Locally Most Powerful Tests Based on Sequential Ranks

Jan Kalina

Center of J. Hájek for Theoretical and Applied Statistics,
KPMS MFF UK,
Sokolovská 83,
186 75 Praha 8, Czech Republic
kalina@karlin.mff.cuni.cz

Abstract. Sequential ranks of data $X_1, X_2, \ldots$ observed sequentially in time are defined as ranks computed from the data observed so far, denoting R_{ii} the rank of X_i among the values $X_1, X_2, \ldots, X_i$ for any i.

This paper studies tests of various hypotheses based on sequential ranks and derives such tests, which are locally most powerful among all tests based on sequential ranks. Such locally most powerful sequential rank test is derived for the hypothesis of randomness against a general alternative, including the regression in location as a special case for the alternative hypothesis. Further, the locally most powerful sequential rank tests are derived for independence of two samples.

The new tests are suitable for the situation when data are observed sequentially in time and the test is carried out each time after obtaining a new observation. While the classical rank tests require to recalculate all values of the ranks each time, the methods based on sequential ranks only require to compute the sequential rank of the only one new observation.

Keywords. Hypotheses tests, sequential analysis, contiguous alternatives.

1 Introduction

Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ represent a random vector with values in $(\mathbb{R}^n, \mathcal{B}^n)$, where $\mathcal{B}$ is the system of Borel sets on $\mathbb{R}$. Arranging $\mathbf{X}$ in ascending order we obtain the vector of order statistics

$$X^n_{(1)} \leq X^n_{(2)} \leq \cdots \leq X^n_{(n)}, \tag{1}$$

where the upper index stresses that these are computed from n random variables. For observed values $(x_1, \ldots, x_n)$ such that no two observations are equal, the rank R_i of the i-th observation is defined by $X_i = X^n_{(R_i)}$ for $i = 1, \ldots, n$ and the vector of ranks will be denoted by $\mathbf{R} = (R_1, \ldots, R_n)^T$.

Sequential ranks $\mathbf{R}^* = (R_{11}, \ldots, R_{nn})^T$ are defined as ranks computed from the data observed so far, denoting R_{kk} the rank of X_k among the values $X_1, \ldots, X_k$ for any $k = 1, \ldots, n$. It holds that $X_i = X^i_{(R_{ii})}$ for any $i = 1, \ldots, n$, in other words R_{ii} denotes the number of values among $r_{11}, \ldots, r_{nn}$ smaller or equal to x_i for $i = 1, \ldots, n$.

The classical theory of rank tests was constructed by Hájek and Šidák (1967). In this paper we study hypotheses tests based on sequential ranks for various situations, which are shown to be locally most powerful among all tests based on sequential ranks. These tests will be called locally most powerful sequential ranks tests.

Mason (1981) considered a linear statistic based on sequential ranks in the form

$$M_n = \sum_{i=1}^{n} (c_{in} - \bar{c}_{i-1,n}) J_n \left(\frac{R_{ii}}{i+1} \right), \tag{2}$$

where

$$\sum_{i=1}^{n} c_{in} = 0 \quad \text{and} \quad \bar{c}_{i-1,n} = \sum_{j=1}^{i-1} \frac{c_{jn}}{i-1}$$

and the scores $J_n(i/(n+1))$ for $i = 1, \ldots, n$ are computed as $J_n(i/(n+1)) = \mathsf{E} J(U^n_{(i)})$ from the generating function J satisfying

$$\int_0^1 J(u)du = 0 \quad \text{and} \quad 0 < \int_0^1 J^2(u)du < \infty.$$

Mason studied M_n for $n \to \infty$ and proved that M_n and the simple linear rank statistic (based on classical ranks) are asymptotically equivalent in the quadratic mean. This asymptotic equivalence is valid under the null hypothesis that $X_1, \ldots, X_n$ are independent identically distributed. Mason followed Hájek and Šidák (1967) to state that they are asymptotically equivalent also under contiguous alternatives of regression in location. Further he studied theoretical properties of M_n, which are based on the independence of sequential ranks, and applied them to study limit theorems for the simple linear rank statistic.

