Statistica Sinica **28** (2018), 1199-1224
doi:https://doi.org/10.5705/ss.202015.0435

SUBSAMPLING FOR GENERAL STATISTICS UNDER LONG RANGE DEPENDENCE WITH APPLICATION TO CHANGE POINT ANALYSIS

Annika Betken and Martin Wendler

Ruhr-Universität Bochum and Universität Greifswald

Abstract: In the statistical inference for long range dependent time series the shape of the limit distribution typically depends on unknown parameters. Therefore, we propose to use subsampling. We show the validity of subsampling for general statistics and long range dependent subordinated Gaussian processes that satisfy mild regularity conditions. We apply our method to a self-normalized change-point test statistic so that we can test for structural breaks in long range dependent time series without having to estimate nuisance parameters. The finite sample properties are investigated in a simulation study. We analyze three data sets and compare our results to the conclusions of other authors.

Key words and phrases: Change-point test, Gaussian processes, long range dependence, subsampling.

1. Introduction

1.1. Long range dependence

While most statistical research is done for independent data or short memory time series, in many applications there are also time series with long memory in the sense of slowly decaying correlations: in hydrology (starting with the work of Hurst (1956)), in finance (e.g. Lo (1989)), in the analysis of network traffic (e.g. Leland et al. (1994)), and in many other fields of research.

As model of dependent time series we consider subordinated Gaussian processes: Let $(\xi_n)_{n\in\mathbb{N}}$ be a stationary sequence of centered Gaussian variables with $\mathrm{Var}(\xi_n) = 1$ and covariance function γ satisfying

$$\gamma(k) := \mathrm{Cov}(\xi_1, \xi_{k+1}) = k^{-D} L_\gamma(k)$$

for $D > 0$ and a slowly varying function L_γ. If $D < 1$, the spectral density f of $(\xi_n)_{n\in\mathbb{N}}$ is not continuous, but has a pole at 0. The spectral density has the form

$$f(x) = |x|^{D-1} L_f(x)$$

for a function L_f which is slowly varying at the origin (see Proposition 1.1.14 in Pipiras and Taqqu (2011)). Let $G : \mathbb{R} \to \mathbb{R}$ be a measurable function such that $\mathrm{E}[G^2(\xi_1)] < \infty$. The stochastic process $(X_n)_{n\in\mathbb{N}}$ given by $X_n := G(\xi_n)$ is called long range dependent if $\sum_{n=0}^{\infty} |\mathrm{Cov}(X_1, X_{n+1})| = \infty$, and short range dependent if $\sum_{n=0}^{\infty} |\mathrm{Cov}(X_1, X_{n+1})| < \infty$.

In limit theorems for the partial sum $S_n = \sum_{i=1}^{n} X_i$, the normalization and the shape of the limit distribution not only depend on the decay of the covariances $\gamma(k)$ as $k \to \infty$, but also on the function G. More precisely, Taqqu (1979) and Dobrushin and Major (1979) independently proved that

$$\frac{1}{L_\gamma(n)^{r/2} n^H} \sum_{i=1}^{n} (X_i - \mathrm{E}[X_i]) \Rightarrow C(r,H) g_r Z_{r,H}(1)$$

if the Hurst parameter $H := \max\{1 - rD/2, 1/2\}$ is greater than 1/2. Here, r denotes the Hermite rank of the function G, $C(r,H)$ is a constant, g_r is the first non-zero coefficient in the expansion of G as a sum of Hermite polynomials, and $Z_{r,H}$ is a Hermite process. For more details on Hermite polynomials and limit theorems for subordinated Gaussian processes we recommend the book of Pipiras and Taqqu (2011). In this case ($rD < 1$), the process $(X_n)_{n\in\mathbb{N}}$ is long range dependent as the covariances are not summable. The limiting random variable $C(r,H)Z_{r,H}(1)$ is Gaussian only if the Hermite rank $r = 1$.

If $rD = 1$, the process $(X_n)_{n\in\mathbb{N}}$ might be short or long range dependent according to the slowly varying function L_γ. If $rD > 1$, the process is short range dependent. In this case, the partial sum $\sum_{i=1}^{n}(X_i - \mathrm{E}[X_i])$ has (with proper normalization) always a Gaussian limit.

There are other models for long memory processes: fractionally integrated autoregressive moving average processes can show long range dependence, see Granger and Joyeux (1980); general linear processes with slowly decaying coefficients were studied by Surgailis (1982).

1.2. Subsampling

In applications the parameters D, r, and the slowly varying function L_γ are unknown and thus the scaling needed in the limit theorems and the shape of the asymptotic distribution are not known. That makes it difficult to use the asymptotic distribution for statistical inference. The situation is even more complicated if one is not interested in partial sums, but in nonlinear statistical functionals: U-statistics can have a limit distribution that is a linear combination of random variables related to different Hermite ranks, see Beutner and Zähle (2014); self-normalized statistics typically converge to quotients of two random variables (e.g. McElroy and Politis (2007)); the change-point test proposed by Berkes et al. (2006) converges to the supremum of a fractional Brownian bridge under the alternative hypothesis.

To deal with the unknown shape of the limit distribution and to avoid the estimation of nuisance parameters, one would like to use nonparametric methods. However, Lahiri (1993) has shown that the popular moving block bootstrap might fail under long range dependence. Another nonparametric approach is subsampling (also called sampling window method), first studied by Politis and Romano (1994), Hall and Jing (1996), and Sherman and Carlstein (1996). The idea is the following: Let $T_n = T_n(X_1, \dots, X_n)$ be a series of statistics converging in distribution to a random variable T. As we typically have just one sample, we observe only one realization of T_n and therefore cannot estimate its distribution If $l = l_n$ is a sequence with $l_n \to \infty$ and $l_n = o(n)$, then T_l also converges in distribution to T and we have multiple (though dependent) realizations $T_l(X_1, \dots, X_l)$, $T_l(X_2, \dots, X_{l+1}), \dots, T_l(X_{n-l+1}, \dots, X_n)$, that can be used to calculate the empirical distribution function.

We do not need to know the limit distribution. In our example (the self-normalized change point test statistic, see Section 3), the shape of the distribution depends on two unknown parameters, but we can still apply subsampling. However, for other statistics, one needs an unknown scaling to achieve convergence. If this is the case, one has to estimate the scaling parameters before applying subsampling.

Under long range dependence the validity of subsampling for the sample mean $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ has been investigated in the literature starting with Hall, Jing and Lahiri (1998) for subordinated Gaussian processes. Nordman and Lahiri (2005) and Zhang et al. (2013) studied linear processes with slowly decaying coefficients. For the case of Gaussian processes, an alternative proof can be found in the book of Beran et al. (2013).

It was noted by Fan (2012) that the proof in Hall, Jing and Lahiri (1998) can be easily generalized to other statistics than the sample mean. Unfortunately, the assumptions on the Gaussian process are restrictive (see also McElroy and Politis (2007)). Their conditions imply that the sequence $(\xi_n)_{n\in\mathbb{N}}$ is completely regular, which might hold for some special cases (see Ibragimov and Rozanov (1978)), but excludes many examples:

Example 1 (Fractional Gaussian Noise). Let $(B_H(t))_{t\in[0,\infty)}$ be a fractional

Brownian motion, that means a centered, self-similar Gaussian process with covariance function

$$\mathrm{E}\left[B_H(t)B_H(s)\right] = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right)$$

for some $H \in (1/2, 1)$. Then, $(\xi_n)_{n\in\mathbb{N}}$ given by $\xi_n = B_H(n) - B_H(n-1)$ is called fractional Gaussian noise. By self-similarity we have

$$\begin{aligned}\operatorname{corr}\left(\sum_{i=1}^{n}\xi_i, \sum_{j=2n+1}^{3n}\xi_j\right) &= \operatorname{corr}\left\{B_H(n), B_H(3n) - B_H(2n)\right\}\\ &= \operatorname{corr}\left\{B_H(1), B_H(3) - B_H(2)\right\}.\end{aligned}$$

As a result, the correlations of linear combinations of observations in the past and future do not vanish if the gap between past and future grows. Thus, fractional Gaussian noise is not completely regular.

Jach, McElroy and Politis (2012) provide a more general result on the validity of subsampling, but under assumptions that are difficult to check in practice (Hermite rank 1, Lipschitz-continuity of G and of the test statistic T_n, see Jach, McElroy and Politis (2016)). The main aim of this paper is to establish the validity of the subsampling method for general statistics T_n without assumptions on the continuity of the statistic, on the function G, and only mild assumptions on the Gaussian process $(\xi_n)_{n\in\mathbb{N}}$. Independently of our research, similar theorems have been proved by Bai, Taqqu and Zhang (2016). We discuss their results after our main theorem in Section 2. In Section 3 we will apply our theorem to a self-normalized, robust change-point statistic. The finite sample properties of this test is dealt with in a simulation study in Section 4. The proof of the main result, and the lemmas needed, can be found in the the supplementary material, Sections S3 and S4.

