Limit theorems for rank statistics detecting gradual changes

Aleš Slabý

Abstract. The purpose of the paper is to investigate weak asymptotic behaviour of rank statistics proposed for detection of gradual changes, linear trends in particular. The considered statistics can be used for various test procedures. The fundaments of the proofs are formed by results of Hušková [4] and Jarušková [5].

Keywords: gradual type of change, location model, rank statistics, weighted suprema, limit theorems

Classification: 62E20, 62G20, 62G30

1. Introduction

The basic underlying problem is testing a sequence of i. i. d. (time ordered) random observations $X_1, \dots, X_n$ having the same common continuous distribution against an alternative that at an unknown time point there is a beginning of gradual type of change in location such that the trend after the change point is linear. Namely, we test

$$H: \quad X_i = \theta + e_i$$

against

$$A: \quad X_i = \theta + \delta \frac{i-m}{n} I_{\{i>m\}} + e_i \quad \text{for some} \ \ 1 \le m < n \ \ \text{and} \ \ \delta \ne 0$$

where $e_1, \dots, e_n$ are i. i. d. with continuous distribution function $F(x, 0)$ provided that $F(x, \theta) = F(x - \theta) = F(x)$. However, statistics studied in the sequel can be used for testing other and more complicated alternatives, see Slabý [7].

Let $R_{1n}, \dots, R_{nn}$ be the ranks corresponding to the observations $X_1, \dots, X_n$. Consider two simple linear rank statistics

$$S_{1k}(\boldsymbol{a}) = \sum_{i=1}^{k} \big(a(R_{in}) - \overline{a}_n\big) \tag{1}$$

The author gives special thanks to Marie Hušková, who was instant in granting hints and consultations. This work was also supported by grants GAČR 201/00/0769 and MSM 113200008.

and

$$S_{2k}(\boldsymbol{a}) = \sum_{i=k+1}^{n} \big(a(R_{in}) - \overline{a}_n\big)\frac{i-k}{n} \tag{2}$$

where $k = 1, \dots, n$, and $a(1), \dots, a(n)$ are scores with properties

$$\frac{1}{n}\sum_{i=1}^{n} \big(a(i) - \overline{a}_n\big)^2 \geq D_1 \tag{3}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \big|a(i) - \overline{a}_n\big|^{2+\eta} \leq D_2 \tag{4}$$

for some finite positive constants D_1, D_2 and η independent of n. Here

$$\overline{a}_n = \frac{1}{n}\sum_{i=1}^{n} a(i). \tag{5}$$

Hušková [4] studied limit behaviour of weighted suprema and L_p-functionals based on $S_{1k}(\boldsymbol{a})$ and, then, proposed corresponding testing procedures for abrupt change in location model setup. Whereas $S_{1k}(\boldsymbol{a})$ is more suitable for testing abrupt changes, analogous weighted suprema based on $S_{2k}(\boldsymbol{a})$ can be better employed to test gradual changes, particularly changes of linear trend in the location model.

The crucial result of Hušková [4] is that limit behaviour of the treated statistics is the same as the limit behaviour of the respective functionals of sums of certain i. i. d. random variables. Limit theorems for i. i. d. random variables, see Csörgő and Horváth [1], can be then extended to ranks. Hušková [4] shows that instead of investigating behaviour of $S_{1k}(\boldsymbol{a})$, it is sufficient to check behaviour of

$$Z_{1k}(\boldsymbol{a}) = \sum_{i=1}^{k} \big(a(1 + [nU_i]) - \overline{a}_n\big) \tag{6}$$

where

$$U_i = F(X_i), \quad i = 1, \dots, n, \tag{7}$$

and $[a]$ denotes the integer part of a. Notice that for $k = 1, \dots, n$ we have

$$E\,S_{1k}(\boldsymbol{a}) = E\,Z_{1k}(\boldsymbol{a}) = 0 \tag{8}$$

and

$$\text{Var } S_{1k}(\boldsymbol{a}) = \frac{n}{n-1} \text{Var} \left\{ Z_{1k}(\boldsymbol{a}) - \frac{k}{n} Z_{1n}(\boldsymbol{a}) \right\} = \frac{k(n-k)}{n} \sigma_n^2(\boldsymbol{a}) \tag{9}$$

where

$$\sigma_n^2(\boldsymbol{a}) = \frac{1}{n-1} \sum_{i=1}^{n} \big(a(i) - \overline{a}_n\big)^2. \tag{10}$$

General formulae for expectation and variance of simple linear rank statistics can be for example found in Hájek and Šidák [3], see Theorem c of Section 3.1.

