Empirical size and power of alternative tests for a change point.

$\pi = 0.5$						$\pi = 0.75$					
$T = 1000$			$T = 3000$			$T = 1000$			$T = 3000$		
Rényi	Hoga	FGP	Rényi	Hoga	FGP	Rényi	Hoga	FGP	Rényi	Hoga	FGP
H_0: Univariate GARCH(1,1)-skewed t, with $(\omega_1, \gamma_1, \beta_1, \nu_1, \lambda_1)$=(0.05, 0.05, 0.9, 16.5, −0.5)											
0.045	0.148	0.104	0.044	0.107	0.105	0.045	0.148	0.104	0.044	0.107	0.105
H_1^{A1}: An increase of 0.04 in the volatility persistence parameter, i.e. $\beta_2 = 0.94$											
0.966	0.602	0.596	1.000	0.736	0.749	0.967	0.758	0.728	1.000	0.934	0.924
H_1^{A2}: A decrease of 0.04 in the volatility persistence parameter, i.e. $\beta_2 = 0.86$											
0.396	0.318	0.291	0.854	0.507	0.465	0.391	0.186	0.137	0.934	0.208	0.162
H_1^{B1}: An increase of 0.04 in the volatility reaction parameter, i.e. $\gamma_2 = 0.09$											
0.794	0.488	0.486	0.968	0.617	0.644	0.799	0.631	0.646	0.988	0.863	0.863
H_1^{B2}: A decrease of 0.04 in the volatility reaction parameter, i.e. $\gamma_2 = 0.01$											
0.338	0.344	0.290	0.851	0.559	0.529	0.337	0.164	0.134	0.914	0.213	0.172
H_1^{C1}: A decrease of 13.5 in the DoF parameter, i.e. $\nu_2 = 3$											
0.193	0.209	0.211	0.461	0.194	0.264	0.268	0.200	0.244	0.648	0.196	0.348
H_1^{C2}: A decrease of 14 in the DoF parameter, i.e. $\nu_2 = 2.5$											
0.445	0.263	0.210	0.833	0.210	0.220	0.576	0.196	0.243	0.907	0.173	0.279
H_0^D: 12 randomly selected returns in the simulated process multiplied by 5											
0.044	0.144	0.115	0.039	0.104	0.099	0.044	0.144	0.115	0.039	0.104	0.099

Note: Empirical size and power, for $a = 5\%$, of three alternative tests (*iii*), (*iv*) and (*v*) under various hypotheses *via* 1000 simulations. We consider two sample sizes: 1000 and 3000, and different locations of the change point at $\lfloor \pi T \rfloor$ with $\pi = 0.5$ and 0.75. For the Rényi-type test, we choose the loss values computed by the FZ0 loss function with 1% VaR and ES estimated by the GARCH(1,1)-skewed t model.

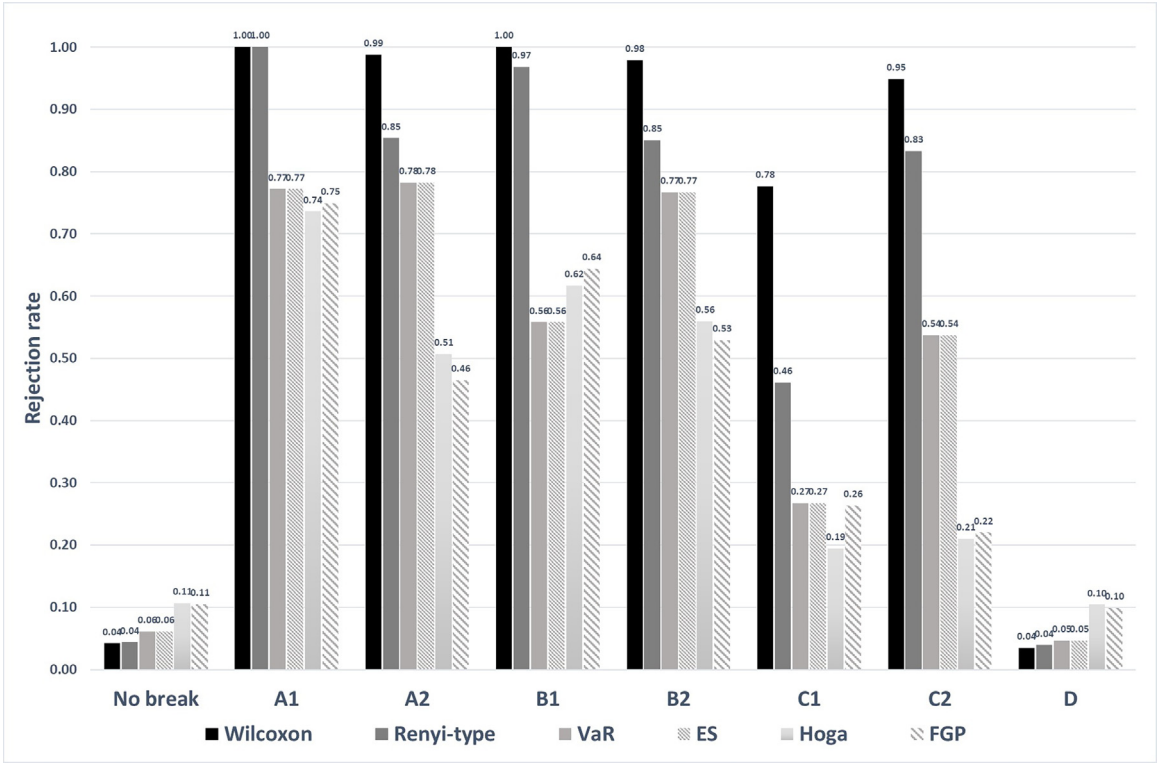

Fig. 1. Size and power of the loss-based Wilcoxon test and alternatives. *Note:* Empirical size and power of the loss-based Wilcoxon test with five alternative tests (i)-(v) under various hypotheses described in Section 4.1 at 5% significance level when the change point occurs at $\lfloor 0.5T \rfloor$. For the Wilcoxon test and Rényi-type test, we use the FZ0 loss function to compute the loss values. For all tests except Hoga and FGP, 1% VaR and ES are estimated by the GARCH(1,1)-skewed t model.

highlight that our test outperforms tests (*iv*) and (*v*) in all cases. We outperform the Rényi-type test (*iii*) when the change point occurs at $\lfloor 0.5T \rfloor$, but when the change point occurs at $\lfloor 0.75T \rfloor$, test (*iii*) has better power properties than our test. This meets our expectation that the Rényi-type test has high power in detecting change points occurring relatively early or late in the sample, but has lower power in the middle. For our simulation setup, we find the Hoga and FGP tests to be oversized under the null hypothesis and to have less power than the Rényi-type loss-based Wilcoxon test.[22]

To offer a visual demonstration, Fig. 1 compares the loss-based Wilcoxon test using the FZ0 loss function with tests (*i*) to (*v*), from the point of view of size and power. The five alternatives are denoted by VaR, ES, Rényi-type, Hoga, and FGP, respectively. For the loss-based Wilcoxon test and alternative tests (*i*)–(*iii*), the VaR and ES are obtained using the GARCH(1,1)-skewed t model. The tests are performed at 5% significance level, and we assume that the change point occurs at $\lfloor 0.5T \rfloor$ under the alternative hypotheses. Based on the empirical sizes of the Hoga and FGP tests under H_0 and H_0^D, it can be concluded that these tests are oversized for the DGP considered. The loss-based Wilcoxon test has higher power than the alternatives for all scenarios of change points corresponding to the different alternative hypotheses. The SN-CUSUM tests work relatively well when volatility changes, but have lower power when the DoF parameter decreases. Overall, our proposed test can identify change points in the risk measures of time series with the correct size and stronger power than all five alternatives considered.[23]

3.3. Simulation study for the location of change point estimator

In this section, we perform two simulations. The first one investigates the accuracy of the estimator for the location of a change point based on the loss-based Wilcoxon test. In the second study, in order to provide new insights of practical relevance and decision support, we evaluate how quickly after the event the estimator can identify the change point (the delay in detection).