2. Main Results

2.1. Statement of the Theorem

For a statistic $T_n = T_n(X_1, \ldots, X_n)$, the subsampling estimator $\hat{F}_{l,n}$ of the distribution function F_{T_n} with $F_{T_n}(t) = P(T_n \leq t)$ is, for $t \in \mathbb{R}$,

$$\hat{F}_{l,n}(t) = \frac{1}{n-l+1}\sum_{i=1}^{n-l+1} 1_{\{T_l(X_i,\ldots,X_{i+l-1})\leq t\}}.$$

Next, we state our assumptions:

Assumption 1. $(X_n)_{n\in\mathbb{N}}$ *is a stochastic process and* $(T_n)_{n\in\mathbb{N}}$ *is a sequence of* *statistics such that $T_n \Rightarrow T$ in distribution as $n \to \infty$ for a random variable T with distribution function F_T.*

This is a standard assumption for subsampling, see for example Politis and Romano (1994). If the distribution does not converge, we cannot expect the distribution of T_l to be close to the distribution of T_n.

Assumption 2. *$X_n = G(\xi_n)$ for a measurable function G and a stationary, Gaussian process $(\xi_n)_{n\in\mathbb{N}}$ with covariance function*

$$\gamma(k) := \mathrm{Cov}(\xi_1, \xi_{1+k}) = k^{-D} L_\gamma(k)$$

such that

1. *$D \in (0,1]$ and L_γ is a slowly varying function with*
$$\max_{\tilde{k}\in\{k+1,\ldots,k+2l'-1\}} \left| L_\gamma(k) - L_\gamma(\tilde{k}) \right| \le K \frac{l'}{k} \min\{L_\gamma(k), 1\}$$
for a constant $K < \infty$ and all $l' \in \{l_k, \ldots, k\}$;
2. *$(\xi_n)_{n\in\mathbb{N}}$ has a spectral density f with $f(x) = |x|^{D-1} L_f(x)$ for a slowly varying function L_f bounded away from 0 on $[0,\pi]$ such that $\lim_{x\to 0} L_f(x) \in (0,\infty]$ exists.*

We do not impose any conditions on the function G: no finite moments or continuity are required, so that our results are applicable for heavy-tailed random variables and robust test statistics. In the next subsection we will show that Assumption 2 holds for some standard examples of long range dependent Gaussian processes.

Assumption 3. *Let $(l_n)_{n\in\mathbb{N}}$ be a non-decreasing sequence of integers such that $l = l_n \to \infty$ as $n \to \infty$ and $l_n = \mathcal{O}(n^{(1+D)/2-\epsilon})$ for some $\epsilon > 0$.*

If the dependence of the underlying process $(\xi_n)_{n\in\mathbb{N}}$ gets stronger, the range of possible values for l gets smaller. A popular choice for the block length is $l \approx C\sqrt{n}$ (see for example Hall, Jing and Lahiri (1998)), which is allowed for all $D \in (0,1]$. Now, we can state our main result:

Theorem 1. *Under Assumptions $1, 2$ and 3 we have*

$$F_{T_n}(t) - \hat{F}_{l,n}(t) \xrightarrow{\mathcal{P}} 0$$

as $n \to \infty$ for all points of continuity t of F_T. If F_T is continuous, then

$$\sup_{t\in\mathbb{R}} \left| F_{T_n}(t) - \hat{F}_{l,n}(t) \right| \xrightarrow{\mathcal{P}} 0.$$

As a result, we have a consistent estimator for the distribution function of T_n. It is possible to build tests and confidence intervals based on this estimator.

If $D > 1$, the process $(\xi_n)_{n\in\mathbb{N}}$ is strongly mixing due to Theorem 9.8 in the book of Bradley (2007). The statements of Theorem 1 hold by Corollary 3.2 in Politis and Romano (1994) for any block length l satisfying $l \to \infty$ and $l = o(n)$.

In a recent article, Bai, Taqqu and Zhang (2016) have shown that subsampling is consistent for long range dependent Gaussian processes without any extra assumptions on the slowly varying function L_f, but with a stronger restriction on the block size l, namely $l = o(n^{2-2H}L_\gamma(n))$. In another article by Bai and Taqqu (2015), the validity of subsampling is shown under the mildest possible assumption on the block length ($l = o(n)$). The condition on the spectral density is slightly stronger than our condition, the case $\lim_{x\to 0} L_f(x) = \infty$ is not allowed.

2.2. Examples for our Assumptions

Example 2 (Fractional Gaussian Noise). The covariance function of fractional Gaussian Noise $(\xi_n)_{n\in\mathbb{N}}$ with Hurst parameter H can be rewritten (with the a Taylor expansion) as

$$\gamma(k) = \frac{1}{2}\left(|k-1|^{2H} - 2|k|^{2H} + |k+1|^{2H}\right) = H(2H-1)\left\{k^{-D} + h(k)k^{-D-1}\right\}$$

for $D = 2 - 2H$ and a function h bounded by a constant $M < \infty$. Hence, $L_\gamma(k) = H(2H-1)(1 + h(k)/k)$, and for all $\tilde{k} \geq k$

$$\left|L_\gamma(k) - L_\gamma(\tilde{k})\right| \leq H(2H-1)\left|\frac{h(k)}{k} - \frac{h(\tilde{k})}{\tilde{k}}\right| \leq H(2H-1)\frac{M}{k} =: K\frac{1}{k},$$

implying part 1 of Assumption 2. For the second part note that the spectral density f corresponding to fractional Gaussian noise is

$$\begin{aligned} f(\lambda) &= C(H)\{1-\cos(\lambda)\}\sum_{k=-\infty}^{\infty}|\lambda + 2k\pi|^{D-3} \\ &= \lambda^{D-1}C(H)\frac{1-\cos(\lambda)}{\lambda^2}\frac{\sum_{k=-\infty}^{\infty}|\lambda + 2k\pi|^{D-3}}{\lambda^{D-3}}, \end{aligned}$$

see Sinai (1976). The slowly varying function

$$L_f(\lambda) = C(H)\frac{1-\cos(\lambda)}{\lambda^2}\frac{\sum_{k=-\infty}^{\infty}|\lambda + 2k\pi|^{D-3}}{\lambda^{D-3}}$$

is bounded away from 0 because this holds for the first factor $\{1-\cos(\lambda)\}/\lambda^2$ and since

$$\frac{\sum_{k=-\infty}^{\infty}|\lambda + 2k\pi|^{D-3}}{\lambda^{D-3}} \geq \frac{|\lambda + 0\pi|^{D-3}}{\lambda^{D-3}} = 1.$$

Example 3 (Gaussian FARIMA processes). Let $(\varepsilon_n)_{n\in\mathbb{Z}}$ be Gaussian white noise with variance $\sigma^2 = \mathrm{Var}(\varepsilon_0)$. Then, for $d \in (0, 1/2)$, a FARIMA(0, d, 0) process $(\xi_n)_{n\in\mathbb{N}}$ is given by

$$\xi_n = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} \varepsilon_{n-j}.$$

According to Pipiras and Taqqu (2011), Section 1.3, it has the specral density

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left|1 - e^{-i\lambda}\right|^{-2d} = |\lambda|^{D-1} \frac{\sigma^2}{2\pi} \left(\frac{|\lambda|}{|1 - e^{-i\lambda}|} \right)^{1-D}$$

with $D = 1 - 2d \in (0, 1)$. As $|1 - e^{-i\lambda}| \leq \lambda$, part 2 of Assumption 2 holds. For part 1 we have, by Corollary 1.3.4 of Pipiras and Taqqu (2011), that

$$\gamma(k) = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(k-d+1)}.$$

Using Stirling's formula,

$$\gamma(k) = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} e^{-2d+1} k^{2d-1} \Big(\frac{k+d}{k}\Big)^{k+d} \Big(\frac{k}{k-d+1}\Big)^{k-d+1} \left\{1 + \mathcal{O}\left(\frac{1}{k}\right)\right\}.$$

A Taylor expansion of $(k+d)\big\{\log(k+d) - \log(k)\big\} + (k-d+1)\big\{\log(k) - \log(k-d+1)\big\}$ gives $\gamma(k) = k^{-D} L_\gamma(k)$ with $L_\gamma(k) = C + \mathcal{O}(1/k)$ for some constant C. Part 1 of Assumption 2 follows in the same way as in Example 2.

