On Robust Nonparametric Repeated Significance Tests for Heavy-tailed Distributions

Joseph Glaz and Vladimir Pozdnyakov

University of Connecticut

2007

Heavy tail distributions

- Distributions with heavy tails have been used in modeling computer network traffic (Crovella, Taqqu and Bestavros 1998 and Willinger, Paxson and Taqqu 1998), telecommunication systems (Crovella and Taqqu 1999), high frequency financial data (Müller, Dacorogna and Pictet 1998) and risk management and insurance data (Bassi, Embrecht and Kafetzaki 1998).

- In these areas of application the data is accumulated sequentially. Repeated significance tests can be used in the design and analysis of continuous monitoring schemes.

- We develop a repeated significance test for independent and identically distributed (iid) observations from a continuous symmetric distribution with heavy tails and infinite variance and possibly no mean, as in the case of the Cauchy distribution.

Repeated Significance Test (RST)

Let $\{X, X_i\}_{i\geq 1}$ be i.i.d random variables with a mean μ and known σ. Let us assume that we want to test $H_0 : \mu = 0$ versus $H_a : \mu \neq 0$.

- **Traditional approach**
  We fixed the sample size N and we reject H_0 if the statistics

$$\frac{|S_N|}{\sigma\sqrt{N}}$$

  is large.

- **Sequential approach**
  Let n_0 be an initial sample size, and N is a target sample size. Consider the following test statistic

$$h_N = \max\left\{\frac{|S_n|}{\sigma\sqrt{n}} : n_0 \leq n \leq N\right\}, \tag{1}$$

  and we reject H_0 if the value of the test statistic is too large. It can be done with help of so-called invariance principles. The usage of these in context of sequential analysis have been discussed in Sen (1981, 1985 and 1991).

Heavy tails?

- **Traditional approach**

  Use *trimmed or truncated* sums.

- **Sequential approach**

  Use *trimmed or truncated* sums?

Heavy tails?

- **Truncated Sums**
  Let $\{X, X_i\}_{i\geq 1}$ be iid random variables with a symmetric continuous distribution and $\mathbf{E}X^2 = \infty$. Let $\{b_n\}_{n\geq 1}$ be a sequence of positive numbers such that $b_n \uparrow$ and

$$n\mathbf{P}\Big(|\mathbf{X}| > \mathbf{b_n}\Big) \sim \gamma_\mathbf{n} \uparrow .$$

  The *truncated sum* S_n are define by

$$S_n = \sum_{i=1}^{n} X_i \mathbf{I}_{|X_i|\leq b_n}. \tag{2}$$

- **Trimmed Sums**
  Let $\{X_{k,n} : 0 \leq k \leq n\}$ be the order statistics of

