SFB 823

Tests for scale changes based on pairwise differences

Discussion Paper

Carina Gerstenberger, Daniel Vogel, Martin Wendler

Nr. 82/2016

Tests for scale changes based on pairwise differences

Carina Gerstenberger, Daniel Vogel, and Martin Wendler

November 13, 2016

In many applications it is important to know whether the amount of fluctuation in a series of observations changes over time. In this article, we investigate different tests for detecting change in the scale of mean-stationary time series. The classical approach based on the CUSUM test applied to the squared centered, is very vulnerable to outliers and impractical for heavy-tailed data, which leads us to contemplate test statistics based on alternative, less outlier-sensitive scale estimators.

It turns out that the tests based on Gini's mean difference (the average of all pairwise distances) or generalized Q_n estimators (sample quantiles of all pairwise distances) are very suitable candidates. They improve upon the classical test not only under heavy tails or in the presence of outliers, but also under normality. An explanation for this counterintuitive result is that the corresponding long-run variance estimates are less affected by a scale change than in the case of the sample-variance-based test.

We use recent results on the process convergence of U-statistics and U-quantiles for dependent sequences to derive the limiting distribution of the test statistics and propose estimators for the long-run variance. We perform a simulation study to investigate the finite sample behavior of the tests and their power. Furthermore, we demonstrate the applicability of the new change-point detection methods at two real-life data examples from hydrology and finance.

Keywords: Asymptotic relative efficiency, Change-point analysis, Gini's mean difference, Long-run variance estimation, Median absolute deviation, Q_n scale estimator, U-quantile, U-statistic

1. Introduction

The established approach to testing for changes in the scale of a univariate time series $X_1, \ldots, X_n$ is a CUSUM test applied to the squares of the centered observations, which may be written as

$$\hat{T}_{\sigma^2} = \max_{1 \leq k \leq n} \frac{k}{\sqrt{n}} \left| \hat{\sigma}_{1:k}^2 - \hat{\sigma}_{1:n}^2 \right|, \tag{1}$$

where $\hat{\sigma}_{i:j}^2$ denotes the sample variance computed from the observations $X_i, \ldots, X_j$, $1 \leq i < j \leq n$. To carry out the test in practice, the test statistic is usually divided by (the square root of) a suitable estimator of the corresponding long-run variance, cf. (8).

This has first been considered by Inclan and Tiao (1994), who derive asymptotics for centered, normal, i.i.d. data. It has subsequently been extended by several authors to broader situations, e.g., Gombay et al. (1996) allow the mean to be unknown and also propose a weighted version of the testing procedure, Lee and Park (2001) extend it to linear processes, and Wied et al. (2012) study the test for sequences that are L_2 NED on α-mixing processes. A multivariate version was considered by Aue et al. (2009).

The test statistic (1) is prone to outliers. This has already been remarked by Inclan and Tiao (1994) and has led Lee and Park (2001) to consider a version of the test using trimmed observations. Outliers may affect the test decision in both directions: A single outlier suffices to make the test reject the null hypothesis at an otherwise stationary sequence, but more often one finds that outliers mask a change, and the test is generally very inefficient at heavy-tailed population distributions. An intuitive explanation is that, while outliers blow up the test statistic $\hat{T}_{\sigma^2}$, they do even more so blow up the long-run variance estimate, by which the test statistic is divided.[1]

Writing the test statistic as in (1) suggests this behavior may be largely attributed to the use of the sample variance as a scale estimator. The recognition of the extreme "non-robustness" of the sample variance and derived methods, in fact, stood at the beginning of the development of the area of robust statistics as a whole (e.g. Box, 1953; Tukey, 1960). Thus, an intuitive way of constructing robust scale change-point tests is to replace the sample variance in (1) by an alternative scale measure.

We consider several popular scale estimators: the mean absolute deviation, the median absolute deviation (MAD), the mean of all pairwise differences (Gini's mean difference) and sample quantiles of all pairwise differences. All of them allow an explicit formula and are computable in a finite number of steps. Particularly the latter two, the mean as well as sample quantiles of the pairwise differences are promising candidates as they are almost as efficient as the standard deviation at normality and, hence, the improvement in robustness is expected to come at practically no loss in terms of power under normality. In fact, as it turns out, these tests can have a better power than the variance-based test also under normality.

The paper is organized as follows: In Section 2 we review the properties of the scale estimators and detail on our choice to particularly consider pairwise-difference-based estimators. Section 3 states the test statistics and long-run variance estimators and contains asymptotic results. Section 4 addresses the question which sample quantile of the pairwise differences is most appropriate for the change-point problem. Section 5 presents simulation results. Section 6 illustrates the behavior of the tests at informative real-life data examples. We draw our conclusions in Section 7. Appendix A contains supplementary material for Section 4. Proofs are deferred to the Appendix B.

---

[1]This applies in principle to any estimation method, bootstrapping or subsampling, not only to the kernel estimation method employed in the present article.

2. Scale Estimators

We use $\mathcal{L}(X)$ to denote the *law*, i.e., the distribution, of any random variable X. We call any function $s : \mathcal{F} \to [0, \infty]$, where $\mathcal{F}$ is the set of all univariate distributions F, a *scale measure* (or, analogously, a *dispersion measure*) if it satisfies

$$s(\mathcal{L}(aX + b)) = |a| s(\mathcal{L}(X)) \qquad \text{for all } a, b \in \mathbb{R}. \tag{2}$$

Although not being a scale measure itself, the variance $\sigma^2 = E(X - EX)^2$ is, in a lax use of the term, often referred to as such, since it is closely related to the standard deviation $\sigma = \sqrt{\sigma^2}$, which is a proper scale measure in the above sense. A *scale estimator* s_n is then generally understood as the scale measure s applied to the empirical distribution F_n corresponding to the data set $\mathbb{X}_n = (X_1, \ldots, X_n)$. However, it is common in many situations to define the finite-sample version of the scale measure slightly different, due to various reasons. For instance, the sample variance $\sigma_n^2 = \{(n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2\}^{1/2}$ has the factor $(n-1)^{-1}$ due to the thus obtained unbiasedness.

A concept that we will refer to repeatedly throughout is asymptotic relative efficiency. Letting s_n be a scale estimator and s the corresponding population value, the asymptotic variance $ASV(s_n) = ASV(s_n; F)$ of s_n at the distribution F is defined as the variance of the limiting normal distribution of $\sqrt{n}(s_n - s)$, when s_n is evaluated at an independent sample $X_1, \ldots, X_n$ drawn from F. Generally, for two consistent estimators a_n and b_n estimating the same quantity $\theta \in \mathbb{R}$, i.e., $a_n \xrightarrow{\text{P}} \theta$ and $b_n \xrightarrow{\text{P}} \theta$, the asymptotic relative efficiency of a_n with respect to b_n at distribution F is defined as

$$ARE(a_n, b_n; F) = ASV(b_n; F) / ASV(a_n; F).$$

In order to make two scale estimators $s_n^{(1)}$ and $s_n^{(2)}$ comparable efficiency-wise, we have to normalize them appropriately, and define their asymptotic relative efficiency at the population distribution F as

$$ARE(s_n^{(1)}, s_n^{(2)}; F) = \frac{ASV(s_n^{(2)}; F)}{ASV(s_n^{(1)}; F)} \left\{ \frac{s^{(1)}(F)}{s^{(2)}(F)} \right\}^2, \tag{3}$$

where $s^{(1)}(F)$ and $s^{(2)}(F)$ denote the corresponding population values of the scale estimators $s_n^{(1)}$ and $s_n^{(2)}$, respectively.

In the following, we review some basic properties of four scale estimators: the mean deviation, the median absolute deviation (MAD), Gini's mean difference and the Q_n scale estimator proposed by Rousseeuw and Croux (1992).

Let $\text{md}(F)$ denote the median of the distribution F, i.e., the center point of the interval $\{x \in \mathbb{R} | F(x-) \leq 1/2 \leq F(x)\}$, where $F(x-)$ denotes the left-hand limit. We define the mean deviation as $d(F) = E|X - \text{md}(F)|$ and its empirical version as $d_n = \frac{1}{n-1} \sum_{i=1}^{n} |X_i - \text{md}(F_n)|$.

The question of whether to prefer the standard deviation or the mean deviation has become known as Eddington–Fisher debate. The tentative winner was the standard deviation after Fisher (1920) showed that its asymptotic relative efficiency with respect to

3

the mean deviation is 114% at normality. However, Tukey (1960) pointed out that it is less efficient than the mean deviation if the normal distribution is slightly contaminated. Thus, the mean deviation appears to be a suitable candidate scale estimator for constructing less outlier-sensitive change-point tests. Gerstenberger and Vogel (2015) argue that, when pondering the mean deviation instead of the standard deviation for robustness reasons, it may be better to use Gini's mean difference $g_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} |X_i - X_j|$, i.e., the mean of the absolute distances of all pairs of observations. The population version is $g(F) = E|X - Y|$, where $X, Y \sim F$ are independent. Gini's mean difference has qualitatively the same robustness under heavy-tailed distributions as the mean deviation, but retains an asymptotic relative efficiency with respect to the standard deviation of 98% at the normal distribution (Nair, 1936).

Both estimators, mean deviation and Gini's mean difference, improve upon the variance and the standard deviation in terms of robustness, but are not robust in a modern understanding of the term. They both have an unbounded influence function and an asymptotic breakdown point of zero. Since robustness is, at least initially, our main motivation, it is natural to consider estimators that have been suggested particularly for that purpose. A common highly robust scale estimator is the median absolute deviation (MAD), popularized by Hampel (1974). The population value $m(F)$ is the median of the distribution of $|X - \mathrm{md}(F)|$ and the sample version $m_n = m_n(\mathbb{X}_n)$ is the median of the values $|X_i - \mathrm{md}(F_n)|$, $1 \leq i \leq n$. The MAD has a bounded influence function (see Huber and Ronchetti, 2009) and an asymptotic breakdown point of 50%. Its main drawback is its poor asymptotic efficiency under normality, which is only 37% as compared to the standard deviation. It is also unsuitable for change-in-scale detection due to other reasons that will be detailed in Sections 4 and 5.

Similarly, to going from the *mean* absolute deviation to the *median* absolute deviation, we may consider the median, or more generally any sample α-quantile, of all pairwise differences. We call this estimator Q_n^α and the corresponding population scale measure Q^α, i.e. $Q^\alpha = U^{-1}(\alpha) = \inf\{x \,|\, \alpha \leq U(x)\}$, where U is the distribution function of $|X - Y|$ for $X, Y \sim F$ independent, and U^{-1} the corresponding quantile function. For the precise definition of Q_n^α, any sensible definition of the sample quantile can be employed. See, e.g., the nine different versions the R function quantile() offers. The asymptotic results we derive later are not affected, and any practical differences turn out to be negligible in the current context. So merely for simplicity, we define $Q_n^\alpha = U_n^{-1}(\alpha)$, where U_n is the empirical distribution function associated with the sample $|X_i - X_j|$, $1 \leq i < j \leq n$. Thus letting $D_1, \ldots, D_{\binom{n}{2}}$ be the elements of $\{|X_i - X_j| \,|\, 1 \leq i < j \leq n\}$ in ascending order, we have $Q_n^\alpha = D_{\alpha\binom{n}{2}}$ if $\alpha\binom{n}{2}$ is integer and $Q_n^\alpha = D_{\lceil \alpha\binom{n}{2} \rceil}$ otherwise.[2]

In case of Gini's mean difference, we observed that the transition from the average distance from the symmetry center to the average pairwise distance led to an increase in efficiency under normality. The effect is even more pronounced for the median distances, we have $ARE(Q_n^{0.5}, \sigma_n, N(0,1)) = 86.3\%$. Rousseeuw and Croux (1993) propose to use the lower quartile, i.e., $\alpha = 1/4$, instead of the median. Specifically, they define the finite-

---

[2]Note that the empirical 1/2-quantile in this sense does not generally coincide with the above definition of the sample median.

sample version as the $\binom{\lfloor n/2 \rfloor + 1}{2}$th order statistic of the $\binom{n}{2}$ values $|X_i - X_j|$, $1 \leq i < j \leq n$. They call this estimator Q_n, and it's become known under this name, which leads us to call the generalized version Q_n^α. The original Q_n has an asymptotic relative efficiency with respect to the standard deviation at normality of 82%. Rousseeuw and Croux (1993) settle for the slightly lesser efficiency to achieve the maximal breakdown point of about 50%. However, this aspect is of much lesser relevance in the change-point context, quite the contrary, the very property of high-breakdown-point robustness is counterproductive for detecting change-points. The original Q_n is unsuited as a substitution for the sample variance in (1), and a larger value of roughly $0.7 < \alpha < 0.9$ is much more appropriate for the problem at hand. We defer further explanations to Section 4, where we discuss how to choose the α appropriately.

