The population version of Spearman's rank correlation coefficient in the case of ordinal discrete random variables

A.M.Hanea[*] D.Kurowicka R.M.Cooke
A.Hanea@ewi.tudelft.nl D.Kurowicka@ewi.tudelft.nl R.M.Cooke@ewi.tudelft.nl
Delft Institute for Applied Mathematics
Delft University of Technology
The Netherlands

Applications in various domains often lead to high dimensional dependence modelling. Whereas independence is a well defined concept, various dependence measures have been studied, e.g. the product moment correlation, Spearman's rank correlation r, Kendall's τ. Our focus is on Spearman's rank correlation. A population version of Spearman's rank correlation has been defined in the case of continuous variables. We propose a correction of the "classical" population version of Spearman's rank correlation coefficient ($\bar{r}$), which can be applied to discrete ordinal variables (i.e. variables which can be written as monotone transforms of uniform variables).

Consider a population distributed according to two variates X and Y. Two members (X_1, Y_1) and (X_2, Y_2) of the population will be called *concordant* if:

$$X_1 < X_2, Y_1 < Y_2 \text{ or } X_1 > X_2, Y_1 > Y_2.$$

They will be called *discordant* if:

$$X_1 < X_2, Y_1 > Y_2 \text{ or } X_1 > X_2, Y_1 < Y_2.$$

The probabilities of concordance and discordance are denoted with P_c, and P_d respectively. The population version of Spearman's r is defined as proportional to the difference between the probability of concordance, and the probability of discordance for two vectors (X_1, Y_1) and (X_2, Y_2), where (X_1, Y_1) has distribution F_{XY} with marginal distribution functions F_X and F_Y and X_2, Y_2 are independent with distributions F_X and F_Y; moreover (X_1, Y_1) and (X_2, Y_2) are independent (e.g., Joe, 1997):

$$r = 3 \cdot (P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]). \tag{1}$$

The above definition is valid only for populations for which the probabilities of $X_1 = X_2$ or $Y_1 = Y_2$ are zero. The main types of such populations are an infinite population with

[*] Correspondence to: A.M.Hanea, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands, Telephone: +31 1527 84563.

both X and Y distributed continuously, or a finite population where X and Y have disjoint ranges (Hoffding, 1947). In order to formulate a population version of Spearman's r for discrete variables, one will have to correct for the probabilities of $X_1 = X_2$ and $Y_1 = Y_2$. This correction is similar to the correction for ties for the sample version of the rank correlation.

In order to define the sample version of Spearman's r we consider N samples of the random vector (X, Y). Suppose the samples for both variables are ranked such that in the rankings of each variable there are sets of u and v tied ranks, respectively. We first define:

$$U = \frac{1}{12}\sum(u^3 - u); V = \frac{1}{12}\sum(v^3 - v).$$

We will denote with d the differences between the ranks of each pair. Then, the rank correlation for samples, when ties are present, r_{st} [†] can be computed as follows (Kendall and Gibbons, 1990):

$$r_{st}(X,Y) = \frac{\frac{1}{6}(N^3 - N) - \sum d^2 - U - V}{\sqrt{[\frac{1}{6}(N^3 - N) - 2U][\frac{1}{6}(N^3 - N) - 2V]}}. \tag{2}$$

Let us now consider the discrete random vectors (X_1, Y_1), (X_2, Y_2), where X_2 and Y_2 are independent with the same marginal distributions as X_1 and Y_1, respectively; moreover (X_1, Y_1) and (X_2, Y_2) are independent. The states of X_i are ranked from 1 to m; the states of Y_i are ranked from 1 to n. The joint probabilities of (X_1, Y_1) and (X_2, Y_2) are given in terms of p_{ij} and q_{ij}, $i = 1, .., m; j = 1, .., n$, respectively.