$$|X_1|, |X_2|, ..., |X_n|,$$

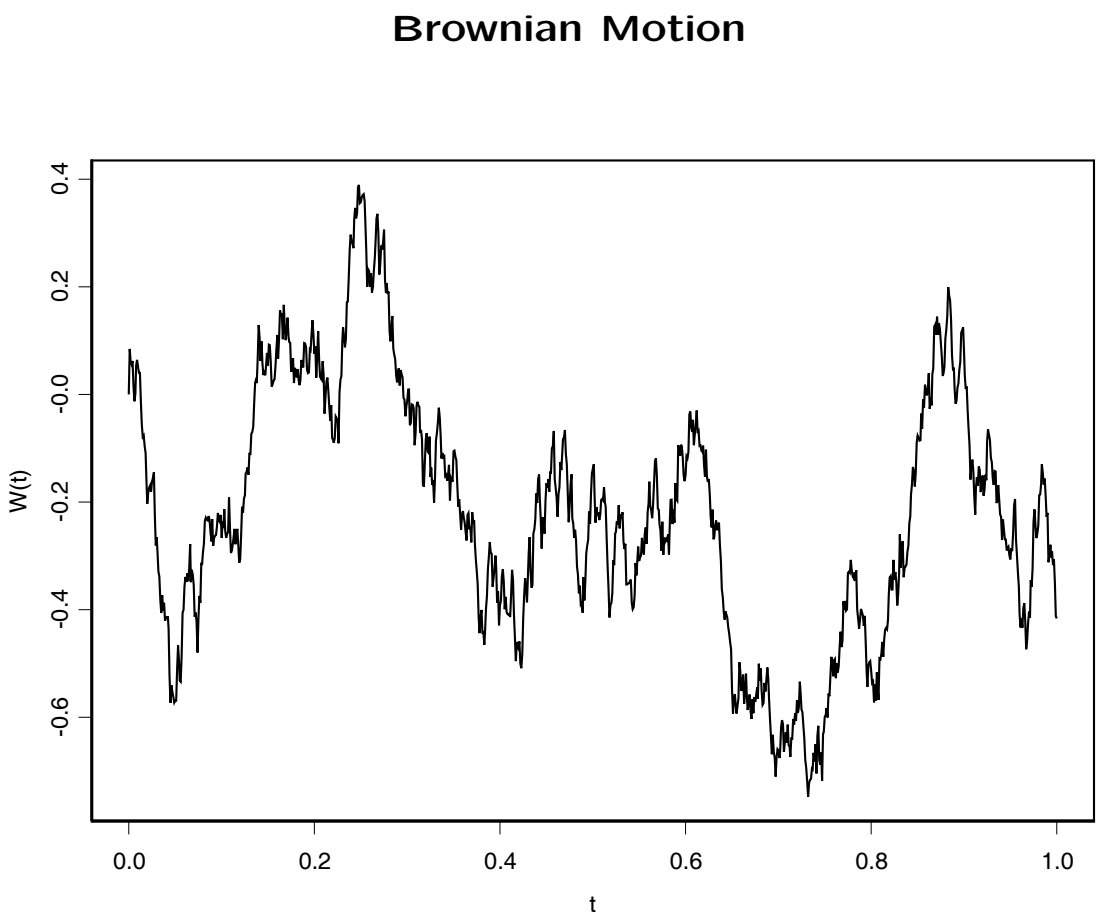
  that is

$$|X_{1,n}| \leq ... \leq |X_{n,n}|.$$

  The *trimmed sums* T_n are defined by

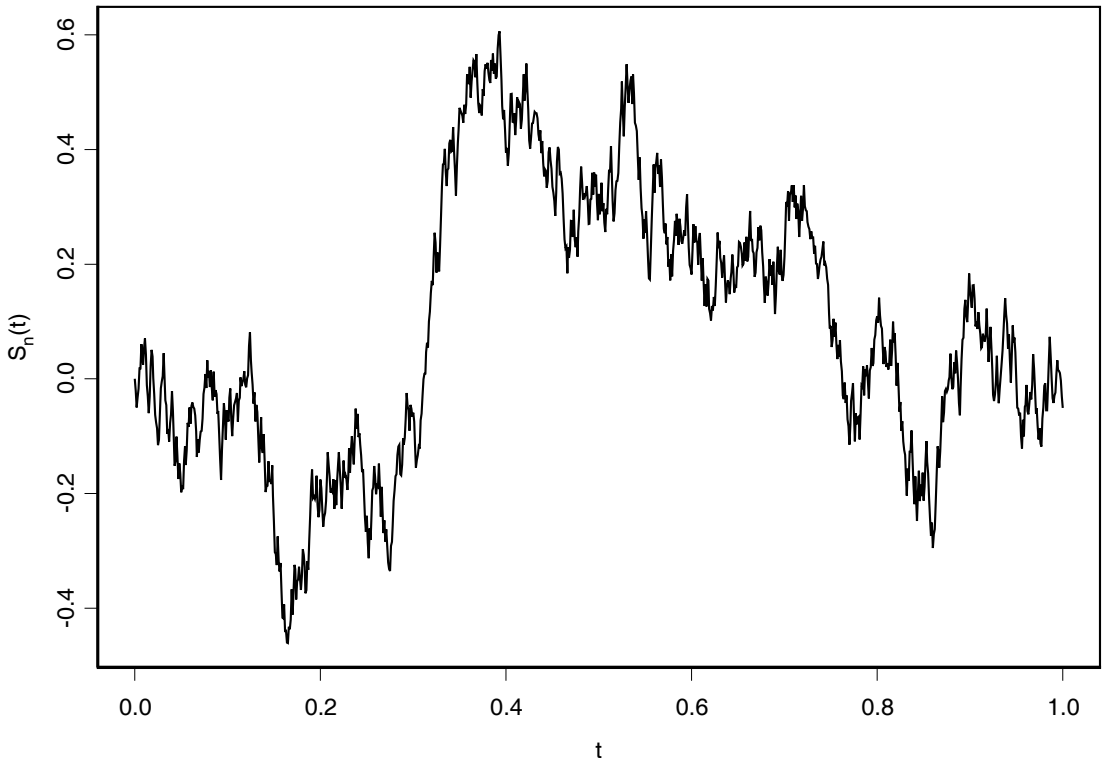
$$T_n = \sum_{i=1}^{n-k(n)} X_{i,n},$$