These five scale measures, the standard deviation, the mean deviation d_n, Gini's mean difference g_n, the median absolute deviation (MAD) m_n, and the α-sample quantile of all pairwise differences Q_n^α, are the ones we restrict our attention to in the present article. They are summarized in Table 1 along with their sample versions. There are of course many more potential scale estimators that satisfy the scale equivariance (2) and more robust proposals in the literature, many of which include a data-adaptive re-weighting of the observations (e.g. Huber and Ronchetti, 2009, Chapter 5). In the present paper we explore the use of these common, easy-to-compute estimators in the change-point setting. They all admit explicit formulas, all can be computed in $O(n \log n)$ time, and the pairwise-difference estimators allow computing time savings for sequentially updated estimates (which are required in the change-point setting) – more so than, e.g., implicitly defined estimators. The two pairwise-difference based estimators, the average and the α-quantile of all pairwise differences, possess promising statistical properties. We will mainly focus on these and derive their asymptotic distribution under no change in the following section.

3. Test statistics, long-run variances and asymptotic results

We first describe the data model employed: a very broad class of data generating processes, incorporating heavy tails and short-range dependence (Section 3.1). We then propose several change-point test statistics based on the scale estimators introduced in the previous section and provide estimates for their long-run variances (Section 3.2). We show asymptotic results for Gini's mean difference and the Q_n^α based tests under the null hypothesis (Sections 3.3 and 3.4, respectively) and discuss methods for an optimal bandwidth selection for the long-run variance estimation (Section 3.5).

3.1. The data model

We assume the data $X_1, \ldots, X_n$ to follow the model

$$X_i = \lambda_i Y_i + \mu, \qquad 1 \leq i \leq n, \tag{4}$$

Table 1: Scale estimators; $\mathrm{md}\,(F)$ denotes the median of the distribution F, F_n its empirical distribution function.

Scale Estimator	Population value	Sample version
Standard deviation	$\sigma\,(F) = \{E\,(X - EX)^2\}^{1/2}$	$\sigma_n = \left\{ \dfrac{1}{n-1} \sum\limits_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \right\}^{1/2}$
Mean deviation	$d\,(F) = E\,\lvert X - \mathrm{md}\,(F)\rvert$	$d_n = \dfrac{1}{n-1} \sum\limits_{i=1}^{n} \lvert X_i - \mathrm{md}(F_n)\rvert$
Gini's mean difference	$g\,(F) = E\,\lvert X - Y\rvert$	$g_n = \dfrac{2}{n(n-1)} \sum\limits_{1 \le i < j \le n} \lvert X_i - X_j\rvert$
Median absolute deviation	$m\,(F) = \mathrm{md}\,(G)$; where G is cdf of $\mathcal{L}\{\lvert X - \mathrm{md}\,(F)\rvert\}$	$m_n = \mathrm{md}\,\lvert X_i - \mathrm{md}(F_n)\rvert$
Q_n^α	$Q^\alpha = U^{-1}\,(\alpha)$; where U is cdf of $\mathcal{L}\{\lvert X - Y\rvert\}$	$Q_n^\alpha = U_n^{-1}(\alpha)$; where U_n is empirical cdf of $\lvert X_i - X_j\rvert$, $1 \le i < j \le n$

where $Y_1, \ldots, Y_n$ are part of the stationary, median-centered sequence $(Y_i)_{i \in \mathbb{Z}}$. We want to test the hypothesis

$$H_0 : \lambda_1 = \ldots = \lambda_n$$

against the alternative

$$H_1 : \exists\, k \in \{1, \ldots, n-1\} : \lambda_1 = \ldots = \lambda_k \neq \lambda_{k+1} = \ldots = \lambda_n.$$

Note that this set-up is completely moment-free. We allow the underlying process to be dependent, more precisely we assume $(Y_i)_{i \in \mathbb{Z}}$ to be near epoch dependent in probability on an absolutely regular process. Let us briefly introduce this kind of short-range dependence condition.

Definition 3.1.   *1. Let $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ be two σ-fields on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We define the absolute regularity coefficient of $\mathcal{A}$ and $\mathcal{B}$ by*

$$\beta\,(\mathcal{A}, \mathcal{B}) = E \sup_{A \in \mathcal{A}} \lvert \mathrm{P}\,(A|\mathcal{B}) - \mathrm{P}\,(A)\rvert.$$

2. Let $(Z_n)_{n \in \mathbb{Z}}$ be a stationary process. Then the absolute regularity coefficients of $(Z_n)_{n \in \mathbb{Z}}$ are given by

$$\beta_k = \sup_{n \in \mathbb{N}} \beta\,(\sigma\,(Z_1, \ldots, Z_n), \sigma\,(Z_{n+k}, Z_{n+k+1}, \ldots)).$$

We say that $(Z_n)_{n \in \mathbb{Z}}$ is absolutely regular, if $\beta_k \to 0$ as $k \to \infty$.

The model class of absolutely regular processes is a common model for short-range dependence. But it does not include important classes of time series, e.g., not all linear processes. Therefore, we will not study absolutely regular processes themselves, but approximating functionals of such processes. In this situation, L_2 near epoch dependent processes are frequently considered. Since we also consider quantile-based estimators with the advantage of moment-freeness, we want to avoid moment assumptions implicitly in the short-range conditions. For this reason, we employ the concept of near epoch dependence in probability, introduced by Dehling et al. (2016). For further information see Dehling et al. (2016, Appendix A).

Definition 3.2. *We call the process $(X_n)_{n \in \mathbb{N}}$ near epoch dependent in probability (or short P-NED) on the process $(Z_n)_{n \in \mathbb{Z}}$ if there is a sequence of approximating constants $(a_l)_{l \in \mathbb{N}}$ with $a_l \to 0$ as $l \to \infty$, a sequence of functions $f_l : \mathbb{R}^{2l+1} \to \mathbb{R}$ and a non-increasing function $\Phi : (0, \infty) \to (0, \infty)$ such that*

$$\mathrm{P}\left(|X_0 - f_l(Z_{-l}, \ldots, Z_l)| > \epsilon\right) \leq a_l \Phi(\epsilon) \tag{5}$$

for all $l \in \mathbb{N}$ and $\epsilon > 0$.

The absolute regularity coefficients β_k and the approximating constants a_l will have to fulfill certain rate conditions that are detailed in Assumption 3.3.

3.2. Change-point test statistics and long-run variance estimates

We test the null hypothesis H_0 against the alternative H_1 by means of CUSUM-type test statistics of the form

$$\hat{T}_s = \max_{1 \leq k \leq n} \frac{k}{\sqrt{n}} \left|s_{1:k} - s_{1:n}\right|.$$

Throughout, we use s_n as generic notation for a scale estimator (where we include, for completeness' sake, the variance), and $s_{1:k}$ denotes the estimator applied to $X_1, \ldots, X_k$. Considering, besides the variance, the four scale estimators introduced in the previous section, we have the test statistics $\hat{T}_{\sigma^2}$, $\hat{T}_d$, $\hat{T}_g$, $\hat{T}_m$, and $\hat{T}_Q(\alpha)$ based on the sample variance, the mean deviation, Gini's mean difference, the median absolute deviation and the Q_n^α scale estimator, respectively. Under the null hypothesis, the sequence $X_1, \ldots, X_n$ is stationary, and can be thought of as being part of a stationary process $(X_i)_{i \in \mathbb{Z}}$ with marginal distribution F (i.e., $X_i = \lambda Y_i + \mu$). Then, under suitable regularity conditions (that are specific to the choice of s_n), the test statistic $\hat{T}_s$ converges in distribution to $D_s \sup_{0 \leq t \leq 1} |B(t)|$, where B is a Brownian bridge. The quantity D_s^2 is referred to as the long-run variance. It depends on the estimator s_n and the data generating process. Expressions for the scale estimators considered here are given below. The distribution of $\sup_{0 \leq t \leq 1} |B(t)|$ is well known and sometimes referred to as Kolmogorov distribution. However, D_s^2 is generally unknown, depends on the distribution of the whole process $(X_i)_{i \in \mathbb{Z}}$ and must be estimated when applying the test in practice.[3]

---

[3]Alternatively, bootstrapping can be employed, we take up this discussion in Section 7.

In the following definitions, let $X, Y \sim F$ be independent. The long-run variances corresponding to the scale estimators under consideration are

$$D_{\sigma^2}^2 = \sum_{h=-\infty}^{\infty} \text{cov} \left\{ (X_0 - EX_0)^2, (X_h - EX_h)^2 \right\}$$

$$D_d^2 = \sum_{h=-\infty}^{\infty} \text{cov} \left(|X_0 - \text{md}(F)|, |X_h - \text{md}(F)| \right),$$

$$D_g^2 = 4 \sum_{h=-\infty}^{\infty} \text{cov} \left(\varphi(X_0), \varphi(X_h) \right), \tag{6}$$

where $\varphi(x) = E \left| x - Y \right| - g(F)$,

$$D_m^2 = \frac{1}{f_Z^2(m(F))} \sum_{h=-\infty}^{\infty} \text{cov} \left(\mathbb{1}_{\{|X_0 - \text{md}(F)| \leq m(F)\}}, \mathbb{1}_{\{|X_h - \text{md}(F)| \leq m(F)\}} \right),$$

where f_Z is the density of $Z = |X - \text{md}(F)|$, and

$$D_Q^2(\alpha) = \frac{4}{u^2(Q^\alpha(F))} \sum_{h=-\infty}^{\infty} \text{cov} \left(\psi(X_0), \psi(X_h) \right), \tag{7}$$

where $\psi(x) = P(|x - Y| \leq Q^\alpha) - \alpha$ and $u(t)$ is the density associated with the distribution function $U(t) = P(|X - Y| \leq t)$ of $|X - Y|$. An intuitive derivation of the expressions for D_g^2 and $D_Q^2(\alpha)$ are given in Appendix B.

The following long-run variance estimators follow the construction principle of heteroscedasticity and autocorrelation consistent (HAC) kernel estimators, for which we use results by de Jong and Davidson (2000). The HAC kernel function (or weight function) W can be quite general, but has to fulfill Assumption 3.1 (a) below, which is basically Assumption 1 of de Jong and Davidson (2000). There is further a bandwidth to choose, which basically determines up to which lag the autocorrelations are included. For consistency of the long-run variance estimate, the bandwidth has to fulfill the rate condition of Assumption 3.1 (b).

Assumption 3.1.

(a) The function $W : [0, \infty) \to [-1, 1]$ is continuous at 0 and at all but a finite number of points and $W(0) = 1$. Furthermore, $|W|$ is dominated by a non-increasing, integrable function and

$$\int_0^\infty \left| \int_0^\infty W(t) \cos(xt) \, dt \right| dx < \infty.$$

(b) The bandwidth b_n satisfies $b_n \to \infty$ as $n \to \infty$ and $b_n / \sqrt{n} \to 0$.