A similar approach can be used in the case of statistics based on $S_{2k}(\boldsymbol{a})$, namely, we consider

$$Z_{2k}(\boldsymbol{a}) = \sum_{i=k+1}^{n} \big(a(1 + [nU_i]) - \overline{a}_n\big) \frac{i-k}{n} \tag{11}$$

instead of $S_{2k}(\boldsymbol{a})$. There is a simple but very important relationship between $Z_{1k}(\boldsymbol{a})$ and $Z_{2k}(\boldsymbol{a})$. Obviously

$$Z_{2k}(\boldsymbol{a}) = -\frac{1}{n} \sum_{i=k}^{n-1} Z_{1i}(\boldsymbol{a}) + \frac{n-k}{n} Z_{1n}(\boldsymbol{a}). \tag{12}$$

The same holds for $S_{1k}(\boldsymbol{a})$ and $S_{2k}(\boldsymbol{a})$, however, since $S_{1n}(\boldsymbol{a}) = 0$ the analog of (12) can be simplified to

$$S_{2k}(\boldsymbol{a}) = -\frac{1}{n} \sum_{i=k}^{n-1} S_{1i}(\boldsymbol{a}). \tag{13}$$

Finally note that for $k = 1, \dots, n$ we obtain

$$E\, S_{2k}(\boldsymbol{a}) = E\, Z_{2k}(\boldsymbol{a}) = 0 \tag{14}$$

and

$$\begin{aligned} \text{Var } S_{2k}(\boldsymbol{a}) &= \frac{n}{n-1} \text{Var} \left\{ Z_{2k}(\boldsymbol{a}) - \frac{(n-k)(n-k+1)}{2n^2} Z_{1n}(\boldsymbol{a}) \right\} \\ &= v(n,k)\, \sigma_n^2(\boldsymbol{a}) \end{aligned} \tag{15}$$

where

$$v(n,k) = \frac{(n-k)(n-k+1)(2(n-k)+1)}{6n^2} - \frac{(n-k)^2(n-k+1)^2}{4n^3}\,. \tag{16}$$

2. Limit theorems

The main results are summarized below in Theorem 1. These results have been mentioned without any proof in Theorem 2 of [7] and employed there in a comparative simulation study.

Theorem 1. *Let $X_1, \dots, X_n$ be i. i. d. random variables with common continuous distribution function F. Let assumptions* (3) *and* (4) *hold.*

(i) *As $n \to \infty$, for arbitrary $y \in \mathbb{R}$,*

$$P\left\{ \sqrt{2\log\log n} \max_{1\le k\le n-1} \frac{1}{\sqrt{v(n,k)}} \frac{1}{\sigma_n(\boldsymbol{a})} \left|S_{2k}(\boldsymbol{a})\right| \le y + 2\log\log n + \log\frac{\sqrt{3}}{4\pi} \right\} \longrightarrow \exp(-2e^{-y}) \tag{17}$$

where $v(n,k)$ is defined in (16).

(ii) *If moreover, as $n \to \infty$,*

$$\frac{n}{G} \longrightarrow \infty \quad \text{and} \quad \frac{n^{2/(2+\eta)}\log n}{G} \longrightarrow 0, \tag{18}$$

then for arbitrary $y \in \mathbb{R}$, as $n \to \infty$,

$$P\left\{ \sqrt{2\log\frac{n}{G}} \max_{G<k<n-G} \frac{1}{\sqrt{w(n,G)}} \frac{1}{\sigma_n(\boldsymbol{a})} \left|S_{2,k+G}(\boldsymbol{a}) - 2S_{2k}(\boldsymbol{a}) + S_{2,k-G}(\boldsymbol{a})\right| \le y + 2\log\frac{n}{G} + \log\frac{\sqrt{3}}{4\pi} \right\} \longrightarrow \exp(-2e^{-y}) \tag{19}$$

where

$$w(n,G) = \frac{G(2G^2+1)}{3n^2} - \frac{G^4}{n^3}\,. \tag{20}$$

The theorem below is a result analogous to Theorem 1 but it assumes certain i. i. d. variables instead of the ranks.