It would be interesting to know, if the sampling window method is also consistent for long range dependent linear processes and general statistics without the assumption of Gaussianity. However, this is beyond the scope of this article.

3. Applications

3.1. Robust, self-normalized change-point test

Our main motivation for considering subsampling procedures to approximate the distribution of test statistics consists in avoiding the choice of unknown parameters. As an example we consider a self-normalized test statistic that can be applied to detect changes in the mean of long range dependent and heavy-tailed time series.

Given observations $X_1, \ldots, X_n$ with $X_i = \mu_i + G(\xi_i)$ we are concerned with a decision on the change-point problem

$$\mathbf{H} : \mu_1 = \cdots = \mu_n$$

against

$$\mathbf{A} : \mu_1 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n \ \text{ for some } k \in \{1, \ldots, n-1\}.$$

Under the hypothesis **H** we assume that the data generating process $(X_n)_{n\in\mathbb{N}}$ is stationary, while under the alternative **A** there is a change in location at an unknown point in time. This problem has been widely studied: Csörgő and Horváth (1997) give an overview of parametric and non-parametric methods that can be applied to detect change-points in independent data.

Many testing procedures are based on Cusum (cumulative sum) test statistics. When applied to data sets generated by long range dependent processes, these change-point tests often falsely reject the hypothesis of no change in the mean (see also Baek and Pipiras (2014)) and are sensitive to outliers in the data.

Testing procedures that are based on rank statistics have the advantage of not being sensitive to outliers in the data. Rank-based tests were introduced by Antoch et al. (2008) for detecting changes in the distribution function of independent random variables. Wilcoxon-type rank tests have been studied by Wang (2008) in the presence of linear long memory time series and by Dehling, Rooch and Taqqu (2013) for subordinated Gaussian sequences.

The normalization of the Wilcoxon change-point test statistic, as proposed in Dehling, Rooch and Taqqu (2013), depends on the slowly varying function L_γ, the LRD parameter D, and the Hermite rank r of the class of functions $1_{\{X_i\leq x\}} - F(x)$, $x \in \mathbb{R}$. Many authors assume $r = 1$ and, while there are well-tried methods to estimate D, estimating L_γ does not seem to be an easy task. For this reason, the Wilcoxon change-point test does not seem to be suitable for applications.

To avoid these issues, Betken (2016) proposed an alternative normalization for the Wilcoxon change-point test. This normalization approach was originally established by Lobato (2001) for decision on the hypothesis that a short range dependent stochastic process is uncorrelated up to a lag of a certain order. The normalization has recently been applied to change-point test statistics: Shao and Zhang (2010) define a self-normalized Kolmogorov-Smirnov test statistic that serves to identify changes in the mean of short range dependent time series; Shao (2011) adopted the normalization so as to define an alternative normalization for a Cusum test that detects changes in the mean of short range dependent as well as long range dependent time series.

To construct a robust test statistic, we introduce the ranks $R_i := \operatorname{rank}(X_i) = \sum_{j=1}^{n} 1_{\{X_j\leq X_i\}}$ for $i = 1, \ldots, n$. It seems natural to transfer the normalization that has been used in Shao (2011) to the Cusum test statistic of the ranks in order to establish a self-normalized version of the Wilcoxon test statistic, which is robust to outliers in the data. The corresponding two-sample test statistic is

$$G_n(k) := \frac{\sum_{i=1}^{k} R_i - (k/n)\sum_{i=1}^{n} R_i}{\Big\{(1/n)\sum_{t=1}^{k} S_t^2(1,k) + (1/n)\sum_{t=k+1}^{n} S_t^2(k+1,n)\Big\}^{1/2}},$$

where

$$S_t(j,k) := \sum_{h=j}^{t}\big(R_h - \bar{R}_{j,k}\big) \quad \text{with } \bar{R}_{j,k} := \frac{1}{k-j+1}\sum_{t=j}^{k} R_t.$$

The self-normalized Wilcoxon change-point test rejects the hypothesis for large values of $\max_{k\in\{\lfloor n\tau_1\rfloor,\ldots,\lfloor n\tau_2\rfloor\}} |G_n(k)|$, where $0 < \tau_1 < \tau_2 < 1$. The proportion of the data that is included in the calculation of the supremum is restricted by τ_1 and τ_2. A common choice for these parameters is $\tau_1 = 1-\tau_2 = 0.15$; see Andrews (1993).

For long range dependent subordinated Gaussian processes $(X_n)_{n\in\mathbb{N}}$, the asymptotic distribution of the test statistic under the hypothesis $\mathbf{H}$ can be derived by the Continuous Mapping Theorem (see Theorem 1 in Betken (2016)):

$$T_n(\tau_1,\tau_2) := \max_{k\in\{\lfloor n\tau_1\rfloor,\ldots,\lfloor n\tau_2\rfloor\}} |G_n(k)| \Rightarrow$$

$$\sup_{\tau_1\le\lambda\le\tau_2} \frac{|Z_r(\lambda) - \lambda Z_r(1)|}{\big[\int_0^{\lambda}\{Z_r(t)-(t/\lambda)Z_r(\lambda)\}^2 dt + \int_0^{1-\lambda}\{Z_r^{\star}(t)-(t/(1-\lambda))Z_r^{\star}(1-\lambda)\}^2 dt\big]^{1/2}}.$$

Here, Z_r is an r-th order Hermite process with Hurst parameter $H := \max\{1 - rD/2, 1/2\}$ and $Z_t^{\star}(r) = Z_r(1) - Z_r(1-t)$. A comparison of $T_n(\tau_1,\tau_2)$ with the critical values of its limit distribution still presupposes determination of these parameters. We can bypass the estimation of D and r by applying the subsampling procedure since Assumption 1 holds.

Under the alternative $\mathbf{A}$ (change in location), we also have to find the quantiles of the distribution under the hypothesis (stationarity). As the block length l is much shorter than the sample size n, most blocks are not contaminated by the change-point so that the distribution of the test statistic does not change much. The accuracy and the power of the test will be investigated by a simulation study in Section 4.

If the distribution of X_i is not continuous, there might be ties in the data and consideration of the ranks $R_i = \sum_{j=1}^{n} 1_{\{X_j \le X_i\}}$ may not be appropriate. We propose to use a modified statistic based on the modified ranks $\tilde{R}_i = \sum_{j=1}^{n}(1_{\{X_j<X_i\}} + (1/2)1_{\{X_j=X_i\}})$ in this case. The convergence of the corresponding self-normalized change point test follows from results of Dehling, Rooch and Wendler (2017), see the supplementary material, Section S1, for details.

The test statistic $T_n(\tau_1,\tau_2)$ is designed for the detection of a single change-point. An extension of the testing procedure that allows for multiple change-points is possible by adapting Shao's testing procedure which takes this problem into consideration (see Shao (2011)). For convenience, we describe the construction of the modified test statistic in the case of two change-points. The general idea consists in dividing the sample given by $X_1, \ldots, X_n$ according to the pair (k_1, k_2) of potential change-point locations and to compute the original test statistic with respect to the subsamples $X_1, \ldots, X_{k_2}$ and $X_{k_1+1}, \ldots, X_n$. We reject the hypothesis for large values of the sum of the corresponding single statistics.

For $\varepsilon \in (0, \tau_2 - \tau_1)$ consider the test statistic $T_n(\tau_1, \tau_2, \varepsilon) := \sup_{(k_1,k_2)\in\Omega_n(\tau_1,\tau_2,\varepsilon)} |G_n(k_1, k_2)|$, where $\Omega_n(\tau_1, \tau_2, \varepsilon) := \{(k_1, k_2) : \lfloor n\tau_1 \rfloor \leq k_1 < k_2 \leq \lfloor n\tau_2 \rfloor,\ k_2 - k_1 \geq \lfloor n\varepsilon \rfloor\}$ and

$$G_n(k_1,k_2) := \frac{\left|\sum_{i=1}^{k_1} R_i^{(1)} - (k_1/k_2)\sum_{i=1}^{k_2} R_i^{(1)}\right|}{\left\{(1/n)\sum_{t=1}^{k_1}\left(S_t^{(1)}(1,k_1)\right)^2 + (1/n)\sum_{t=k_1+1}^{k_2}\left(S_t^{(1)}(k_1{+}1,k_2)\right)^2\right\}^{1/2}}$$

$$+ \frac{\left|\sum_{i=k_1+1}^{k_2} R_i^{(2)} - \{(k_2-k_1)/(n-k_1)\}\sum_{i=k_1+1}^{n} R_i^{(2)}\right|}{\left\{(1/n)\sum_{t=k_1+1}^{k_2}\left(S_t^{(2)}(k_1+1,k_2)\right)^2 + (1/n)\sum_{t=k_2+1}^{n}\left(S_t^{(2)}(k_1+1,n)\right)^2\right\}^{1/2}},$$

with

$$R_i^{(1)} := \sum_{j=1}^{k_2} 1_{\{X_j \leq X_i\}}, \quad R_i^{(2)} := \sum_{j=k_1+1}^{n} 1_{\{X_j \leq X_i\}},$$

$$S_t^{(h)}(j,k) := \sum_{i=j}^{t}\left(R_i^{(h)} - \bar{R}_{j,k}^{(h)}\right) \quad \text{with} \quad \bar{R}_{j,k}^{(h)} := \frac{1}{k-j+1}\sum_{t=j}^{k} R_t^{(h)}.$$

The distribution of the test statistic converges to a limit $T(r, \tau_1, \tau_2, \varepsilon)$ (see the supplementary material, Section S2), so subsampling can be applied. The critical values corresponding to the asymptotic distribution of the test statistic are reported in Table 1.

