Janusz L. Wywiał

University of Economics in Katowice

ON PREDICTION OF TIME SERIES ON THE BASIS OF RANK CORRELATION COEFFICIENTS

1. Basic properties of rank correlation coefficient

Let ranks of the series of observations of a time series be denoted by: $a_{k,t}$, $t = 1, \dots, k$. Let $l_{k,t}$ be number of the implication of the type: $a_{k,t} > a_{k,h}$, if $t > h$, where $t = 2, \dots, k$, $h = 1, \dots, t-1$. Let $l'_{k,t}$ be number of the following implication: $a_{k,t} < a_{k,h}$, if $t > h$, where $t = 2, \dots, k$, $h = 1, \dots, t-1$. The Kendall's (1958) rank correlation coefficient is as follows:

$$Q_k = \frac{2\left(L_k - L'_k\right)}{k(k-1)}, \tag{1}$$

where:

$$L_k = \sum_{t=1}^{k} l_{k,t}\,, \qquad L'_k = \sum_{t=1}^{k} l'_{k,t}\,.$$

Moreover, the statistic Q_k can be rewritten as follows, see Höffding (1947):

$$Q_k = 2G_k - 1 \tag{2}$$

where

$$G_k = \frac{2L_k}{k(k-1)}, \qquad 0 \le G_k \le 1.$$

The values of the statistics L_k, $L_k^{'}$ and Q_k will be denoted by l_k, $l_k^{'}$ and q_k, respectively. When all permutations $(a_{k,1}, a_{k,2}, \ldots, a_{k,k})$ are equally probable than $E(Q_4) = \rho_4 = 0$. In this case the expression (3) show the distribution of the rank correlation coefficient under the assumption that $k = 4$.

Now let us assume that the permutation $(a_{4,1}, a_{4,2}, a_{4,3}, a_{4,4})$ is chosen with probability proportional to the values $(1 + l_4)^3$. In this case the probabilities $P(a_{4,1}, \ldots, a_{4,4})$ are given in the ninth column of Table 1. This leads to the distribution of the rank correlation given by the expression (4) and $E(Q_4) = \rho_4 = 0.374$ and $D(Q_4) = 0.506$. Let us note that more about the conception of determining distribution of the statistics Q_k and L_k can be found in the paper by Höffding (1947).

Table 1

Rank permutation and distributions of the rank correlation coefficient

No.	$a_{4,1}, \ldots, a_{4,4}$	$a_{3,1}, \ldots, a_{3,3}$	$a_{2,1}, a_{2,2}$	l_4	$l_4^{'}$	$l_4 - l_4^{'}$	q_4	$P(a_{4,1}, \ldots, a_{4,4})$
1	1,2,3,4	1,2,3	1,2	6	0	6	1	0,1574
2	1,2,4,3	1,2,3	1,2	5	1	4	2/3	0,0991
3	1,3,2,4	1,3,2	1,2	5	1	4	2/3	0,0991
4	1,3,4,2	1,2,3	1,2	4	2	2	1/3	0,0574
5	1,4,2,3	1,3,2	1,2	4	2	2	1/3	0,0574
6	1,4,3,2	1,3,2	1,2	3	3	0	0	0,0294
7	2,1,3,4	2,1,3	2,1	5	1	4	2/3	0,0991
8	2,1,4,3	2,1,3	2,1	4	2	2	1/3	0,0574
9	2,3,1,4	2,3,1	2,1	4	2	2	1/3	0,0574
10	2,3,4,1	1,2,3	1,2	3	3	0	0	0,0294
11	2,4,1,3	2,3,1	1,2	3	3	0	0	0,0294
12	2,4,3,1	1,3,2	1,2	2	4	−2	−1/3	0,0124
13	3,1,2,4	3,1,2	2,1	4	2	2	1/3	0,0574
14	3,1,4,2	2,1,3	2,1	3	3	0	0	0,0294
15	3,2,1,4	3,2,1	2,1	3	3	0	0	0,0294
16	3,2,4,1	2,1,3	2,1	2	4	−2	−1/3	0,0124
17	3,4,1,2	2,3,1	1,2	2	4	−2	−1/3	0,0124
18	3,4,2,1	2,3,1	1,2	2	4	−2	−1/3	0,0124
19	4,1,2,3	3,1,2	2,1	3	3	0	0	0,0294
20	4,1,3,2	3,1,2	2,1	2	4	−2	−1/3	0,0124
21	4,2,1,3	3,2,1	2,1	2	4	−2	−1/3	0,0124
22	4,2,3,1	3,1,2	2,1	1	5	−4	−2/3	0,0037
23	4,3,1,2	3,2,1	2,1	1	5	−4	−2/3	0,0037
24	4,3,2,1	3,2,1	2,1	0	6	−6	−1	0,0005