We propose the following long-run variance estimators for the three moment-based scale measures: For the variance we take a weighted sum of empirical autocorrelations of the centered squares of the data, i.e.,

$$\hat{D}_{\sigma^2}^2 = \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} \left\{(X_i - \bar{X}_n)^2 - \sigma_n^2\right\} \left\{(X_{i+|k|} - \bar{X}_n)^2 - \sigma_n^2\right\}, \quad (8)$$

where $\bar{X}_n$ and σ_n^2 denote the sample mean and the sample variance, respectively, computed from the whole sample, cf. Table 1. Similar expressions have been considered, e.g., by Gombay et al. (1996), Lee and Park (2001) and Wied et al. (2012). There are, of course, other possibilities regarding the exact definition of the long-run variance estimate (e.g., use factor $1/n$ or $1/(n-k)$), the choice of which turns out to have an negligible effect. The strongest impact has the choice of the bandwidth b_n (see Section 3.5).

For the mean deviation, we propose

$$\hat{D}_d^2 = \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} (|X_i - \mathrm{md}(F_n)| - d_n)\left(|X_{i+|k|} - \mathrm{md}(F_n)| - d_n\right),$$

where $md(F_n)$ and d_n denote the sample median and the sample mean deviation, respectively, of the whole sample (cf. Table 1). For Gini's mean difference we consider

$$\hat{D}_g^2 = 4 \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} \hat{\varphi}_n(X_i)\hat{\varphi}_n(X_{i+|k|}),$$

where

$$\hat{\varphi}_n(x) = \frac{1}{n} \sum_{i=1}^{n} |x - X_i| - g_n$$

is an empirical version of $\varphi(x)$ in (6). For the long-run variance estimates for the quantile-based scale measures m_n and Q_n^α, we need estimates for the densities f_Z and u, respectively, for which we use kernel density estimates

$$\hat{f}(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{|X_i - \mathrm{md}(F_n)| - t}{h_n}\right),$$

$$\hat{u}(t) = \frac{2}{n(n-1)h_n} \sum_{1 \leq i < j \leq n} K\left(\frac{|X_i - X_j| - t}{h_n}\right).$$

The density kernel K and the bandwidth h_n have to fulfill the following conditions.

Assumption 3.2. *Let $K : \mathbb{R} \to \mathbb{R}$ be symmetric (i.e. $K(x) = K(-x)$), Lipschitz-continuous function with bounded support which is of bounded variation and integrates to 1. Let the bandwidth h_n satisfy $h_n \to 0$ and $n\,h_n^{8/3} \to \infty$, as $n \to \infty$.*

If we were interested in an accurate estimate of the whole density function, the estimate could be further improved by incorporating the fact that its support is the positive half-axis, but since we are only interested in density estimates at one specific point, this is not necessary. Define

$$\hat{D}_m^2 = \frac{1}{\hat{f}_Z(m_n)} \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} \hat{\xi}_n(X_i)\hat{\xi}_n(X_{i+|k|})$$

with

$$\hat{\xi}_n(x) = \mathbb{1}_{\{|x-\mathrm{md}(F_n)|\leq m_n\}} - 1/2,$$

and

$$\hat{D}_Q^2(\alpha) = \frac{4}{\hat{u}(Q_n^\alpha)} \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} \hat{\psi}_n(X_i)\hat{\psi}_n(X_{i+|k|})$$

with

$$\hat{\psi}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{|x-X_i|\leq Q_n^\alpha\}} - \alpha.$$

Keep in mind that in the expressions above, the HAC bandwidth b_n and the kernel density bandwidth h_n play distinctively different roles: b_n increases as n increases, whereas h_n decreases to zero. Also, the kernels W and K serve different purposes: W is an HAC-kernel and is scaled such that $W(0) = 1$, while K is a density kernel and it is scaled such that it integrates to 1.

Below, in Sections 3.3 and 3.4, we give sufficient conditions for the convergence of the studentized test statistics $\hat{D}_g^{-1}\hat{T}_g$ and $\hat{D}_Q^{-1}(\alpha)\hat{T}_Q(\alpha)$, respectively, since the corresponding estimators, as outlined in Section 2, exhibit the best statistical properties, and these tests indeed show the best performance, as demonstrated in Section 5. The variance-based test statistic $\hat{D}_{\sigma^2}^{-1}\hat{T}_{\sigma^2}$, or versions of it, has been considered by several authors. It is treated for L_2 NED on α-mixing processes by Wied et al. (2012). As for the mean-deviation-based test statistic $\hat{D}_d^{-1}\hat{T}_d$, the convergence can be shown by similar techniques as for $\hat{D}_{\sigma^2}^{-1}\hat{T}_{\sigma^2}$: the same $(2+\delta)$ moment condition as for Gini's mean difference along with corresponding rate for the short-range dependence conditions (Assumption 3.3) are required. Additionally, a smoothness condition around $\mathrm{md}(F)$ is necessary to account for the estimation of the central location. For the MAD-based test statistic $\hat{D}_m^{-1}\hat{T}_m$, no moment conditions are required, but smoothness conditions on F at $\mathrm{md}(F)$ as well as $m(F) = |X - \mathrm{md}(F)|$, $X \sim F$. However, it turns out that the MAD does not provide a workable change-point test. Roughly speaking, the estimate is rather coarse, and the convergence to the limit distribution too slow to yield usable critical values. But even for large n or with the use of bootstrapping methods, the test is dominated in terms of power by the other tests considered.

3.3. Gini's mean difference

We assume the following condition on the data generating process.

Assumption 3.3. *Let* $(X_i)_{i\in Z}$ *be a stationary process that is PNED on an absolutely regular sequence* $(Z_n)_{n\in\mathbb{Z}}$*. There is a* $\delta > 0$ *such that*

(a) *the PNED approximating constants* a_l *and the absolute regularity coefficients* β_k *satisfy*

$$a_l \Phi\left(l^{-6}\right) = O\left(l^{-6\frac{2+\delta}{\delta}}\right) \ \ as \ l \to \infty \qquad and \qquad \sum_{k=1}^{\infty} k\beta_k^{\frac{\delta}{2+\delta}} < \infty, \qquad (9)$$

where Φ *is defined in Definition 3.2, and*

(b) *there is a positive constant* M *such that* $E|X_0|^{2+\delta} \leq M$ *and* $E\left|f_l\left(Z_{-l}, \ldots, Z_l\right)\right|^{2+\delta} \leq M$ *for all* $l \in \mathbb{N}$*.*

Then we have the following result about the asymptotic distribution of the test statistic $\hat{D}_g^{-1}\hat{T}_g$ based on Gini's mean difference under the null hypothesis. The proof is given in Appendix B.

Theorem 3.1. *Under Assumptions 3.1, 3.2, and 3.3, we have* $\hat{D}_g^{-1}\hat{T}_g \xrightarrow{d} \sup_{0\leq\lambda\leq 1}|B(\lambda)|$*, where* $(B(\lambda))_{0\leq\lambda\leq 1}$ *is a standard Brownian bridge.*

3.4. Q_n^α

Since Q_n^α is a quantile-based estimator, we require no moment condition, and it suffices that the short-range dependence condition (9) is satisfied for "$\delta = \infty$".

Assumption 3.4. *Let* $(X_i)_{i\in Z}$ *be a stationary process that is PNED of an absolutely regular sequence* $(Z_n)_{n\in\mathbb{Z}}$ *such that the PNED approximating constants* a_l *and the absolute regularity coefficients* β_k *satisfy*

$$a_l \Phi(l^{-6}) = O(l^{-6}) \ \ as \ l \to \infty \qquad and \qquad \sum_{k=1}^{\infty} k\beta_k < \infty,$$

where Φ *is defined in Definition 3.2.*

However, instead of the moment condition we require a smoothness condition on F.

Assumption 3.5. *The distribution F has a Lebesgue density f such that*

(a) f *is bounded,*

(b) *the support of f, i.e.,* $\overline{\{x|f(x) > 0\}}$*, is a connected set, and*

(c) *the real line can be decomposed in finitely many intervals such that f is continuous and (non-strictly) monotonic on each of them.*

We are now ready to state the following result concerning the asymptotic distribution of the Q_n^α-based change-point test statistic. The proof is given in Appendix B.

Theorem 3.2. *Under Assumptions 3.1, 3.2, 3.4, and 3.5, we have, for any fixed $0 < \alpha < 1$ that* $\hat{D}_Q^{-1}(\alpha)\hat{T}_Q(\alpha) \xrightarrow{d} \sup_{0\leq\lambda\leq 1}|B(\lambda)|$*, where* $(B(\lambda))_{0\leq\lambda\leq 1}$ *is a standard Brownian bridge.*

11

3.5. Data-adaptive bandwidth selection

The rate conditions of Assumptions 3.1 on the HAC bandwidth b_n to achieve consistency of the long-run variance estimate are rather mild. However, the question remains how to choose b_n optimally for a given sequence of observations of length n. The answer depends on the degree of serial dependence present in the sequence. Loosely speaking, choosing it too small results in a size distortion, choosing it too large will render the tests conservative and less powerful.

A general approach is to assume a parametric time-series model, minimize the mean squared error in terms of the parameters of the model and then plug-in estimates for the parameters. For instance, to estimate the long-run variance of an AR(1) process with auto-correlation parameter ρ Andrews (1991, Section 6) gives an optimal bandwidth of

$$b_n = 1.447 n^{1/3} \left(\frac{4\rho^2}{(1-\rho^2)^2} \right)^{1/3} \tag{10}$$

when the Bartlett kernel is used.

When employing alternative methods to estimate the long-run variance, i.e., block sub-sampling or block bootstrap, one faces the similar challenge of selecting an appropriate blocklength, and techniques that have been derived for data-adaptive blocklength selection may be "borrowed" for bandwidth selection. For instance, Carlstein (1986) obtains a very similar expression as (10) for an optimal blocklength in the AR(1) setting for a block sub-sampling scheme.

Politis and White (2004) (see also Patton et al. (2009)) derive optimal blocklengths for block bootstrap methods and obtain also an expression similar to (10). This technique is adapted, e.g., by Kojadinovic et al. (2015) for a change-point test based on Spearman's rho. Furthermore, Bücher and Kojadinovic (2016a) identified the asymptotically optimal blocklength for the multiplier bootstrap of U-statistics and developed a method to estimate it.

All such methods require the estimation of unknown parameters as, e.g., in (10) the autocorrelation. Alongside exploring alternatives to the sample variance for robust scale estimation, it is advisable to consider alternatives to the sample autocorrelation. A recent comparison of robust autocorrelation estimators is given by Dürre et al. (2015), who recommend to use Gnanadesikan-Kettenring-type estimators. Such estimates require robust scale estimates, for which also the Q_n has been proposed (Ma and Genton, 2000).

Hall et al. (1995) have suggested an alternative approach without estimating autocorrelations: If the rate, but not the constant, in an expression as (10) is known, the blocklength can be chosen by sub-sampling. For U-statistics, the rate is $n^{1/3}$ in the case of the Bartlett kernel (Dehling and Wendler, 2010) and $n^{1/5}$ for flat-top kernels (Bücher and Kojadinovic, 2016b). Results on the optimal blocklength for quantiles and U-quantiles are not known to us, but we suspect that the rates for the optimal blocklength are the same as for partial sums, because our variance estimator uses a linearization.

Table 2: Power of change-point tests at asymptotic 5% level, based on sample variance/Gini's mean difference/Q_n for independent, centered normal observations. Standard deviation changes from 1 in first half to λ in second half.

sample size	standard deviation λ in second half				
	1.0	1.2	1.5	2.0	3.0
$n = 60$	.04/.04/.44	.04/.07/.32	.12/.24/.18	.23/.51/.06	.36/.82/.01
$n = 500$	.04/.04/.08	.63/.64/.52	1/1/1	1/1/1	1/1/1

4. The choice of α for Q_n^α

In this section, we address the question which α to choose best when employing the Q_n^α scale estimator. The theoretical U-quantile results of the previous setting apply to Q_n^α for any $0 < \alpha < 1$. The original Q_n proposed by Rousseeuw and Croux (1993) corresponds roughly to $\alpha = 1/4$, which is mainly motivated by breakdown considerations. A high-breakdown-estimator is in fact counterproductive for change-point detection purposes, and higher values of α are more appropriate for the problem at hand. A high-breakdown-point estimator is designed, in some sense, to ignore a large portion of outliers, no matter how they are distributed, spatially or temporarily. The perception of robustness in the change-point setting is conceptually different: we want to safeguard against a few outliers or several but evenly distributed over the observed sequence, as they may be generated by a heavy-tailed distribution. A subsequence of outliers on the other hand, which even exhibits some common characteristics differing from the rest of the data, constitutes a structural change, which we want to detect rather than ignore. To illustrate this point, consider the following example: we have a sample where the first 60% of the observations alternate between 1 and -1 and the second 40% alternate between 100 and -100. This constitutes a very noticeable change in the scale, but the $Q_n^{0.25}$-based change-point test of the CUSUM type is not able to detect this change. Similar counter-examples can be constructed for analogous change-point tests which compare the first part to the second part of the sample. However, this very artificial example should not be given to much attention. Any quantile-based method is generally suitable for continuous distributions only.

