Metodološki zvezki, Vol. 3, No. 1, 2006, 9-19

A Distance Function for Ranked Variables: A Proposal for a New Rank Correlation Coefficient

Antonio Mango[1]

Abstract

Rank correlation coefficients, RCC, are usually based on differences between matched pairs of modalities of two related variables in ordinal scale or on the number of inversions existing between couples of paired modalities, like ρ - *Spearman* and τ - *Kendall*, respectively. Here a way to build a RCC based on determinants of all second order minors of a $n \times 2$ data matrix in ordinal scale is proposed. Some possibilities of use of these determinants will be pointed out and one that seems particularly interesting will be considered showing the properties of the coefficient obtained and comparing them with the corresponding of *Kendall's* and *Spearman's* ones.

1 Introductory remarks

A little more than one hundred years ago *T.G.Fechner*, (Fechner,1887), introduced a RCC based on the number of inversions of a variable with respect to an ordered one.

At the beginning of this century, *C.Spearman*, (Spearman, 1904), introduced a RCC based on the difference between paired ranks; later on, *M.G.Kendall*, (Kendall, 1962), starting from *Fechner's* idea, proposed a RCC in which the number of existing inversions, among all couples of paired modalities of two ordered variables, was taken into account.

Here I propose a criterion for the construction of a RCC where the determinants of all second order minors of a $n \times 2$ data matrix enter as elements.

2 A distance function among vectors

We have already met in Statistics the use of determinants of 2×2 matrices in the analysis of dependence among variables both for qualitative and quantitative scales, I will show that the determinant of a 2×2 matrix can be used to build a distance function.

Given the following column vectors:

$$\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} r \\ s \end{bmatrix}$$

[1] Dipartimento di Matematica e Statistica, Universita' degli Studi di Napoli Federico II, 80126 Napoli; mango@unina.it

we assume that the inequality:

$$\left|det\begin{bmatrix} x & a \\ y & b \end{bmatrix}\right| \leq \left|det\begin{bmatrix} x & r \\ y & s \end{bmatrix}\right| + \left|det\begin{bmatrix} a & r \\ b & s \end{bmatrix}\right| \tag{2.1}$$

holds, any real x, y, a, b, r and s.

If true, the absolute value of a 2×2 matrix determinant should be a distance function between the vectors; this hypothesis is not true in general because there exist at least three vectors that don't verify it:

$$\left|det\begin{bmatrix} 1 & 2 \\ 5 & 3 \end{bmatrix}\right| - \left|det\begin{bmatrix} 1 & 4 \\ 5 & 1 \end{bmatrix}\right| - \left|det\begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}\right| = 22,$$

it will be shown that 2.1 is $true$ under the condition that the sum of the elements of the three column vectors is constant, the simplest way to get it is to divide the elements of each column vector for their sum.

Inequality 2.1 may be rewritten as:

$$\left|det\begin{bmatrix} \frac{x}{x+y} & \frac{a}{a+b} \\ \frac{y}{x+y} & \frac{b}{a+b} \end{bmatrix}\right| \leq \left|det\begin{bmatrix} \frac{x}{x+y} & \frac{r}{r+s} \\ \frac{x}{x+y} & \frac{s}{r+s} \end{bmatrix}\right| + \left|det\begin{bmatrix} \frac{a}{a+b} & \frac{r}{r+s} \\ \frac{b}{a+b} & \frac{s}{r+s} \end{bmatrix}\right|$$

which turned into:

$$\left|\frac{a}{a+b} - \frac{x}{x+y}\right| \leq \left|\frac{a}{a+b} - \frac{r}{r+s}\right| + \left|\frac{x}{x+y} - \frac{r}{r+s}\right|$$

verifies the triangular property of the Euclidean metrics.

I propose to calculate these distances for all minors of a $n \times 2$ matrix, whose columns are permutations of the first n integers, and sum them to obtain a distance measure for these column vectors.

3 Determinants of matrices as elements for indices construction

Let us consider a $n \times 2$ data matrix at ordinal scale level, extract all second order minors, say $\mathbf{T}_j$, $j = 1, 2, \ldots, \frac{n(n-1)}{2}$, and calculate all determinants $\det(\mathbf{T}_j)$.

We have null determinants in the case of perfect concordance of ranks, positive and negative determinants in other cases.