3.2. Data examples

We revisit some data sets from the literature. We use the self-normalized Wilcoxon change-point test combined with subsampling and compare our findings to the conclusions of other authors.

The plot in Figure 1 depicts the annual volume of discharge from the Nile river at Aswan in 10^8 m^3 for the years 1871 to 1970. The data set has been analyzed for the detection of a change-point by numerous authors under differing assumptions concerning the data generating random process and by usage of

Table 1. Simulated critical values for the distribution of $T(1,\tau_1,\tau_2,\varepsilon)$ when $[\tau_1,\tau_2] = [0.15, 0.85]$ and $\varepsilon = 0.15$. The sample size is 1,000, the number of replications is 10,000.

	10%	5%	1%
$H = 0.501$	17.79	19.76	24.13
$H = 0.6$	19.80	22.38	27.68
$H = 0.7$	22.08	24.95	30.46
$H = 0.8$	24.24	27.61	34.04
$H = 0.9$	26.50	30.11	37.78
$H = 0.999$	28.28	32.32	41.24

diverse methods. Amongst others, Cobb (1978), Macneill, Tang and Jandhyala (1991), Wu and Zhao (2007), and Shao (2011) provided statistically significant evidence for a decrease of the Nile's annual discharge toward the end of the 19th century. The construction of the Aswan Low Dam between 1898 and 1902 serves as a popular explanation for an abrupt change in the data.

The value of the self-normalized Wilcoxon test statistic computed with respect to the data is given by $T_n(\tau_1,\tau_2) = 13.48729$. For a level of significance of 5%, the self-normalized Wilcoxon change-point test rejects the hypothesis for every possible value of $H \in (1/2, 1)$. Furthermore, we approximate the distribution of the self-normalized Wilcoxon test statistic by the sampling window method with block size $l = \lfloor \sqrt{n} \rfloor = 10$. The subsampling-based test decision also indicates the existence of a change-point in the mean of the data, even if we consider the 99%-quantile of $\hat{F}_{l,n}$.

Previous analysis of the Nile data done by Wu and Zhao (2007) and Balke (1993) suggests that the change in the discharge volume occurred in 1899. We applied the self-normalized Wilcoxon test and the sampling window method to the corresponding pre-break and post-break samples. Neither of these two approaches leads to rejection of the hypothesis, so that it seems reasonable to consider both samples as stationary. Based on the whole sample, local Whittle estimation with bandwidth parameter $m = \lfloor n^{2/3} \rfloor$ suggests the existence of long range dependence characterized by an Hurst parameter $\hat{H} = 0.962$, whereas the estimates for the pre-break and post-break samples given by $\hat{H}_1 = 0.517$ and $\hat{H}_2 = 0.5$, respectively, should be considered as indication of short range dependent data. In this regard, our findings support the conjecture of spurious long memory caused by a change-point and therefore agree with the results of Shao (2011).

The second data set consists of the seasonally adjusted monthly deviations of the temperature (degrees C) for the northern hemisphere during the years 1854

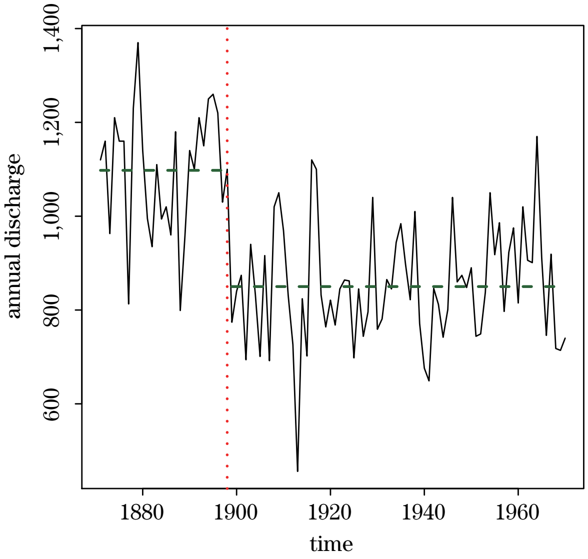

Figure 1. Measurements of the annual discharge of the river Nile at Aswan in 10^8 m^3 for the years 1871-1970. The dotted line indicates the location of the change-point; the dashed lines designate the sample means for the pre-break and post-break samples.

to 1989 from the monthly averages over the period 1950 to 1979. The data results from spatial averaging of temperatures measured over land and sea. At first sight, the plot in Figure 2 may suggest an increasing trend as well as an abrupt change of the temperature deviations. Statistical evidence for a positive deterministic trend implies affirmation of the conjecture that there has been global warming during the last decades.

The question of whether the Northern hemisphere temperature data acts as an indicator for global warming of the atmosphere is a controversial issue. Deo and Hurvich (1998) provided some indication for global warming by fitting a linear trend to the data. Beran and Feng (2002) considered a more general stochastic model by the assumption of so-called semiparametric fractional autoregressive (SEMIFAR) processes. Their method did not deliver sufficient statistical evidence for a deterministic trend. Wang (2007) applied another method for the detection of gradual change to the global temperature data and did not detect a trend, either. He offers an alternative explanation for the occurrence of a trend-like behavior by pointing out that it may have been generated by stationary long range dependent processes. In contrast, it is shown in Shao (2011) that the existence of a change-point in the mean yields yet another explanation for the performance of the data.

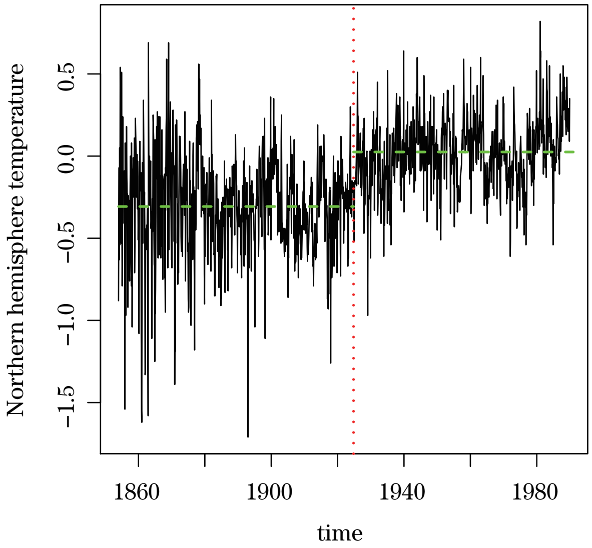

Figure 2. Monthly temperature of the northern hemisphere for the years 1854-1989 from the data base held at the Climate Research Unit of the University of East Anglia, Norwich, England. The temperature anomalies (in degrees C) are calculated with respect to the reference period 1950-1979. The dotted line indicates the location of the potential change-point; the dashed lines designate the sample means for the pre-break and post-break samples.

The value of the self-normalized Wilcoxon test statistic for this data set is $T_n(\tau_1, \tau_2) = 18.98636$. Consequently, this test would reject the hypothesis of stationarity for every value of $H \in (1/2, 1)$ at a level of significance of 1%. An application of the sampling window method with respect to the self-normalized Wilcoxon test statistic based on comparison of $T_n(\tau_1, \tau_2)$ with the 99%-quantile of the sampling distribution $\hat{F}_{l,n}$ yields a test decision in favor of the alternative hypothesis for any choice of the block length $l \in \{\lfloor n^\gamma \rfloor | \ \gamma = 0.3, 0.4, \ldots, 0.9\} = \{9, 19, 40, 84, 177, 371, 778\}$. All in all, both testing procedures provide strong evidence for the existence of a change in the mean.