The point may also be illustrated by Monte Carlo simulation results for continuous distributions. Consider the following data situation: starting from an i.i.d. sequence of standard normal observations, we multiply the second half by some value λ. Table 2 shows rejection frequencies at the asymptotic 5% level based on the original Q_n for sample sizes $n = 60$ and $n = 500$ and several values of λ (bandwidth as in the simulations in Section 5, i.e., $b_n = 2n^{1/3}$). The numbers of Table 2 suggest that the original Q_n is very unsuitable for change-point detection in small samples. For $n = 60$, it strongly exceeds the size, and the power is actually lower under alternatives than under the null.

In principle, when the data is large enough, the $Q_n^{0.25}$ is able to pick up scale changes in the continuous setting, i.e., the 1/4th quantile of all pairwise differences is larger than
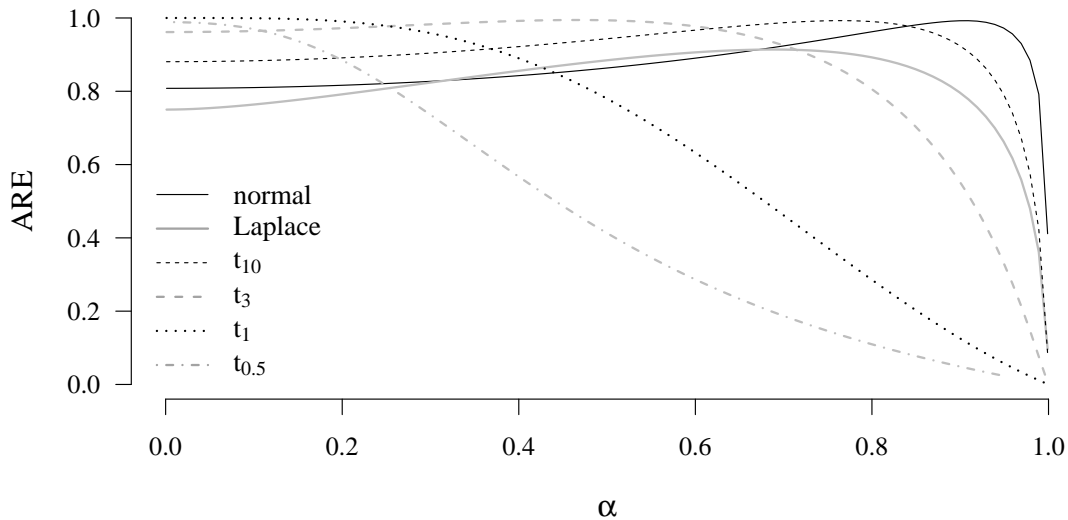
Figure 1: Asymptotic relative efficiencies (AREs) of Q_n^α at normal, Laplace, and several t distributions wrt respective maximum-likelihood estimators of scale.

the 1/4th quantile of all pairs of the first half of the observations. For $n = 500$, this difference is sufficiently pronounced so that the test more or less works satisfactorily. However, for $n = 60$, this difference is relatively small compared to the increased long-run variance (by which the test statistic is divided) when λ is large. This largely accounts for the decrease of power as λ increases. Additionally, the Q_n-based test grossly exceeds the size under the null. This can be largely attributed to a general bad "small sample behavior" of the test statistic. We observe a similar effect for the MAD-based test, cf. Section 5 and the median-based change-point test for location considered by Vogel and Wendler (2016). Thus the $Q_n^{0.25}$ is particularly unsuitable for a quick detection of strong changes.

The problems can be overcome by considering Q_n^α for larger values of α than $1/4$, and using such a scale estimator indeed leads to a workable change-point test also for $n = 60$. As a guideline for a suitable choice of α, we may look at the asymptotic efficiencies. Figure 1 plots the asymptotic relative efficiency of Q_n^α at several scale families with respect to the respective maximum-likelihood estimator for scale. The solid line, showing the asymptotic relative efficiency of Q_n^α with respect to the standard deviation at normality, is also depicted in Rousseeuw and Croux (1992). Since we are also interested in efficiency at heavy-tailed distributions, we further include several members of the t_ν-family (for $\nu = 1/2, 1, 3, 10$) and the Laplace distribution, cf. Table 3. The mathematical derivations for this plot, i.e., the asymptotic variances of the Q_n^α and the MLE of the scale parameter of the t-distribution, are given in Appendix A.

At normality, the Q_n^α is asymptotically most efficient for $\alpha = 0.9056$ with an asymptotic relative efficiency of 99% with respect to the standard deviation, but it shows a very slow decay as α decreases to zero. For the t-distributions considered, the maximal

asymptotic efficiency of the Q_n^α is achieved for smaller values of α, e.g., for the t_3 distribution, the optimum is attained at $\alpha = 0.4661$. However, within the range $0 < \alpha < 0.73$, the asymptotic relative efficiency with respect to the maximum-likelihood estimator is above 90%. The t_ν-distributions with $\nu = 1/2$ and $\nu = 1$, also depicted in Figure 1, are extremely heavy-tailed. We consider t_ν distributions with $\nu = 3$ and $\nu = 5$ as more realistic data models, and these are also included in the simulation results presented in Section 5.

Altogether, Figure 1 suggests that $\alpha \approx 3/4$ may be a suitable choice as far as asymptotic efficiency is concerned. We ran simulations with many different values of α and found that Q_n^α generally performed best within the range $0.7 < \alpha < 0.9$. In the tables in Section 5, we report results for $\alpha = 0.8$.

5. Simulation Results

In a simulation study we want to investigate the empirical size and power of the change-point tests based on the test statistics introduced in Section 3.2. We consider the following general model: We assume the process $(Y_i)_{i\in\mathbb{Z}}$ to follow an $AR(1)$ process

$$Y_i = \rho Y_{i-1} + \epsilon_i, \qquad i \in \mathbb{Z},$$

for some $-1 < \rho < 1$ where the ϵ_i are an i.i.d. sequence with a mean-centered distribution. Then the data $X_1, \ldots, X_n$ are generated as

$$X_i = \begin{cases} Y_i, & 1 \le i \le \lfloor \theta n \rfloor, \\ \lambda Y_i, & \lfloor \theta n \rfloor < i \le n, \end{cases}$$

for some $0 < \theta < 1$. Thus we have four parameters, λ, θ, ρ, and n, which regulate the size of the change, the location of the change, the degree of the serial dependence of the sequence and the sample size, respectively. The tables below report results for $\lambda = 1, 1.5, 2$, $\theta = 1/4, 1/2, 3/4$, $\rho = 0, 0.8$ and $n = 60, 120, 240, 500$. A fifth "parameter" is the distribution of the ϵ_i, for which we consider the standard normal $N(0,1)$, the standard Laplace $L(0,1)$, the normal scale mixture $NM(\gamma, \varepsilon)$ with $\gamma = 3$, $\varepsilon = 0.01$ and t_ν-distributions with $\nu = 3$ and $\nu = 5$. The densities of these distributions are summarized in Table 3. The latter four of them are, to varying degrees, heavier-tailed than the normal. The normal mixture distribution captures the notion that the majority of the data stems from the normal distribution, except for some small fraction, which stems from a normal distribution with a γ times larger standard deviation. This type of contamination model has been popularized by Tukey (1960), who argues that $\gamma = 3$ is a realistic choice in practice, and further pointed out that in this case the mean deviation d_n is more efficient scale estimator than the standard deviation for values of ϵ as low as 1%. Concerning the long-run variance estimation, we take the following choices for bandwidths and kernels,

$$K(t) = \frac{3}{4}(1-t^2)\mathbb{1}_{[-1,1]}(t), \quad W(t) = (1-t^2)^2\mathbb{1}_{[-1,1]}(t), \quad h_n = I_n\, n^{-1/3}, \quad b_n = 2n^{1/3},$$

15

Table 3: Densities of innovation distributions considered in the simulation set-up. $B(\cdot, \cdot)$ denotes the beta function.

distribution	density $f(x)$	parameters	kurtosis
normal $N(\mu, \sigma^2)$	$\dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\dfrac{(x-\mu)^2}{2\sigma^2} \right\}$	$\mu \in \mathbb{R},$ $\sigma^2 > 0$	0
normal mixture $NM(\gamma, \epsilon)$	$\dfrac{\epsilon}{\sqrt{2\pi}\gamma} \exp\left\{ \dfrac{-x^2}{2\gamma^2} \right\} + \dfrac{1-\epsilon}{\sqrt{2\pi}} \exp\left\{ \dfrac{-x^2}{2} \right\}$	$0 \leq \epsilon \leq 1,$ $\gamma \geq 1$	$\dfrac{3\epsilon(1-\epsilon)(\lambda^2-1)^2}{(\epsilon\lambda^2+1-\epsilon)^2}$
Laplace $L(\mu, \alpha)$	$\dfrac{1}{2\alpha} \exp\left\{ \dfrac{-|x-\mu|}{\alpha} \right\}$	$\mu \in \mathbb{R},$ $\alpha > 0$	3
t_ν	$\sqrt{\nu}B(\nu/2, 1/2)\left(1 + x^2/\nu\right)^{-\frac{\nu+1}{2}}$	$\nu > 0$	$\dfrac{6}{\nu-4}\ (\nu > 4)$

where I_n denotes the sample interquartile range of the data points the kernel density estimator is applied to. The kernel K above is known as Epanechnikov kernel, and W as quartic kernel. The results of the long-run variance estimation generally differ very little with respect to the choice of kernels. The HAC-bandwidth b_n tends to have the largest influence. We fix it in our simulations to $b_n = 2n^{1/3}$, which is large enough to account for the strong serial dependence setting of $\rho = 0.8$, but is certainly far from optimal in the independence setting, and the results can be improved upon by using one of the data-adaptive bandwidth selection methods outlined in Section 3.5. However, choosing the same, albeit fixed, bandwidth for all change-point tests allows a fair comparison and puts the focus on the impact of the different estimators.

For each setting we generate 1000 repetitions. Tables 4–8 report empirical rejection frequencies (in %) at the asymptotic 5% significance level, i.e., we count how often the test statistics exceed 1.358, i.e., the 95%-quantile of the limiting distribution of the studentized test statistics under the null.

Analysis of size. Table 4 reports rejection frequencies of the change-point tests based on the variance σ_n^2 (Var), the mean deviation d_n (MD), Gini's mean difference g_n (GMD), the median absolute deviation m_n (MAD), the original Q_n as considered by Rousseeuw and Croux (1993), i.e., the $\binom{\lfloor n/2 \rfloor + 1}{2}$-th order statistic of all pairwise differences, and the $Q_n^{0.8}$, i.e., the $0.8\binom{n}{2}$-th order statistic of all pairwise differences. We notice that the Q_n and the MAD heavily exceed the size. This effect wears off as n increases, but rather slowly. The Q_n shows an acceptable size behavior for $n = 500$, the MAD not even for this sample size. This behavior can be attributed to a discretization effect. A similar observations is made at the median-based change-point test for location, which is discussed in detail in Vogel and Wendler (2016, Section 4). Due to the size distortion, the

Table 4: *Test size.* Rejection frequencies (%) of change-point tests based on six different scale estimators at stationary sequences. Asymptotic 5% significance level; sample sizes $n = 60, 120, 240, 500$; five different innovation distributions. HAC bandwidth for long-run variance estimation: $b_n = 2n^{1/3}$.