Let us consider a special case of a two-sample problem. Let us say that the first sample $X_1, \ldots, X_m$ is observed and then the second sample $Y_1, \ldots, Y_n$ is observed and let us denote $N = m+n$. Then (2) has the complicated form

$$-\frac{n}{N} J_N\left(\frac{R_{11}}{2}\right) + J_N\left(\frac{R_{m+1,m+1}}{m+2}\right) + \frac{m}{m+1} J_N\left(\frac{R_{m+2,m+2}}{m+3}\right) + \cdots + \frac{m}{N-1} J_N\left(\frac{R_{NN}}{N+1}\right).$$

Standard monographs on sequential nonparametrics Sen (1981) or Gosh and Sen (1991) study sequential hypotheses tests based on statistics computed from classical ranks. These are recalculated after each new observation is added. Hypothesis tests based on sequential ranks are not studied in these monographs. Mason's result are proven even without the local asymptotic normality. A possible alternative approach would be to study the local asymptotic normality for $n \to \infty$. Based on Le Cam's theory, if the local asymptotic normality is true under the null hypothesis, then it is valid also under contiguous alternatives.

In this paper we study locally most powerful tests based on sequential ranks. Section 2 presents useful properties of sequential ranks. The paper considers the test of the null hypothesis of randomness against a general alternative in Section 3 and against regression in location in Section 4. Section 5 derives the locally most powerful sequential rank test for independence of two samples. These tests correspond to intuition and are convenient for computation.

2 Some Properties of Sequential Ranks

The mean and variance of sequential ranks are equal to

$$\mathsf{E} R_{ii} = \frac{i+1}{2}, \quad \mathsf{var}\, R_{ii} = \frac{i^2-1}{12}.$$

These are actually the mean and variance of classical ranks computed from $X_1, \ldots, X_i$ for any i. It follows from Barndorff-Nielsen (1963) that for any $i \neq j$ it holds

$$\mathsf{cov}\,(R_{ii}, R_{jj}) = 0.$$

The computational complexity of computing the classical ranks from data $X_1, \ldots, X_n$ has the order $\mathcal{O}(n \log n)$. To compute the sequential ranks $R_{11}, \ldots, R_{nn}$ has the computational complexity also $\mathcal{O}(n \log n)$, because computing R_{ii} has the complexity of order $\mathcal{O}(\log i)$ for any i.

Actually there exists a one-to-one mapping between the vector of ranks and the vector of sequential ranks for any fixed $n < \infty$. We describe the **algorithm** to find the original data $X_1, X_2, \ldots, X_n$ in the correct order based on arranged values (1) and the observed values of the sequential ranks $R_{11} = r_{11}, \ldots, R_{nn} = r_{nn}$.

1. Initialize $(s_1, \ldots, s_n)^T$ as $(s_1, \ldots, s_n)^T := (r_{11}, \ldots, r_{nn})^T$.
2. For $t \in \{1, \ldots, n\}$ perform the following:
   (a) $p := \max\{j;\ s_j = \min\{s_1, \ldots, s_n\}\}$;
   (b) $X_p := X_{(t)}$;
   (c) $s_p := \infty;\ s_{p+1} := s_{p+1} - 1;\ \ldots;\ s_n := s_n - 1.$

In this notation we put $\infty - 1 := \infty$.

3 Test Against a General Alternative

Let us assume the i.i.d. random variables $X_1, X_2, \ldots$ to be observed sequentially. Let X_1 have a density with respect to the Lebesgue measure, so that for any $n < \infty$ there is a zero probability of two observations attaining the same value. Let $c_1, c_2, \ldots$ denote a known sequence of regression constants. The joint density of the random vector $(X_1, \ldots, X_n)^T$ with fixed n under the null hypothesis will be denoted by $p(x_1, \ldots, x_n)$ and under the alternative hypothesis depending on a parameter $\Delta > 0$ by $q^n_\Delta(x_1, \ldots, x_n)$.

A family of densities $d(x, \theta)$ with values of θ in an open interval containing 0 will be considered with the assumptions $II.4.8.A_1$ of Hájek and Šidák (1967). These ensure that $\dot{d}(x, \theta)$ denoting the partial derivative with respect to θ exists for almost every θ at every point x such that $d(x, \theta)$ is absolutely continuous in θ.