We can resort to these determinants as intermediate scores and use:

1) the absolute value of the sum of their algebraic values,
2) the sum of their absolute values,
3) the square root of the sum of their second power,

to build RCCs whose variable components are:

S_1	$=$	$\left\lvert\sum_{j=1}^{\frac{1}{2}n(n-1)} \det(\mathbf{T}_j)\right\rvert$
S_2	$=$	$\sum_{j=1}^{\frac{1}{2}n(n-1)} \lvert\det(\mathbf{T}_j)\rvert$
S_3	$=$	$\sqrt{\sum_{j=1}^{\frac{1}{2}n(n-1)} (\det(\mathbf{T}_j))^2}.$

.

The purpose of this work is to introduce an index that fulfills the fundamental property of ordering the greater part of the couples of permutations of the first n integers on the base of their mutual degree of correlation. The presence of this property can be insured by its metrics structure, and by its *diversity*.

The quantity S_2 has been selected for the construction of such an index, say δ, because it gives the best results under this condition with respect to quantities S_1 and S_3.

As a consequence of the emphasis here given to this property, since the metric structure of the index has already been verified for the index δ, the concept of *diversity* will be deepened, in order to verify that the best index under this aspect is δ with respect to δ_1, δ_2, ρ and τ.

4 The property of diversity

Two indices are different if they give different results when applied to a same data set. An index has a greater *diversity* than another if its set of results is larger than the set of the other when applied to a same data set.

In the case of perfect diversity of an index, we would observe a one-to-one correspondence among the elements of a sample universe and its image in R.

In literature on rank correlation we can find a considerable number of coefficients, among which ρ of Spearman and τ of Kendall, are different and do not show a perfect diversity.

Generally we agree to associate 1 to the case of perfect rank positive correlation and -1, to the case of perfect rank negative correlation, and other real numbers r, where $-1 < r < 1$, to other cases. But how do we distinguish between two "other" cases?

For example, it is not an easy task to say at first sight, among the following three cases, which is the one with higher correlation:

a

2	1
3	2
4	3
1	4
5	5

b

1	1
2	2
5	3
3	4
4	5

c

2	1
3	2
4	3
5	4
1	5

As a result, let us consider the values of ρ and of τ for these couples of permutations, we have:

$index \backslash perm$	a	b	c
ρ	0.4	0.7	0
τ	0.4	0.6	0.2

we see that in case a the indices agree , in case b results $\rho > \tau$ and, in case c, $\rho < \tau$. This means evidently that each index follows a proper definition of rank correlation.

Two main aspects arise in the comparison of the universes of two different indices applied to couples of permutations of the first n integers: The one pertains the lack of an identical progression with respect to the same sequence of couples of permutations, as seen; the other pertains to the frequency distribution, say *DFindex*, whose elements are couples (index,frequency).

For example, for $n = 4$, we have:

$DF\rho = \{\{-1, 1\}, \{-0.8, 3\}, \{-0.6, 1\}, \{-0.4, 4\}, \{-0.2, 2\}, \{0, 2\}, \{0.2, 2\}$
$, \{0.4, 4\}, \{0.6, 1\}, \{0.8, 3\}, \{1, 1\}\}$
$DF\tau = \{\{-1, 1\}, \{-0.6, 3\}, \{-0.3, 5\}, \{0, 6\}, \{0.3, 5\}, \{0.6, 3\}, \{1, 1\}\}$

and observe that $DF\rho$ has a wider variety of indices than $DF\tau$, this can be seen evidencing the frequency distributions of the frequencies for the two indices, say $index\,(c_i, f_i)$ where c_i represents numerousness of the class, that is the number of times an index has appeared in the universe, and f_i the frequency of indices that have appeared c_i times:

$$\boxed{\begin{array}{l} \rho\,(c_i, f_i) = \{\{1, 4\}, \{2, 3\}, \{3, 2\}, \{4, 2\}\} \\ \hline \tau\,(c_i, f_i) = \{\{1, 2\}, \{3, 2\}, \{5, 2\}, \{6, 1\}\} \end{array}}. \tag{4.1}$$

It is expected that, in the case of complete diversity of an index, we should have:

$$index\,(c, f) = \{\{1,\ n!\}\}$$

As a measure of the degree of diversity for an index, we will choose the *Shannon* relative entropy index:

$$H_{rel} = \frac{-\sum_{i=1}^{k} p_i \log p_i}{\log k}$$

where $p_i = \frac{c_i \times f_i}{n!}$ and k is the number of classes c_i.

We observe that for the distributions in4.1 the *Shannon* relative entropy index gives:

$$H_{rel} = \begin{array}{l} 0.71 \text{ for } \rho\,(c_i, f_i) \\ 0.86 \text{ for } \tau\,(c_i, f_i) \end{array}$$

this means that more permutations may be distinguished by ρ than by τ.