We consider a univariate GARCH(1,1) skewed t process as the DGP of the returns series, as demonstrated in (15). This model is used to build the risk measures. For the simulations, we consider a sample size of $T = 1000$ and we assume that the change occurs at $\lfloor 0.5T \rfloor$. We consider one of the scenarios listed in Section 3.1 where after the change point, the volatility persistence parameter β_2 increases from 0.90 to 0.94. We apply our proposed loss-based Wilcoxon test and the Rényi-type test, as well as two benchmark tests to identify the location of the change in the risk measures. Figure 2 shows the estimated locations of the change points obtained *via* these tests. The values on the x-axis represent the difference between the estimated change point and the actual change point. In these histograms, the red lines indicate the 5%, 95% quantiles and the median of the estimated locations. The 90% confidence interval (around from −25 to 50) of the estimated location of change point obtained *via* the loss-based Wilcoxon test is narrower than the one generated by the other tests. This leads to

[22] In Table S.5 of the Supplemental Appendix, we show that the loss-based Wilcoxon test has strong power in detecting the change point in the series of VaR and ES estimated by historical simulations. In Figure S.2 we present the power curves of this test for three (semi)parametric models with alternative tests to show the detection power in terms of the marginal change in parameters. Figure S.3 compares our test with alternative tests in terms of size and power for AR(1) and ARCH(1) processes, which are the DGPs used by FGP. Our results are consistent with the results in Tables 2 and 4.

[23] In Figure S.4 of the Supplemental Appendix we present the results when the change point occurs at $\lfloor 0.75T \rfloor$.

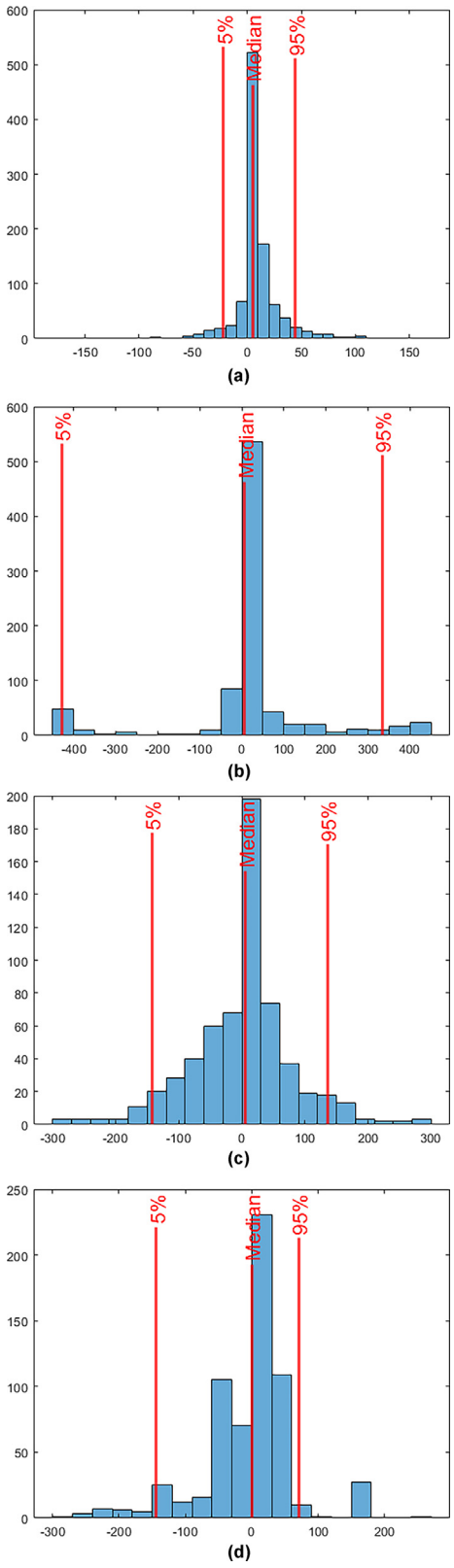

Fig. 2. The histogram of estimated relative location of change point $\hat{k}_W(1000)$ for $\pi = 0.5$. *Note:* The estimated location of change point relative to the actual change point obtained *via* the loss-based Wilcoxon, Rényi-type, Hoga and FGP tests at 5% significance level with $T = 1000$ when the change point occurs at $\lfloor 0.5T \rfloor$. In the simulation, we use the GARCH(1,1)-skewed t model to generate the return process and to obtain 1% VaR and ES estimates. For the Wilcoxon test, we use the FZ0 loss function to compute the loss values. The red lines indicate the 5% and 95% quantiles and the median of the estimators. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the conclusion that the loss-based Wilcoxon test has more accurate detection than the benchmark tests.

In order to evaluate the delay in the detection of a novel change point, we propose an expanding window procedure, as detailed below, following Smith & Timmermann (2021). Starting with a change point detection based on the first half of observations, the detection window is sequentially expanded by five days forward and the detection procedure is re-implemented until the end of the sample is reached. We consider the same DGP setup as the one used in the in-sample simulation with the change point occurring in the middle of the sample. We use a sample size of $T = 3000$ to ensure the asymptoticity of the test, and we sequentially apply the Rényi-type loss-based Wilcoxon test, which is more powerful for a change point occurring towards the end of sample. Figure 3 shows the histogram of the delay of detection for a new change point obtained by the Rényi-type Wilcoxon test. The blue bars indicate the frequencies of delays for a novel change point, with the red line indicating the 95% quantile of detection delays. Following Harvey & Liu (2020), the testing performance is gauged by two metrics: the test power (1−Type II error rate) and the test size (Type I error rate). Thus, to evaluate the probability of a false discovery, we repeat the above procedure under the null hypothesis of no changes in risk measures. In Fig. 3, the orange bars represent the frequencies of false detections under the null hypothesis. By mitigating the overlapping areas of the blue and orange bars, we have the corrected performance of the detection delay by considering both false and missed discoveries. Practically, we require roughly 50 to 100 observations (two to four trading months) to identify a novel change. This is reasonable given that the object of the forecasts is VaR and ES at 1% level.

4. Empirical application

In this section, we apply our proposed Wilcoxon change point test to S&P 500 index daily log returns. The index data is collected from Datastream and spans the period from January 2, 1990 to December 31, 2019, in total 7559 observations. We apply the proposed loss-based Wilcoxon test to detect change points in the 1% VaR and ES risk measures estimated by the GARCH(1,1)-skewed t model. Based on our simulations above that consider the Wilcoxon tests based on loss functions with different degrees of positive homogeneity, it can be concluded that our test is not sensitive to the choice of loss function. As such, in the empirical section we only use the FZ0 loss function to compute loss values, following Patton et al. (2019) and Dimitriadis & Schnaitmann (2021). In order to see the usefulness of change point detection, we compare i) the average FZ0 loss values obtained when the change points are taken into consideration with ii) the average FZ0 loss values when change points are ignored. We will show that by detecting change points, superior risk estimates are obtained, which will highlight the practical relevance of detecting change points in VaR and ES series.

In order to find change points in the risk measures, we first compute the loss-based Wilcoxon test statistic W_T. Then, we bootstrap the return process 1000 times *via* the stationary bootstrap method with the optimal block length (Patton et al., 2009; Politis & White, 2004), obtain the empirical distribution of the Wilcoxon statistic and get the 95% critical values. If the test statistic W_T is larger than the critical value, we reject the null hypothesis of no change. In such cases, a change point is detected, and we follow the binary segmentation method discussed by Inclan & Tiao (1994) and Ye et al. (2012) to find further change points. Specifically, the data can be split into sub-periods according to the locations of the detected change points until no further change point can be found. The detailed algorithm and procedure of detecting

