A Robust Nonparametric RST

Joseph Glaz and Vladimir Pozdnyakov

Department of Statistics

University of Connecticut

Outline

1. Introduction

2. A Nonparametric Repeated Significance Test

3. A Robust Nonparametric Repeated Significance Test

4. Evaluating the Power Function of the RST

5. Concluding Remarks and Future Work

Introduction

Repeated significance tests were introduced in Armitage (1958).

Major developments in the theory and applications of RST:

Dias and Garcia (1999)

Hu (1988)

Jennison and Turnbull (2000)

Lai and Siegmund (1977 and 1979)

Lalley (1983)

Lerche (1986)

Selke and Siegmund (1983)

Sen (1981, 1985 and 1991)

Siegmund (1982 and 1985)

Takahashi (1990)

Whitehead (1997)

Woodroofe (1979 and 1982)

Woodroofe and Takahashi (1982).

Distributions with heavy tails have been used in modeling

1. computer network traffic:

Crovella, Taqqu and Bestavros (1998)

Willinger, Paxson and Taqqu 1998)

2. telecommunication systems:

Crovella and Taqqu (1999)

3. **high frequency financial data:**

M$\ddot{u}$ller, Dacorogna and Pictet (1998)

4. risk management and insurance data:

Bassi, Embrecht and Kafetzaki (1998)

Invariance Theorem

Donsker's Theorem

Let $\{X, X_i\}_{i\geq 1}$ be i.i.d random variables, $EX = \mu$, $VarX = \sigma^2$ and

$S_n = X_1 + X_2 + ... + X_n$. A Functional Central Limit Theorem (Donsker's Theorem, see Billingsley (1995, p. 520)) implies that if $S_n(t)$ is the linear interpolation between points

$$\Big(0, 0\Big), \Big(\frac{1}{n}, \frac{S_1 - \mu}{\sigma\sqrt{n}}\Big), ..., \Big(1, \frac{S_n - n\mu}{\sigma\sqrt{n}}\Big)$$

then

$$S_n(t) \xrightarrow{d} W$$

in the sense $\mathcal{C}[0,1]$ with uniform metric ρ where W is standard Brownian motion on $[0,1]$.

A Nonparametric Repeated Significance Test

Let $X_1, X_2, ..., X_n, ...$ be a sequence of independent and identically distributed (iid) observations from a continuous distribution F with median $-\infty < \theta < \infty$. We are interested in testing

$H_0 : \theta = 0$ vs $H_a : \theta \neq 0$ via a nonparametric sequential procedure. Suppose that we want to be sure that at most N observations will be needed to reach a decision. Assuming that $\sigma^2 = Var(X_1) < \infty$, Sen (1981 and 1985) develops the following *repeated significance test* (**RST**), described below.

Let $S_n = \sum_{i=1}^{n} X_i$ and define the stopping rule

$$\tau = \min\left\{n_0 \leq n \leq N; |S_n| \geq b\sigma\sqrt{n}\right\},$$

where n_0 is the initial sample size, N is the target sample size and $b > 0$ is a constant. The repeated significance test stops and rejects H_0 if and only if $\tau \leq N$. The power function of the RST, $\beta(\theta)$, is given by:

$$\begin{aligned} \beta(\theta) &= P_\theta(\tau \leq N) = 1 - P_\theta(\tau > N) \\ &= 1 - P_\theta\left(|S_n| < b\sigma\sqrt{n}; n_0 \leq n \leq N\right). \end{aligned}$$

For a specified significance level $\alpha > 0$, if $n_0/N \to t_0$ as $N \to \infty$, it follows from Donsker's Theorem that

$$\max\left\{\frac{|S_n|}{\sigma\sqrt{n}}; n_0 \le n \le N\right\} \xrightarrow{d} \sup\left\{\frac{W(t)}{\sqrt{t}}; t_0 \le t \le 1\right\}$$

and

$$\beta(0) = P_0\left(\max\left\{\frac{|S_n|}{\sigma\sqrt{n}}; n_0 \le n \le N\right\} \ge b\right) \to \alpha,$$

where $W(t)$ is a standard Brownian motion on the interval $[0, 1]$ and $b = b_{t_0}(\alpha)$ is the constant that characterizes the continuation region, given by the square root boundary, corresponding to the prescribed significance level α.