5 Definition of the rank correlation coefficient δ and its properties

The definition of $RCC\ \delta$ is bounded to the quantity S_2, for simplicity S, which is the sum of the absolute values of the determinants appearing in columns from (4) to (6) in Table 1, where

$$det \begin{bmatrix} x_j^{'} & y_j^{'} \\ x_i & y_i \end{bmatrix} = det \begin{bmatrix} \frac{x_j}{x_j+x_i} & \frac{y_j}{y_j+y_i} \\ \frac{x_i}{x_j+x_i} & \frac{y_i}{y_j+y_i} \end{bmatrix} = \frac{x_j y_i - y_j x_i}{(x_j + x_i)\,(y_j + y_i)},$$

and:

$$S = \frac{|x_1 y_2 - x_2 y_1|}{(x_1 + x_2)\,(y_1 + y_2)} + \frac{|x_1 y_3 - x_3 y_1|}{(x_1 + x_3)\,(y_1 + y_3)} + \cdots + \frac{|x_{n-1} y_n - x_n y_{n-1}|}{(x_{n-1} + x_n)\,(y_{n-1} + y_n)}, \tag{5.1}$$

Table 1: Determinants of the second order minors with constant sum of the elements of each column from a data matrix of two ordinal variables.

$subj.$	X	Y	$det\begin{bmatrix} x'_1 & y'_1 \\ x'_i & y'_i \end{bmatrix}$	$\cdot$	$det\begin{bmatrix} x'_j & y'_j \\ x'_i & y'_i \end{bmatrix}$	$\cdot$	$det\begin{bmatrix} x'_{n-1} & y'_{n-1} \\ x'_i & y'_i \end{bmatrix}$
(1)	(2)	(3)	(4)	$\cdot$	(5)	$\cdot$	(6)
1	x_1	y_1		$\cdot$		$\cdot$	
2	x_2	y_2	$x'_1y'_2 - x'_2y'_1$	$\cdot$		$\cdot$	
3	x_3	y_3	$x'_1y'_3 - x'_2y'_3$	$\cdot$		$\cdot$	
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$		$\cdot$	
i	x_i	y_i	$x'_1y'_i - x'_iy'_1$	$\cdot$	$x'_jy'_i - x'_iy'_j$	$\cdot$	
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	
n	x_n	y_n	$x'_1y'_n - y'_1x'_n$	$\cdot$	$x'_jy'_n - x'_ny'_j$	$\cdot$	$x'_{n-1}y'_n - x'_ny'_{n-1}$

or

$$S = \sum_{i=1}^{n-1}\sum_{j=2}^{n}\left|\frac{x_i}{x_i + x_j} - \frac{y_i}{y_i + y_j}\right|. \tag{5.2}$$

S is an absolute index of cograduation, we will normalize it, obtaining S_{rel}, and make it to vary within the interval $[-1, 1]$, obtaining δ.

We first need to find the maximum value of S, say $S_{\max}$, which will be done by putting in Table 1:

$$x_i = i$$

and

$$y_i = (n - i + 1), \tag{5.3}$$

which represent the case of perfect discordance between the two variables, and then apply the S definition. In this way the determinants bear the same sign, then the sum of their absolute values corresponds to the absolute value of their sum. A very simple formula to obtain $S_{\max}$ is the following:

$$S_{\max} = \sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{i-j}{i+j}. \tag{5.4}$$

When the variables show tied ranks to get $S_{\max}$ it is necessary to reorganize one variable in increasing order and the other in decreasing order and apply the S definition.

The normalized value of S, say S_{rel}, will be then:

$$S_{rel} = \frac{S}{S_{\max}} \tag{5.5}$$

it follows δ:

$$\delta = 1 - 2S_{rel}$$

or also:

$$\delta = \frac{S_{\max} - 2S}{S_{\max}}.$$

6 Some sampling distribution properties

6.1 Triangular inequality

We have already seen that the triangular property holds for δ, while for ρ and τ does not.

6.2 Diversity

Beginning with a frequency distribution of an index in the universe, which we will call *Index* whose generic couple is $(index_i, f_i)$, we can build a frequency distribution of the frequencies of Index, which we can indicate with FI whose generic couple is (f_i, f_{f_i}). From here we go to the frequency distribution CN which we shall call *Numerousness of the classes of Index*, or simply *Numerousness of Index*, whose generic couple is

$$\left(c_i = f_i, f_{c_i} = \frac{f_i \times f_{c_i}}{n!} \times 100\right)$$

Hence c_i represents the number of times an index can be repeated in the universe and f_{c_i} represents the percentage of indices of the universe which are repeated a number of times equal to c_i.