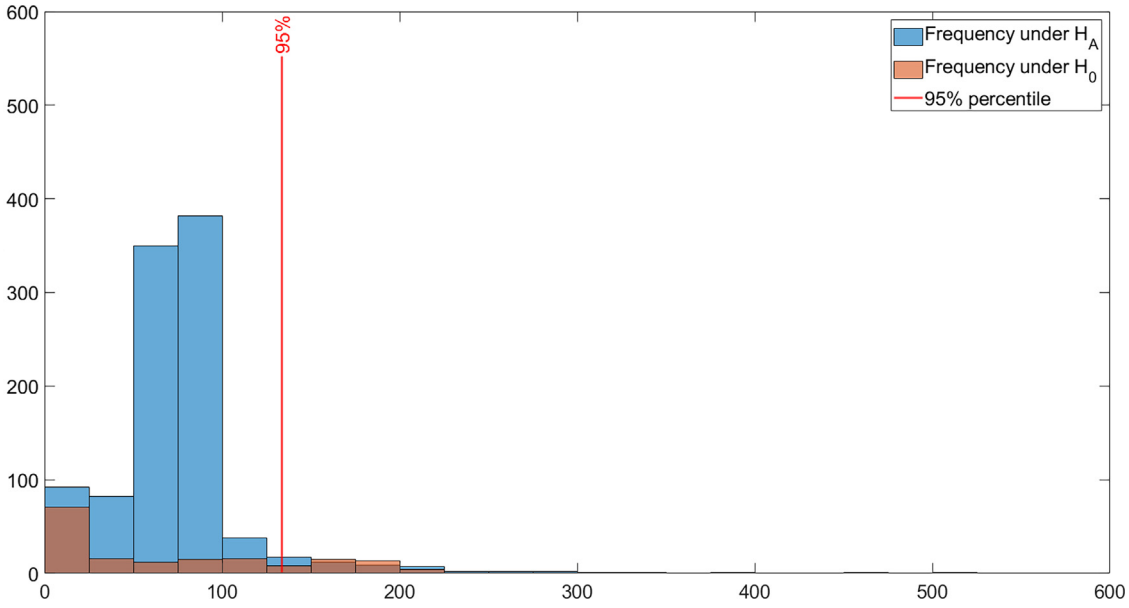

Fig. 3. The histogram of the delay of detection for a new change point by the Rényi-type Wilcoxon test. *Note:* The blue histogram is for the delay of detection for a new change point obtained *via* the Rényi-type Wilcoxon test at 5% significance level with $T = 3000$. The orange histogram is for the false detections under the null hypothesis. In the simulation, we use the GARCH(1,1)-skewed t model to generate the return process and to build 1% VaR and ES estimates. For the Rényi-type Wilcoxon test, we use the FZ0 loss function to compute the loss values. The red line indicates the 95% quantile of the estimated delay. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

multiple change points can be found in Figure S.5 of the Supplemental Appendix.

Based on our test, the earliest change point we detect in the estimated risk measures occurred in June 1992 (following the early 1990s recession in the United States). The second change point occurred in December 1996 (the start of the dot-com bubble). Then another change point is identified in June 2003 (after the burst of the dot-com bubble), and the following change points are in July 2007 (the beginning of the subprime mortgage crisis), September 2008 (the bankruptcy of Lehman Brothers), July 2009, and January 2012 (the start and end of the European debt crisis). We also successfully detect change points associated with the 2015–16 stock market selloff[24] and the 2018 cryptocurrency crash[25]. Figure 4 presents the returns as well as the risk estimates, highlighting the detected change points. Additionally, we apply this test for other estimation approaches (GAS-Hybrid and historical simulations) and compare the empirical results with alternative tests applied for the same sample (more details can be found in Tables S.6 and S.7 of the Supplemental Appendix).

Table 5 reports the GARCH(1,1)-skewed t parameter estimates and standard errors obtained by the QMLE method for each subperiod, the average VaR and ES estimates, and the average loss values. Firstly, it can be seen that the volatility parameters and the DoF estimates experience large changes across the sub-periods, which leads to change points in the VaR and ES processes as well. For instance, after the burst of the dot-com bubble, we can observe a decline in the level of the volatility. Moreover, we can see a large reduction in the value of the DoF parameter from 11.1 to 6.5 during the European debt crisis period. Secondly, during a crisis or a crash period, VaR and ES are high in absolute values, as can be seen in the 2007–2008 financial crisis and the European debt crisis. The average loss values are also found to be generally higher during crisis periods than during stable periods. Finally, we calculate the average FZ0 losses for each subsample based on the parameter values estimated within sub-periods (denoted by "Loss") and average FZ0 losses based on the parameter values estimated within the whole sample period (denoted by "Loss_NC"), respectively. When change points are taken into consideration, the FZ0 loss values are typically lower than the ones computed when the change points are ignored (this can be seen comparing "Loss" and "Loss_NC" in Table 5), which means that the risk values estimated based on the change points are superior to the risk values that do not take the change points into account. According to our findings, it can be concluded that risk management practitioners can improve on the risk estimates by first identifying change points in the loss series of risk measures and then computing model parameter values based on the identified change points.

Another essential concern of risk managers is how quickly a test is able to identify a novel change point in risk measure estimates, which would enable them to adjust the parameters of their risk model in a timely manner. To address this issue, we implement an expanding window procedure to highlight the speed of change detection based on the test proposed in Section 3.3. Starting with the initial window of the first 3000 observations of our data, the detection window is sequentially expanded forward by one month and the detection procedure is repeated until the end of the sample is reached. We apply the proposed loss-based Wilcoxon test and the binary segmentation method to detect multiple change points in the 1% VaR and ES risk measures estimated by the GARCH(1,1)-skewed t model. To further support the decision-making of risk managers, we implement a "traffic light" approach built on the loss-based Wilcoxon test for the detection of change points in the risk measure estimates. According to this, "Green" means that there is no change point on that day, "Yellow" means that there is a change point on that day detected at 10% significance level, and "Red" means that there is a change point on that day detected at 5% significance level.

Figure 5 shows the change point dates estimated *via* the expanding window procedure above, at 5% and 10% significance levels, respectively. The vertical solid line denotes the initial window for estimating risk measures and implementing the test for change

[24] Between August 2015 and early 2016, the S&P 500 and DJIA dropped more than 10% twice.

[25] The S&P 500 index dropped almost 20% between September and December 2018.

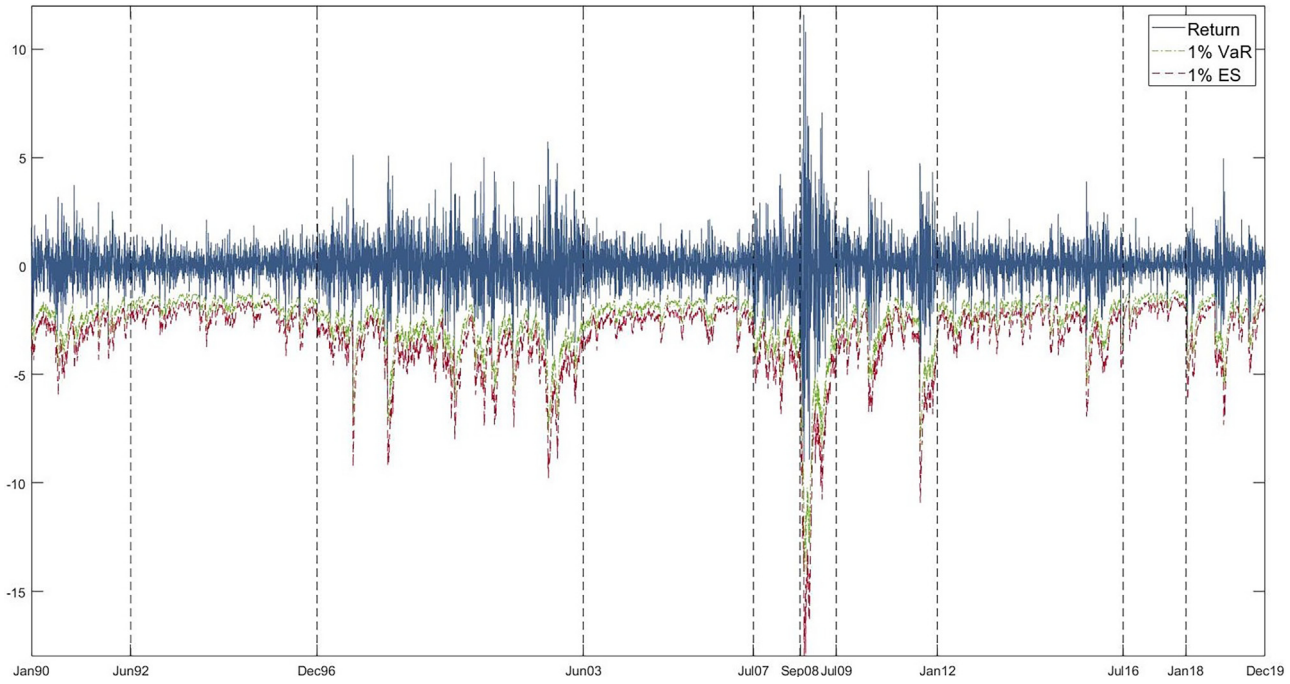

Fig. 4. Daily S&P 500 index returns and risk estimates at 1% level. *Note:* Daily S&P 500 index returns and 1% VaR and ES estimates obtained by the GARCH skewed t model. The vertical dash lines are at the estimated change points.