Estimator	independence						$AR(1)$ with $\rho = 0.8$					
	Var	MD	GMD	MAD	Q_n	$Q_n^{0.8}$	Var	MD	GMD	MAD	Q_n	$Q_n^{0.8}$
$n = 60$												
$N(0,1)$	3	5	3	27	43	6	8	7	13	24	15	14
$L(0,1)$	3	4	3	23	38	8	7	6	11	23	12	12
$NM(3,0.01)$	4	6	4	26	46	7	6	6	11	22	14	10
t_5	3	4	1	24	40	8	7	6	11	23	13	13
t_3	2	5	2	27	43	11	4	4	8	24	12	11
$n = 120$												
$N(0,1)$	3	3	3	21	29	4	6	6	8	19	8	7
$L(0,1)$	2	2	2	15	23	5	3	4	7	17	7	7
$NM(3,0.01)$	2	2	2	19	27	4	4	4	6	18	8	7
t_5	1	2	2	19	23	4	4	4	7	17	8	7
t_3	2	4	2	16	25	8	3	3	5	19	6	6
$n = 240$												
$N(0,1)$	2	3	2	15	11	2	3	3	5	15	5	4
$L(0,1)$	2	3	2	12	10	3	3	3	5	13	5	5
$NM(3,0.01)$	2	3	3	14	11	4	4	4	6	14	5	6
t_5	3	4	3	13	12	3	2	4	5	14	5	4
t_3	2	3	3	13	13	6	2	3	4	15	4	5
$n = 500$												
$N(0,1)$	3	3	3	10	5	4	4	4	4	11	5	5
$L(0,1)$	3	5	4	9	6	5	5	6	5	10	4	4
$NM(3,0.01)$	2	3	3	12	6	3	4	4	5	12	3	3
t_5	3	4	3	11	7	4	4	4	4	10	5	4
t_3	1	3	2	11	7	5	3	4	3	12	4	4

MAD and the Q_n are excluded from any further power analysis. The size exceedance also takes places to a lesser degree for the $Q_n^{0.8}$ at $n = 60$ and for GMD under dependence at $n = 60$. This must be taken into account when comparing the power under alternatives. Based on the simulation results for size and power, the $Q_n^{0.8}$ can be seen to provide a sensible change-point test for scale, but some caution should be taken for sample sizes below $n = 100$. All other tests keep the size for the situations considered.

Analysis of power. Tables 5–8 lists empirical rejection frequencies under various alternatives. Tables 5 and 6 show results for serial independence ($\rho = 0$) for change-sizes $\lambda = 1.5$ and $\lambda = 2$, respectively. Tables 7 and 8 contain figures for the serial dependence setting ($\rho = 0.8$), again for change-sizes $\lambda = 1.5$ and $\lambda = 2$, respectively. All tables show results for $\theta = 1/4, 1/2$, and $3/4$ and $n = 60, 120, 240$, and 500. We make the following observations:

Table 5: *Test power.* Rejection frequencies (%) at asymptotic 5% level. Change-point tests based on variance (Var), mean deviation (MD), Gini's mean difference (GMD), and $Q_n^{0.8}$. **Independent observations**; scale changes by **factor** $\lambda = 1.5$. Several sample sizes, marginal distributions, and change locations. HAC bandwidth: $b_n = 2n^{1/3}$.

Change location:	$[n/4]$				$[n/2]$				$[n3/4]$			
Estimator:	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
$n = 60$												
$N(0,1)$	4	3	10	11	9	8	21	22	5	4	8	9
$L(0,1)$	2	2	5	8	4	4	10	14	2	3	5	8
$NM(3,0.01)$	4	4	8	12	4	4	10	20	5	4	10	10
t_5	2	4	6	9	4	6	11	12	3	4	6	8
t_3	1	2	3	8	3	5	8	12	3	4	4	10
$n = 120$												
$N(0,1)$	10	12	22	22	43	43	57	49	26	21	34	24
$L(0,1)$	4	6	11	8	13	21	27	21	11	11	15	11
$NM(3,0.01)$	6	11	17	18	30	37	47	46	22	19	30	20
t_5	4	8	11	10	18	28	34	28	11	12	19	13
t_3	2	5	7	7	8	19	22	20	7	11	13	10
$n = 240$												
$N(0,1)$	36	46	60	57	90	88	93	92	72	64	77	69
$L(0,1)$	11	21	26	22	44	62	65	56	34	39	44	32
$NM(3,0.01)$	25	37	49	49	72	86	89	89	57	59	69	65
t_5	12	26	31	30	48	73	74	70	34	48	50	44
t_3	5	18	18	20	19	52	49	51	15	32	32	32
$n = 500$												
$N(0,1)$	92	94	97	96	100	100	100	100	99	97	98	99
$L(0,1)$	40	70	72	64	88	96	95	94	75	83	84	79
$NM(3,0.01)$	66	90	93	94	95	100	100	100	89	96	98	99
t_5	39	78	77	78	85	99	98	99	73	90	90	91
t_3	10	53	46	57	40	89	84	93	34	70	64	74

(1) All tests have better power at independent sequences than dependent sequences. Note that $\rho = 0.8$ constitutes a scenario of rather strong serial dependence.

(2) All tests loose power as the tails of the innovation distribution increase, but the loss is much more pronounced for the variance than for the other estimators. The distributions listed in the tables are in ascending order according to their kurtosis. The kurtoses of $N(0,1)$, $NM(3,1\%)$, $L(0,1)$, t_5, and t_3 are 0, 1.63, 3, 6, and ∞, respectively. General formulae are given in Table 3.

(3) The tests generally have a higher power for $\theta = 3/4$ than for $\theta = 1/4$. This may seem odd at first glance since in both cases the change occurs equally far away from the center of the sequence. However, since we always consider changes from a smaller ($\lambda = 1$) to a larger scale ($\lambda = 1.5$ or 2), a sequence with $\theta = 1/4$ has a higher overall

Table 6: *Test power.* Rejection frequencies (%) at asymptotic 5% level. Same set-up as Table 5 except scale changes by **factor** $\lambda = 2.0$.

Change location:	$[n/4]$				$[n/2]$				$[n3/4]$			
Estimator:	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
$n = 60$												
$N(0,1)$	2	4	18	21	18	18	46	39	7	5	14	9
$L(0,1)$	1	1	7	12	7	8	21	22	5	3	8	10
$NM(3,0.01)$	3	4	16	21	15	16	40	32	6	4	14	9
t_5	2	3	9	12	8	12	29	24	4	3	8	11
t_3	1	2	5	10	5	7	18	19	4	3	8	12
$n = 120$												
$N(0,1)$	17	27	53	52	79	86	94	84	58	57	75	40
$L(0,1)$	6	11	22	22	34	58	68	47	25	30	42	20
$NM(3,0.01)$	11	20	43	45	65	80	88	81	47	49	66	37
t_5	6	14	26	26	43	68	75	63	31	40	49	28
t_3	3	8	14	16	18	46	51	45	15	29	35	22
$n = 240$												
$N(0,1)$	72	89	97	94	100	100	100	100	99	99	100	98
$L(0,1)$	24	57	66	54	86	98	98	94	76	86	89	77
$NM(3,0.01)$	52	83	92	92	94	100	100	100	91	98	99	97
t_5	26	70	76	77	85	99	99	98	77	92	93	90
t_3	9	45	45	49	46	92	89	94	43	76	75	78
$n = 500$												
$N(0,1)$	100	100	100	100	100	100	100	100	100	100	100	100
$L(0,1)$	81	100	100	98	100	100	100	100	99	100	100	100
$NM(3,0.01)$	91	100	100	100	100	100	100	100	100	100	100	100
t_5	76	100	100	100	98	100	100	100	97	100	100	100
t_3	30	94	89	98	75	100	99	100	72	98	98	100

variability and hence yields a larger long-run variance estimate than a sequence with $\theta = 3/4$. Since we divide the test statistics by the (root of the) long-run variance estimates, this implies a difference in power.

(4) The GMD-based test turns out to have the overall best performance, with the $Q_n^{0.8}$ and also MD not trailing far behind. In the independence setting, the $Q_n^{0.8}$ is the best for the t_3 distribution.

(5) It is interesting to note that all the competitors, GMD, $Q_{0.8}$, and MD, dominate the variance-based test also under normality. The explanation is that a multiplicative change in scale, as we consider here, tends to blow up the long-run variance estimate $\hat{D}_{\sigma^2}$ much more than the corresponding estimates for the other estimators. To illustrate this, consider a sequence of i.i.d. $N(0,1)$-variates $X_1, \ldots, X_{\lfloor n/2 \rfloor}$ followed by a sequence of i.i.d. $N(0,4)$-variates $X_{\lfloor n/2 \rfloor + 1}, \ldots, X_n$. For large n, the quantity that $\hat{D}_{\sigma^2}^2$ estimates when applied to $X_1, \ldots, X_n$ can be seen to be $ASV(\sigma_n^2; NM(\gamma = 2, \epsilon = 1/2)) = E(Y^4) - E(Y^2)^2$ for $Y \sim NM(\gamma = 2, \epsilon = 1/2)$,

Table 7: *Test power.* Rejection frequencies (%) at asymptotic 5% level. Change-point tests based on variance (Var), mean deviation (MD), Gini's mean difference (GMD), and $Q_n^{0.8}$. **AR(1) process with** $\rho = 0.8$; scale changes by **factor** $\lambda = 1.5$. Several sample sizes, marginal distributions, and change locations. HAC bandwidth: $b_n = 2n^{1/3}$.

Change location:	[n/4]				[n/2]				[n3/4]			
Estimator:	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
n = 60												
$N(0,1)$	11	10	20	20	17	13	30	25	16	11	24	19
$L(0,1)$	10	8	19	18	11	8	21	20	12	7	18	18
$NM(3,0.01)$	9	9	19	20	15	12	25	25	15	8	23	20
t_5	9	9	18	20	14	12	23	24	12	9	20	19
t_3	8	6	15	17	10	10	21	23	11	7	18	17
n = 120												
$N(0,1)$	8	10	18	19	20	22	36	30	16	14	25	19
$L(0,1)$	7	8	15	15	18	17	30	24	12	11	21	13
$NM(3,0.01)$	10	11	18	18	18	20	32	24	15	14	24	16
t_5	8	8	16	15	15	16	28	22	14	11	22	16
t_3	5	6	12	11	11	14	22	20	8	7	16	14
n = 240												
$N(0,1)$	12	16	26	23	32	36	49	43	27	24	34	28
$L(0,1)$	9	12	20	17	27	30	40	32	20	19	27	22
$NM(3,0.01)$	10	14	22	22	29	32	44	40	24	22	34	26
t_5	8	11	18	18	26	30	40	32	20	21	30	23
t_3	4	8	13	12	15	22	29	22	14	14	21	18
n = 500												
$N(0,1)$	25	32	42	41	72	73	80	74	56	51	60	54
$L(0,1)$	18	27	34	32	53	60	65	63	43	40	47	42
$NM(3,0.01)$	23	33	43	38	64	68	75	71	50	48	56	51
t_5	17	28	36	29	52	61	67	64	41	40	46	45
t_3	8	18	22	21	30	45	48	45	25	32	36	31

whereas $\hat{D}_d^2$ estimates $ASV(d_n; NM(2,1/2)) = E(Y^2) - (E|Y|)^2$ with Y as before. Compared to a stationary sequence of $N(0,1)$-variates, the former quantity is blown up by the factor 19.25, the latter only by 2.93.

6. Data Example

We consider two data examples: a hydrological and a financial time series. According to our knowledge, neither has been analyzed in a change-point context. The first data set consists of the annual maximum discharge (in cubic meters per second) of the river Rhine at Cologne, Germany, in the years 1817 to 2009 ($n = 193$). The time series is plotted in Figure 2, top row. Figure 3 depicts a normal q-q plot, which reveals that the marginal distribution is fairly normal. Furthermore, the autocorrelation function of the data and

Table 8: *Test power.* Rejection frequencies (%) at asymptotic 5% level. Same set-up as Table 7 except scale changes by **factor** $\lambda = 2.0$.