According to Shao (2011) the change-point is located around October 1924. Based on the whole sample local Whittle estimation with bandwidth $m = \lfloor n^{2/3} \rfloor$ provides an estimator $\hat{H} = 0.811$. The estimated Hurst parameters for the pre-break and post-break sample are $\hat{H}_1 = 0.597$ and $\hat{H}_2 = 0.88$, respectively. Neither subsampling with respect to the self-normalized Wilcoxon test statistic nor comparison of the value of $T_n(\tau_1, \tau_2)$ with the corresponding critical values of its limit distribution, provides evidence for another change-point in the pre-break or post-break sample.

Computation of the test statistic that allows for two change-points yields $T_n(\tau_1, \tau_2, \varepsilon) = 17.88404$ (for $\tau_1 = 1 - \tau_2 = \varepsilon = 0.15$). If compared to the values in Table 1, the test statistic only surpasses the critical value corresponding to $H = 0.501$ and a significance level of 10%, but does not exceed any of the other values. Subsampling with respect to the test statistic $T_n(\tau_1, \tau_2, \varepsilon)$ does not support the conjecture of two changes, either. In fact, subsampling leads to a rejection of the hypothesis when the block length is $l = \lfloor n^{0.7} \rfloor = 177$ (based on a comparison of $T_n(\tau_1, \tau_2, \varepsilon)$ with the 95%-quantile of the corresponding sampling distribution $\hat{F}_{l,n}$), but yields a test decision in favor of the hypothesis for block lengths $l \in \{\lfloor n^{\gamma} \rfloor | \ \gamma = 0.5, 0.6, 0.8, 0.9\} = \{40, 84, 371, 778\}$ and for comparison with the 90%-quantile of $\hat{F}_{l,n}$.

It seems safe to conclude that the appearance of long memory in the post-break sample is not caused by another change-point in the mean. The pronounced difference between the local Whittle estimators $\hat{H}_1$ and $\hat{H}_2$ suggests a change in the dependence structure of the times series. Another explanation could be a gradual change of the temperature in the post-break period. We conjecture that our test has only low power in the case of a gradual change, because the denominator of our self-normalized test statistic is inflated as the ranks systematically deviate from the mean rank of the first and second part. When using subsampling, the trend also appears in subsamples so that we fail to approximate the distribution under the hypothesis.

As pointed out by one of the referees, the Northern hemisphere temperature data does not seem to be second-order stationary as the variance in the first part of the time series seems higher. A change in variance should also result in a loss of power,. The ranks in the part with the higher variance are more extreme, so that the distance to the mean rank of this part is larger. This leads to a higher value of the denominator of our self-normalized test statistic, and consequently to a lower value of the ratio.

The third data set consists of the arrival rate of Ethernet data (bytes per 10 milliseconds) from a local area network (LAN) measured at Bellcore Research and Engineering Center in 1989. For more information on the LAN traffic monitoring we refer to Leland and Wilson (1991) and Beran (1994). Figure 3 reveals that the observations are strongly right-skewed. As the self-normalized Wilcoxon test is based on ranks, we do not expect that this affects our analysis.

Coulon, Chabert and Swami (2009) examined this data set for change-points before. The method proposed in their paper is based on the assumption that a FARIMA model holds for segments of the data. The number of different sections

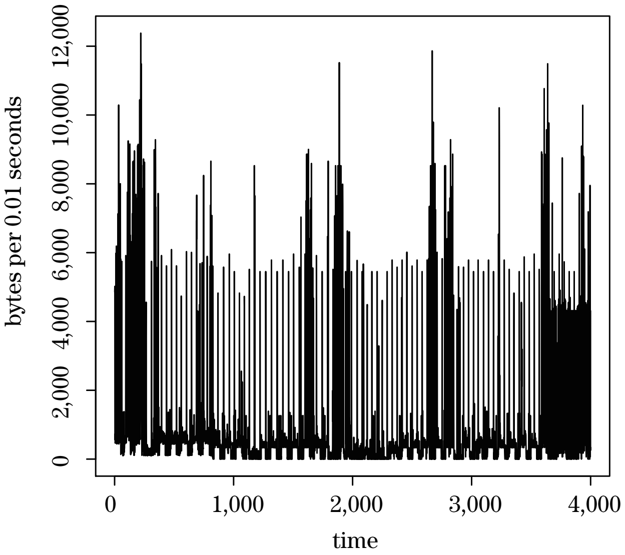

Figure 3. Ethernet traffic in bytes per 10 milliseconds from a LAN measured at Bellcore Research Engineering Center.

and the location of the change-points are chosen by a model selection criterion. The algorithm proposed by Coulon, Chabert and Swami (2009) detects multiple changes in the parameters of the corresponding FARIMA time series.

In contrast, an application of the self-normalized Wilcoxon change-point test does not provide evidence for a change-point in the mean: the value of the test statistic is given by $T_n(\tau_1, \tau_2) = 3.270726$. Even for a level of significance of 10%, the self-normalized Wilcoxon change-point test does not reject the hypothesis for any value $H \in (1/2, 1)$. Subsampling with respect to the self-normalized Wilcoxon test statistic does not lead to a rejection of the hypothesis , either (for any choice of block length $l \in \{\lfloor n^{\gamma} \rfloor | \ \gamma = 0.3, 0.4, \ldots, 0.9\} = \{12, 27, 63, 144, 332, 761, 1745\}$ and for comparison with the 90%-quantile of the corresponding sampling distribution $\hat{F}_{l,n}$).

Taking into consideration that the data set contains ties (the value 0 appears several times), we also applied the self-normalized Wilcoxon test statistic based on the modified ranks $\tilde{R}_i$ and used subsampling with respect to this statistic. Both approaches did not lead to a rejection of the hypothesis.

An application of the test statistic constructed for the detection of two changes yields a value of $T_n(\tau_1, \tau_2, \varepsilon) = 15.24527$ when $\varepsilon = \tau_1 = 1 - \tau_2 = 0.15$. This does not lead to a rejection of the hypothesis for any value of the parameter H. Subsampling based on comparison of $T_n(\tau_1, \tau_2, \varepsilon)$ with the 90%-quantile of the corresponding sampling distribution $\hat{F}_{l,n}$ does not provide evidence for the assertion of multiple changes for any block lenght $l \in \{\lfloor n^{\gamma} \rfloor |\, \gamma = 0.5, 0.6, 0.7, 0.8\} = \{63, 144, 332, 761\}$ in the data, either.

These results do not coincide with the analysis of Coulon, Chabert and Swami (2009). This may be due to the fact that our methods differ considerably from the testing procedures applied before. The change-point estimation algorithm proposed in Coulon, Chabert and Swami (2009) is not robust to skewness or heavy-tailed distributions and decisively relies on the assumption of FARIMA time series. This seems to contradict observations made by Bhansali and Kokoszka (2001) as well as Taqqu and Teverovsky (1997) who stress that the model that fits the Ethernet traffic data is very unlikely to be FARIMA.

Estimation of the Hurst parameter by the local Whittle procedure with bandwidth parameter $m = \lfloor n^{2/3} \rfloor$ yields an estimate of $\hat{H} = 0.845$, so it indicates the existence of long range dependence. This is consistent with the results of Leland et al. (1994) and Taqqu and Teverovsky (1997).

In the three data examples, we find that the results obtained by subsampling and by parameter estimation are in good accordance with each other. The methods take into account long range dependence or heavy tails, but still detect a change in location in the first two examples. For the third data example our analysis supports the hypothesis of stationarity.

4. Simulations

We investigated the finite sample performance of the subsampling procedure with respect to the self-normalized Wilcoxon test and with respect to the classical Wilcoxon change-point test. We compared these results to the performance of the tests when the test decision is based on critical values obtained from the asymptotic distribution of the test statistic. We considered subordinated Gaussian time series $(X_n)_{n\in\mathbb{N}}$, $X_n = G(\xi_n)$, where $(\xi_n)_{n\in\mathbb{N}}$ was fractional Gaussian noise (introduced in Examples 1 and 2) with Hurst parameter $H \in \{0.6, 0.7, 0.8, 0.9\}$ and covariance function

$$\gamma(k) \sim k^{-D}\left(1 - \frac{D}{2}\right)(1 - D)\,,$$

where $D = 2 - 2H$. Initially, we took $G(t) = t$, so that $(X_n)_{n\in\mathbb{N}}$ has normal marginal distributions. We also considered the transformation

$$G(t) = \left\{\frac{\beta k^2}{(\beta-1)^2(\beta-2)}\right\}^{-1/2}\left[k\{\Phi(t)\}^{-1/\beta} - \frac{\beta k}{\beta - 1}\right]$$

(with Φ denoting the standard normal distribution function) so as to generate Pareto-distributed data with parameters $k, \beta > 0$ (referred to as Pareto(β, k)). In both cases, the Hermite rank r of $1_{\{G(\xi_i) \leq x\}} - F(x), x \in \mathbb{R}$, is $r = 1$ and

$$\left| \int_{\mathbb{R}} J_1(x) dF(x) \right| = \frac{1}{2\sqrt{\pi}};$$

see Dehling, Rooch and Taqqu (2013).