Theorem 2. *Let $X_1, \dots, X_n$ be i. i. d. random variables with $EX_1 = 0$, $\operatorname{Var} X_1 = 1$ and $E|X_1|^{2+\eta} < \infty$ for some $\eta > 0$. For $k = 0, \dots, n$ denote*

$$\widehat{S}_k = \sum_{i=k+1}^{n} X_i \frac{i-k}{n}\,. \tag{21}$$

(i) *As $n \to \infty$, for arbitrary $y \in \mathbb{R}$,*

$$P\left\{ \sqrt{2\log\log n} \max_{1\le k\le n-1} \frac{1}{\sqrt{\widehat{v}(n,k)}} \left|\widehat{S}_k\right| \le y + 2\log\log n + \log\frac{\sqrt{3}}{4\pi} \right\} \longrightarrow \exp(-2e^{-y}) \tag{22}$$

where

$$\widehat{v}(n,k) = \frac{(n-k)(n-k+1)(2(n-k)+1)}{6n^2}\,. \tag{23}$$

(ii) *If moreover condition* (18) *holds then for arbitrary $y \in \mathbb{R}$, as $n \to \infty$,*

$$P\left\{ \sqrt{2\log\frac{n}{G}} \max_{G<k<n-G} \frac{1}{\sqrt{\widehat{w}(n,G)}} \left|\widehat{S}_{k+G} - 2\widehat{S}_k + \widehat{S}_{k-G}\right| \le y + 2\log\frac{n}{G} + \log\frac{\sqrt{3}}{4\pi} \right\} \longrightarrow \exp(-2e^{-y}) \tag{24}$$

where

$$\widehat{w}(n,G) = \frac{G(2G^2+1)}{3n^2}\,. \tag{25}$$

Corollary. *Convergence* (22) *remains true if $\widehat{v}(n,k)$ is replaced with $v(n,k)$ defined in* (16)*. Similarly, convergence* (24) *remains true in the case that $\widehat{w}(n,G)$ is replaced with $w(n,G)$ defined in* (20)*.*

The following theorem, which can be considered to be an extension of Theorem 3 in [4], poses the crucial step in the proof of Theorem 1. It is an interlink between Theorem 1 and Theorem 2.

Theorem 3. *Under assumptions of Theorem 1, as $n \to \infty$,*

$$\max_{1\le k\le n-1} \frac{1}{\sqrt{v(n,k)}} \left|D_k(\boldsymbol{a})\right| = \mathcal{O}_P\left(\max\left\{n^{-1/2}, n^{-\eta/(2+\eta)}\right\}(1+\log n)\right) \tag{26}$$

where

$$D_k(\boldsymbol{a}) = S_{2k}(\boldsymbol{a}) - \left(Z_{2k}(\boldsymbol{a}) - \frac{(n-k)(n-k+1)}{2n^2} Z_{1n}(\boldsymbol{a})\right).$$

Moreover, if assumption (18) *holds then*

$$\max_{G<k<n-G} \frac{1}{\sqrt{w(n,G)}} \left|D_{k+G}(\boldsymbol{a}) - 2D_k(\boldsymbol{a}) + D_{k-G}(\boldsymbol{a})\right| = o_P\left((\log n)^{-1/2}\right). \tag{27}$$

Remarks.

1. Compare our Theorem 1 to Theorem 2 of [4] and notice that mutual relationship between behaviour of maxima of cumulative and moving sums is different in our case. Also, note that a printing error occurred in Theorem 2 of [4] and condition (1.17) should accord with our condition (18). The error was dragged in [7] as well. In fact condition (18) can be slightly relaxed as follows

$$\frac{n}{G} \longrightarrow \infty \quad \text{and} \quad \frac{n^{2/(2+\eta)} \log(n/G)}{G} \longrightarrow 0.$$

However, (27) then holds with $o_P\left((\log(n/G))^{-1/2}\right)$ on the right-hand side.