Table 5
GARCH(1,1)-skewed t estimation results.

Panel A: Whole sample estimation results					
Parameters	ω	γ	β	ν	λ
Estimates	0.009 (0.002)	0.087 (0.007)	0.910 (0.008)	6.479 (0.495)	−0.078 (0.014)
Panel B: Subsample estimation results					
	1990/01–1992/05	1992/05–1996/12	1996/12–2003/06	2003/06–2007/07	2007/07–2008/09
ω	0.031 (0.012)	0.013 (0.006)	0.084 (0.022)	0.021 (0.007)	0.153 (0.075)
γ	0.037 (0.013)	0.039 (0.013)	0.092 (0.017)	0.031 (0.014)	0.032 (0.036)
β	0.924 (0.023)	0.925 (0.024)	0.862 (0.024)	0.924 (0.023)	0.881 (0.032)
ν	7.343 (1.992)	5.567 (0.817)	9.943 (2.264)	11.137 (4.338)	14.078 (11.435)
λ	0.001 (0.002)	−0.008 (0.039)	−0.042 (0.037)	−0.076 (0.030)	−0.081 (0.066)
VaR	−2.237	−1.491	−3.270	−1.719	−3.424
ES	−2.805	−1.952	−4.000	−2.097	−4.095
Loss	1.038	0.747	1.431	0.702	1.244
Loss_NC	1.135	0.736	1.475	0.815	1.400
	2008/09–2009/07	2009/07–2012/01	2012/01–2016/07	2016/07–2018/01	2018/01–2019/12
ω	0.012 (0.154)	0.036 (0.013)	0.075 (0.016)	0.059 (0.052)	0.042 (0.012)
γ	0.063 (0.059)	0.119 (0.028)	0.170 (0.034)	0.047 (0.089)	0.181 (0.042)
β	0.930 (0.072)	0.860 (0.024)	0.723 (0.041)	0.692 (0.276)	0.773 (0.038)
ν	11.378 (9.630)	6.736 (1.683)	8.019 (1.819)	3.814 (0.684)	6.189 (1.830)
λ	−0.047 (0.067)	−0.151 (0.042)	−0.089 (0.035)	0.113 (0.059)	−0.211 (0.062)
VaR	−7.044	−3.223	−2.074	−1.071	−2.532
ES	−8.517	−4.138	−2.601	−1.517	−3.299
Loss	1.973	1.270	0.916	0.748	1.181
Loss_NC	2.109	1.362	1.011	0.827	1.435

Note: Estimated parameter values and standard errors for ω, β, γ, ν, and λ in the GARCH(1,1)-skewed t model: $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma u_{t-1}^2$, $u_t \sim i.i.d.$ *skewed* t (ν, λ) for the S&P 500 index. Panel A shows the estimated values and standard errors of parameters in the GARCH(1,1)-skewed t model over the whole sample period. Panel B presents the estimated parameter values and standard errors in 10 sub-periods. We also report the average VaR and ES at 1% level and the associated average FZ0 loss values using the parameters estimated within the sub-periods (Loss) and the average FZ0 loss using parameters estimated over the whole sample period without consideration of change points (Loss_NC).

points detection. The solid line with slope denotes the points at which a change point could first be detected with a delay of zero. The red and yellow bubbles on the plot mark the change point dates as estimated in the expanding window procedure at 5% and 10% significance levels, respectively, with horizontal bands of bubbles indicating that the change points are detected in subsequent windows. In the figure, to add clarity, we add green bubbles only for the change points which are confirmed by red bubbles in subsequent windows. Thus, the length of the green bands shows the delay in detecting change points at 10% level, whilst the joint length of the green and yellow bands shows the delay in detecting change points at 5% level. In this study, we define the delay as the number of months between the point when a change point is first detected and the change point. The isolated bubbles outside the

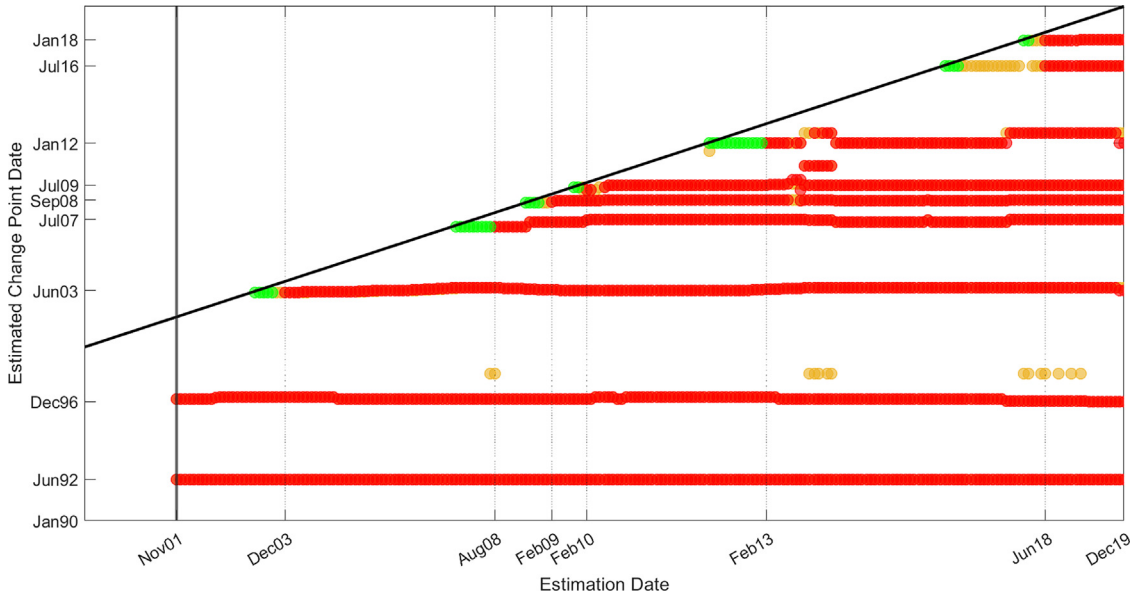

Fig. 5. Traffic light plot for the loss-based Wilcoxon test for 1% (VaR, ES) of daily S&P 500 index returns. *Note:* The 1% VaR and ES estimates of daily S&P 500 index returns are obtained by the GARCH skewed t model. The vertical solid line denotes the end point of the initial estimation period, the x-axis shows the estimation date, and the y-axis shows the dates of the estimated change points. The red (yellow) bubbles mark the change point dates estimated at 5% (10%) level. The length of green bands shows the delay in the change detection at 10% level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