$X_1 \setminus Y_1$	1	2	...	n	
1	p_{11}	p_{12}	$\cdots$	p_{1n}	p_{1+}
2	p_{21}	p_{22}	$\cdots$	p_{2n}	p_{2+}
...	...	...	...	...	...
m	p_{m1}	p_{m2}	$\cdots$	p_{mn}	p_{m+}
	p_{+1}	p_{+2}	$\cdots$	p_{+n}	

$X_2 \setminus Y_2$	1	2	...	n	
1	q_{11}	q_{12}	$\cdots$	q_{1n}	p_{1+}
2	q_{21}	q_{22}	$\cdots$	q_{2n}	p_{2+}
...	...	...	...	...	...
m	q_{m1}	q_{m2}	$\cdots$	q_{mn}	p_{m+}
	p_{+1}	p_{+2}	$\cdots$	p_{+n}	

Table 1: Joint distribution of (X_1, Y_1) (left); Joint distribution of (X_2, Y_2) (right)

In Table 1, p_{i+}, $i = 1, ..., m$ represent the margins of X_1 and X_2, and the margins of Y_1 and Y_2 are denoted p_{+j}, $j = 1...n$. One can rewrite each q_{ij} as $q_{ij} = p_{i+}p_{+j}$, for all $i = 1, ..., m$, and $j = 1, ..., n$. Using this terminology we calculate:

$$P_c - P_d = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}\left(\sum_{k\neq i}\sum_{l\neq j} sign(k-i)(l-j)q_{kl}\right)\right) \tag{3}$$

Spearman's rank correlation coefficient of two discrete variables can be calculated as in the following theorem:

[†] The index s indicates that this is a sample version, and the t comes from "ties".

Theorem 1. *Consider a population distributed according to two variates X and Y. Two members (X_1, Y_1) and (X_2, Y_2) of the population are distributed as in Table 1. Let $P_c - P_d$ be given by formula (3). Then the population version of Spearman's rank correlation coefficient of X and Y is:*

$$\bar{r} = \frac{P_c - P_d}{\sqrt{\left(\sum\limits_{j>i} p_{i+}p_{j+} - \sum\limits_{k>j>i} p_{i+}p_{j+}p_{k+}\right) \cdot \left(\sum\limits_{j>i} p_{+i}p_{+j} - \sum\limits_{k>j>i} p_{+i}p_{+j}p_{+k}\right)}}$$

Proof: We start from the sample version of the rank correlation when ties are present, given in formula (2). Dividing by N (the number of samples), we give an interpretation of this formula in terms of frequencies.

$$\bar{r}(X,Y) = \frac{-\sum \bar{d}^2 - \bar{U} - \bar{V}}{\sqrt{(-2\bar{U})(-2\bar{V})}}. \tag{4}$$

Without loss of generality we will consider $n = m$. In order to recalculate the numerator of formula (4) we will denote $R_X(i), i = 1, .., m$ the ranks of variable X and $R_Y(j), j = 1, .., m$ the ranks of variable Y. We obtain:

$$R_X(i) = 1 + \sum_{k=1}^{i-1} \frac{p_{k+}}{2} - \sum_{k=i+1}^{m} \frac{p_{k+}}{2} \qquad \text{and} \qquad R_Y(j) = 1 + \sum_{l=1}^{j-1} \frac{p_{+l}}{2} - \sum_{l=j+1}^{m} \frac{p_{+l}}{2}$$

If we calculate $-\bar{U}$, we obtain:

$$\begin{aligned} -\bar{U} &= -\frac{1}{12}\left(p_{1+}^3 - p_{1+} + p_{2+}^3 - p_{2+} + \ldots + p_{m+}^3 - p_{m+}\right) \\ &= \frac{1}{12}\left((p_{1+} + p_{2+} + \ldots + p_{m+})^3 + \left(p_{1+}^3 + p_{2+}^3 + \ldots + p_{m+}^3\right)\right) \\ &= \frac{1}{4}\left(\sum_{i=1}^{m}\sum_{i \neq j} p_{i+}^2 p_{j+} + 2\sum_{k>j>i} p_{i+}p_{j+}p_{k+}\right) \end{aligned} \tag{5}$$

In the same manner we obtain the following for $-\bar{V}$:

$$-\bar{V} = \frac{1}{4}\left(\sum_{i=1}^{m}\sum_{i \neq j} p_{+i}^2 p_{+j} + 2\sum_{k>j>i} p_{+i}p_{+j}p_{+k}\right) \tag{6}$$