2. Of course, convergence (19) and (24) also remains true if $w(n,G)$ and $\widehat{w}(n,G)$ is replaced with $(2/3)\cdot(G^3/n^2)$.

3. Corollary of Theorem 2 can be further extended to

$$\widetilde{S}_k = \sum_{i=k+1}^{n} \left(X_i - \overline{X}_n\right) \frac{i-k}{n} \tag{28}$$

where $\overline{X}_n = n^{-1}\sum_1^n X_i$. The assumption of zero mean can be relaxed in this case. The limit behaviour in question holds as well if the assumption of unit variance is relaxed and (21) or (28) is standardized by appropriate estimate of variance. See [5] for the discussion. Note in this context that the problem of estimation of mean and variance does not arise in the case of ranks because (5) and (10) are known.

4. The use of the above functionals of $S_{2k}(\boldsymbol{a})$ for testing changes in the location model as well as applicability of the limit theorems is — besides others — discussed in [7].

3. Proofs

Proof of Theorem 3. Using (12) and (13) we can write

$$D_k(\boldsymbol{a}) = -\frac{1}{n}\sum_{i=k}^{n-1} \Delta_i(\boldsymbol{a})$$

where

$$\Delta_i(\boldsymbol{a}) = S_{1i}(\boldsymbol{a}) - \left(Z_{1i}(\boldsymbol{a}) - \frac{i}{n} Z_{1n}(\boldsymbol{a})\right).$$

It follows that

$$\max_{1\le k\le n-1}\frac{1}{\sqrt{v(n,k)}}\,|D_k(\boldsymbol{a})| = \frac{1}{n}\max_{1\le k\le n-1}\frac{1}{\sqrt{v(n,k)}}\left|\sum_{i=k}^{n-1}\Delta_i(\boldsymbol{a})\right| \le \frac{1}{n}\max_{1\le k\le n-1}\sqrt{\frac{(n-k)^3}{v(n,k)}}\max_{1\le i\le n-1}\sqrt{\frac{n}{i\,(n-i)}}\,|\Delta_i(\boldsymbol{a})| \tag{29}$$

and

$$\begin{aligned}\max_{G<k<n-G}\frac{1}{\sqrt{w(n,G)}}\,|D_{k+G}(\boldsymbol{a}) - 2D_k(\boldsymbol{a}) + D_{k-G}(\boldsymbol{a})| \\ = \max_{G<k<n-G}\frac{1}{\sqrt{w(n,G)}}\,\frac{1}{n}\left|\sum_{i=k-G}^{k-1}\Delta_i(\boldsymbol{a}) - \sum_{i=k}^{k+G-1}\Delta_i(\boldsymbol{a})\right| \\ \le \frac{1}{\sqrt{w(n,G)}}\,\frac{2G}{\sqrt{n}}\max_{1\le i\le n-1}\frac{1}{\sqrt{n}}\,|\Delta_i(\boldsymbol{a})|.\end{aligned} \tag{30}$$

Theorem 3 in [4] implies that

$$\max_{1\le i\le n-1}\sqrt{\frac{n}{i\,(n-i)}}\,|\Delta_i(\boldsymbol{a})| = \mathcal{O}_P\left(\max\left\{n^{-1/2}, n^{-\eta/(2+\eta)}\right\}(1+\log n)\right) \tag{31}$$

and

$$\max_{1\le i\le n-1}\frac{1}{\sqrt{n}}\,|\Delta_i(\boldsymbol{a})| = \mathcal{O}_P\left(\max\left\{n^{-1/2}, n^{-\eta/(2+\eta)}\right\}\right). \tag{32}$$

Since

$$\frac{v(n,k)}{(n-k)^3} = \frac{1}{n^2}\left(1+\frac{1}{n-k}\right)\left(\frac{1}{3}+\frac{1}{6(n-k)}-\frac{n-k+1}{4n}\right)$$

we can easily see that $v(n,k)/(n-k)^3$ is increasing in k independently of n. It follows that $v(n,k)/(n-k)^3 = \mathcal{O}\left(n^{-2}\right)$ for arbitrary $1\le k\le n-1$ and hence

$$\max_{1\le k\le n-1}\sqrt{\frac{(n-k)^3}{v(n,k)}} = \mathcal{O}\left(n\right).$$

This along with (29) and (31) yields behaviour (26).