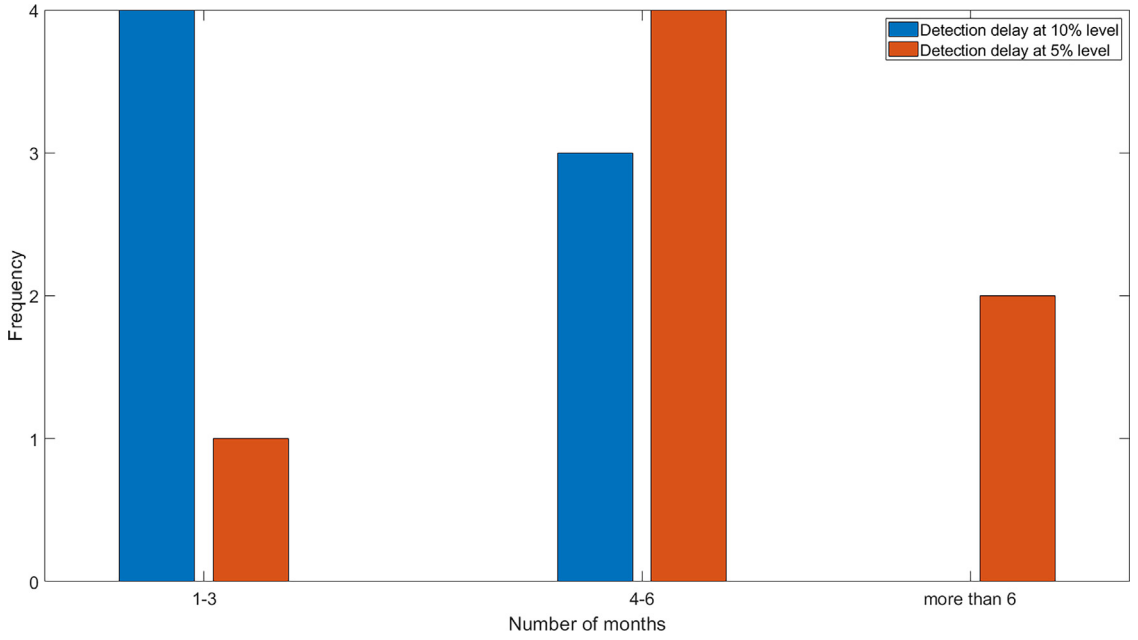

Fig. 6. Histograms of the delay of change detection in the 1% (VaR, ES) of daily S&P 500 index returns. *Note:* The 1% VaR and ES estimates of daily S&P 500 index returns are obtained by the GARCH skewed t model. This plot displays the histograms the delay in detecting change points, for the loss-based Wilcoxon test at 5% and 10% significance levels.

horizontal bands are treated as "false discoveries". Figure 6 shows the distribution of the delay in detecting change points (until first detection) at 5% and 10% significance levels. Overall, at 5% level, the length of delay is up to six months in about 70% of the cases, consistent with the results shown in Fig. 3.

5. Conclusions

We propose a new test, named the loss-based Wilcoxon test, to detect change points in the series of VaR and ES risk measures considered jointly. Our test is based on the Wilcoxon test (Dehling et al., 2013) applied to the FZ loss functions proposed by Fissler & Ziegel (2016). The framework of our test is general and can accommodate for any type of (semi)parametric estimation methods for VaR and ES. We perform extensive simulations based on various types of change point scenarios, including different locations for the change points and different changes in the volatility and DoF parameters. Our results show that the proposed test has better size under the null hypothesis and higher power properties under the considered alternative hypotheses, compared with five different alternative tests. We present an application of the loss-based Wilcoxon test on the S&P 500 index returns. The empirical results show that the test can detect the change points associated with well-known financial events.

Declaration of Competing Interest

None.

Appendix A. Proof of Theorem 2.1

Proof. In general, the Hoeffding decomposition can be applied to a U-statistic with a kernel $h(x,y)$, so that we have:

$$h(x,y) = \psi_0 + h_1(x) + h_2(y) + g(x,y),$$

where $\psi_0 = \mathbb{E}[h(X,Y)]$, $h_1(x) = \mathbb{E}[h(x,Y) - \psi_0]$, $h_2(y) = \mathbb{E}[h(X,y) - \psi_0]$ and $g(x,y) = h(x,y) - h_1(x) - h_2(y) - \psi_0$.

We have the properties for these three terms:

$$\mathbb{E}[h_1(X)] = \mathbb{E}[h_2(X)] = 0, \tag{A.1}$$

and

$$\mathbb{E}[g(x,Y)] = \mathbb{E}[g(X,y)] = 0. \tag{A.2}$$

The proof of Theorem 2.1 is based on a lemma introduced below.

Lemma A.1 (Dehling & Wendler, 2010)**.** *Let h be a $\mathscr{P}$-Lipschitz-continuous kernel with $2+\delta$ moments for some $\delta > 0$, $\{X_n\}_{n\in\mathbb{N}}$ be a stationary strong mixing process with $\mathbb{E}[|X_1|^\gamma] < \infty$ for some $\gamma > 0$ and $\alpha(T) = O(T^{-\rho})$ with $\rho > \frac{3\gamma\delta+2\delta+5\gamma+2}{2\gamma\delta}$, then for $V_T(g) = \frac{2}{\sqrt{T}(T-1)}\sum_{1\le i<j\le T} g(X_i,X_j)$, we have*

$$\mathbb{E}\left[TV_T^2(g)\right] \le \frac{4}{T(T-1)^2}\sum_{1\le i_1<i_2\le T}\sum_{1\le i_3<i_4\le T}|\mathbb{E}[g(X_{i_1},X_{i_2})g(X_{i_3},X_{i_4})]|$$
$$\le \frac{4}{T^3}\sum_{i_1,i_2,i_3,i_4=1}^{T}|\mathbb{E}[g(X_{i_1},X_{i_2})g(X_{i_3},X_{i_4})]| = O(T^{-\eta})$$

where $\eta = \min\left\{\rho\frac{2\gamma\delta}{3\gamma\delta+\delta+5\gamma+2} - 1, 1\right\} > 0.$

The proof of this lemma can be found in Dehling & Wendler (2010) as the proof for Lemma 3.6.

Recall $h_W(L_i,L_j)$ is antisymmetric with $\psi_0 = 0$. In order to prove the asymptotic normality of this U-process, we use the Hoeffding decomposition for the kernel $h_W(L_i,L_j)$:

$$h_W(L_i,L_j) = h_1(L_i) + h_2(L_j) + g(L_i,L_j).$$

Thus, based on (6), we have the decomposed U-process:

$$\frac{1}{T^{3/2}}U_T(\tau) = \frac{1}{T^{3/2}}\sum_{i=1}^{\lfloor\tau T\rfloor}\sum_{j=\lfloor\tau T\rfloor+1}^{T}\left[h_1(L_i) + h_2(L_j) + g(L_i,L_j)\right]$$
$$= \frac{1}{T^{3/2}}\left[(T-\lfloor\tau T\rfloor)\sum_{i=1}^{\lfloor\tau T\rfloor}h_1(L_i) + \lfloor\tau T\rfloor\sum_{j=\lfloor\tau T\rfloor+1}^{T}h_2(L_j) + \sum_{i=1}^{\lfloor\tau T\rfloor}\sum_{j=\lfloor\tau T\rfloor+1}^{T}g(L_i,L_j)\right]$$

By Lemma A.1, we have that for a given $\tau\in[0,1]$, the upper boundary of the variance of $\frac{1}{T^{3/2}}\sum_{i=1}^{\lfloor\tau T\rfloor}\sum_{j=\lfloor\tau T\rfloor+1}^{T}g(L_i,L_j)$:

$$\frac{1}{T^3}\sum_{\substack{i_1=1:\lfloor\tau T\rfloor\\ i_2=\lfloor\tau T\rfloor+1:T}}\sum_{\substack{i_3=1:\lfloor\tau T\rfloor\\ i_4=\lfloor\tau T\rfloor+1:T}}\left|\mathbb{E}\left[g(L_{i_1},L_{i_2})g(L_{i_3},L_{i_4})\right]\right|$$
$$\le \frac{1}{T^3}\sum_{1\le i_1<i_2\le T}\sum_{1\le i_3<i_4\le T}|\mathbb{E}[g(L_{i_1},L_{i_2})g(L_{i_3},L_{i_4})]|$$
$$< \frac{1}{T^3}\sum_{i_1,i_2,i_3,i_4=1}^{T}|\mathbb{E}[g(L_{i_1},L_{i_2})g(L_{i_3},L_{i_4})]| = O(T^{-\eta}). \tag{A.3}$$

Hence, the variance of $\frac{1}{T^{3/2}}\sum_{i=1}^{\lfloor\tau T\rfloor}\sum_{j=\lfloor\tau T\rfloor+1}^{T}g(L_i,L_j)$ vanishes as T increases.

By (A.2) and (A.3), we have

$$\frac{1}{T^{3/2}}\sup_{0\le\tau\le1}\left|\sum_{i=1}^{\lfloor\tau T\rfloor}\sum_{j=\lfloor\tau T\rfloor+1}^{T}g(L_i,L_j)\right| \to 0$$

in probability.