  where $k(n) \sim \gamma_n$. It is quite intuitive that under certain conditions the truncated and trimmed sums will be very close.


Brownian Motion
W(t)
0.4
0.2
-0.0
-0.2
-0.4
-0.6
0.0
0.2
0.4
0.6
0.8
1.0
t

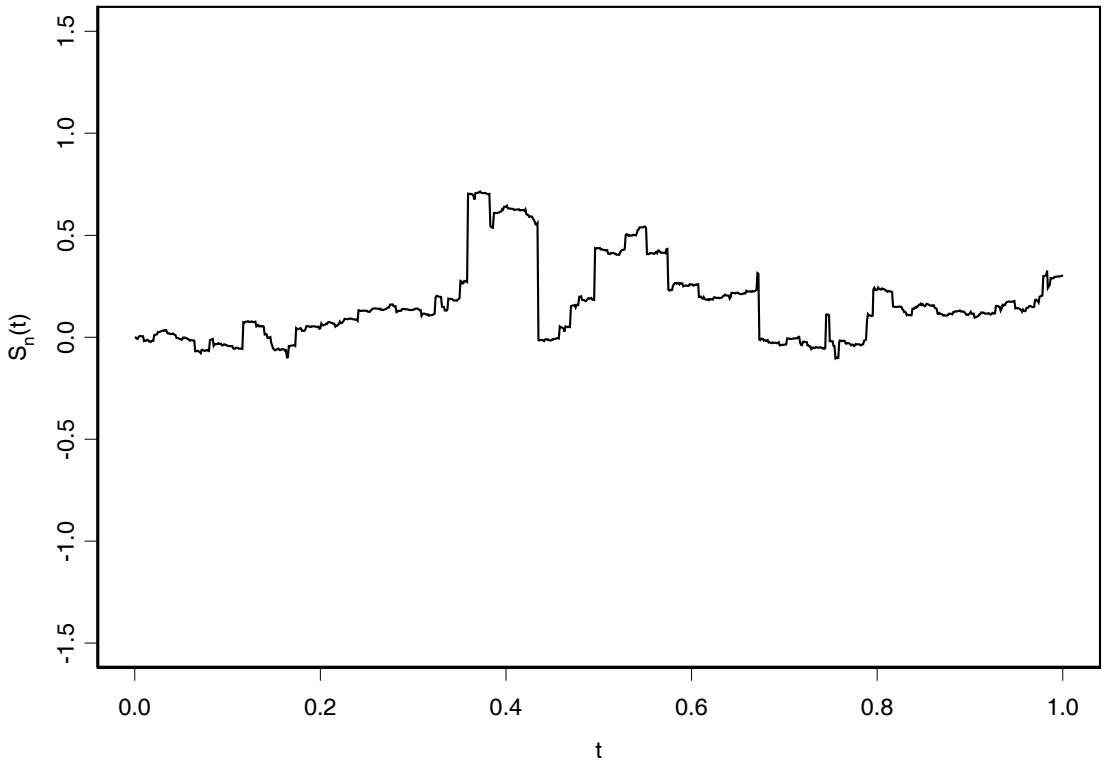
Uniform $U[-10, 10]$ random walk normalized