Change location:	[n/4]				[n/2]				[n3/4]			
Estimator:	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
$n = 60$												
$N(0,1)$	13	11	27	26	24	18	41	39	16	10	27	22
$L(0,1)$	14	12	26	21	20	15	38	34	16	8	25	20
$NM(3,0.01)$	12	11	27	24	22	16	41	39	18	10	29	20
t_5	10	10	22	25	18	14	36	34	15	8	25	18
t_3	8	8	18	21	15	12	31	29	15	9	23	21
$n = 120$												
$N(0,1)$	12	16	30	29	36	38	59	51	28	25	41	26
$L(0,1)$	10	11	24	23	26	31	50	38	25	21	38	24
$NM(3,0.01)$	11	13	27	26	33	38	56	47	27	22	40	24
t_5	10	12	25	24	26	29	48	41	24	20	37	24
t_3	7	8	18	19	20	24	40	33	15	14	28	18
$n = 240$												
$N(0,1)$	20	32	52	44	70	77	87	80	63	60	74	54
$L(0,1)$	14	22	37	37	58	70	79	66	52	51	64	42
$NM(3,0.01)$	16	29	45	43	67	78	85	74	57	55	68	51
t_5	14	24	36	35	54	68	78	67	48	49	60	44
t_3	8	17	26	23	33	49	59	54	35	37	47	30
$n = 500$												
$N(0,1)$	56	80	88	85	98	99	100	99	95	94	97	96
$L(0,1)$	37	64	74	70	92	97	97	96	88	89	93	87
$NM(3,0.01)$	45	76	83	77	95	99	99	99	91	91	94	93
t_5	34	65	74	73	88	96	97	98	84	87	90	86
t_3	18	45	51	50	64	88	88	87	60	72	76	67

the squared mean-centered data are plotted. They indicate a weak serial dependence. We thus apply the change-point tests from the previous section with long-run variance estimation settings as before except we set the HAC-bandwidth to $b_n = 4$. The latter is in consistency with the other data example. The results are very similar for smaller bandwidths.

The change-point test processes $\hat{D}_s^{-1}(k/\sqrt{n}|s_{1:k} - s_{1:n}|)_{2 \le k \le n}$ show a fair agreement for $s_n = \sigma_n^2, d_n, g_n$, and $Q_n^{0.8}$, cf. Figure 2, middle and bottom row. All attain their maxima at 1919 with p-values ranging from 0.021 to 0.046, i.e., they confirm the existence of a change in scale around 1920, which is suggested by a visual inspection of the series. This change coincides with the implementation of a variety of structural river works upstream from Cologne, particularly along the Rhine's tributaries Main and Neckar in the early 1920s. For illustration purposes, we also plot the corresponding curves for the original Q_n ($\alpha = 1/4$) and the MAD in the bottom row of Figure 2. Both are very rugged, distinctively different from the other curves, and yield p-values above 5%.

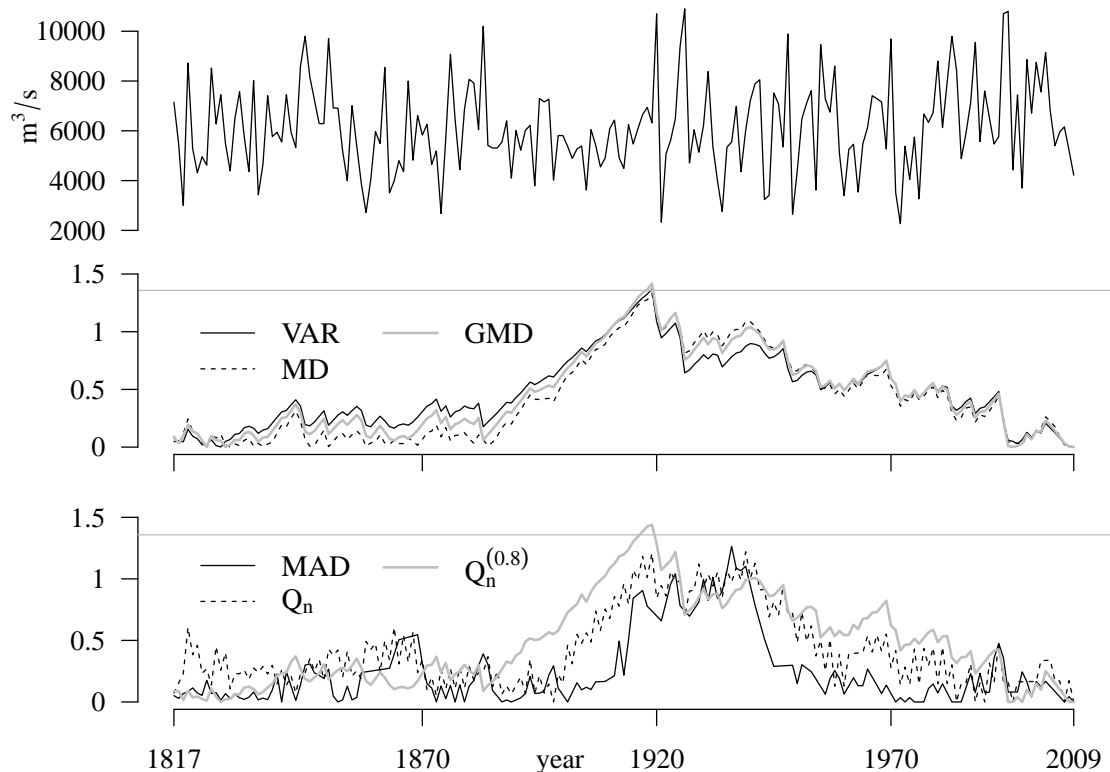The second example consists of the log returns of the daily closings of the Volkswagen

Figure 2: Top row: annual maximum discharge (m^3/s) of the river Rhine at Cologne, Germany, between 1817 and 2009. Middle and bottom rows: change-point processes $\hat{D}_s^{-1}\left(k/\sqrt{n}|s_{1:k} - s_{1:n}|\right)_{2 \leq k \leq n}$ for estimators $s_n = \sigma_n^2$, d_n, g_n (middle) and m_n, Q_n, and $Q_n^{0.8}$ (bottom); HAC kernel bandwidth: $b_n = 4$.

share, traded at the German stock exchange in Frankfurt, within the last three quarters of the year 2001 ($n = 196$). The impact of 9/11 on the volatility of the series is clearly visible, cf. Figure 4, top row. In the "normal" situation of the first example, all tests were in fair agreement (except the previously discarded MAD and Q_n). This example differs in a variety of features, which shall illustrate the differences of the tests. The normal q-q plot shows heavier than normal tails, and the auto-correlation function of the squared sequence reveals some serial dependence (cf. Figure 5). Furthermore, there is a short period of strong oscillation from July 10–12, 2001 (the reason for which is not known to us). Removing these three dates from the series, all tests consistently detect a change with p-values of 1% and less and place it at the beginning of September. However, *with* those three days, the variance-based change-point test process $\hat{D}_{\sigma^2}^{-1}\left(k/\sqrt{n}|\sigma_{1:k}^2 - \sigma_{1:n}^2|\right)_{2 \leq k \leq n}$ attains its maximum at July 5 and yields a p-value of 0.31. This is an example where few outliers mask an apparent change. The tests based on MD and GMD behave in principle similarly, but provide some mild evidence for a change with p-values of about 0.10. The $Q_n^{0.8}$, however, gives a p-value of 0.03, and the maximum is attained at September 7.
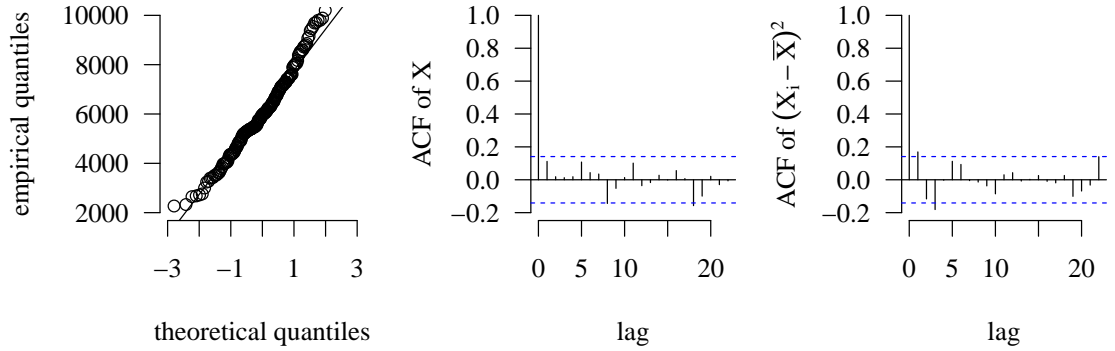
22

Figure 3: Hydrology data from Figure 2: Normal q-q plot, ACF of data, and ACF of squared centered data.

The curves for the MAD and Q_n are plotted as well. Both curves, again, look distinctively different from the others, and the respective tests do not provide strong evidence for a change. The tests were carried out, as before, with $b_n = 4$ and the other long-run variance estimation parameters as in Section 5. All simulations and data analyses were performed in R (R Core Team, 2015).

7. Conclusion and Outlook

We have studied the problem of detecting changes in scale beyond the established sum-of-squares methodology. We have considered test statistics based on alternative scale measures, which have a better outlier-resistance and a better efficiency at heavy-tailed distributions than the sample variance: the mean deviation, the median absolute deviation (MAD), Gini's mean difference, and the Q_n^α. The MAD and the original Q_n (i.e., Q_n^α with $\alpha = 1/4$) may be confidently discarded for the purpose of change-point detecting, whereas the mean deviation, Gini's mean difference, and the Q_n^α for $0.7 < \alpha < 0.9$ provide very good alternatives, that improve upon the classical test not only under heavy tails but also under normality. We found Gini's mean difference and the Q_n^α, which are both based on pairwise differences, to altogether outperform the mean deviation. Our general recommendation is to use the Gini's mean difference based test in case of normal or near-normal data situations, and to use the $Q_n^{0.8}$ or $Q_n^{0.75}$ if the occurrence of gross errors is suspected.