Source: Own calculations.

$$P(Q_4 = q \mid \rho_4 = 0) = \begin{cases} 1/24 & dla \quad q = 1 \\ 1/8 & dla \quad q = 2/3 \\ 5/24 & dla \quad q = 1/3 \\ 1/4 & dla \quad q = 0 \\ 5/24 & dla \quad q = -1/3 \\ 1/8 & dla \quad q = -2/3 \\ 1/24 & dla \quad q = -1 \end{cases}, \tag{3}$$

$$P(Q_4 = q \mid \rho_4 = 0.374) = \begin{cases} 0{,}157 & dla \quad q = 1 \\ 0{,}298 & dla \quad q = 2/3 \\ 0{,}287 & dla \quad q = 1/3 \\ 0{,}176 & dla \quad q = 0 \\ 0{,}074 & dla \quad q = -1/3 \\ 0{,}007 & dla \quad q = -2/3 \\ 0{,}001 & dla \quad q = -1 \end{cases}, \tag{4}$$

2. Rank prediction

Let us start consideration with the following example.

Example 1. On the basis of the lines no. 15, 21,23 and 24 of Table 1 we calculate the following probability:

$$P(a_{3,1} = 3, a_{3,2} = 2, a_{3,3} = 1, \rho_4 = 0) = P(a_{4,1} = 3, a_{4,2} = 2, a_{4,3} = 1, a_{4,4} = 4, \rho_4 = 0) + P(a_{4,1} = 4, a_{4,2} = 2, a_{4,3} = 1, a_{4,4} = 3, \rho_4 = 0) + P(a_{4,1} = 4, a_{4,2} = 3, a_{4,3} = 1, a_{4,4} = 2, \rho_4 = 0) + P(a_{4,1} = 4, a_{4,2} = 3, a_{4,3} = 2, a_{4,4} = 1, \rho_4 = 0) = 4/24 = 1/6.$$

For instance:

$$P(a_{4,1} = 4 \mid a_{3,1} = 3, a_{3,2} = 2, a_{3,3} = 1, \rho_4 = 0) = \frac{P(a_{4,1} = 4, a_{4,2} = 3, a_{4,3} = 2, a_{4,4} = 1, \rho_4 = 0)}{P(a_{3,1} = 3, a_{3,2} = 2, a_{3,3} = 1, \rho_4 = 0)} = \frac{1/24}{1/6} = \frac{1}{4}.$$

So,

$$P(a_{4,1} = a \mid a_{3,1} = 3, a_{3,2} = 2, a_{3,3} = 1, \rho_4 = 0) = \frac{1}{4} \text{ for a = 1,2,3,4.}$$

Now on the basis of the line no. 11 of Table 1 we have:

$$P(a_{3,1}=2,\ a_{3,2}=3,\ a_{3,3}=1,\ Q_4=0,\ \rho_4=0) =$$
$$= P(a_{4,1}=2,\ a_{4,2}=4,\ a_{4,3}=1,\ a_{4,4}=3,\ Q_4=0,\ \rho_4=0) = 1/24.$$

Hence:

$$P(a_{4,4}=3|a_{3,1}=2, a_{3,2}=3,\ a_{3,3}=1,\ Q_4=0,\ \rho_4=0) =$$
$$= \frac{P\left(a_{4,1}=2, a_{4,2}=4, a_{4,3}=1, a_{4,4}=3, Q_4=0, \rho_4=0\right)}{P\left(a_{3,1}=2, a_{3,2}=3, a_{3,3}=1, Q_4=0, \rho_4=0\right)} = \frac{1/24}{1/24} = 1.$$

So, it means that $a_{4,4}=3$ under the condition that $a_{3,1}=2$, $a_{3,2}=3$, $a_{3,3}=1$, $\rho_4=0$ and ,$Q_4=0$ with probability one. In this case let us suppose that in the periods $t=1,2,3$ the observations of the time series are $y_1=10.1$, $y_2=10.8$ and $y_3=9.2$. The ranks of this values are $(2,3,1)$. According to the obtained results the ranks of all four values of the time series are the elements of the sequence: $(a_{4,1},a_{4,2},a_{4,3},a_{4,4}) = (2,4,1,3)$. Hence, in the fourth period the predicted value of the time series has the rank 3. This leads to the conclusion that the predicted value of the time series in the fourth period is between values $y_1=10.1$ and $y_2=10.8$. Hence $10.1 < y_4 < 10.8$.