For a fixed $n < \infty$, the null hypothesis of interest

$$H_0 : p(x_1, \ldots, x_n) = \prod_{i=1}^{n} d(x_i, 0)$$

will be tested against the alternative formulated in a general way as

$$H_1 : q^n_\Delta(x_1, \ldots, x_n) = \prod_{i=1}^{n} d(x_i, \Delta c_i), \quad \Delta > 0. \tag{3}$$

The null hypothesis can be also formulated as $H_0 : \Delta = 0$. We now describe the locally most powerful sequential rank test of H_0 against the general alternative H_1, which can be described by steps analogous with Hájek and Šidák (1967). A special case for testing H_0 against regression in location will be formulated in the next section.

Theorem 1. *Let the condition II.4.8.A_1 of Hájek and Šidák (1967) be fulfilled. Then the test with the critical region*

$$\sum_{i=1}^{n} c_i \mathsf{E} \frac{\dot{d}(X^i_{(R_{ii})}, 0)}{d(X^i_{(R_{ii})}, 0)} \geq k_\alpha \tag{4}$$

is the locally most powerful sequential rank test for H_0 against (3) at level α.

4 Test of H_0 Against Regression in Location

The alternative hypothesis of regression in location is a special case of (3) in the form

$$H_1 : q^n_\Delta = \prod_{i=1}^{n} f(x_i - \Delta c_i), \quad \Delta > 0. \tag{5}$$

Let us define

$$F(x) = \int_{-\infty}^{x} f(y) dy,$$

its inverse $F^{-1}(u) = \inf\{x;\ F(x) \geq u\}$ and the score function φ by

$$\varphi(u, f) = \frac{f'(F^{-1}(u))}{f(F^{-1}(u))}, \quad 0 < u < 1. \tag{6}$$

Let us assume a random sample of the total number of n uniform random variables $U^n_1, \ldots, U^n_n$ and denote their i-th order statistic by $U^n_{(i)}$.

Let us assume f to be absolutely continuous with

$$\int_{-\infty}^{\infty} |f'(x)| dx < \infty \tag{7}$$

to have well-defined scores for a fixed $n < \infty$

$$a_n(i,f) = \mathsf{E}\varphi(U_{(i)}^n, f) = \mathsf{E}\left[-\frac{f'}{f}(F^{-1}(U_{(i)}^n))\right] = \mathsf{E}\left[-\frac{f'}{f}(X_{(i)}^n)\right], \quad i = 1,\ldots,n. \tag{8}$$

Theorem 2. *The locally most powerful sequential rank test of H_0 against the system $\{q_\Delta^n,\ \Delta > 0\}$ with q_Δ^n defined by (5) at level α has the critical region*

$$\sum_{i=1}^{n} c_i a_i(R_{ii}, f) \geq k_\alpha$$

assuming (7).

5 Test of Independence

Let the i.i.d. random variables $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are observed sequentially in such a way that the pair $(X_i, Y_i)^T$ is observed at the same time for $i = 1, \ldots, n$. We introduce the notation $\mathbf{R}^* = (R_{11}, \ldots, R_{nn})^T$ and $\mathbf{Q}^* = (Q_{11}, \ldots, Q_{nn})^T$ for sequential ranks in the first and second sample, respectively.

The null hypothesis

$$H_0^* : p(x_1, y_1, \ldots, x_n, y_n) = \prod_{i=1}^{n} f(x_i) g(y_i)$$

will be tested against the alternative that $X_i = X_i^* + \Delta Z_i$ and $Y_i = Y_i + \Delta Z_i$, where X_i^*, Y_i^* and Z_i^* are mutually independent, and X_i^* and Y_i^* have a specified distribution. This alternative can be formally expressed as

$$H_1 : q_\Delta^n(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^{n} h_\Delta(x_i, y_i), \quad \Delta \in \mathbb{R}, \tag{9}$$

where

$$h_\Delta(x,y) = \int_{-\infty}^{\infty} f(x - \Delta z) g(y - \Delta z) dM(z)$$

with an arbitrary distribution function M. The score function (6) and scores (8) are used also in this context.