by its sample standard deviation, $n = 1000$



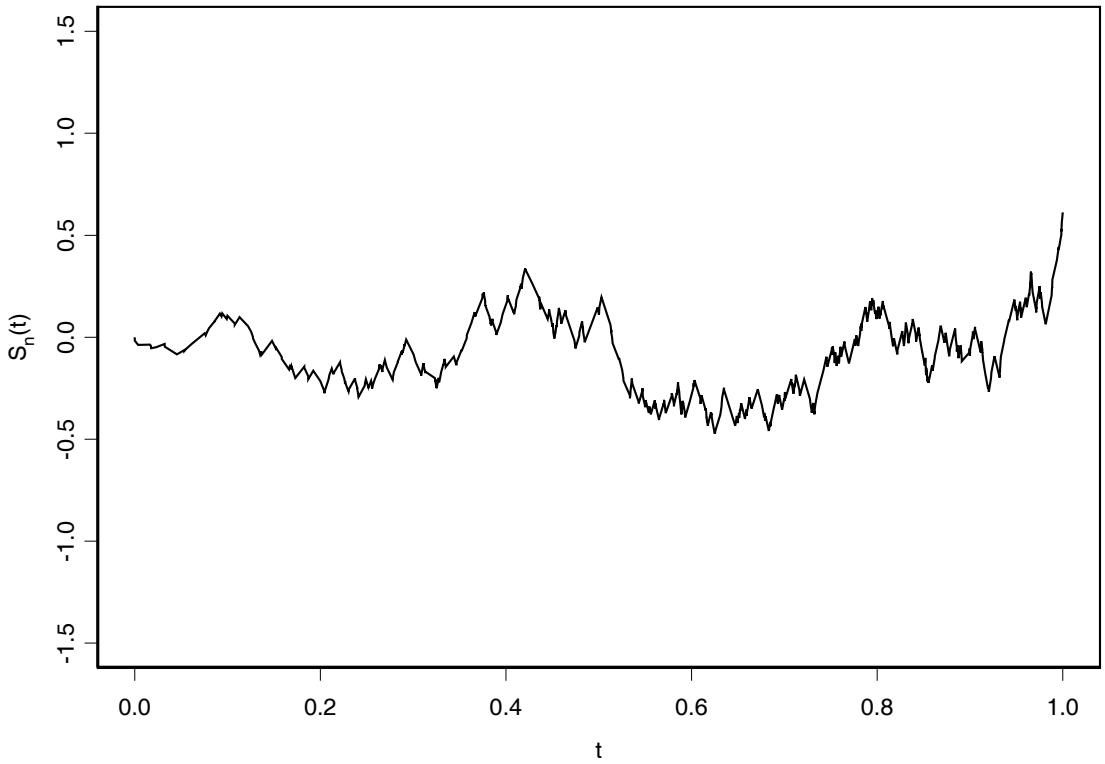
Cauchy random walk normalized

by its sample standard deviation, $n = 1000$



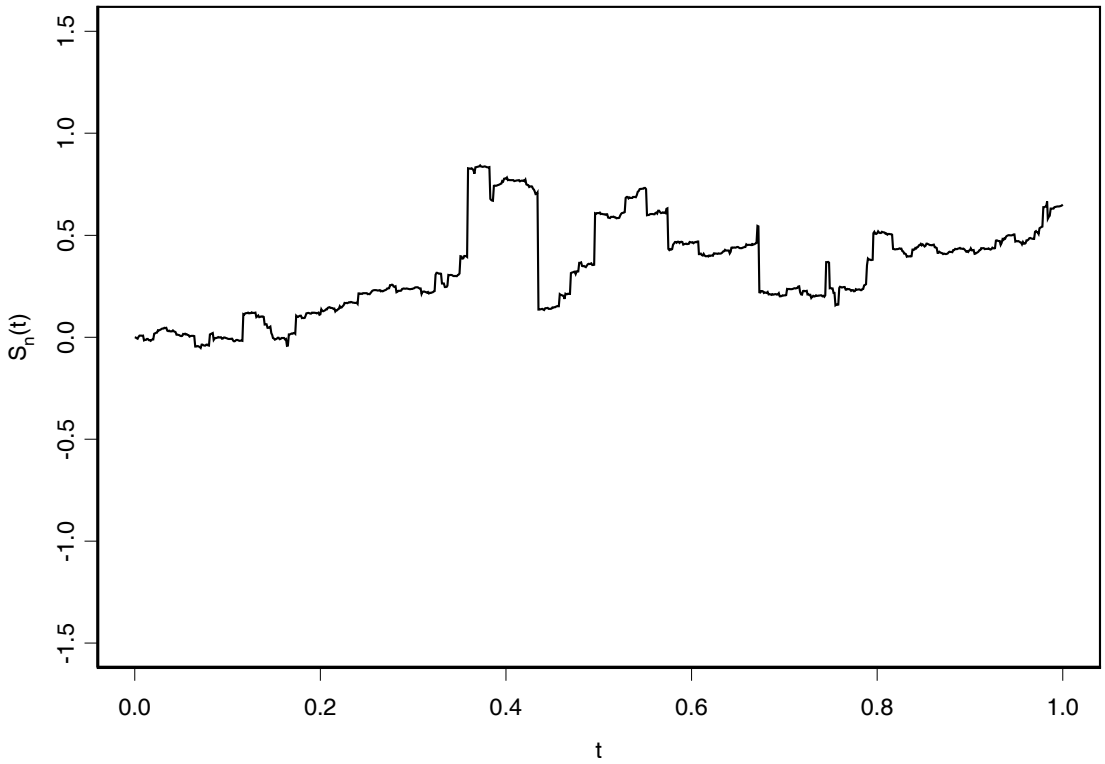