Example 2. Similarly like in the previous example the following probabilities can be calculated on the basis of the lines no. 15 and 21 of Table 1:

$$P\left(a_{4,4}=a \mid a_{3,1}=3, a_{3,2}=2, a_{3,3}=1, |Q_4| \le \frac{1}{3}, \rho_4=0\right) = \begin{cases} 0.5 & for \quad a=3 \\ 0.5 & for \quad a=4 \end{cases}$$

because

$$P\left(a_{4,4}=4 \mid a_{3,1}=3, a_{3,2}=2, a_{3,3}=1, |Q_4| \le \frac{1}{3}, \rho_4=0\right) =$$
$$= \frac{P\left(a_{4,1}=3, a_{4,2}=2, a_{4,3}=1, a_{4,4}=4, |Q_4|=\frac{1}{3}, \rho_4=0\right)}{P\left(a_{3,1}=3, a_{3,2}=2, a_{3,3}=1, |Q_4| \le \frac{1}{3}, \rho_4=0\right)} = \frac{1/24}{2/24} = \frac{1}{2}$$

$$P\left(a_{4,4}=3 \mid a_{3,1}=3, a_{3,2}=2, a_{3,3}=1, |Q_4| \le \frac{1}{3}, \rho_4=0\right) =$$
$$= \frac{P\left(a_{4,1}=4, a_{4,2}=2, a_{4,3}=1, a_{4,4}=3, |Q_4| \le \frac{1}{3}, \rho_4=0\right)}{P\left(a_{3,1}=3, a_{3,2}=2, a_{3,3}=1, |Q_4| \le \frac{1}{3}, \rho_4=0\right)} = \frac{1/24}{2/24} = \frac{1}{2}$$

The obtained result can be interpreted as follows. Under the assumptions that $\rho_4 = 0$ and $|Q_4| \le 1/3$ and $a_{3,1} = 3$, $a_{3,2} = 2$, $a_{3,3} = 1$, the probability that forth rank is equal to *3 (4)* is equal to *0.5*.

The generalization of the derived in Examples 1 and 2 results is as follows.

$$P\left(a_{k,k} = b \mid a_{k-1,1} = a_1, a_{k-1,2} = a_2, \ldots, a_{k-1,k-1} = a_{k-1}, |Q_k| \leqslant q_0, \rho_k = \rho_0\right) =$$
$$= \frac{P\left(a_{k,k} = b_k, a_{k,k-1} = b_{k-1}, a_{k-1,k-2} = b_{k-2}, \ldots, a_{k,1} = b_1, |Q_k| \leqslant q_0, \rho_k = \rho_0\right)}{P\left(a_{k-1,1} = a_1, a_{k-1,2} = a_2, \ldots, a_{k-1,k-1} = a_{k-1}, |Q_k| \leqslant q_0, \rho_k = \rho_0\right)} \tag{5}$$

Example 3. Now, let us consider the problem of the prediction of two ranks. For instance on the basis of the lines no. 6, 10 and 11 of Table 1 we have:

$$P\left(a_{4,3} = a, a_{4,4} = b \mid a_{2,1} = 1, a_{2,2} = 2, Q_4 = 0, \rho_4 = 0\right) = \begin{cases} 1/3 & for \quad a = 3, b = 2 \\ 1/3 & for \quad a = 4, b = 1 \\ 1/3 & for \quad a = 1, b = 3 \end{cases}$$

because:

$$P\left(a_{4,3} = 3, a_{4,4} = 2 \mid a_{2,1} = 1, a_{2,2} = 2, Q_4 = 0, \rho_4 = 0\right) =$$
$$= \frac{P\left(a_{4,1} = 1, a_{4,2} = 4, a_{4,3} = 3, a_{4,4} = 2, Q_4 = 0, \rho_4 = 0\right)}{P\left(a_{2,1} = 1, a_{2,2} = 2, Q_4 = 0, \rho_4 = 0\right)} = \frac{1/24}{3/24} = \frac{1}{3}$$
$$P\left(a_{4,3} = 4, a_{4,4} = 1 \mid a_{2,1} = 1, a_{2,2} = 2, Q_4 = 0, \rho_4 = 0\right) =$$
$$= \frac{P\left(a_{4,1} = 2, a_{4,2} = 3, a_{4,3} = 4, a_{4,4} = 1, Q_4 = 0, \rho_4 = 0\right)}{P\left(a_{2,1} = 1, a_{2,2} = 2, Q_4 = 0, \rho_4 = 0\right)} = \frac{1/24}{3/24} = \frac{1}{3}$$
$$P\left(a_{4,3} = 1, a_{4,4} = 3 \mid a_{2,1} = 1, a_{2,2} = 2, Q_4 = 0, \rho_4 = 0\right) =$$
$$= \frac{P\left(a_{4,1} = 2, a_{4,2} = 4, a_{4,3} = 1, a_{4,4} = 3, Q_4 = 0, \rho_4 = 0\right)}{P\left(a_{2,1} = 1, a_{2,2} = 2, Q_4 = 0, \rho_4 = 0\right)} = \frac{1/24}{1/8} = \frac{1}{3}$$