Under these conditions, the critical values of the asymptotic distribution of the self-normalized Wilcoxon test statistic were reported in Table 2 in Betken (2016). The limit of the Wilcoxon change-point test statistic can be found in Dehling, Rooch and Taqqu (2013), the corresponding critical values can be taken from Table 1 in Betken (2016).

The frequencies of rejections of both testing procedures are reported in Table 2 and Table 3 for the self-normalized Wilcoxon change-point test, and in Table 4 and Table 5 for the classical Wilcoxon test (without self-normalization). The calculations are based on $5,000$ realizations of time series with sample size $n = 300$ and $n = 500$. We chose block lengths $l = l_n = \lfloor n^{\gamma} \rfloor$ with $\gamma \in \{0.4, 0.5, 0.6\}$. As level of significance we chose 5%, comparing the values of the test statistic with the 95%-quantile of its asymptotic distribution and the 95%-quantile of the empirical distribution function $\hat{F}_{l,n}$, respectively.

For the usual testing procedures the estimation of the Hermite rank r, the slowly varying function L_{γ} and the integral $\int J_1(x) dF(x)$ was neglected. For every simulated time series we estimate the Hurst parameter H by the local Whittle estimator $\hat{H}$ proposed in Künsch (1987). This estimator is based on an approximation of the spectral density by the periodogram at the Fourier frequencies. It depends on the spectral bandwidth parameter $m = m(n)$ which denotes the number of Fourier frequencies used for the estimation. If the bandwidth m satisfies $1/m + m/n \longrightarrow 0$ as $n \longrightarrow \infty$, the local Whittle estimator is a consistent estimator for H; see Robinson (1995). For convenience we always chose $m = \lfloor n^{2/3} \rfloor$. The critical values corresponding to the estimated values of H were determined by linear interpolation.

Under the alternative **A** we analyzed the power of the testing procedures by considering different choices for the height of the level shift (denoted by h) and the location $[n\tau]$ of the change-point. In the tables the columns that are superscribed by "$h = 0$" correspond to the frequency of a type 1 error.

For the self-normalized Wilcoxon change-point test based on the asymptotic distribution, the empirical size almost equals the level of significance of 5% for

Table 2. Rejection rates of the *self-normalized* Wilcoxon change-point test obtained by subsampling with block length $l = n^{0.4}, n^{0.5}, n^{0.6}$, and by comparison with asymptotic critical values for fractional Gaussian noise of length n with Hurst parameter H.

				$\tau = 0.25$		$\tau = 0.5$	
fGn	n	method	$h = 0$	$h = 0.5$	$h = 1$	$h = 0.5$	$h = 1$
$H = 0.6$	300	subspl. $l = 9$	0.041	0.263	0.700	0.502	0.952
		subspl. $l = 17$	0.064	0.313	0.742	0.570	0.964
		subspl. $l = 30$	0.070	0.322	0.705	0.555	0.943
		asymptotic	0.044	0.209	0.521	0.424	0.861
	500	subspl. $l = 12$	0.053	0.396	0.859	0.697	0.994
		subspl. $l = 22$	0.060	0.421	0.861	0.720	0.995
		subspl. $l = 41$	0.069	0.411	0.829	0.697	0.991
		asymptotic	0.049	0.303	0.687	0.577	0.958
$H = 0.7$	300	subspl. $l = 9$	0.057	0.155	0.412	0.291	0.759
		subspl. $l = 17$	0.070	0.171	0.423	0.313	0.763
		subspl. $l = 30$	0.077	0.177	0.403	0.314	0.737
		asymptotic	0.053	0.108	0.268	0.228	0.611
	500	subspl. $l = 12$	0.056	0.183	0.513	0.382	0.856
		subspl. $l = 22$	0.059	0.193	0.508	0.382	0.854
		subspl. $l = 41$	0.065	0.192	0.476	0.387	0.819
		asymptotic	0.048	0.133	0.359	0.302	0.730
$H = 0.8$	300	subspl. $l = 9$	0.070	0.126	0.251	0.223	0.526
		subspl. $l = 17$	0.067	0.117	0.234	0.208	0.494
		subspl. $l = 30$	0.073	0.114	0.218	0.201	0.466
		asymptotic	0.048	0.081	0.144	0.141	0.362
	500	subspl. $l = 12$	0.066	0.121	0.295	0.217	0.591
		subspl. $l = 22$	0.068	0.114	0.278	0.210	0.567
		subspl. $l = 41$	0.069	0.119	0.257	0.205	0.532
		asymptotic	0.053	0.085	0.198	0.163	0.462
$H = 0.9$	300	subspl. $l = 9$	0.093	0.126	0.208	0.209	0.462
		subspl. $l = 17$	0.074	0.097	0.161	0.169	0.397
		subspl. $l = 30$	0.073	0.095	0.145	0.165	0.367
		asymptotic	0.057	0.065	0.106	0.125	0.308
	500	subspl. $l = 12$	0.079	0.105	0.194	0.185	0.461
		subspl. $l = 22$	0.067	0.091	0.166	0.162	0.416
		subspl. $l = 41$	0.063	0.087	0.146	0.152	0.391
		asymptotic	0.051	0.068	0.120	0.128	0.350

normally distributed data (see Table 2). The sampling window method yields rejection rates that slightly exceed this level. For Pareto(3, 1) time series both testing procedures lead to similar results and tend to reject the hypothesis too often when there is no change. With regard to the empirical power, it is notable that for fractional Gaussian noise time series the sampling window method yields

Table 3. Rejection rates of the *self-normalized* Wilcoxon change-point test obtained by subsampling with block length $l = n^{0.4}, n^{0.5}, n^{0.6}$, and by comparison with asymptotic critical values for Pareto(3, 1)-transformed fractional Gaussian noise of length n with Hurst parameter H.

				$\tau = 0.25$		$\tau = 0.5$	
Pareto	n	method	$h = 0$	$h = 0.5$	$h = 1$	$h = 0.5$	$h = 1$
$H = 0.6$	300	subspl. $l = 9$	0.041	0.847	0.977	0.990	1.000
		subspl. $l = 17$	0.067	0.871	0.946	0.990	1.000
		subspl. $l = 30$	0.070	0.831	0.946	0.979	1.000
		asymptotic	0.056	0.820	0.912	0.984	0.999
	500	subspl. $l = 12$	0.055	0.947	0.997	0.999	1.000
		subspl. $l = 22$	0.066	0.946	0.994	0.999	1.000
		subspl. $l = 41$	0.071	0.921	0.976	0.996	1.000
		asymptotic	0.061	0.920	0.970	0.996	1.000
$H = 0.7$	300	subspl. $l = 9$	0.057	0.571	0.821	0.990	0.994
		subspl. $l = 17$	0.064	0.527	0.738	0.876	0.990
		subspl. $l = 30$	0.077	0.527	0.738	0.842	0.975
		asymptotic	0.070	0.529	0.702	0.856	0.982
	500	subspl. $l = 12$	0.066	0.693	0.904	0.949	0.999
		subspl. $l = 22$	0.068	0.684	0.893	0.942	0.998
		subspl. $l = 41$	0.072	0.632	0.838	0.921	0.994
		asymptotic	0.076	0.663	0.820	0.940	0.995
$H = 0.8$	300	subspl. $l = 9$	0.070	0.355	0.574	0.703	0.931
		subspl. $l = 17$	0.068	0.284	0.454	0.666	0.905
		subspl. $l = 30$	0.073	0.284	0.454	0.633	0.857
		asymptotic	0.072	0.297	0.428	0.640	0.875
	500	subspl. $l = 12$	0.064	0.401	0.609	0.738	0.948
		subspl. $l = 22$	0.063	0.379	0.581	0.714	0.933
		subspl. $l = 41$	0.064	0.345	0.509	0.688	0.903
		asymptotic	0.069	0.369	0.510	0.715	0.920
$H = 0.9$	300	subspl. $l = 9$	0.093	0.253	0.396	0.597	0.832
		subspl. $l = 17$	0.071	0.168	0.254	0.532	0.772
		subspl. $l = 30$	0.073	0.168	0.254	0.482	0.729
		asymptotic	0.073	0.165	0.236	0.499	0.738
	500	subspl. $l = 12$	0.073	0.256	0.405	0.585	0.839
		subspl. $l = 22$	0.064	0.219	0.340	0.547	0.802
		subspl. $l = 41$	0.065	0.190	0.296	0.503	0.762
		asymptotic	0.068	0.199	0.296	0.529	0.782

considerably better power than the test based on asymptotic critical values. If Pareto(3, 1)-distributed time series are considered, the empirical power of the subsampling procedure is still better than the empirical power that results from using asymptotic critical values. However, in this case, the deviation of the

Table 4. Rejection rates of the *classical* Wilcoxon change-point test obtained by subsampling with block length $l = n^{0.4}, n^{0.5}, n^{0.6}$, and by comparison with asymptotic critical values for fractional Gaussian noise of length n with Hurst parameter H.