The critical values $b_{t_0}(\alpha)$ for different choices of α and t_0 can obtained

from DeLong (1981).

If σ is unknown, one can replace it by the sample standard deviation since it converges almost surely to σ (Sen 1981).

Robust Nonparametric Repeated

Significance Tests

Let $X_1, X_2, ..., X_n, ...$ be a sequence of independent and identically distributed observations from a continuous distribution F symmetric about the median $-\infty < \theta < \infty$.

Assume that F is from a class of heavy tail distributions with an infinite variance and possibly no mean, as in the case of the Cauchy distribution. We extend the approach in Sen (1981 and 1985) and derive a RST for testing

$H_0 : \theta = 0$ vs $H_a : \theta \neq 0$.

Main obstacles needed to overcome in deriving a RST for data from heavy tail distributions:

- A random walk based on increments from a distribution with an

  infinite second moment does not converge to a Brownian motion

- New invariance principles are needed for the random walks

  that will be used with the RST

To overcome the first difficulty we employ truncated sums. We first consider the approximation for $H_0 : \theta = 0$. Let $\{X, X_i\}_{i \geq 1}$ be i.i.d. random variables with a symmetric continuous distribution and $EX^2 = \infty$. Let $\{d_n\}_{n \geq 1}$ be an increasing sequence of positive numbers such that

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

The *truncated sums* S_n^* we will consider are defined by

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

Denote by B_n the variance of S_n^*. The main theoretical difficulty is that the sequence of the truncated sums is not a process with independent increments. Therefore, the classical weak invariance principle is not applicable here.

However, in the case of symmetric distributions, the truncated sums form a martingale, and this observation allows one to prove the analog of Donsker's theorem for the truncated sums. The following result is true:

Theorem (Pozdnyakov 2002) If the random variable X belongs to the Feller class

$$\limsup_{t\to\infty} \frac{t^2 P(|X|>t)}{E\Big(X^2 \mathsf{I}_{|X|\le t}\Big)} < \infty,$$

the average number of the excluded variables

$$nP\Big(|X|>d_n\Big) \sim \gamma_n \nearrow \infty,$$

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

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

If B_n is an unknown sequence then estimates of B_n could be used. It follows from Egorov and Pozdnyakov (1997) that under the additional condition

$$\frac{nP\Big(|X| > d_n\Big)}{\ln\ln(n)} \to +\infty,$$

the iterated logarithm law for the truncated sums S_n^* is valid:

$$\limsup_{n\to\infty} \frac{S_n^*}{\sqrt{2B_n \ln\ln(B_n)}} = 1 \text{ a.s..}$$

Under the same conditions the strong law of large numbers for the truncated sums of squares holds

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i^2 \mathsf{I}_{|X_i|\le d_n}}{B_n} = 1 \text{ a.s..}$$

To construct an effective RST procedure we propose the use of the following sample variance version of truncated sums B_n:

$$A_n = \sum_{i=1}^{n} X_i^2 \mathsf{I}_{|X_i|\le d_n} - \frac{S_n^{*2}}{\sum_{i=1}^{n} \mathsf{I}_{|X_i|\le d_n}}.$$

It is not difficult to show that the LIL and the SLLN for the truncated sums imply that A_n and B_n are a.s. equivalent:

$$\frac{A_n}{B_n} \to 1 \text{ a.s.}$$

and therefore we can use A_n as a normalizing sequence.

Let

$$\tau = \min\left\{n_0 \leq n \leq N; |S_n^*| \geq b_n \sqrt{A_n}\right\}$$

be the stopping rule, where n_0 and N are the initial and the target sample size, respectively and $b_n > 0$ is a constant.