Truncated Cauchy random walk normalized

by its sample standard deviation, $n = 1000$



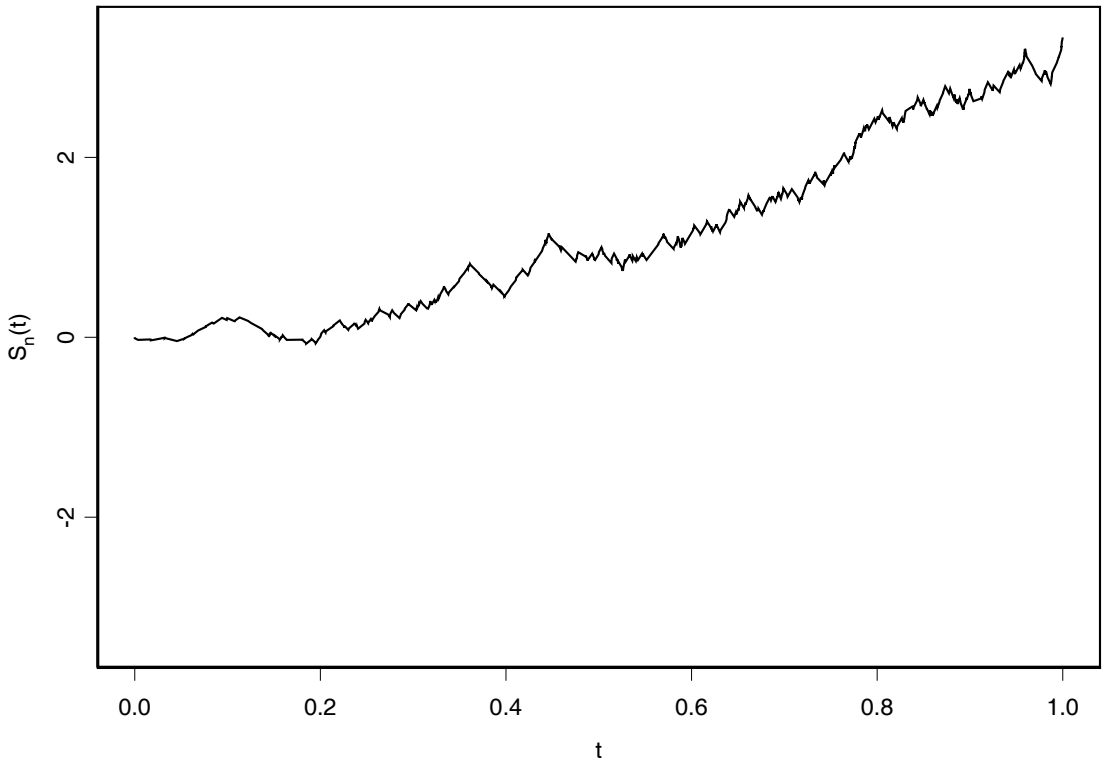
Non-centered ($\mu = .2$) Cauchy random walk