We can also rewrite $\sum \bar{d}^2$ as follows:

$$\begin{aligned} \sum \bar{d}^2 &= \sum_{i,j=1}^{m} p_{ij}\left(R_X(i) - R_Y(j)\right)^2 \\ &= \frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\left(\sum_{k=1}^{i-1} p_{k+} - \sum_{k=i+1}^{m} p_{k+}\right) + \left(\sum_{l=j+1}^{m} p_{+l} - \sum_{l=1}^{j-1} p_{+l}\right)\right)^2 \end{aligned}$$

$$\begin{aligned}\sum \bar{d}^2 &= \frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{k=1}^{i-1} p_{k+} - \sum_{k=i+1}^{m} p_{k+}\right)^2 + \frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{l=j+1}^{m} p_{+l} - \sum_{l=1}^{j-1} p_{+l}\right)^2 \\ &+ \frac{1}{2}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{k=1}^{i-1} p_{k+} - \sum_{k=i+1}^{m} p_{k+}\right)\left(\sum_{l=j+1}^{m} p_{+l} - \sum_{l=1}^{j-1} p_{+l}\right)\end{aligned} \tag{7}$$

One can recalculate the first term of the above sum and, using formula (5), show the following:

$$\begin{aligned}\frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{k=1}^{i-1} p_{k+} - \sum_{k=i+1}^{m} p_{k+}\right)^2 &= \frac{1}{4}\sum_{k=1}^{m} p_{k+}^2 \cdot \sum_{i=1}^{m}\sum_{k\neq j} p_{ij} - \frac{1}{2}\sum_{k<j<l} p_{k+}p_{l+}p_{j+} \\ &+ \frac{1}{2}\sum_{k<l<j} p_{k+}p_{l+}p_{j+} + \frac{1}{2}\sum_{j<l<k} p_{k+}p_{l+}p_{j+} \qquad (8) \\ &= \frac{1}{4}\left(\sum_{k=1}^{m}\sum_{k\neq j} p_{k+}^2 p_{j+} + 2\sum_{k<l<j} p_{k+}p_{l+}p_{j+}\right) \\ &= -\bar{U}\end{aligned}$$

Recalculating the second sum and using (6) we obtain:

$$\frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{l=j+1}^{m} p_{+l} - \sum_{l=1}^{j-1} p_{+l}\right)^2 = \frac{1}{4}\left(\sum_{l=1}^{m}\sum_{l\neq j} p_{+l}^2 p_{+j} + 2\sum_{l<k<j} p_{+k}p_{+l}p_{+j}\right) = -\bar{V} \tag{9}$$

The last term of (7) can be also rewritten as:

$$-\frac{1}{2}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{k\neq i}\sum_{l\neq j} sign(k-i)(l-j)q_{kl}\right) \tag{10}$$

Using relations (8), (9) and (10) in equation (7) we can write:

$$\sum \bar{d}^2 = -\bar{U} - \bar{V} - \frac{1}{2}(P_c - P_d)$$

Therefore formula (4) becomes:

$$\bar{r}(X,Y) = \frac{\frac{1}{2}(P_c - P_d)}{\sqrt{(-2\bar{U})(-2\bar{V})}}.$$

If we further calculate $-\bar{U}$, we obtain:

$$-\bar{U} = \frac{1}{4}\left(\sum_{i=1}^{m}\sum_{i\neq j} p_{i+}^2 p_{j+} + 2\sum_{k>j>i} p_{i+}p_{j+}p_{k+}\right) = \frac{1}{4}\left(\sum_{j>i} p_{i+}p_{j+} - \sum_{k>j>i} p_{i+}p_{j+}p_{k+}\right) \tag{11}$$

Using (11) and a similar formula calculated for $-\bar{V}$, we obtain:

$$\bar{r} = \frac{P_c - P_d}{\sqrt{\left(\sum_{j>i} p_{i+}p_{j+} - \sum_{k>j>i} p_{i+}p_{j+}p_{k+}\right)\cdot\left(\sum_{j>i} p_{+i}p_{+j} - \sum_{k>j>i} p_{+i}p_{+j}p_{+k}\right)}}$$

□

One can express $\sum_{j>i} p_{i+}p_{j+}$ as $P(X_1 < X_2)$ where X_1, X_2 are independent and have the same distribution; and, similarly $\sum_{k>j>i} p_{i+}p_{j+}p_{k+}$ as $P(X_1 < X_2 < X_3)$ where X_1, X_2 and X_3 are independent with the same distribution. When $m, n \to \infty$, the denominator of the formula given in Theorem 1 goes to 1/3. Hence, in this case, $\bar{r}$ is equivalent to r for continuous variables.