Numerousness is strictly related to *diversity* which consists of both *richness* (how many) and *evenness* (how distributed), we will, through the data reported in Tables 2, 3, 4, 5, and 6 in which the CN distributions for the indices δ, δ_1, δ_2, ρ and τ for some n values are shown, measure the degree of diversity of these indices.

Table 2: Percent frequency of δ for classes of numerousness for some n.

				δ					
$n = 4$		5		6		7		8	
c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}
1	41.7	1	21.7	1	9.72	1	4.50	1	8.8
2	41.7	2	63.4	2	58.61	2	66.40	2	45.3
4	16.6	4	10.0	3	0.83	3	0.98	3	8.6
		6	4.9	4	22.22	4	20.87	4	19.6
				6	5.84	6	5.24	5	3.2
				8	1.11	8	2.06	6	5.2
				12	1.67	16	0.63	7	1.5
								8	1.9
								9	0.5
								10	0.6
								$11 \div 24$	0.9

We can observe that, for $n = 4$, δ furnishes 8.8 percent of indices which are observed only $once$ and 45.3 percent that are observed $twice$, while these percentages are respectively, with reference to δ_1, 0.1 and 0.0, with reference to δ_2, 16.7 and 25, with reference

Table 3: Percent frequency of δ_1 for classes of numerousness for some n.

.

δ_1

$n=4$		5		6		7		8	
c	f_c	c	f_c	c	f_c	c	f_c	c	f_c
1	91.7	1	55.0	1	7.5	1	0.4	1	0.1
2	8.3	2	28.3	2	11.7	2	1.0	2	0.0
		3	10.0	3	20.0	3	1.2	3	0.1
		4	6.7	4	23.3	4	0.6	4	0.1
				5	19.4	5	1.4	5	0.1
				6	5.0	6	2.9	6	0.1
				7	3.9	7	1.7	7	0.2
				8	6.7	8	4.2	8	0.1
				9	2.5	9	5.0	9	0.1
						10	6.3	10	0.3
						$11 \div 32$	75.3	$11 \div 153$	98.8

Table 4: Percent frequency of δ_2 for classes of numerousness for some n.

δ_2

$n=4$		5		6		7		8	
c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}
1	16.7	1	1.7	1	2	1	0.04	1	0.01
2	25.0	3	$5,0$	5	2	6	0.24	7	0.03
3	25.0	4	13.3	6	2	10	0.40	15	0.07
4	33.3	6	35.0	9	2	14	0.56	22	0.11
		7	11.7	12	2	26	1.03	47	0.23
		10	33.3	14	2	29	1.15	54	0.27
				16	2	35	1.39	70	0.35
				20	4	46	1.83	94	0.47
				21	2	54	2.14	124	0.62
				23	3	55	2.18	129	0.64
				$24 \div 42$	14	$70 \div 184$	89.05	$178 \div 1066$	97.21

to ρ, still 16.7 and 25.0, the same as for δ_2 and, with reference to τ, 8.3 and 25, this fact means that δ produces a greater diversity of indices than the others.

Diversity indices must include contemporarily both richness and evenness, for this purpose we have chosen *Shannon relative entropy index* to point out the different degree of diversity of the indices under analysis:

$$H_{rel} = \frac{H}{\ln k} = \frac{\sum_{i=1}^{S} -p_i \ln p_i}{\ln k}$$

Table 5: Percent frequency of ρ for classes of numerousness for some n.

ρ									
$n=4$		5		6		7		8	
c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}
1	16.7	1	1.7	1	0.3	1	0.04	1	0.01
2	25.0	3	5.0	5	1.4	6	0.24	7	0.03
3	25.0	4	13.0	6	1.7	10	0.40	15	0.07
4	33.3	6	35.0	9	2.5	14	0.56	22	0.11
		7	11.7	12	3.3	26	1.03	47	0.23
		10	33.0	14	3.9	29	1.15	54	0.27
				16	4.4	35	1.39	70	0.35
				20	11.1	46	1.83	94	0.47
				21	5.8	54	2.14	124	0.62
				23	6.9	55	2.18	129	0.64
				$24 \div 42$	59.2	$70 \div 184$	89.05	$178 \div 1066$	97.21

Table 6: Percent frequency of τ for classes of numerousness for some n.