normalized by its sample standard deviation, $n = 1000$



Truncated non-centered ($\mu = .2$) Cauchy random walk

normalized by its sample standard deviation, $n = 1000$



RST: Statement

Let $\{X, X_i; i \geq 1\}$ be iid observations from a continuous distribution F symmetric about $-\infty < \theta < \infty$ and $\mathbf{E}(X^2) = \infty$. We are interested in developing a repeated significance test with initial sample size n_0 and target sample size N for testing

$$H_0 : \theta = 0 \text{ vs } H_a : \theta \neq 0. \tag{3}$$

In other words, we consider shift models $F(x - \theta), -\infty < \theta < \infty$ and test $\theta = 0$. Moreover, assume that F belongs to the domain of attraction of a stable distribution with exponent $0 < \gamma < 2$, i.e.

$$\mathbf{E}\left(X^2 \mathrm{I}_{(|X| \leq t)}\right) \sim t^{2-\gamma} L(t), \tag{4}$$

where L is a slowly varying function.

RST: usage of truncated sums

We would like the repeated significance test to be based on a sequence of partial sums. The problem we encounter here is that the sequence of partial sums from a distribution with an infinite second moment does not converge to a Brownian motion. To overcome this difficulty we employ partial sums of truncated random variables. Let $\{d_n; n \geq 1\}$ be an increasing sequence of positive numbers such that

$$nP\left(|X| > d_n\right) \sim \gamma_n, \tag{5}$$

where $\gamma_n \nearrow \infty$, as $n \to \infty$. Define the truncated partial sums:

$$S_n = \sum_{i=1}^{n} X_i \mathrm{I}_{(|X_i| \leq d_n)}. \tag{6}$$

Denote by

$$B_n = Var(S_n). \tag{7}$$

RST: Invariance Principle

The main theoretical difficulty we are facing here is that the sequence of truncated partial sums, $\{S_n; n \geq 1\}$, does not have independent increments and therefore the classical invariance principle (Donsker Theorem, Billingsley 1995, p. 520) is not applicable here. Let $\{W(t); 0 \leq t < \infty\}$ be the standard Brownian motion. The following invariance principle will be used to construct the continuation region for the proposed repeated significance test:

Theorem 1 *(Pozdnyakov 2003) If the random variable X belongs to the Feller class, i.e.*

$$\limsup_{t\to\infty} \frac{t^2\mathbf{P}(|X| > t)}{\mathbf{E}\Big(X^2\mathbf{I}_{|X|\leq t}\Big)} < \infty, \tag{8}$$

the average number of the excluded variables

$$n\mathbf{P}\Big(|X| > d_n\Big) \sim \gamma_n \nearrow \infty \tag{9}$$

and $\lim_{n\to\infty} B_n/B_{n+1} = 1$, *then* $S_n(t) \xrightarrow{d} W(t)$ *in the sense* $(\mathcal{C}[0,1], \rho)$, *where* $S_n(t)$ *is the linear interpolation between points*

$$\Big(0,0\Big), \Big(\frac{B_1}{B_n}, \frac{S_1}{\sqrt{B_n}}\Big), ..., \Big(1, \frac{S_n}{\sqrt{B_n}}\Big).$$

RST: design

Let

$$\tau = \min\left\{n_0 \leq n \leq N; |S_n| \geq b\sqrt{A_n}\right\} \tag{10}$$

be the stopping time associated with the repeated significance test, where n_0 is the initial sample size and N is the target sample size, and

$$A_n = \sum_{i=1}^{n} X_i^2 \mathrm{I}_{(|X_i| \leq d_n)} - \frac{S_n^2}{\sum_{i=1}^{n} \mathrm{I}_{(|X_i| \leq d_n)}}. \tag{11}$$

the sample version of variance which is a.s. equivalent to B_n.