The RST stops and rejects H_0 if and only if $\tau \leq N$.

The power function of the RST is given by

$$\begin{aligned} \beta(\theta) &= P_\theta(\tau \leq N) = 1 - P_\theta(\tau > N) \\ &= 1 - P_\theta\left(|S_n^*| < b_n \sqrt{A_n}; n_0 \leq n \leq N\right). \end{aligned}$$

In view of Pozdnyakov (2002), if $B_{n_0}/B_N \to t_0$ and $N \to \infty$, then under H_0,

$$\max\left\{\frac{|S_n^*|}{\sqrt{A_n}}; n_0 \leq n \leq N\right\} \xrightarrow{d} \sup\left\{\frac{W(t)}{\sqrt{t}}; t_0 \leq t \leq 1\right\}$$

and consequently,

$$\beta(0) = P_0\left(\max_{n_0 \leq n \leq N}\left\{\frac{|S_n^*|}{\sqrt{A_n}}\right\} \geq b_n\right) \to \alpha,$$

where the constant $b_n = b_n(\alpha)$ is the critical value that determines the continuation region of the RST.

We now present an approximation for $b_n(\alpha)$ for the class of stable distributions.

Let $\{X, X_i\}_{i\geq 1}$ be i.i.d. random variables with a symmetric continuous distribution and $EX^2 = \infty$. Assume also that X belongs to the domain of attraction of the stable distribution with an exponent $0 < \gamma < 2$, i.e.

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

where $L(t)$ is a slowly varying function. Feller (1971, p. 313) shows that the above condition is equivalent to

$$\lim_{t\to\infty} \frac{t^2 P(|X| > t)}{E\Big(X^2 \mathsf{I}_{(|X|\leq t)}\Big)} = \frac{2-\gamma}{\gamma}.$$

Hence, the random variable X belongs to the Feller class.

Let $\{X, X_i\}_{i \geq 1}$ be i.i.d. random variables with a symmetric stable continuous distribution and $EX^2 = \infty$. To apply the invariance principle for truncated sums in Pozdnyakov (2002) the average number of excluded terms, $nP(|X| > d_n)$, has to approach infinity. For $0 < \gamma < 2$ we have that

$$P(|X| > t) \sim \frac{2-\gamma}{\gamma} t^{-\gamma} L(t),$$

where $L(t)$ is a slowly varying function. Hence, for $d_n = dn^\delta$, where $d, \delta > 0$, we have that

$$nP(|X| > dn^\delta) \sim \frac{2-\gamma}{\gamma} d^{-\gamma} n^{1-\gamma\delta} L(dn^\delta).$$

Therefore the average number of the excluded terms

$nP(|X| > dn^{\delta}) \nearrow \infty$ whenever $1 - \gamma\delta > 0$. If $0 < \delta < 1/2$ this is guaranteed for all $0 < \gamma < 2$.

To approximate the constant $b_n(\alpha)$ associated with a specified significance level α of the RST one has to evaluate

$t_0 = \lim \frac{B_{n_0}}{B_N}$ as n_0, $N \to \infty$.

Let n_0, $N \to \infty$ such that $n_0/N \to c$, where $0 < c < 1$. Since,

$$B_n \sim n d^{2-\gamma} n^{(2-\gamma)\delta} L(dn^\delta)$$

and

$$\lim_{\substack{n_0, N \to \infty \\ \frac{n_0}{N} \to c}} \frac{L(dn_0^\delta)}{L(dN^\delta)} = 1,$$

it follows that

$$\lim_{\substack{n_0, N \to \infty \\ \frac{n_0}{N} \to c}} \frac{B_{n_0}}{B_N} = c^{1+(2-\gamma)\delta}.$$

Therefore, if $b_n = bn^{\delta}$, $0 < \gamma < 2$, $0 < \delta < 1/2$, $n_0, N \to \infty$ and