Thus, by the Lemma of Slutsky, it is enough to show that the sum of the first two terms

$$\left[\frac{T-\lfloor\tau T\rfloor}{T^{3/2}}\sum_{i=1}^{\lfloor\tau T\rfloor}h_1(L_i) + \frac{\lfloor\tau T\rfloor}{T^{3/2}}\sum_{j=\lfloor\tau T\rfloor+1}^{T}h_2(L_j)\right]_{0\le\tau\le1}$$

converges in distribution to the limit process of Theorem 2.1. Because the kernel $h_W(L_i,L_j)$ is antisymmetric, we have that $h_2(L_j) = -h_1(L_j)$. Thus, we can rewrite the representation as

$$\frac{T-\lfloor\tau T\rfloor}{T^{3/2}}\sum_{i=1}^{\lfloor\tau T\rfloor}h_1(L_i) - \frac{\lfloor\tau T\rfloor}{T^{3/2}}\sum_{i=\lfloor\tau T\rfloor+1}^{T}h_1(L_i)$$
$$= \frac{1}{T^{1/2}}\sum_{i=1}^{\lfloor\tau T\rfloor}h_1(L_i) - \frac{\lfloor\tau T\rfloor}{T^{3/2}}\sum_{i=1}^{T}h_1(L_i).$$

To obtain the limit of the process, we state the theorem below, which is a direct consequence of Theorem 4 in Borovkova et al. (2001) and Theorem 3.1 in Davidson & De Jong (2000).

Theorem A.1. *Let $\{Y_k\}_{k\in\mathbb{Z}}$ be a L_2 near-epoch dependent (NED) with respect to a strong mixing process. Also, suppose that $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[|Y_i|^{4+\delta}] \le \infty$ for some $\delta > 0$. Then, as $T\to\infty$,*

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{T}Y_i \xrightarrow{d} \mathcal{N}(0,\sigma^2),$$

where $\sigma^2 = Var(Y_1) + 2\sum_{k=2}^{\infty}Cov(Y_1,Y_k)$.

The proof of the theorem follows immediately from Borovkova et al. (2001) and Davidson & De Jong (2000).

Based on Assumption 2.1 (B), applying Theorem A.1 on the partial sum process and using similar arguments as in Chapter 4 of Csörgő & Horváth (1997) and Donsker's theorem, it can be shown that $\frac{1}{T^{1/2}}\sum_{i=1}^{\lfloor\tau T\rfloor}h_1(L_i) - \frac{\lfloor\tau T\rfloor}{T^{3/2}}\sum_{i=1}^{T}h_1(L_i)$ converges to a limit process $\{\sigma_W[W(\tau)-\tau W(1)]\}_{0\le\tau\le1}$, where $\{W(\tau)\}_{0\le\tau\le1}$ is a Wiener process, and

$$\sigma_W^2 = Var(h_1(L_1)) + 2\sum_{k=2}^{\infty}Cov(h_1(L_1),h_1(L_k)).$$

Additionally, we have that $h_1(x) = \frac{1}{2} - F(x)$. Thus,

$$\sigma_W^2 = Var(F(L_1)) + 2\sum_{k=2}^{\infty}Cov(F(L_1),F(L_k)).$$

By the Lemma of Slutsky, we obtain that as $T\to\infty$, $\frac{1}{T^{3/2}}U_T(\tau)$ converges in distribution to $\{\sigma_W B(\tau)\}_{0\le\tau\le1}$, where $B(\tau) = W(\tau) - \tau W(1)$ is a Brownian bridge and $\sigma_W^2 = Var(F(L_1)) + 2\sum_{k=2}^{\infty}Cov(F(L_1),F(L_k))$. □

Appendix B. Proof of Theorem 2.2

In the following, we discuss the $\mathscr{P}$-Lipschitz-continuity property for the kernel $h_W(X,Y)$.

Definition B.1. **($\mathscr{P}$-Lipschitz-continuity)** Let $\{X_t\}_{t\in\mathbb{N}}$ be a stationary process. A kernel h is called $\mathscr{P}$-Lipschitz-continuous if there is a constant $a > 0$ with

$$\mathbb{E}\left[\left|h(X,Y) - h(X',Y)\right|\mathbf{1}[|X-X'|\le\epsilon]\right] \le a\epsilon,$$

for every $\epsilon > 0$, every pair X and Y with the common distribution $\mathscr{P}_{X_1,X_k}$ for $k \in \mathbb{N}$ with $k > 1$ or $\mathscr{P}_{X_1} \times \mathscr{P}_{X_1}$ and X' and Y also with one of these common distributions.

To prove Theorem 2.2, we need the following preliminary result.

Proposition B.1. *If Assumption 2.1 (B) holds, then the antisymmetric kernel $h_W(X,Y) = \mathbf{1}[X \le Y] - \frac{1}{2}$ for the test statistic is $\mathscr{P}$-Lipschitz-continuous.*

The proof of this proposition can be found below.

Proof. The kernel $h_W(X,Y)$ is $\mathscr{P}$-Lipschitz-continuous, if there is a constant $a > 0$, so that for all $\epsilon > 0$ and every common distribution of X, X' and Y,

$$\mathbb{E}\Big[|h_W(X,Y) - h_W(X',Y)|\mathbf{1}\Big[|X - X'| \le \epsilon\Big]\Big] < a\epsilon.$$

For random variables X, X' and Y, we have:

$$\mathbb{E}\Big[|h_W(X,Y) - h_W(X',Y)|\mathbf{1}\Big[|X - X'| \le \epsilon\Big]\Big]$$
$$= \mathbb{E}\Big[|\mathbf{1}[X \le Y] - \mathbf{1}\Big[X' \le Y\Big]|\mathbf{1}\Big[|X - X'| \le \epsilon\Big]\Big].$$

We have:

$$\mathbb{E}\Big[|\mathbf{1}[X \le Y] - \mathbf{1}\Big[X' \le Y\Big]|\mathbf{1}\Big[|X - X'| \le \epsilon\Big]\Big] \le P(-\epsilon \le X - X' \le \epsilon).$$

Based on Assumption 2.1 (B) on the continuous distribution function, there exists a constant $a = 2\sup(f)$ that satisfies the following:

$$P(X' - \epsilon \le X \le X' + \epsilon) = F(X' + \epsilon) - F(X' - \epsilon)$$
$$= \int_{X'-\epsilon}^{X'+\epsilon} f(t)dt \le \frac{a}{2} \cdot 2\epsilon = a\epsilon.$$

Thus, based on Definition B.1, the antisymmetric kernel of the Wilcoxon test statistic $h_W(X,Y)$ is $\mathscr{P}$-Lipschitz-continuous. □

Now, we can turn our attention to the proof of Theorem 2.2.

Proof. In order to obtain the asymptotic behavior of the bootstrapped U-process, we use the Hoeffding decomposition for the bootstrapped kernel $h_W(L_i^*, L_j^*)$:

$$h_W(L_i^*, L_j^*) = h_1(L_i^*) + h_2(L_j^*) + g(L_i^*, L_j^*).$$

Thus, we have the decomposed bootstrapped U-process:

$$\frac{1}{T^{3/2}}U_T^*(\tau) = \frac{1}{T^{3/2}}\sum_{i=1}^{\lfloor \tau T \rfloor}\sum_{j=\lfloor \tau T \rfloor+1}^{T}\Big[h_1(L_i^*) + h_2(L_j^*) + g(L_i^*, L_j^*)\Big]$$
$$= \frac{1}{T^{3/2}}\Bigg[(T - \lfloor \tau T \rfloor)\sum_{i=1}^{\lfloor \tau T \rfloor} h_1(L_i^*) + \lfloor \tau T \rfloor \sum_{j=\lfloor \tau T \rfloor+1}^{T} h_2(L_j^*)$$
$$+ \sum_{i=1}^{\lfloor \tau T \rfloor}\sum_{j=\lfloor \tau T \rfloor+1}^{T} g(L_i^*, L_j^*)\Bigg]. \tag{B.1}$$

In the following, we are going to use the result below:

Lemma B.1 (Hwang & Shin, 2015)**.** *Let h be a $\mathscr{P}$-Lipschitz-continuous kernel with $2+\delta$ moments for some $\delta > 0$, $\{X_n^*\}_{n\in\mathbb{N}}$ be a stationary bootstrapped strong mixing process with $\mathbb{E}[|X_1^*|^\gamma] < \infty$ for some $\gamma > 0$ and $\alpha(T) = O(T^{-\rho})$ with $\rho > \frac{3\gamma\delta+2\delta+5\gamma+2}{2\gamma\delta}$, then for $V_T^*(g) = \frac{2}{\sqrt{T}(T-1)}\sum_{1\le i<j\le T} g(X_i^*, X_j^*)$:*

$$\mathbb{E}\Big[TV_T^{*2}(g)\Big] = O(T^{-\eta}),$$

where $\eta = \min\left\{\rho\frac{2\gamma\delta}{3\gamma\delta+\delta+5\gamma+2} - 1, 1\right\} > 0.$

The proof of this lemma can be found in Hwang & Shin (2015) as the proof for Lemma 2.