By condition (18) we have

$$\frac{1}{\sqrt{w(n,G)}}\,\frac{2G}{\sqrt{n}} = \mathcal{O}\left((n/G)^{1/2}\right) = o\left(n^{1/2-1/(2+\eta)}(\log n)^{-1/2}\right).$$

This along with (30) and (32) yields (27).

Proof of Theorem 2. Assertion (i) of Theorem 2 coincides with Theorem 1 in [5]. The corollary presents a result which is half a step to Theorem 2 of the paper and is proven ibidem. Theorem 2 of [5] shows the convergence for $\widetilde{S}_k$ defined in (28) and standardized by $\sqrt{v(n,k)}$.

To prove assertion (ii) assume at first that $X_1, \dots, X_n$ are normally distributed and define a zero-mean standardized Gaussian process

$$\xi_n(t) = \frac{1}{\sqrt{\widehat{w}(n,G)}}\, Y_{[G(t+1)]}, \quad 0 \le t \le \frac{n}{G} - 2,$$

where

$$(33)\quad Y_k = \widehat{S}_{k+G} - 2\widehat{S}_k + \widehat{S}_{k-G} = \sum_{i=k-G+1}^{k} X_i \frac{i-k+G}{n} + \sum_{i=k+1}^{k+G-1} X_i \frac{k+G-i}{n}.$$

For $k \ge G$ and $k + l \le n - G$ we have

$$\operatorname{Cov}(Y_k, Y_{k+l}) = \operatorname{Cov}(Y_G, Y_{G+l}) = R_{n,G}(l).$$

Hence, if $n/G \to \infty$ then $\xi_n(t)$ converges to a zero-mean standardized stationary Gaussian process $\{\xi(t), t \ge 0\}$ with autocovariance function

$$(34)\qquad \rho(t) = \lim_{n/G\to\infty} \frac{R_{n,G}([Gt])}{\widehat{w}(n,G)}\,.$$

Now investigate properties of $\rho(t)$. For $0 \le l \le G-2$ we get

$$\begin{aligned} R_{n,G}(l) &= \sum_{i=l+1}^{G} \frac{i}{n}\cdot\frac{i-l}{n} + \sum_{i=G+1}^{G+l} \frac{2G-i}{n}\cdot\frac{i-l}{n} + \sum_{i=G+l+1}^{2G-1} \frac{2G-i}{n}\cdot\frac{2G+l-i}{n} \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{G-l}(i+l)\,i + \sum_{i=1}^{l}(G-i)(i+G-l) + \sum_{i=1}^{G-l-1}(G-l-i)(G-i)\right). \end{aligned}$$

Plugged in (34) it implies

$$\begin{aligned} \rho(t) &= \frac{3}{2}\left(\int_0^{1-t}(s+t)\,s\,\mathrm{d}s + \int_0^{t}(1-s)(1+s-t)\,\mathrm{d}s + \int_0^{1-t}(1-s)(1-s-t)\,\mathrm{d}s\right) \\ &= 1 - \frac{3}{2}\,t^2 + \frac{3}{4}\,t^3 \end{aligned}$$

for $t \in \langle 0, 1)$. Further if $l \ge 2G$ then $R_{n,G}(l) = 0$ and hence $\rho(t) = 0$ for $t > 2$. Realize that

$$\max_{t\ge 0}\left\{\xi_n(t) - \xi(t)\right\} = o_P(1)$$

and apply Lemma 1 of [5] or the original Theorem 12.3.5 of [6] for $\xi(t)$ to obtain convergence (24) for normally distributed $X_1, \dots, X_n$.