Theorem 3. *The locally most powerful sequential rank test of H_0^* against the system $\{q_\Delta^n,\ \Delta > 0\}$ with q_Δ^n defined by (9) at level α has the critical region*

$$\sum_{i=1}^{n} a_i(R_{ii}, f)\, a_i(Q_{ii}, g) \geq k_\alpha. \tag{10}$$

Proof. Under the alternative with a fixed Δ, Q^n_Δ will denote the probability distribution corresponding to the density (9). Starting to express the most powerful sequential rank test based on Neyman-Pearson lemma, we obtain

$$\lim_{\Delta\to 0}\frac{1}{\Delta^2}\left[(n!)^2 Q^n_\Delta(R_{11},\dots,R_{nn},Q_{11},\dots,Q_{nn})-1\right]= \tag{11}$$

$$=\sigma^2(n!)^2\sum_{i=1}^n \int_{\mathbf{R}^*=\mathbf{r}^*,\ \mathbf{Q}^*=\mathbf{q}^*}\!\!\cdots\int \left[\frac{f'(x_i)g'(y_i)}{f(x_i)g(y_i)}\prod_{k=1}^n f(x_k)g(y_k)\right]dx_1\cdots dx_n dy_1\cdots dy_n,$$

where σ^2 is a constant corresponding to the arbitrary variables $Z_1,\dots,Z_n$ and $\mathbf{r}^*=(r_{11},\dots,r_{nn})^T$ and $\mathbf{q}^*=(q_{11},\dots,q_{nn})^T$ are observed values of the sequential ranks of the two samples.

Further we will express the conditional expectation of a measurable function t of random variables $X_1,\dots,X_n$ conditional on the sequential ranks $\mathbf{r}^*=(r_{11},\dots,r_{nn})^T$ as

$$\mathsf{E}\left[t(X_1,\dots,X_n)|R_{11}=r_{11},\dots,R_{nn}=r_{nn}\right]=\mathsf{E}t(X^1_{(R_{11})},\dots,X^n_{(R_{nn})}).$$

Modifying further steps of Hájek and Šidák (1967), we express (11) as

$$\sigma^2\sum_{i=1}^n\left[n!\int_{\mathbf{R}^*=\mathbf{r}^*}\!\!\cdots\int\frac{f'(x_i)}{f(x_i)}\prod_{k=1}^n f(x_k)dx_1\cdots dx_n\right]\left[n!\int_{\mathbf{Q}^*=\mathbf{q}^*}\!\!\cdots\int\frac{g'(y_i)}{g(y_i)}\prod_{k=1}^n g(y_k)dy_1\cdots dy_n\right]=$$

$$=\sigma^2\sum_{i=1}^n\mathsf{E}\left[\frac{f'}{f}(X^i_{(R_{ii})})\,|R_{11},\dots,R_{nn}\right]\mathsf{E}\left[\frac{g'}{g}(Y^i_{(Q_{ii})})\,|Q_{11},\dots,Q_{nn}\right]=$$

$$=\sigma^2\sum_{i=1}^n\mathsf{E}\left[-\frac{f'}{f}(F^{-1}(U^i_{(R_{ii})}))\right]\mathsf{E}\left[-\frac{g'}{g}(G^{-1}(U^i_{(Q_{ii})}))\right],$$

where F and G are distribution functions corresponding to densities f and g, respectively. This test statistic is further equivalent to

$$\sum_{i=1}^n\mathsf{E}\varphi(U^i_{(R_{ii})},f)\,\mathsf{E}\varphi(U^i_{(Q_{ii})},g).$$

This equals the test sequential-rank statistic (10), which arranges the sequential ranks $r_{11},\dots,r_{nn}$ in the same way as the most powerful test among all tests based on sequential ranks. This concludes the proof.

Example 1. As a special case we can express the test statistic of (10) for the linear score function φ; such test statistic

$$\sum_{i=1}^n\frac{R_{ii}Q_{ii}}{(i+1)^2}$$

can be interpreted as the analogy of Spearman correlation coefficient based on sequential ranks.

Acknowledgements

This work is supported by Center of J. Hájek for Theoretical and Applied Statistics, project LC06024 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

Barndorff-Nielsen, O. (1963): On the limit behaviour of extreme order statistics. *Annals of Mathematical Statistics*, **34**, 992–1002.

Gosh, B.K. and Sen, P.K. (1991): *Handbook of sequential analysis.* Marcel Dekker, Inc., New York.

Hájek, J. and Šidák, Z. (1967): Theory of rank tests. Academia, Prague.

Hájek, J., Šidák, Z., Sen, P.K. (1999): *Theory of rank tests. Second edition.* Academic Press, San Diego.

Mason, D.M. (1981): On the use of a statistic based on sequential ranks to prove limit theorems for simple linear rank statistics. *Annals of Statistics*, **9**, 424–436.

Sen, P.K. (1981): *Sequential nonparametrics: Invariance principles and statistical inference.* John Wiley & Sons, Inc., New York.