We have confined our investigation here to a number of popular and explicit scale estimators. The choice of the estimators is based upon a prior assessment of the efficiency and robustness (where we understand the latter as "high efficiency over a broad range of distributions" rather than "high breakdown-point") of various scale estimators, and we suspect that the tests may not be substantially improved by plugging in other scale estimators. Nevertheless, the question arises, what other alternative are possible and might be used for the problem at hand. For an overview of approaches to robust scale estimation we refer the reader to Huber and Ronchetti (2009, Chapter 5). For a numeri-
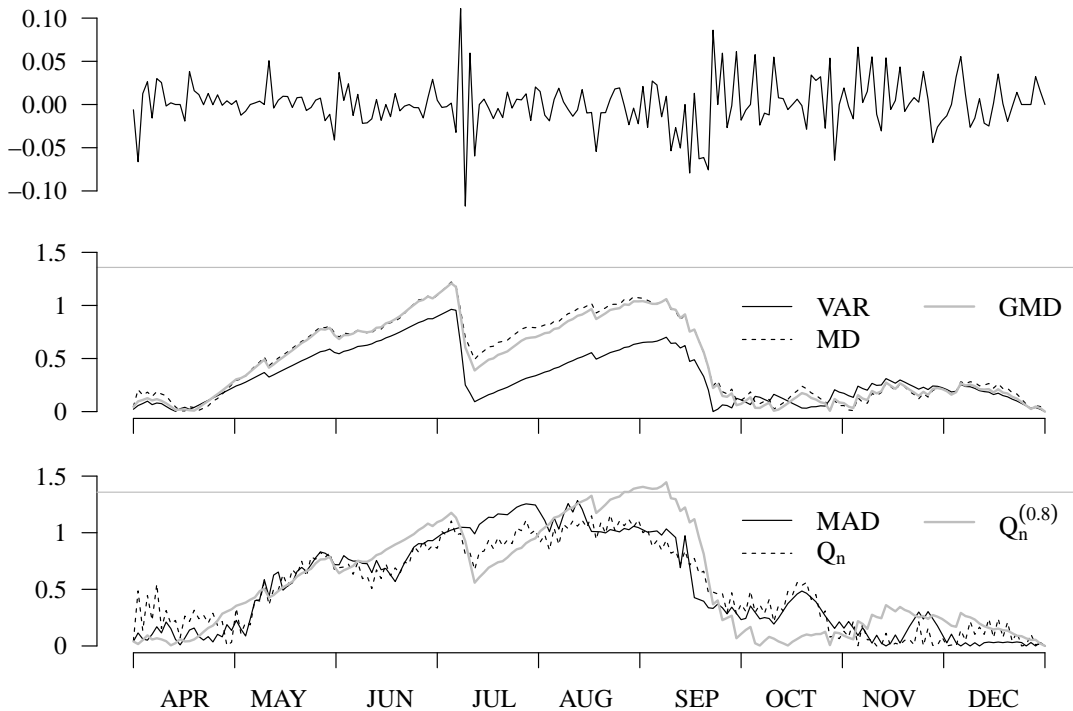
Figure 4: Top row: daily log returns of VW share from April to December 2001. Middle and bottom rows: change-point processes $\hat{D}_s^{-1}\left(k/\sqrt{n}|s_{1:k} - s_{1:n}|\right)_{2 \le k \le n}$ for estimators $s_n = \sigma_n^2$, d_n, g_n (middle) and m_n, Q_n, and $Q_n^{0.8}$ (bottom); HAC kernel bandwidth: $b_n = 4$.

cal comparison of scale estimators, see, e.g., Lax (1985). One common way to safeguard against outliers is to use truncation or Winsorization. Lee and Park (2001) consider a version of the σ^2-based test based with truncated observations. Their simulation results suggest quite a substantial loss in power under normality, whereas we observe the opposite for our robustification approach. There is another conceptual advantage of pairwise differences: there is no location to estimate. At skewed distributions, taking, e.g., the mean or the median leads to distinctively different scale estimators. This ambiguity does not exist for pairwise-difference-based estimators.

Gini's mean difference and the Q_n^α are based on the kernel $g(x, y) = |x - y|$ of order two. We have noted that both have a considerably higher efficiency under normality as compared to the respective estimators based on distances to the central location. So, may kernels of higher order even be better? A general observation seems to be that this is not the case, see, e.g., Rousseeuw and Croux (1992, Section 4). Higher-order-kernels require a higher computational effort, but tend to provide neither better efficiency nor robustness.

A crucial part of all change-point tests for dependent data is the estimation of the
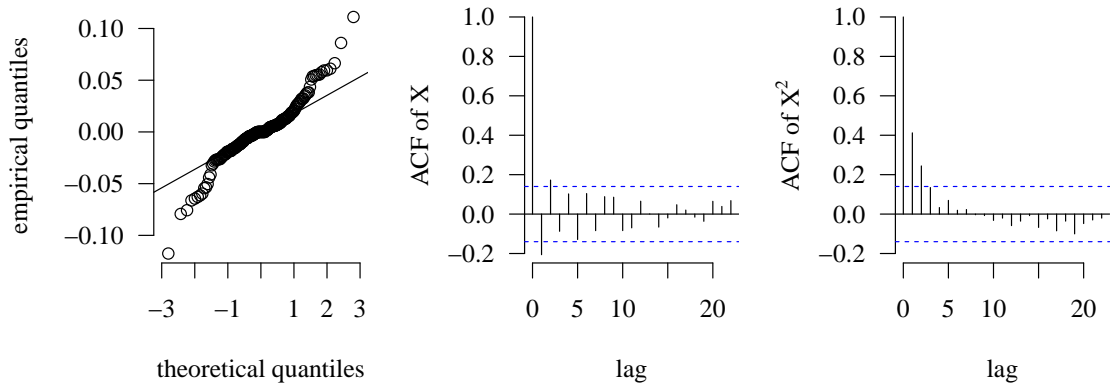
Figure 5: Financial data from Figure 4: Normal q-q plot, ACF of data, and ACF of squared data.

long-run variance. We have proposed estimators based on HAC kernel estimation, which is common in the change-point context. An alternative estimation technique is block sub-sampling, see, e.g., Dehling et al. (2013) and the references therein. An entirely different approach, which avoids any unknown scaling constants in the limit distribution of the test statistic is the self-normalization approach as proposed by Shao and Zhang (2010).

Yet an alternative way of assessing critical values is to estimate the distribution of the test statistic using bootstrapping. A variety of bootstrap procedures have been proposed for dependent data, e.g., the block bootstrap (Künsch, 1989), the stationary bootstrap (Politis and Romano, 1994), the tapered block bootstrap (Paparoditis and Politis, 2001), or the dependent wild (or multiplier) bootstrap (Shao, 2010).

The block bootstrap for U-statistics (such as Gini's mean difference) has been studied by Dehling and Wendler (2010). Recently, Bücher and Kojadinovic (2016b) showed the consistency of the dependent multiplier bootstrap for U-statistics and also established the validity of dependent multiplier bootstrap procedures for change-point test statistics based on this class of statistics. For quantiles, the block bootstrap was investigated by Sun and Lahiri (2006) and the multiplier bootstrap by Doukhan et al. (2015). For U-quantiles, such as the Q_n^α, we are unaware of any work concerning bootstrap methods.

Finally, the consideration of high-breakdown-point estimators such as the MAD and the original Q_n, which have turned out to be rather unsuited for the problem at hand, prompts the question of what kind of robustness can be expected and is desired in a change-point context, and if this can be mathematically formalized and quantified.

Acknowledgement

for the stock exchange data set.

A. Supplementary material for Section 4

A.1. The population value and the asymptotic variance of Q_n^α

General expressions. Let $X, Y \sim F$ be independent and let F possess a Lebesgue density f. Generally, $Q^\alpha = Q^\alpha(F) = U^{-1}(\alpha)$, where U is the distribution function of $|X - Y|$ and U^{-1} the corresponding quantile function, i.e., $U^{-1}(\alpha) = \inf\{t \mid \alpha \leq U(t)\}$.

For the asymptotic variance of the Q_n^α, we analyze (7) for independent sequences $X_1, \ldots, X_n \sim F$:

$$ASV(Q_n^\alpha; F) = \frac{4 \, E\psi^2(X)}{u^2(Q^\alpha)}, \qquad X \sim F,$$

where

$$E\psi^2(X) = \int_R \{F(x + Q^\alpha) - F(x - Q^\alpha)\}^2 \, f(x) dx - \alpha^2,$$

and u is the density of $|X - Y|$, i.e.,

$$u(x) = 2 \int_{\mathbb{R}} f(t) f(t - x) dt, \qquad x \geq 0. \tag{11}$$

The corresponding cdf U is then given by

$$U(x) = \int_0^x u(y) dy, \qquad x \geq 0. \tag{12}$$

Specific expressions. For distributions F, where the convolution with itself admits a tractable form, more explicit expressions for the above terms can be given. For instance, we have for the standard normal distribution $N(0,1)$ with density ϕ and cdf Φ:

$$Q^\alpha = \sqrt{2}\Phi^{-1}\left\{(\alpha + 1)/2\right\}, \quad u(x) = \sqrt{2}\phi(x/\sqrt{2}), \; (x \geq 0).$$

For the t_1 or Cauchy distribution we obtain

$$Q^\alpha = 2\tan(\pi\alpha/2), \quad u(x) = 4\pi^{-1}(x^2 + 4)^{-1}, \; (x \geq 0),$$

and

$$ASV(Q_n^\alpha; t_1) = \frac{\pi}{4}\{(Q^\alpha)^2 + 4\}^2 \left\{ \int_{\mathbb{R}} \frac{\pi^2[\arctan(x + Q^\alpha) - \arctan(x - Q^\alpha)]^2}{1 + x^2} dx - \alpha^2 \right\}.$$

For the Laplace distribution $L(0,1)$, cf. Table 3, we get

$$u(x) = 1/2(1 + x)e^{-x}, (x \geq 0), \quad U(x) = 1 - e^{-x} - x/2e^{-x}, (x \geq 0),$$

Q^α is obtained by numerically solving $U(x) = \alpha$, and

$$ASV(Q_n^\alpha; L(0,1)) = 16 \frac{1 - (1 + Q^\alpha)e^{-Q^\alpha} - 1/6\,e^{-Q^\alpha}\{1 - e^{-Q^\alpha}\}^2 - \alpha^2}{(1 + Q^\alpha)^2\,e^{-2Q^\alpha}}.$$

A.2. The asymptotic variance of the t_ν maximum-likelihood estimator for scale.

In Figure 1, the asymptotic relative efficiencies of the Q_n^α at various distributions with respect to the respective scale maximum likelihood estimators are displayed as a function of α. The asymptotic variance of the normal scale MLE (the standard deviation) at $N(0,1)$ is $1/2$, the asymptotic variance of the Laplace scale MLE (the mean deviation) at $L(0,1)$ is 1. Lemma A.1 below yields the asymptotic variance of the t_ν scale MLE. The authors believe that this result is known in the literature, but were not able to find a reference for exactly this result and would be grateful for any information regarding an earlier reference.

Consider, for fixed $\nu > 0$, the t_ν scale family, i.e., the parametric family of densities $\{f_{\nu,s} \mid s > 0\}$ with

$$f_{\nu,s}(x) = \frac{c_\nu}{s}\left(1 + \frac{x^2}{\nu s^2}\right)^{-(\nu+1)/2}, \quad x \in \mathbb{R}, \qquad \text{with } c_\nu = \sqrt{\nu}B(\nu/2, 1/2),$$

where B denotes the beta function. Let $F_{\nu,s}$ denote the corresponding distribution function. Then, the maximum likelihood estimator $\hat{s}_n$ for s at an i.i.d. sample $X_1, \ldots, X_n \sim F_{\nu,s}$ satisfies $\sqrt{n}(\hat{s}_n - s) \xrightarrow{d} N(0, I(\nu,s)^{-1})$, where $I(\nu,s) = E\left[\{\partial/(\partial s) \log f_{\nu,s}(X)\}^2\right]$ with $X \sim F_{\nu,s}$ denotes the Fisher information about s contained in X.

Lemma A.1. $I(\nu,s) = 2\nu(\nu+3)^{-1}s^{-2}$.

Proof. $I(\nu,s) = s^{-2}I(\nu,1)$

$$= \frac{c_\nu}{s^2}\int_\mathbb{R}\left(\frac{(\nu+1)x^2}{\nu+x^2} - 1\right)^2\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}dx = \frac{c_\nu}{s^2}\int_\mathbb{R}(x^2-1)^2\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+5}{2}}dx$$

Substituting $y = \sqrt{(\nu+4)/\nu}\,x$, we transform this integral to

$$\frac{c_\nu}{s^2}\sqrt{\frac{\nu}{\nu+4}}\int_\mathbb{R}\left\{\left(\frac{\nu}{\nu+4}\right)^2 y^4 - 2\frac{\nu}{\nu+4}y^2 + 1\right\}\frac{1}{c_{\nu+4}}f_{\nu+4,1}(y)dy$$

For $Y \sim t_d$, we have $EY^4 = 3d^2(d-2)^{-1}(d-4)^{-1}$, $d > 4$, and $EY^2 = d/(d-2)$, $d > 2$. Applying this with $d = \nu + 4$, we obtain

$$s^{-2}I(\nu,1) = \frac{1}{s^2}\frac{c_\nu}{c_{\nu+4}}\sqrt{\frac{\nu}{\nu+1}}\frac{2(\nu+1)}{\nu+2}.$$

With $c_\nu = \Gamma(\frac{\nu+1}{2})/(\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2}))$ we arrive at the expression given in Lemma A.1. $\qquad\square$

Remark: If we consider (more realistically) the location-scale t_ν-family, where location and scale are estimated simultaneously, the scale maximum-likelihood estimator $\hat{s}_n$ has the same asymptotic variance $(\nu+3)s^2/(2\nu)$ as computed above. Due to the symmetry of the distribution, the maximum-likelihood estimators of location and scale are uncorrelated by an invariance argument and the 2×2 Fisher information matrix is diagonal. For a discussion on the uniqueness of the solutions to the t_ν likelihood equations in the location-and-scale case and the scale-only case in the univariate as well as the multivariate setting, see Kent and Tyler (1991) and the references therein.