				$\tau = 0.25$		$\tau = 0.5$	
fGn	n	method	$h = 0$	$h = 0.5$	$h = 1$	$h = 0.5$	$h = 1$
$H = 0.6$	300	subspl. $l = 9$	0.066	0.200	0.232	0.386	0.591
		subspl. $l = 17$	0.054	0.223	0.411	0.439	0.784
		subspl. $l = 30$	0.059	0.264	0.529	0.663	0.870
		asymptotic	0.026	0.096	0.160	0.223	0.727
	500	subspl. $l = 12$	0.063	0.285	0.436	0.569	0.856
		subspl. $l = 22$	0.058	0.345	0.663	0.627	0.952
		subspl. $l = 41$	0.062	0.397	0.789	0.683	0.975
		asymptotic	0.036	0.148	0.256	0.378	0.897
$H = 0.7$	300	subspl. $l = 9$	0.052	0.080	0.088	0.162	0.302
		subspl. $l = 17$	0.049	0.095	0.158	0.206	0.466
		subspl. $l = 30$	0.051	0.120	0.227	0.267	0.593
		asymptotic	0.035	0.067	0.228	0.167	0.66
	500	subspl. $l = 12$	0.042	0.104	0.153	0.249	0.539
		subspl. $l = 22$	0.039	0.131	0.267	0.287	0.689
		subspl. $l = 41$	0.046	0.160	0.373	0.343	0.789
		asymptotic	0.030	0.079	0.259	0.225	0.714
$H = 0.8$	300	subspl. $l = 9$	0.028	0.030	0.031	0.054	0.092
		subspl. $l = 17$	0.029	0.038	0.048	0.075	0.179
		subspl. $l = 30$	0.034	0.057	0.088	0.070	0.272
		asymptotic	0.077	0.153	0.421	0.245	0.673
	500	subspl. $l = 12$	0.023	0.031	0.036	0.064	0.162
		subspl. $l = 22$	0.028	0.044	0.070	0.097	0.273
		subspl. $l = 41$	0.039	0.071	0.129	0.137	0.391
		asymptotic	0.050	0.112	0.439	0.226	0.714
$H = 0.9$	300	subspl. $l = 9$	0.009	0.010	0.006	0.016	0.020
		subspl. $l = 17$	0.009	0.014	0.009	0.021	0.060
		subspl. $l = 30$	0.015	0.029	0.028	0.011	0.153
		asymptotic	0.360	0.484	0.739	0.524	0.830
	500	subspl. $l = 12$	0.008	0.006	0.003	0.015	0.026
		subspl. $l = 22$	0.011	0.009	0.011	0.029	0.086
		subspl. $l = 41$	0.021	0.021	0.032	0.058	0.197
		asymptotic	0.319	0.439	0.743	0.511	0.845

rejection rates is rather small. While the empirical size is not much affected by the Hurst parameter H, the empirical power is lower for $H = 0.8, 0.9$.

Considering the classical Wilcoxon test (without self-normalization), for both procedures the empirical size is in most cases not close to the nominal level of significance (5%), ranging from 1.1% to 20.8% using subsampling and from

Table 5. Rejection rates of the *classical* Wilcoxon change-point test obtained by subsampling with block length $l = n^{0.4}, n^{0.5}, n^{0.6}$, and by comparison with asymptotic critical values for Pareto(3, 1)-transformed fractional Gaussian noise of length n with Hurst parameter H.

				$\tau = 0.25$		$\tau = 0.5$	
Pareto	n	method	$h = 0$	$h = 0.5$	$h = 1$	$h = 0.5$	$h = 1$
$H = 0.6$	300	subspl. $l = 9$	0.170	0.949	0.742	0.991	0.923
		subspl. $l = 17$	0.130	0.963	0.861	0.996	0.991
		subspl. $l = 30$	0.109	0.962	0.871	0.998	0.998
		asymptotic	0.108	0.938	0.985	0.998	1.000
	500	subspl. $l = 12$	0.163	0.991	0.916	1.000	0.993
		subspl. $l = 22$	0.132	0.997	0.976	1.000	0.999
		subspl. $l = 41$	0.114	0.997	0.989	1.000	1.000
		asymptotic	0.128	0.988	0.999	1.000	1.000
$H = 0.7$	300	subspl. $l = 9$	0.224	0.785	0.568	0.939	0.796
		subspl. $l = 17$	0.175	0.802	0.680	0.955	0.949
		subspl. $l = 30$	0.140	0.789	0.708	0.959	0.976
		asymptotic	0.179	0.833	0.969	0.974	0.999
	500	subspl. $l = 12$	0.208	0.921	0.763	0.989	0.956
		subspl. $l = 22$	0.167	0.931	0.862	0.992	0.996
		subspl. $l = 41$	0.143	0.925	0.891	0.994	0.998
		asymptotic	0.191	0.940	0.994	0.996	1.000
$H = 0.8$	300	subspl. $l = 9$	0.203	0.508	0.326	0.743	0.565
		subspl. $l = 17$	0.160	0.496	0.347	0.776	0.808
		subspl. $l = 30$	0.137	0.484	0.364	0.791	0.881
		asymptotic	0.204	0.729	0.925	0.918	0.993
	500	subspl. $l = 12$	0.190	0.639	0.445	0.865	0.770
		subspl. $l = 22$	0.160	0.649	0.513	0.886	0.929
		subspl. $l = 41$	0.137	0.626	0.556	0.890	0.961
		asymptotic	0.212	0.805	0.963	0.948	0.999
$H = 0.9$	300	subspl. $l = 9$	0.128	0.150	0.077	0.320	0.336
		subspl. $l = 17$	0.097	0.128	0.071	0.403	0.550
		subspl. $l = 30$	0.092	0.125	0.077	0.481	0.677
		asymptotic	0.309	0.712	0.901	0.848	0.966
	500	subspl. $l = 12$	0.112	0.159	0.089	0.402	0.436
		subspl. $l = 22$	0.100	0.161	0.101	0.518	0.680
		subspl. $l = 41$	0.095	0.170	0.106	0.571	0.771
		asymptotic	0.270	0.726	0.911	0.851	0.975

2.6% to 36.0% using asymptotic critical values. In general, the sampling window method becomes more conservative for higher values of the Hurst parameter H, while the test based on the asymptotic distribution becomes more liberal. Under the alternative, the usual application of the Wilcoxon test yields better power than the sampling window method, especially for high values of H. But it should be emphasized that this comparison is problematic because the rejection frequencies under the hypothesis differ.

We conclude that the self-normalized Wilcoxon change-point test is more reliable than the classical change-point test. The reason can be that in the scaling of the classical test, the estimator $\hat{H}$ of the Hurst parameter enters as a power of the sample size n. Thus, a small error in this estimation can lead to a large error in the value of the test statistic. By using the sampling window method for the self-normalized version, we avoid the estimation of unknown parameters so that the performance is similar to the performance of the classical testing procedure which compares the values of the test statistic with the corresponding critical values.

In most cases covered by our simulations the choice of the block length for the subsampling procedure does not have a big impact on the frequency of a type 1 error. Considering the self-normalized Wilcoxon change-point test, an increase of the block length tends to go along with a decrease in power, especially for big values of the Hurst parameter H and Pareto-distributed random variables. For smaller values of H the effect is not pronounced. We recommend using a block length $\lfloor n^{0.4} \rfloor$ or $\lfloor n^{0.5} \rfloor$ for the self-normalized change-point test as the choice $l = \lfloor n^{0.6} \rfloor$ implies worse properties in most cases.

An application of the subsampling testing procedure to the classical (non-self-normalized) Wilcoxon test for different choices of the block length shows the opposite effect on the rejection rate under the alternative: an increase of the block length results in a higher frequency of rejections. Here, the block length $\lfloor n^{0.6} \rfloor$ leads to better results in many cases, but we do not recommend to use this test, but rather to self-normalize the test statistic.