Similarly, like in Example 1 let us assume that $y_1 = 10.1$ and $y_2 = 10.8$. On the basis of the above results we can write that $10.1 < y_4 < y_3 < 10.8$ with probability 1/3, $y_3 < 10.1 < y_4 < 10.8$ with probability 1/3, $y_4 < 10.1$ and $y_3 > 10.8$ with probability 1/3.

The obtained result can be straightforward generalized into the case of the prediction of *m*-ranks provided that the k-ranks and values of Q_{k+m} and ρ_{k+m} are fixed as follows:

$$P\left(a_{k,m+1} = b_{m+1}; a_{k,m+2} = b_{m+2}; \ldots; a_{k,n} = b_n \mid a_{m,1} = a_1; a_{m,2} = a_2; \ldots\right.$$
$$\left. \ldots; a_{m,m} = a_m; |Q_{k+m}| \leqslant q; \rho_{k+m} = \rho_0\right) =$$
$$= \frac{P\left(a_{k,1} = b_1; \ldots; a_{k,m+1} = b_{m+1}; \ldots; a_{k,k} = b_k; |Q_{k+m}| \leqslant q_0; \rho_{k+m} = \rho_0\right)}{P\left(a_{m,1} = a_1; a_{m,2} = a_2; \ldots; a_{m,m} = a_m; |Q_{k+m}| \leqslant q_0; \rho_{k+m} = \rho_0\right)} \tag{6}$$

Example 4. Similarly, like in the example 2 the expressions (4), (5) and the lines no. 15, 21 of Table 1 lead to the following:

$$P\left(a_{4,4}=b \mid a_{3,1}=3, a_{3,2}=2, a_{3,3}=1, \mid Q_4 \mid \leqslant 0.4, \rho_4=0.374\right)=\begin{array}{lll} 0.703 & for & b=3 \\ 0.297 & for & b=4 \end{array}$$

because

$$\begin{aligned} &P\,(a_{3,1}=3,\ a_{3,2}=2,\ a_{3,3}=1,\ |Q_4| \le 0.4,\ \rho_4=0.374) = \\ &\quad = P\,(a_{4,1}=3,\ a_{4,2}=2,\ a_{4,3}=1,\ a_{4,4}=4,\ |Q_4| \le 0.4,\ \rho_4=0.374) + \\ &\quad + P\,(a_{4,1}=4,\ a_{4,2}=2,\ a_{4,3}=1,\ a_{4,4}=3,\ |Q_4| \le 0.4,\ \rho_4=0.374) = \\ &\quad = 0.0124 + 0.0294 = 0.0418, \end{aligned}$$

$$P\left(a_{4,4}=4 \mid a_{3,1}=3, a_{3,2}=2, a_{3,3}=1, \mid Q_4 \mid \leqslant 0.4, \rho_4=0.374\right)=$$

$$=\frac{P\left(a_{4,1}=3, a_{4,2}=2, a_{4,3}=1, a_{4,4}=4, \left|Q_4\right| \leqslant 0.4, \rho_4=0.374\right)}{P\left(a_{3,1}=3, a_{3,2}=2, a_{3,3}=1, \mid Q_4 \mid \leqslant 0.4, \rho_4=0.374\right)}=0.703\,.$$

Hence, under the assumptions that *(3,2,1)* are the ranks of the time series in three periods and $\rho_4 = 0.374$ and $|Q_4| \le 0.4$, the probability that rank of the time series in the fourth period is equal to *3* (*4*) is equal to *0.703* (*0.297*).

Let us assume that $y_1 = 5.8$, $y_2 = 5.1$, $y_3 = 4.9$. The above results let us infer that $y_4 > 5.8$ with probability 0.703 else $5.2 < y_4 < 5.8$ with probability 0.297.