B. Proofs

Before stating the proofs of Theorems 3.1 and 3.2, we want to provide an intuitive explanation for the expressions (6) and (7) of the long-run variances for Gini's mean difference and the Q_n^α, respectively. Let $(X_i)_{i \in \mathbb{Z}}$ be a stationary sequence with marginal distribution F, and furthermore $X, Y \sim F$ be independent. Gini's mean difference is a U-statistic with kernel $g(x, y) = |x - y|$. The main tool for deriving asymptotics for U-statistics is the Hoeffding decomposition. By letting

$$g_0 = Eg(X, Y), \quad g_1(x) = Eg(X, Y) - g_0, \quad g_2(x, y) = g(x, y) - g_1(x) - g_1(y) - g_0,$$

we can decompose the U-statistic

$$G_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} g(X_i, X_j)$$

into

$$\sqrt{n}(G_n - g_0) = \frac{2}{\sqrt{n}} \sum_{i=1}^{n} g_1(X_i) + \frac{2}{\sqrt{n}(n-1)} \sum_{1 \leq i < j \leq n} g_2(X_i, X_j).$$

The first term on the right-hand side is called the linear part and the second term is the degenerate part of the U-statistic. Under appropriate regularity conditions, the degenerate part can be seen to vanish asymptotically, and hence the asymptotic variance of G_n is determined by that of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(X_i)$, which, by a central limit theorem for dependent sequences, can be seen to be $\sum_{i=-\infty}^{\infty} E\{g_1(X_0)g_1(X_h)\}$. The fact that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_1(X_i)$ appears twice in the Hoeffding decomposition explains the factor 4 in (6) and (7). The asymptotic variance of G_n carries over to the "change-point process" $\{kn^{-1/2}(G_k - G_n)\}_{2 \leq k \leq n}$. A detailed proof for general U-statistics under the dependence scenario considered here is given by Dehling et al. (2016).

The estimator Q_n^α is a U-quantile, i.e., instead of taking the first sample moment of $|X_i - X_j|$, $1 \leq i < j \leq n$, we consider the sample α-quantile of these $\binom{n}{2}$ values. An essential step in the asymptotic analysis of U-quantiles is to relate the empirical quantile function U_n^{-1} of $|X_i - X_j|$, $1 \leq i < j \leq n$, to the corresponding empirical distribution function U_n by means of a generalized Bahadur representation:

$$Q_n^\alpha = U_n^{-1}(\alpha) = Q^\alpha + \frac{\alpha - U_n(Q^\alpha)}{u(Q^\alpha)} + R_n,$$

where Q^α and u are as in Section A.1, and R_n is remainder term, which, under appropriate regularity conditions, converges to zero sufficiently fast. Then, recalling that

$$U_n(t) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \mathbb{1}_{\{|X_i - X_j| \leq t\}}, \qquad t \in \mathbb{R},$$

yields

$$\sqrt{n}(Q_n^\alpha - Q^\alpha) = \frac{\sqrt{n}}{u(Q^\alpha)} \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \left(\alpha - \mathbb{1}_{\{|X_i - X_j| \leq Q^\alpha\}}\right) + \sqrt{n}R_n.$$

This explains on the one hand the appearance of the density $u(x)$ in the long-run variance (7) and on the other hand relates the U-quantile Q^α to a U-statistic with kernel $h(x, y) = \mathbb{1}_{\{|X_i - X_j| \leq Q^\alpha\}}$, which is then further treated by the Hoeffding decomposition as described above. The function $\psi(x)$ in (7) can be seen to be the linear kernel $g_1(x)$ associated with the kernel $h(x, y)$.

Proof of Theorem 3.1. The theorem is a corollary of Corollary 2.8 of Dehling et al. (2016). There are four conditions imposed there, labeled Assumptions 2.2, 2.3, 2.4 and 2.6. Assumption 2.2 and 2.3 in Dehling et al. (2016) are essentially our Assumption 3.3 (a) and (b). Assumption 2.3 in Dehling et al. (2016) is a condition on the moments on the kernel $g(X_i, X_j)$, which in our case $g(x, y) = |x - y|$ directly translate into moment condition on the data itself. Assumption 2.6 concerns the long-run variance and is identical to our Assumption 3.1. Thus it remains Assumption 2.4 of Dehling et al. (2016), which is as follows:

Assumption B.1. *There are positive constants $L, \epsilon_0 > 0$ such that for all $\epsilon \in (0, \epsilon_0)$*

$$E \left(\sup_{|x-X| \leq \epsilon, |y-Y| \leq \epsilon} |g(x, y) - g(X, Y)| \right)^2 \leq L\epsilon,$$

where X, Y are independent with the same distribution as X_0.

This condition is also known as the variation condition. It can be understood as a form of Lipschitz continuity of the kernel g with respect to F. Since in our case $g(x, y) = |x-y|$, g itself is Lipschitz continuous, this condition is fulfilled for any distribution F. $\square$

Proof of Theorem 3.2. This is a corollary of Corollary 2.5 (B) of Vogel and Wendler (2016). This statement required six conditions, which are labeled Assumptions 1 to 6. Assumptions 1, 4, and 5 in Vogel and Wendler (2016) coincide with our Assumptions 3.4, 3.2, and 3.1, respectively. They concern the serial dependence, the kernel density estimation, and the HAC kernel estimation, respectively. Furthermore, Assumption 6 in Vogel and Wendler (2016) is Assumption B.1 above, which is fulfilled since the kernel $g(x, y) = |x - y|$ is Lipschitz continuous. Assumption 2 in Vogel and Wendler (2016) is a similar variation condition for the kernel $h_t(x, y) = \mathbb{1}_{\{|x-y| \leq t\}}$, which is required to hold not only for $t = Q^\alpha$, but, slightly stronger, for all t in some neighborhood of Q^α. The boundedness of f (our Assumption 3.5(a)) is sufficient for this. Thus it remains to show Assumption 3 of Vogel and Wendler (2016), which is the following smoothness condition on the distribution function $U(t) = P(|X - Y| \leq t)$, $t \geq 0$, with $X, Y \sim F$ independent.

Assumption B.2. *Let U be differentiable in a neighborhood of $Q^\alpha = U^{-1}(\alpha)$ with $u(t) = U'(t)$. Furthermore,*

(a) there are constants $c, C > 0$ such that $c \leq u(t) \leq C$ in a neighborhood of Q^α and

(b) $\left| \dfrac{U(t) - \alpha}{t - Q^\alpha} - u(Q^\alpha) \right| = O\left(|t - Q^\alpha|^{1/2}\right)$ *for* $t \to Q^\alpha$.

It remains to show that this condition is implied by Assumption 3.5. Recall the representations (12) and (11) for U and the corresponding density u, respectively. Assumptions 3.5 (a) and (b), i.e., f is bounded and its support "has no gaps", ensure that $u(t)$ stays away from 0 and ∞ at any point strictly between $\inf\{x|u(x) > 0\}$ and $\sup\{x|u(x) > 0\}$, i.e., Assumption B.2 (a). From Assumptions 3.5 (a) and (c), we find that $|u(x) - u(y)| \leq 2m||f||_\infty^2|x - y|$, where m denotes the number of intervals. Hence u is Lipschitz continuous. Furthermore, for x, y close enough such that u is monotonic between both (and without loss of generality we assume $u(x) \leq u(y)$), we have $u(x) \leq |(U(y) - U(x))/(y - x)| \leq u(y)$ and hence

$$\left| \frac{U(y) - U(x)}{y - x} - u(x) \right| = O\left(|u(x) - u(y)|\right) = O(|x - y|) \quad \text{as } x \to y$$

and hence B.2(b) holds. This completes the proof. $\qquad\square$

References

D. W. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991.

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

G. E. P. Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

A. Bücher and I. Kojadinovic. A dependent multiplier bootstrap for the sequential empirical copula process under strong mixing. *Bernoulli*, 22(2):927–968, 2016a.

A. Bücher and I. Kojadinovic. Dependent multiplier bootstraps for non-degenerate U-statistics under mixing conditions with applications. *Journal of Statistical Planning and Inference*, 170:83–105, 2016b.

E. Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Annals of Statistics*, 14(3):1171–1179, 1986.

R. M. de Jong and J. Davidson. Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, 68(2):407–424, 2000.

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

H. Dehling, R. Fried, O. S. Sharipov, D. Vogel, and M. Wornowizki. Estimation of the variance of partial sums of dependent processes. *Statistics & Probability Letters*, 83(1):141–147, 2013.

H. Dehling, D. Vogel, M. Wendler, and D. Wied. Testing for changes in Kendall's tau. *Econometric Theory*, available online, 2016.

P. Doukhan, G. Lang, A. Leucht, and M. H. Neumann. Dependent wild bootstrap for the empirical process. *Journal of Time Series Analysis*, 36(3):290–314, 2015.

A. Dürre, R. Fried, and T. Liboschik. Robust estimation of (partial) autocorrelation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(3):205–222, 2015.

R. A. Fisher. A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean square. *Monthly Notices of the Royal Astronomical Society*, 80: 758–770, 1920.

C. Gerstenberger and D. Vogel. On the efficiency of Gini's mean difference. *Statistical Methods & Applications*, 24(4):569–596, 2015.

E. Gombay, L. Horváth, and M. Husková. Estimators and tests for change in variances. *Statistics & Risk Modeling*, 14(2):145–160, 1996.

P. Hall, J. L. Horowitz, and B.-Y. Jing. On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574, 1995.

F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley, 2nd edition, 2009.

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

J. T. Kent and D. E. Tyler. Redescending M-estimates of multivariate location and scatter. *Annals of Statistics*, 19(4):2102–2119, 1991.

I. Kojadinovic, J.-F. Quessy, and T. Rohmer. Testing the constancy of Spearman's rho in multivariate time series. *Annals of the Institute of Statistical Mathematics*, 65(5):292–954, 2015.

H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241, 1989.

D. A. Lax. Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, 80(391):736–741, 1985.

S. Lee and S. Park. The cusum of squares test for scale changes in infinite order moving average processes. *Scandinavian Journal of Statistics*, 28(4):625–644, 2001.

Y. Ma and M. G. Genton. Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis*, 21(6):663–684, 2000.

U. S. Nair. The standard error of Gini's mean difference. *Biometrika*, 28(3/4):428–436, 1936.

E. Paparoditis and D. N. Politis. Tapered block bootstrap. *Biometrika*, 88(4):1105–1119, 2001.

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

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

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

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/.

P. J. Rousseeuw and C. Croux. Explicit scale estimators with high breakdown point. In Y. Dodge, editor, *L1-Statistical analysis and related methods*, North-Holland. 77–92, 1992.

P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

X. Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489): 218–235, 2010.

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

S. Sun and S. N. Lahiri. Bootstrapping the sample quantile of a weakly dependent sequence. *Sankhyā: The Indian Journal of Statistics*, 68(1):130–166, 2006.

J. W. Tukey. *A survey of sampling from contaminated distributions*. In I. Olkin et al., editor, Contributions to Probability and Statistcs. Essays in Honor of Harold Hotteling. Stanford University Press, 1960.

D. Vogel and M. Wendler. Studentized sequential u-quantiles under dependence with applications to change-point analysis. *Bernoulli*, available online, 2016.

D. Wied, M. Arnold, N. Bissantz, and D. Ziggel. A new fluctuation test for constant variances with applications to finance. *Metrika*, 75(8):1111–1127, 2012.