RST: design

The repeated significance test stops and rejects H_0, given in equation (3), if and only if $\tau \leq N$. For all values of θ, the power function for the repeated significance test is given by:

$$\pi(\theta) = \mathbf{P}_\theta \left(\tau \leq N\right) = 1 - \beta(\theta), \tag{12}$$

where

$$\beta(\theta) = \mathbf{P}_\theta \left(|S_n| < b\sqrt{A_n}; n_0 \leq n \leq N \right), \tag{13}$$

is the probability of type II error function and $\left\{ b_n = b\sqrt{A_n}; n_0 \leq n \leq N \right\}$ is a sequence of constants that determine the continuation region of the repeated significance test. The significance level of this test is given by:

$$\begin{aligned} \alpha &= \pi(0) = \mathbf{P}_0 \left(\tau \leq N\right) = 1 - \beta(0) \\ &= \mathbf{P}_0 \left\{ \max_{n_0 \leq n \leq N} \left(\frac{|S_n|}{\sqrt{A_n}} \right) \geq b \right\}. \end{aligned} \tag{14}$$

RST: Design

To implement this test we need an accurate approximation for $b = b(\alpha, n_0, N)$ (the critical value of the test statistic associated with the repeated significance test), for a specified value of $0 < \alpha < 1$. The test statistic and the continuation region associated with this repeated significance test depend on the truncating levels d_n and the ratio of the initial and target sample sizes, which can be determined via the following result.

Proposition 1 *Assume that F belongs to the domain of attraction of a continuous symmetric stable distribution with exponent $0 < \gamma < 2$. Then the following results are true:*
1) F belongs to the Feller class.
2) The average number of the excluded terms $n\mathbf{P}(|X| > dn^{\delta}) \nearrow \infty$ whenever $1 - \gamma\delta > 0$. In particular, any $0 < \delta < 1/2$ guarantees it for all $0 < \gamma < 2$.
3) If $1 - \gamma\delta > 0$ and $\lim_{n_0, N \to \infty}(n_0/N) = c < 1$, then

$$\max_{n_0 \le n \le N} \frac{|S_n|}{\sqrt{A_n}} \xrightarrow{d} \sup_{[t_0 \le t \le 1]} \frac{|W(t)|}{\sqrt{t}},$$

where

$$t_0 = c^{1+(2-\gamma)\delta}. \tag{15}$$

RST: design

In view of Proposition 1, set the truncating levels $d_n = dn^{\delta}$, where $d > 0$ and $0 < \delta < 1/2$. Let $c = n_0/N$, be the ratio of the initial and target sample sizes of the repeated significance test.

It follows from Proposition 1 that $b = b(\alpha, n_0, N)$ can be approximated by $b_{t_0}(\alpha)$ by solving

$$\mathbf{P}\left(\sup_{[t_0,1]} \frac{|W(t)|}{\sqrt{t}} > b_{t_0}(\alpha)\right) = \alpha,$$

where t_0 is given in equation (15). We use De Long (1981) to evaluate $b_{t_0}(\alpha)$.

RST: example

Let $\{X_i; i \geq 1\}$ be a sequence of iid observations from a distribution F in the domain of attraction of a Cauchy distribution with location parameter θ and scale parameter 1. Assume that $H_0 : \theta = 0$ is true. We consider here truncation levels $d_n = n^{1/4}, n_0 \leq n \leq N$. In the following table, the performance of the proposed repeated significance test is evaluated in terms of accuracy of achieving an assigned significance level α, for given values of n_0 and N. The theoretical critical values $b_{t_0}(\alpha)$ and the corresponding targeted significance levels have been obtained from De Long (1981). The achieved significance levels were evaluated from a simulation with $10,000$ trials.