As shown in Lemma B.1, the variance of the last term in (B.1) vanishes as T increases:

$$Var^*\left(\frac{1}{T^{3/2}}\sum_{i=1}^{\lfloor \tau T \rfloor}\sum_{j=\lfloor \tau T \rfloor+1}^{T} g(L_i^*, L_j^*)\right) \xrightarrow{P} 0.$$

Thus, by Lemma of Slutsky and the property of kernel shown in (A.2), it is enough to show that

$$\left[\frac{T - \lfloor \tau T \rfloor}{T^{3/2}}\sum_{i=1}^{\lfloor \tau T \rfloor} h_1(L_i^*) + \frac{\lfloor \tau T \rfloor}{T^{3/2}}\sum_{j=\lfloor \tau T \rfloor+1}^{T} h_2(L_j^*)\right]_{0\le\tau\le1}$$

converges in distribution to the limit process of

$$\left[\frac{T - \lfloor \tau T \rfloor}{T^{3/2}}\sum_{i=1}^{\lfloor \tau T \rfloor} h_1(L_i) + \frac{\lfloor \tau T \rfloor}{T^{3/2}}\sum_{j=\lfloor \tau T \rfloor+1}^{T} h_2(L_j)\right]_{0\le\tau\le1}.$$

Because the kernel $h_W(L_i^*, L_j^*)$ is antisymmetric, we have that $h_2(L_j^*) = -h_1(L_j^*)$. Thus, we can rewrite the representation as:

$$\frac{T - \lfloor \tau T \rfloor}{T^{3/2}}\sum_{i=1}^{\lfloor \tau T \rfloor} h_1(L_i^*) - \frac{\lfloor \tau T \rfloor}{T^{3/2}}\sum_{i=\lfloor \tau T \rfloor+1}^{T} h_1(L_i^*)$$
$$= \frac{1}{T^{1/2}}\sum_{i=1}^{\lfloor \tau T \rfloor} h_1(L_i^*) - \frac{\lfloor \tau T \rfloor}{T^{3/2}}\sum_{i=1}^{T} h_1(L_i^*).$$

To obtain the limit of the process, we restate the theorems of Calhoun (2018).

Theorem B.1. *Let $\{Y_k\}_{k\in\mathbb{Z}}$ be a L_2 near-epoch dependent (NED) with respect to a strong mixing process. Additionally, suppose that $\mu_{nt} - \bar{\mu}_n$ is uniformly bounded, where $\mu_{nt} = \mathbb{E}[Y_{nt}]$ and $\bar{\mu}_n = n^{-1}\sum_{t=1}^{n}\mu_{nt}$. Then we have:*

$$\sup_{x\in\mathbb{R}}\left|P^*(\sqrt{n}(\bar{Y}_n^* - \mathbb{E}^*[\bar{Y}_n^*]) \le x) - P(\sqrt{n}(\bar{Y}_n - \mathbb{E}[\bar{Y}_n]) \le x)\right| \xrightarrow{P} 0,$$

where $\bar{Y}_n = \frac{1}{n}\sum_{t=1}^{n} Y_{nt}$, and $\bar{Y}_n^* = \frac{1}{n}\sum_{t=1}^{n} Y_{nt}^*$.

Theorem B.2. *Suppose that the conditions of Theorem B.1 hold and let d be any distance function that metricizes weak convergence. Then we have:*

$$P^*(d(Z_n^*, \sigma W) > \delta) \xrightarrow{P} 0, \tag{B.2}$$

for all positive δ, where $Z_n^*(\tau) = \frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor \tau n \rfloor}(Y_{nt}^* - \mathbb{E}^*[\bar{Y}_n^*])$, and σW denotes a Brownian motion scaled by the positive constant σ. If, in addition, $\sup_{t=1,\dots,n}|\mu_{nt} - \bar{\mu}_n| = o(1/\sqrt{n})$ and

$$n^{-1}\sum_{s,t=1}^{\lfloor \gamma n \rfloor} Cov(Y_{ns}, Y_{nt}) \to \sigma^2\gamma$$

for all $\gamma \in [0,1]$, then

$$P^*(d(Z_n, \sigma W) > \delta) \xrightarrow{P} 0, \tag{B.3}$$

for any positive δ, where $Z_n(\tau) = \frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor \tau n \rfloor}(Y_{nt} - \bar{\mu}_n)$.

Since both (B.2) and (B.3) hold, the distribution of bootstrapped values Z_n^* can be used to approximate the distribution of Z_n, because they have the same distribution asymptotically.

The assumptions listed in Theorem B.1 of Calhoun (2018) are satisfied under Assumptions 2.1 and 2.2 in our study. Applying Theorem B.1 and Theorem B.2 for $h_1(L_t)$, we have:

$$\sup_{x\in\mathbb{R}}\left|P^*\left(\frac{\lfloor \tau T \rfloor}{T^{3/2}}\sum_{i=1}^{T} h_1(L_i^*) \le x\right) - P\left(\frac{\lfloor \tau T \rfloor}{T^{3/2}}\sum_{i=1}^{T} h_1(L_i) \le x\right)\right| \xrightarrow{P} 0,$$

$$\sup_{x\in\mathbb{R}}\left|P^*\left(\frac{1}{T^{1/2}}\sum_{i=1}^{\lfloor \tau T \rfloor} h_1(L_i^*) \le x\right) - P\left(\frac{1}{T^{1/2}}\sum_{i=1}^{\lfloor \tau T \rfloor} h_1(L_i) \le x\right)\right| \xrightarrow{P} 0,$$

As such, based on the Lemma of Slutsky, we have:

$$\sup_{x\in\mathbb{R}}\left|P^*\left(\frac{1}{T^{1/2}}\sum_{i=1}^{\lfloor\tau T\rfloor}h_1(L_i^*)-\frac{\lfloor\tau T\rfloor}{T^{3/2}}\sum_{i=1}^{T}h_1(L_i^*)\le x\right)\right.$$
$$\left.-P\left(\frac{1}{T^{1/2}}\sum_{i=1}^{\lfloor\tau T\rfloor}h_1(L_i)-\frac{\lfloor\tau T\rfloor}{T^{3/2}}\sum_{i=1}^{T}h_1(L_i)\le x\right)\right|\xrightarrow{P}0,$$

Thus, we obtain the convergence in probability in (11):

$$\sup_{x\in\mathbb{R}}|P^*(T^{-3/2}U_T^*(\tau)\le x)-P(T^{-3/2}U_T(\tau)\le x)|\xrightarrow{P}0.$$

□

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2023.03.033.

References

Andersson, E., Bock, D., & Frisén, M. (2006). Some statistical aspects of methods for detection of turning points in business cycles. *Journal of Applied Statistics, 33*(3), 257–278.

Andreou, E., & Ghysels, E. (2002). Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics, 17*(5), 579–600.

Aue, A., Hörmann, S., Horváth, L., & Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics, 37*(6B), 4046–4087.

Aue, A., & Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis, 34*(1), 1–16.

Bali, T. G., & Theodossiou, P. (2007). A conditional-SGT-VaR approach with alternative GARCH models. *Annals of Operations Research, 151*(1), 241–267.

Barassi, M., Horváth, L., & Zhao, Y. (2020). Change-point detection in the conditional correlation structure of multivariate volatility models. *Journal of Business and Economic Statistics, 38*(2), 340–349.