One class of discrete distributions can be obtained as monotone transforms of uniform variables. These distributions can be constructed by specifying the marginal distributions and a copula, i.e. a distribution on the unit square with uniform marginals (Nelsen, 1999). Each term p_{ij} from Table 1 (left) can be written in terms of the chosen copula, as follows:

$$\begin{aligned} p_{ij} &= C\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) + C\left(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\right) \\ &- C\left(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) - C\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\right) \end{aligned} \tag{12}$$

Each copula can be parametrized by its rank correlation r, so we will use the notation C_r instead of C. Further, we will establish the relation between the rank correlation of the discrete variables and the rank correlation of the underlying uniforms.

Theorem 2. *Let C_r be a copula and (X, Y) a random vector distributed as in Table1(left), where each p_{ij} is given by formula (12) . Then the rank correlation of X and Y is denoted $\bar{r}_C$ and it has the same expression as $\bar{r}$, where:*

$$P_c - P_d = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} \left(p_{i+} + p_{(i+1)+}\right)\left(p_{+j} + p_{+(j+1)}\right) C_r\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) - \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} p_{i+}p_{+j}$$

Moreover, if C_r is a positively ordered copula (Nelsen, 1999), then $\bar{r}_C$ is an increasing function of the rank correlation of the underlying uniforms.

Idea of the Proof. For simplicity, all further calculations will be done for the case $n = m$. In order to prove the expression of $P_c - P_d$ from the theorem, we first acquire an intermediate result:

$$P_c - P_d = \sum_{i,j=1}^{m-1} p_{ij}(p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+})(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m}) - \sum_{i,j=1}^{m-1} p_{i+}p_{+j} \tag{13}$$

We start from equation (3) and rewrite the double sum in terms of p_{ij}, p_{i+}, p_{+j} with $i, j = 1, .., m-1$. Collecting like terms and performing a number of calculations we obtain equation (13). Now we can use the expression for p_{ij} from (12) to rewrite the first part of the equation (13). After algebraic manipulations of the terms we obtain:

$$P_c - P_d = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} \left(p_{i+} + p_{(i+1)+}\right)\left(p_{+j} + p_{+(j+1)}\right) C_r\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) - \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} p_{i+}p_{+j}$$

Let C_r be a positively ordered copula. Then $\bar{r}_C$ is a linear combination, with positive coefficients, of positively ordered copulas. Hence the rank correlation of two discrete variables is an increasing function of the rank correlation of the underlying uniforms.

□

If we look at the limiting case, when $m, n \to \infty$, the expression for $P_c - P_d$ given in Theorem 2 is equivalent to:

$$\iint 4 f_X(x) f_Y(y) C_r(F_X(x), F_Y(y)) d_x d_y - 1.$$

where f_X, f_Y are the marginal densities of X and Y, respectively, and F_X and F_Y are the marginal distributions of X and Y, respectively. If we denote $F_X(X) = U$ and $F_Y(Y) = V$, then:

$$\lim_{m,n\to\infty} (P_c - P_d) = 4 \iint C_r(u,v) d_u d_v - 1$$

which is equal to $P_c - P_d$ for continuous variables (Nelsen, 1999).

The class of discrete ordinal distributions which are obtained as monotone transforms of uniform variables using formula (12) can be used in various uncertainty analysis models such as: dependence trees, vines, and Bayesian belief nets (Kurowicka and Cooke, 2006). In these models, the marginal distributions of the variables must be transformed to uniforms on $[0, 1]$ and the dependence structure must be defined (via rank correlations) with respect to the uniform variates. The rank correlation of two discrete variables and the rank correlation of their underlying uniforms are not equal. Theorem 2 describes the relationship between them.

There are still open issues related to this topic, and one of them is the rank correlation between a discrete variable and a continuous one.

References

Hoffding, W. (1947). On the distribution of the rank correlation coefficient r when the variates are not independent. *Biometrika*, 34:183–196.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman & Hall, London.

Kendall, M. and Gibbons, J. (1990). *Rank Correlation Methods.* First published in Great Britain 1948. Oxford University Press.

Kurowicka, D. and Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling.* Wiley.

Nelsen, R. (1999). *An Introduction to Copulas.* Lecture Notes inStatistics. Springer-Verlag, New York.