Example 5. Let *(1,2)* are the ranks of the time series in two periods. We are going to predict the ranks of the time series in next two periods under the assumptions that $Q_4 > 0.5$ and $\rho_4 = 0.374$. Under these assumptions, the expression (6) and the lines no. 1, 2, 3 of Table 1 lead to the following:

$$P\left(a_{4,3}=a, a_{4,4}=b \mid a_{2,1}=1, a_{2,2}=2, Q_4>0.5, \rho_4=0.374\right)=$$

$$=\begin{cases} 0.279 & for \quad a=2 \quad b=4 \\ 0.442 & for \quad a=3 \quad b=4 \\ 0.279 & for \quad a=4 \quad b=3 \end{cases}$$

because for instance

$$P\left(a_{4,3}=2, a_{4,4}=4 \mid a_{2,1}=1, a_{2,2}=2, Q_4>0.5, \rho_4=0.374\right)=$$

$$=\frac{P\left(a_{4,1}=1, a_{4,2}=3, a_{4,3}=2, a_{4,4}=4, Q_4>0.5, \rho_4=0.374\right)}{P\left(a_{2,1}=1, a_{2,2}=2, Q_4>0.5, \rho_4=0.374\right)}=\frac{0.0991}{0.3556}=0.279\,.$$

Hence, if in the first two periods the time series increases then in two next periods the ranks of the time series are *2* and *4* with probability *0.279* or 4 and 3 with probability *0.279* or *3* and 4 with probability *0.442.* Hence, the ranks of the time series can be determined by the sequence *(1,3,2,4)* with probability *0.279* or

the sequence *(1,2,4,3)* with probability *0.279* or the sequence *(1,2,3,4)* with probability *0.442* provided that the distribution of the Kendall's rank coefficient Q_4 is defined by the expression (4) and $Q_4 > 0.5$.

Hence, when $y_1 = 5.1$ and $y_2 = 6.2$ are observed then $5.1 < y_3 < 6.2 < y_4$ with probability 0.279, $y_3 > y_4 > 6.2$ with probability 0.279 and $y_4 > y_3 > 6.2$ with probability 0.442.

The rank coefficient distribution Q_k can be estimated on the basis of a time series observations. In order to do it the time series representing by the sequence: $(y_1, y_2, \ldots, y_t, \ldots, y_N)$ can be divided into segments:

$$(y_1, \ldots, y_k), (y_{k+1}, \ldots, y_{2k}), \ldots, (y_{hk+1}, \ldots, y_{(h+1)k}), \ldots, (y_{(H-1)k+1}, \ldots, y_N)$$

where $N=Hk$. Next, the each segment is transformed into sequence of ranks denoted by

$$\left(a_{1k}^{(1)}, \ldots, a_{kk}^{(1)}\right), \left(a_{1k}^{(2)}, \ldots, a_{kk}^{(2)}\right), \ldots, \left(a_{1k}^{(h)}, \ldots, a_{kk}^{(h)}\right), \ldots, \left(a_{1k}^{(H)}, \ldots, a_{kk}^{(H)}\right).$$

Finally, the frequencies of the each permutation of the ranks is calculated. This let us evaluate the distribution of the permutation as well as the distribution of the rank coefficient. Let us note that the presented procedure can be useful in the case when the time series is long and segments rather short. The large size of the segment leads very quickly to enormous number of permutations of the ranks sequence.

Conclusions

The proposed method can be useful especially to prediction of a stationary time series treated as the sequence of independent and identically distributed random variables. In this case we can assume that all permutations of the ranks are equally probable, so $\rho_k = 0$. It seems that the same procedure can be useful in the case when random variables is not stationary. In this case the probability distribution of the rank permutations should be estimated.

The accuracy of the considered prediction method can be based on the ex-post analysis of frequency of exact prediction of the rank of future observations of the time series. The quality of the proposed prediction procedure can be assessed on the basis of simulation analysis of the actual empirical or artificial time series. But it needs the separate research. Other correlation coefficients like the well known Spearman's rank correlation coefficient should be involved in such a research.

Acknowledgements

The author is grateful to Reviewers for valuable comments.

References

Höffding (1947), On the Distribution of the Rank Correlation Coefficient τ When the Variates Are Not Independent, „Biometrika", Vol. 34, No. 3/4, pp. 183-196

Kendall M.G. (1958), Rank Correlation Methods, C. Griffin and Company, London.

ON PREDICTION OF TIME SERIES ON THE BASIS OF RANK CORRELATION COEFFICIENTS

Summary

In the paper the problem of prediction of a time series is considered. Time series observations can be measured on order scale. On the basis of observed ranks of values of the variables observed in the past periods a forecast of the rank of the observation in the future period is determined. The proposed method results from the derivation of the distribution of the well known Kendall's rank coefficient. The paper was inspired by a lecture of Jean H.P. Paelinck who gave it at the University of Economics in Katowice when he received the title of doctor horis causa of the University in 1987.