Barendse, S., & Patton, A. J. (2022). Comparing predictive accuracy in the presence of a loss function shape parameter. *Journal of Business and Economic Statistics, 40*(3), 1057–1069.

Basel Committee on Banking Supervision (2019). *Minimum capital requirements for market risk*. Bank for International Settlements. URL: https://www.bis.org/bcbs/publ/d457.pdf

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

Berkes, I., Horváth, L., & Kokoszka, P. (2003). Estimation of the maximal moment exponent of a GARCH (1, 1) sequence. *Econometric Theory, 19*(4), 565–586.

Betken, A. (2016). Testing for change-points in long-range dependent time series by means of a self-normalized Wilcoxon test. *Journal of Time Series Analysis, 37*(6), 785–809.

Bonaccolto, G., Caporin, M., & Maillet, B. B. (2022). Dynamic large financial networks via conditional expected shortfalls. *European Journal of Operational Research, 298*(1), 322–336.

Borovkova, S., Burton, R., & Dehling, H. (2001). Limit theorems for functionals of mixing processes with applications to U-statistics and dimension estimation. *Transactions of the American Mathematical Society, 353*(11), 4261–4318.

Calhoun, G. (2018). Block bootstrap consistency under weak assumptions. *Econometric Theory, 34*(6), 1383–1406.

Chen, B., & Hong, Y. (2016). Detecting for smooth structural changes in GARCH models. *Econometric Theory, 32*(3), 740–791.

Chen, Q., & Fang, Z. (2019). Improved inference on the rank of a matrix. *Quantitative Economics, 10*(4), 1787–1824.

Clements, M. P., & Hendry, D. F. (1996). Intercept corrections and structural change. *Journal of Applied Econometrics, 11*(5), 475–494.

Csörgő, M., & Horváth, L. (1988). Invariance principles for changepoint problems. *Journal of Multivariate Analysis, 27*(1), 151–168.

Csörgő, M., & Horváth, L. (1997). *Limit theorems in change-point analysis*. John Wiley & Sons Chichester.

Davidson, J., & De Jong, R. M. (2000). The functional central limit theorem and weak convergence to stochastic integrals II: Fractionally integrated processes. *Econometric Theory, 16*(5), 643–666.

Dehling, H., Rooch, A., & Taqqu, M. S. (2013). Non-parametric change-point tests for long-range dependent data. *Scandinavian Journal of Statistics, 40*(1), 153–173.

Dehling, H., Rooch, A., Taqqu, M. S., et al., (2017). Power of change-point tests for long-range dependent data. *Electronic Journal of Statistics, 11*(1), 2168–2198.

Dehling, H., & Wendler, M. (2010). Central limit theorem and the bootstrap for U-statistics of strongly mixing data. *Journal of Multivariate Analysis, 101*(1), 126–137.

Dette, H., & Gösmann, J. (2020). A likelihood ratio approach to sequential change point detection for a general class of parameters. *Journal of the American Statistical Association, 115*(531), 1361–1377.

Diebold, F. X., & Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics, 105*(1), 131–159.

Dimitriadis, T., & Bayer, S. (2019). A joint quantile and expected shortfall regression framework. *Electronic Journal of Statistics, 13*(1), 1823–1871.

Dimitriadis, T., & Schnaitmann, J. (2021). Forecast encompassing tests for the expected shortfall. *International Journal of Forecasting, 37*(2), 604–621.

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

Fan, L., Glynn, P. W., & Pelger, M. (2018). Change-point testing and estimation for risk measures in time series. arXiv preprint arXiv:1809.02303

Fearnhead, P., & Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association, 114*(525), 169–183.

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

Francq, C., & Zakoïan, J.-M. (2015). Risk-parameter estimation in volatility models. *Journal of Econometrics, 184*(1), 158–173.

Gerlach, R., & Wang, C. (2020). Semi-parametric dynamic asymmetric laplace models for tail risk forecasting, incorporating realized measures. *International Journal of Forecasting, 36*(2), 489–506.

Gerstenberger, C. (2018). Robust Wilcoxon-type estimation of change-point location under short-range dependence. *Journal of Time Series Analysis, 39*(1), 90–104.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review, 35*(3), 705–730.

Harvey, C. R., & Liu, Y. (2020). False (and missed) discoveries in financial economics. *The Journal of Finance, 75*(5), 2503–2553.

Hoga, Y. (2017). Testing for changes in (extreme) VaR. *The Econometrics Journal, 20*(1), 23–51.

Horváth, L., Kokoszka, P., & Wang, S. (2021). Monitoring for a change point in a sequence of distributions. *The Annals of Statistics, 49*(4), 2271–2291.

Horváth, L., Liu, Z., & Lu, S. (2022). Sequential monitoring of changes in dynamic linear models, applied to the US housing market. *Econometric Theory, 38*(2), 209–272.

Horváth, L., Liu, Z., Rice, G., & Wang, S. (2020a). Sequential monitoring for changes from stationarity to mild non-stationarity. *Journal of Econometrics, 215*(1), 209–238.

Horváth, L., Miller, C., & Rice, G. (2020b). A new class of change point test statistics of Rényi type. *Journal of Business and Economic Statistics, 38*(3), 570–579.

Hušková, M., & Kirch, C. (2008). Bootstrapping confidence intervals for the change–point of time series. *Journal of Time Series Analysis, 29*(6), 947–972.

Hwang, E., & Shin, D. W. (2015). Stationary bootstrap for U-statistics under strong mixing. *Communications for Statistical Applications and Methods, 22*(1), 81–93.

Inclan, C., & Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association, 89*(427), 913–923.

Ji, Q., Zhang, D., & Zhao, Y. (2020). Searching for safe-haven assets during the COVID-19 pandemic. *International Review of Financial Analysis, 71*, 101526.

Lazar, E., & Xue, X. (2020). Forecasting risk measures using intraday data in a generalized autoregressive score framework. *International Journal of Forecasting, 36*(3), 1057–1072.

Leung, M., Li, Y., Pantelous, A. A., & Vigne, S. A. (2021). Bayesian value-at-risk backtesting: The case of annuity pricing. *European Journal of Operational Research, 293*(2), 786–801.

Loschi, R. H., Iglesias, P. L., Arellano-Valle, R. B., & Cruz, F. R. B. (2007). Full predictivistic modeling of stock market data: application to change point problems. *European Journal of Operational Research, 180*(1), 282–291.

Meng, X., & Taylor, J. W. (2020). Estimating value-at-risk and expected shortfall using the intraday low and range data. *European Journal of Operational Research, 280*(1), 191–202.

Mikosch, T., & Stărică, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics, 86*(1), 378–390.

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

Pástor, L., & Stambaugh, R. F. (2001). The equity premium and structural breaks. *The Journal of Finance, 56*(4), 1207–1239.

Patton, A., Politis, D. N., & White, H. (2009). Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. *Econometric Reviews, 28*(4), 372–375.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics, 38*(4), 796–809.

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

Pesaran, M. H., & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics, 137*(1), 134–161.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association, 89*(428), 1303–1313.

Politis, D. N., & White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews, 23*(1), 53–70.

Qu, Z. (2008). Testing for structural change in regression quantiles. *Journal of Econometrics, 146*(1), 170–184.

Quaedvlieg, R. (2021). Multi-horizon forecast comparison. *Journal of Business and Economic Statistics, 39*(1), 40–53.

Shao, X., & Zhang, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association, 105*(491), 1228–1240.

Smith, S. C., & Timmermann, A. (2021). Break risk. *The Review of Financial Studies, 34*(4), 2045–2100.

Stock, J. H., & Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics, 14*(1), 11–30.

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

Vogelsang, T. J. (1999). Sources of nonmonotonic power when testing for a shift in mean of a dynamic time series. *Journal of Econometrics, 88*(2), 283–299.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the *k* largest observations. *Journal of the American Statistical Association, 73*(364), 812–815.

Ye, W., Liu, X., & Miao, B. (2012). Measuring the subprime crisis contagion: Evidence of change point analysis of copula functions. *European Journal of Operational Research, 222*(1), 96–103.

Zhang, T., & Lavitas, L. (2018). Unsupervised self-normalized change-point testing for time series. *Journal of the American Statistical Association, 113*(522), 637–648.