By virtue of results of Einmahl [2] there are i. i. d. random variables $X_k^{(N)}$ with standard normal distribution such that

$$\max_{1\le k\le n} k^{-1/(2+\eta)} \left| \sum_{i=1}^{k} \left(X_i - X_i^{(N)}\right) \right| = \mathcal{O}_P(1).$$

Hence, by (33) and (25) and according to condition (18) we have

(35)
$$\max_{G<k<n-G} \frac{1}{\sqrt{\widehat{w}(n,G)}} \left|Y_k - Y_k^{(N)}\right| \le \frac{1}{\sqrt{\widehat{w}(n,G)}} \frac{2G}{n} \max_{1\le k\le n} \left| \sum_{i=1}^{k} \left(X_k - X_k^{(N)}\right) \right|$$
$$= \mathcal{O}_P\left(\frac{n^{1/(2+\eta)}}{\sqrt{G}}\right) = o_P\left(\left(\log\frac{n}{G}\right)^{-1/2}\right)$$

where $Y_k^{(N)}$ are obtained by replacing X_i with $X_i^{(N)}$ in (33). It concludes the proof of assertion (ii). Its extension in the corollary is straightforward since (34) and (35) hold true for $w(n,G)$ as well.

Proof of Theorem 1. According to (3), (4) and (7)

$$\frac{a(1+[nU_i]) - \overline{a}_n}{\sigma_n(\boldsymbol{a})}, \quad i = 1, \dots, n, \tag{36}$$

are i. i. d. random variables with zero mean, with essentially unit variance, and with finite absolute moment of order $2+\eta$, $\eta > 0$. Thus, by Corollary to Theorem 2 we have

$$\max_{1\le k\le n-1} \frac{1}{\sqrt{v(n,k)}} \frac{1}{\sigma_n(\boldsymbol{a})} \left|Z_{2k}(\boldsymbol{a})\right| = \mathcal{O}_P\left(\sqrt{\log\log n}\right) \tag{37}$$

and the rate in (37) cannot be improved.

It can be easily seen that $4n^4(n-k)^{-2}(n-k+1)^{-2}v(n,k)$ is decreasing for $0 < k < n$ and

$$\max_{1\le k\le n-1} \frac{1}{\sqrt{v(n,k)}} \frac{(n-k)(n-k+1)}{2n^2} = \mathcal{O}\left(n^{-1/2}\right)$$

that is attained for $k = 1$. Thus, by central limit theorem we have

$$\max_{1\le k\le n-1} \frac{1}{\sqrt{v(n,k)}} \frac{1}{\sigma_n(\boldsymbol{a})} \frac{(n-k)(n-k+1)}{2n^2} \left|Z_{1n}(\boldsymbol{a})\right| = \mathcal{O}_P(1).$$

This along with (37) implies that

$$\max_{1\le k\le n-1} \frac{1}{\sqrt{v(n,k)}} \frac{1}{\sigma_n(\boldsymbol{a})} \left| Z_{2k}(\boldsymbol{a}) - \frac{(n-k)(n-k+1)}{2n^2} Z_{1n}(\boldsymbol{a}) \right| = \mathcal{O}_P\left(\sqrt{\log\log n}\right) \tag{38}$$

and properly standardized l. h. s. in (37) has the same limit distribution as the standardized l. h. s. in (38). Now apply Theorem 3 to show convergence (17). Proof of convergence (19) is analogous.

References

[1] Csörgő M., Horváth L., *Limit Theorems in Change-Point Analysis*, John Wiley & Sons, 1997.

[2] Einmahl U., *A useful estimate in the multidimensional invariance principle*, Probab. Theory Related Fields **76** (1987), 81–101.

[3] Hájek J., Šidák Z., *Theory of Rank Tests*, Academia, Praha, 1967.

[4] Hušková M., *Limit theorems for rank statistics*, Statist. Probab. Lett. **32** (1997), 45–55.

[5] Jarušková D., *Testing appearance of linear trend*, J. Statist. Plann. Inference **70** (1998), 263–276.

[6] Leadbetter M.R., Lindgren G., Rootzén H., *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, Heidelberg, 1983.

[7] Slabý A., *Behaviour of Some Rank Statistics for Detecting Changes*, in Measuring Risk in Complex Stochastic Systems (proc. conf.), Berlin, 1999 (Franke J., Härdle W., Stahl G., Eds.), Springer-Verlag, Berlin, 2000, pp. 161–172.

Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Prague 8, Czech Republic

(*Received* May 5, 2000, *revised* January 19, 2001)