$n_0/N \to c$, $0 < c < 1$ we get from Pozdnyakov (2002) that

$$\max\left\{\frac{|S_n|}{\sqrt{A_n}}; n_0 \leq n \leq N\right\} \xrightarrow{d} \sup_{[c^{1+(2-\gamma)\delta},1]} \frac{|W(t)|}{\sqrt{t}}.$$

The constant $b_n(\alpha)$ can be approximated by $b_{t_0}(\alpha)$ by solving

$$P\left(\sup_{[c^{1+(2-\gamma)\delta},1]}\left\{\frac{|W(t)|}{\sqrt{t}}\right\} \geq b_{t_0}(\alpha)\right) = \alpha,$$

using the approach in De Long (1981).

We now present numerical results for evaluating the approximation for the critical value $b_n(\alpha)$ for data

from a Cauchy distribution.

Let us assume that X has the Cauchy distribution, i.e. $\gamma = 1$.

We consider the truncation level $d_n = n^{1/4}$, i.e. $\delta = 1/4$. In table 1 simulation results are presented. For each case the number of simulations performed is 10000. The theoretical critical values and the corresponding significance levels are taken from De Long (1981).

$\mathbf{n}_0$	$\mathbf{N}$	$\mathbf{t}_0$	$\mathbf{b}_{t_0}(\boldsymbol{\alpha})$	**theoretical** α	**simulated** α
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**

Table 1. Simulation Results for Probability of Type I Error

Approximating the Power Function of the RST

Let $X, X_1, X_2, ..., X_n, ...$ be a sequence of independent and identically distributed observations from a continuous distribution F symmetric about the median $-\infty < \theta < \infty$ and $E\left(X^2\right) = \infty$.

The power function of the RST is given by:

$$\begin{aligned} \beta(\theta) &= P_\theta(\tau \leq N) = 1 - P_\theta(\tau > N) \\ &= 1 - P_\theta\left(|S_n^*| < b_n\sqrt{A_n}; n_0 \leq n \leq N\right). \end{aligned}$$

To approximate the power function an additional assumption has to be made:

$$E\left(X^2 I_{(|X| \leq t)}\right) \sim Kt^{2-\gamma},$$

where the constants $K > 0$ and $0 < \gamma < 2$ are known.

For $n \geq 1$ let

$$S_n^{*0} = \sum_{i=1}^{n} (X_i - \theta)\, I_{(|X_i - \theta| \, < \, dn^{\delta})}.$$

It follows that

$$\left| \frac{S_n^* - (S_n^{*0} + n\theta)}{\sqrt{B_n}} \right| \xrightarrow{P} 0.$$

In view of that and the fact that

$$\frac{B_n}{B_N} \sim \left(\frac{n}{N}\right)^{1+\delta(2-\gamma)} = t$$

and

$$\frac{S_n^{*0} + n\theta}{\sqrt{B_n}} \sim \frac{S_n^{*0}}{\sqrt{B_n}} + \theta \frac{N^{[1-\delta(2-\gamma)]/2}}{K^{1/2} b^{(2-\gamma)/2}} t^{1/[1+\delta(2-\gamma)]},$$

one can approximate S_n^* with a Brownian motion with a nonlinear drift:

$$W(t) + \theta \frac{N^{[1-\delta(2-\gamma)]/2}}{K^{1/2} b^{(2-\gamma)/2}} t^{1/[1+\delta(2-\gamma)]}.$$

Consequently, the power function of the RST can be approximated by

$$1 - P\left(\begin{array}{c} \left| W(t) + \theta \frac{N^{[1-\delta(2-\gamma)]/2}}{K^{1/2} b^{(2-\gamma)/2}} t^{1/[1+\delta(2-\gamma)]} \right| \\ < b_{t_0}(\alpha)\sqrt{t},\ t_0 \le t \le 1 \end{array} \right).$$