RST: example

Table 1. Simulated Significance Levels

n_0	N	t_0	$b_{t_0}(\alpha)$	$targeted\ \alpha$	$simulated\ \alpha$
100	303	1/4	2.7	.0503	.0541
100	303	1/4	3.3	.0098	.0094
100	754	1/12.5	2.6	.0989	.1012
30	91	1/4	2.7	.0503	.0638
30	91	1/4	3.3	.0098	.0167
30	226	1/12.5	2.6	.0989	.1119

Crossing boundary problem: connection to PDEs

- Significance level calculation: Standard BM

- Power calculation: BM with drift

- Expected time of crossing under alternative

Adaptive target sample size

Now we introduce a sequential test that is a repeated significance test with a random target sample size. This random target sample size adapts itself to the distribution of the observed data and it depends on the gross rate of the sample variance of the sequence of test statistics employed by the repeated significance test. The advantage in using a repeated significance test with an adaptive target sample size is that the resulting sequential test is fully nonparametric. Its implementation does not require the knowledge of the asymptotic tail behavior of the distribution of the observed data, a condition needed for developing the repeated significance test in Glaz and Pozdnyakov (2004).

RST with adaptive target sample size

Let A_n be a sample variance of T_n. Define a stopping time $\mathcal{N}$ by

$$\mathcal{N} = \inf\{k \geq n_0 : \frac{A_k}{A_{n_0}} \geq \frac{1}{t_0}\}, \tag{16}$$

where $0 < t_0 < 1$ is a design parameter. A repeated significance test with adaptive target sample size is defined as follows. At time $k \geq n_0$ observe T_k. Stop and reject H_0, if k is the smallest integer such that $A_k/A_{n_0} < 1/t_0$ and $|T_k| \geq b\sqrt{A_k}$. Otherwise, we stop monitoring at time $\mathcal{N}$ and accept H_a. The following result is of major importance for implementing the repeated significance test with adaptive target sample size.

Main theoretical result

Theorem 2 *Assume that the functional central limit theorem for the sequence* $\{T_n\}$ *holds, and the sequence of* $B_n \sim Var(T_n)$ *satisfies*

$$B_n \uparrow \infty, \quad \frac{B_{n-1}}{B_n} \to 1, \ \textit{as } n \uparrow \infty. \tag{17}$$

If the sample variance A_n *satisfies*

$$\frac{A_n}{B_n} \to 1 \quad \textit{a.s.}, \tag{18}$$

then

$$P\left(\max_{n_0 \le k \le \mathcal{N}} \left|\frac{T_k}{\sqrt{A_k}}\right| > b\right) \longrightarrow \alpha(t_0, b) \ \textit{as } n_0 \to \infty. \tag{19}$$

Simulation

Table 2. Simulated Significance Levels and Expected Stopping Times,

$$n_0 = 100,\ d_n = n^{1/4}$$

t_0^{-1}	b	$Normal$	$Cauchy^{1/2}$	$Cauchy$	$Cauchy^2$
4	3.3	.010	.010	.009	.008
		391 ± 52	319 ± 38	276 ± 36	260 ± 42
	2.7	.051	.047	.047	.046
		382 ± 69	313 ± 49	272 ± 43	256 ± 48
7.5	3.4	.010	.012	.010	.009
		729 ± 104	544 ± 70	439 ± 62	397 ± 69
	2.8	.051	.048	.048	.045
		711 ± 141	533 ± 92	429 ± 77	391 ± 79

Simulation: comparison to classical RST

Table 3. Simulated Significance Levels and Power,

$n_0 = 100$, $1/t_0 = 7.5$, $b = 2.8$, $\alpha = .0513$ and $d_n = n^{1/4}$ (for ARST)

	Normal		*Cauchy*	
θ	$CRST$	$ARST$	$CRST$	$ARST$
0	.0521	.0531	.0069	.0473
.05	.1872	.1885	.0082	.0751
.10	.6306	.6283	.0099	.1531
.15	.9445	.9357	.0111	.2847
.20	.9981	.9968	.0176	.4769
.25	1.0	.9999	.0204	.6682
.30	1.0	1.0	.0339	.8196

Simulation: Comparison to RST based on Cauchy scores

Table 3. Simulated Significance Levels and Power, Cauchy+U[-5;5]

$n_0 = 100$, $1/t_0 = 7.5$, $b = 2.8$, $\alpha = .0513$ and $d_n = n^{1/2}$ (for ARST)

θ	$CSRST$	$ARST$
0	.0532	.0507
.25	.0896	.1602
.50	.2239	.5082
.75	.4407	.8590
1.0	.7240	.9798