An alternative way of choosing the block length is to apply the data-driven block selection rule proposed by Götze and Račkauskas (2001) and Bickel and Sakov (2008). Although the algorithm had originally been implemented for applications of the m-out-of-n bootstrap to independent and identically distributed data, it also lead to satisfactory simulation results in applications to long range dependent time series (see Jach, McElroy and Politis (2012)). Another general approach to the selection of the block size in the context of hypothesis testing is given by Algorithm 9.4.2 in Politis, Romano and Wolf (1999).

Supplementary Materials

In the online supplement, additional information about the change point test for long range dependent data with ties can be found (see Section S1). More details on the test for multiple change points is given in Section S2. The technical lemmas in Section S3 are needed for the proof of Theorem 1, in Section S4.

Acknowledgment

We thank the referee for his careful reading of the article and his thoughtful comments which led to a significant improvement of the article. We also thank Norman Lambot for reading the article, thereby helping to reduce the number of misprints. The research was supported by the German National Academic Foundation and Collaborative Research Center SFB 823 *Statistical modelling of nonlinear dynamic processes*.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**, 821–856.

Antoch, J., Hušková, M., Janic, A. and Ledwina, T. (2008). Data driven rank test for the change point problem. *Metrika* **68**, 1–15.

Baek, C. and Pipiras, V. (2014). On distinguishing multiple changes in mean and long-range dependence using local Whittle estimation. *Electronic Journal of Statistics* **8**, 931–964.

Bai, S. and Taqqu, M. S. (2015). Canonical correlation between blocks of long-memory time series and consistency of subsampling. *arXiv:1512.00819*.

Bai, S., Taqqu, M. S. and Zhang, T. (2016). A unified approach to self-normalized block sampling. *Stochastic Processes and their Applications* **126**, 2465–2493.

Balke, N. S. (1993). Detecting level shifts in time series. *Journal of Business & Economic Statistics* **11**, 81–92.

Beran, J. (1994). *Statistics for Long-Memory Processes.* Chapman & Hall.

Beran, J. and Feng, Y. (2002). SEMIFAR models - a semiparametric framework for modelling trends, long-range dependence and nonstationarity. *Computational Statistics & Data Analysis* **40**, 393–419.

Beran, J., Feng, Y., Ghosh, S. and Kulik, R. (2013). *Long-Memory Processes.* Springer-Verlag Berlin Heidelberg.

Berkes, I., Horváth, L., Kokoszka, P. and Shao, Q.-M. (2006). On discriminating between long-range dependence and changes in mean. *The Annals of Statistics* **34**, 1140–1165.

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**, 185–809.

Beutner, E. and Zähle, H. (2014). Continuous mapping approach to the asymptotics of U- and V-statistics. *Bernoulli* **20**, 846–877.

Bhansali, R. J. and Kokoszka, P. S. (2001). Estimation of the long-memory parameter: a review of recent developments and an extension. *Lecture Notes-Monograph Series*, 125–150.

Bickel, P. J. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica* **18**, 967–985.

Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*. Kendrick press.

Cobb, G. W. (1978). The problem of the Nile: conditional solution to a changepoint problem. *Biometrika* **65**, 243–251.

Coulon, M., Chabert, M. and Swami, A. (2009). Detection of multiple changes in fractional integrated ARMA processes. *IEEE Transactions on, Signal Processing* **57**, 48–61.

Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley Chichester; New York.

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

Dehling, H., Rooch, A. and Wendler, M. (2017). Two-sample u-statistic processes for long-range dependent data. *Statistics* **51**, 84 –104.

Deo, R. S. and Hurvich, C. M. (1998). Linear trend with fractionally integrated errors. *Journal of Time Series Analysis* **19**, 379–397.

Dobrushin, R. L. and Major, P. (1979). Non-central limit theorems for non-linear functionals of Gaussian fields. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **50**, 27–52.

Fan, Z. (2012). *Statistical Issues and Developments in Time Series Analysis and Educational Measurement*. BiblioBazaar.

Götze, F. and Račkauskas, A. (2001). Adaptive choice of bootstrap sample sizes. *Lecture Notes-Monograph Series* **36**, 286–309.

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* **1**, 15–29.

Hall, P. and Jing, B. (1996). On sample reuse methods for dependent data. *Journal of the Royal Statistical Society Series B Statistical Methodology* **58**, 727–737.

Hall, P., Jing, B.-Y. and Lahiri, S. N. (1998). On the sampling window method for long-range dependent data. *Statistica Sinica* **8**, 1189–1204.

Hurst, H. E. (1956). Methods of using long-term storage in reservoirs. *ICE Proceedings* **5**, 519–543.

Ibragimov, I. A. and Rozanov, Y. A. (1978). *Gaussian Random Processes*. Springer, New York.

Jach, A., McElroy, T. and Politis, D. N. (2012). Subsampling inference for the mean of heavy-tailed long-memory time series. *Journal of Time Series Analysis* **33**, 96–111.

Jach, A., McElroy, T. and Politis, D. N. (2016). Corrigendum to 'subsampling inference for the mean of heavy-tailed long-memory time series' by A. Jach, T. S. McElroy and D. N. Politis. *Journal of Time Series Analysis* **37**, 713–720.

Künsch, H. R. (1987). Statistical aspects of self-similar processes. in *Proceedings of the First World Congress of the Bernoulli Society* **1**, 67–74. VNU Science Press Utrecht, The Netherlands.

Lahiri, S. N. (1993). On the moving block bootstrap under long range dependence. *Statistics & Probability Letters* **18**, 405–413.

Leland, W. E., Taqqu, M. S., Willinger, W. and Wilson, D. V. (1994). On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking* **2**, 1–15.

Leland, W. E. and Wilson, D. V. (1991). High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection. in *INFOCOM'91. Proceedings. Tenth Annual Joint Conference of the IEEE Computer and Communications Societies. Networking in the 90s., IEEE* IEEE 1360–1366.

Lo, A. W. (1989). Long-term memory in stock market prices. *Technical Report National Bureau of Economic Research.*

Lobato, I. N. (2001). Testing that a dependent process is uncorrelated. *Journal of the American Statistical Association* **96**, 1066–1076.

Macneill, I. B., Tang, S. M. and Jandhyala, V. K. (1991). A Search for the source of the Nile's change-points. *Environmetrics* **2**, 341–375.

McElroy, T. and Politis, D. (2007). Self-normalization for heavy-tailed time series with long memory. *Statistica Sinica* **17**, 199.

Nordman, D. J. and Lahiri, S. N. (2005). Validity of the sampling window method for long-range dependent linear processes. *Econometric Theory* **21**, 1087–1111.

Pipiras, V. and Taqqu, M. S. (2011). *Long-Range Dependence and Self-Similarity.* Cambridge University Press.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 2031–2050.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling.* Springer, New York.

Robinson, P. M. (1995). Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics*, 1630–1661.

Shao, X. (2011). A simple test of changes in mean in the possible presence of long-range dependence. *Journal of Time Series Analysis* **32**, 598–606.

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

Sherman, M. and Carlstein, E. (1996). Replicate histograms. *Journal of the American Statistical Association* **91**, 566–576.

Sinai, Y. G. (1976). Self-similar probability distributions. *Theory of Probability & Its Applications* **21**, 64–80.

Surgailis, D. (1982). Zones of attraction of self-similar multiple integrals. *Lithuanian Mathematical Journal* **22**, 327–340.

Taqqu, M. S. (1979). Convergence of integrated processes of arbitrary Hermite rank. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **50**, 53–83.

Taqqu, M. S. and Teverovsky, V. (1997). Robustness of Whittle-type estimators for time series with long-range dependence. *Communications in Statistics. Stochastic Models* **13**, 723–757.

Wang, L. (2007). Gradual changes in long memory processes with applications. *Statistics* **41**, 221–240.

Wang, L. (2008). Change-point detection with rank statistics in long-memory time-series models. *Australian & New Zealand Journal of Statistics* **50**, 241–256.

Wu, W. B. and Zhao, Z. (2007). Inference of trends in time series. *Journal of the Royal Statistical Society: Series B Statistical Methodology* **69**, 391–410.

Zhang, T., Ho, H.-C., Wendler, M. and Wu, W. B. (2013). Block sampling under strong dependence. *Stochastic Processes and Their Applications* **123**, 2323–2339.

Fakultät für Mathematik, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany.

E-mail: annika.betken@rub.de

Institut für Mathematik und Informatik, University Greifswald, Walther-Rathenau-Straße 47, 17489 Greifswald, Germany.

E-mail: martin.wendler@uni-greifswald.de

(Received December 2015; accepted March 2017)