The power computation boils down to computing:

$$P\left(|W(t) + ct^{\rho}| < b\sqrt{t} \text{ for all } t \in [t_0, 1] \right),$$

where $1/2 < \rho < 1$. Since

$$\begin{aligned} & P\left(-b\sqrt{t} - ct^{\rho} < W(t) < b\sqrt{t} - ct^{\rho} \text{ for all } t \in [t_0, 1] \right) \\ & = \int_{-b\sqrt{t_0} - ct_0^{\rho}}^{b\sqrt{t_0} - ct_0^{\rho}} P\left(\begin{array}{c} -b\sqrt{t} - ct^{\rho} < W(t) < \\ b\sqrt{t} - ct^{\rho} \text{ for all } t \in [t_0, 1] \Big| W(t_0) = x \end{array} \right) \times \\ & \qquad P\Big(W(t_0) \in dx \Big), \end{aligned}$$

computing the power function is equivalent to solving the following problem. Consider the domain

$$D = \Big\{(x, y) : -t_0 \leq y \leq 1 - t_0,$$
$$-b\sqrt{t_0 + y} - c(t_0 + y)^{\rho} < x < b\sqrt{t_0 + y} - c(t_0 + y)^{\rho}\Big\}.$$

Let $\tau_D(x, y)$ be the first time when the "degenerated" two-dimensional diffusion

$(x_t, y_t) = (x + W(t), y + t)$ exits from the domain D, where (x, y) belongs to the interior of the domain D.

What is the probability that a Brownian motion starting at point x and at time y will stay inside the curved boundaries, i.e.

$$P\Big(y_{\tau_D(x,y)} = 1 - t_0\Big)?$$

The generating operator of the diffusion (x_t, y_t) is given by

$$\frac{1}{2}\frac{\partial^2}{\partial x^2} + \frac{\partial}{\partial y}.$$

By Venttsel (1996, p. 333) the function

$$v(x,y) = P\Big(y_{\tau_D(x,y)} = 1 - t_0\Big)$$

is the unique solution of the PDE

$$\frac{1}{2}\frac{\partial^2 v}{\partial x^2}(x,y) + \frac{\partial v}{\partial y}(x,y) = 0 \quad (x,y) \in D,$$

that satisfies the following boundary conditions:

1. $v(\pm b(t_0 + y)^{1/2} - c(t_0 + y)^{\rho}, y) \equiv 0$,

2. $v(x, 1 - t_0) \equiv 1$.

We can solve this parabolic equation numerically which in turn will yield an approximation for the power function of the RST.

Numerical Results

In Table 2 we present approximations for the power function of the RST for the Cauchy distribution for various choices of θ computed via the Brownian motion approximation and by simulations .

The initial sample size n is 100. The target sample size N_0 is 303. These choices correspond to the first row of **Table 1**. However, in this case we choose a higher truncating level

$d_n = dn^\delta = 5n^{1/4}$. The multiplier $d = 5$ is taken in order to get a good approximation by the Brownian motion with the nonlinear

drift. Note that the multiplier d does not have an effect on the approximation for the probability of type I error .

For the Cauchy distribution $K = 2/\pi$ and $\gamma = 1$.

θ	BM approx (4 refinm)	BM approx (5 refinm)	Sim (1000 sim)
0	.0508	.0505	.044 (.0503)
.25	.1918	.1913	.183
.5	.5958	.5962	.557
.75	.9185	.9186	.907
1	.9948	.9948	.991

Table 2. Approximations for the Power Function .

Concluding Remarks and Future Work

1. **Evaluating $\mathrm{E}(\tau)$ and $\mathrm{Var}(\tau)$.**

2. **Approximating $\mathrm{P}(\tau \leq \mathrm{n})$.**

3. **Approximating the P-value of the RST**

4. **Confidence interval for θ, after the RST rejects $\mathrm{H_0}$.**

5. **Multivariate RST**