τ									
$n=4$		5		6		7		8	
c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}	c_i	f_{c_i}
1	8.3	1	1.7	1	0.3	1	0.1	1	0.01
3	25.0	4	6.7	5	1.4	6	0.2	7	0.09
5	41.7	9	15.0	14	3.9	20	0.8	27	0.36
6	25.0	15	25.0	29	8.1	49	1.9	76	0.95
		20	33.3	49	13.6	98	3.9	174	2.16
		22	18.3	71	19.6	169	6.7	343	4.27
				90	25.0	259	10.3	602	2.49
				101	28.1	359	14.2	961	11.95
						455	18.1	1415	17.60
						531	21.1	1940	24.13
						573	22.7	$2493 \div 3836$	31.01

where k is the number of classes c_i, and p_i is the probability to extract an index from the class i^{th}:

$$p_i = \frac{c_i \times f_i}{n!}$$

in this way the value of H_{rel} must diminish when the number of these classes decreases and, furthermore, when the concentration of indices increases in the classes, as we can see in Tables $2-6$, the following results are obtained:

n	ρ	τ	δ	δ_1	δ_2
4	$0,57$	0.65	0.37	0.06	0.60
5	0.48	0.65	0.24	0.70	0.55
6	0.71	0.61	0.21	0.34	0.76
5	0.73	0.62	0.13	0.51	0.80
8	0.78	0.65	0.12	0.69	0.85

We can see that the values of H_{rel} related to δ are decreasing when n increasing, and are the lowest with respect to δ_1,δ_2, ρ and τ.

After these results we continue our discussion without considering indices δ_1 and δ_2 as we have chosen δ to compare with the most used indices ρ and τ.

6.3 Variance

The variance sampling distribution of ρ and of τ are:

$$var(\rho) = \frac{1}{n-1}$$

$$var(\tau) = \frac{2(2n+5)}{9n(n-1)}$$

it results, Kendall (1962):

$$var(\rho) \geq var(\tau)$$

and specifically:

$$lim_{n\to\infty}\frac{var(\rho)}{var(\tau)} = 2.25$$

Due to the difficulties of obtaining simple expressions for δ and its sampling distribution parameters, comparisons among parameters with those of ρ and τ will be done for definite values of n:

n	$var\ (\rho)$	$var\ (\tau)$	$var\ (\delta)$
4	0.33	0.241	0.064
5	0.25	0.167	0.045
6	0.2	0.126	0.035
7	0.17	0.102	0.029
8	0.14	0.083	0.024
∞	0	0	0

thus we observe that $var\ (\delta)$ is a decreasing function of n just like $var\ (\rho)$ and $var\ (\tau)$, and it is always lesser than the other two.

6.4 Symmetry

The symmetry of the sampling distribution of δ is ensured, as it can be seen, by the empirical values of the odd mean moments for some n and r, $\mu_{S,r} = \frac{\sum_{i=1}^{n!}(S_i-\mu_S)^r}{n!S_{\max}^r}$, since, as n and r become larger, the parameters decrease:

n	$\mu_{S,3}$	$\mu_{S,5}$
4	-0.01079	-0.00443
5	-0.00590	-0.00194
6	-0.00376	-0.00099
7	-0.00266	-0.00059
8	-0.00198	-0.00038

6.5 Normality

Parameters $\beta_r = \mu_{2r}/\mu_2^r$, and $\gamma_2 = 2\mu_2/({}^1S)^2$, where ${}^1S$ is the simple mean deviation index, of the sampling distribution of δ, τ and ρ, for increasing values of n, are bounded above and in particular those of γ_2 converge to the normal distribution corresponding value as shown in the following table:

	ρ			τ			δ			ρ	τ	δ
n	β_2	β_3	β_4	β_2	β_3	β_4	β_2	β_3	β_4	γ_2	γ_2	γ_2
4	1..85	5.16	10.28	2.37	7.58	27.72	2.78	12.16	65.13	3.18	2.67	3.03
5	2.07	5.40	15.86	2.53	9.11	39.6	2.82	13.74	90.40	3.12	2.83	3.01
6	2.23	6.45	21.46	2.62	10.10	48.34	2.82	13.95	97.25	3.00	2.90	3.01
7	2.28	7.12	25.91	2.68	10.79	54.90	2.84	14.10	100.31	3.03	2.95	3.01
8	2.42	8.05	31.78	2.73	11.30	60.11	2.86	14.26	102.57	3.09	2.97	3.03
∞	3	15	105	3	15	105	3	15	105	3.14	3.14	3.14

7 Final remarks

The mathematical and statistical properties taken in examination for the different indices shown in the preceding sections allow us to assume that the greatest part of these are on a basis of parity to represent rank correlation indices.

I have written this paper because I have encountered both in my professional practice and in my research the necessity of putting in order couples of variables in a group and to distinguish the greatest number of these on the basis of their degree of correlation.

Besides the other properties like normality and variability which put index δ in a favorable position in comparison to the others, the two properties dealing with the metrics structure of an index and with its degree of diversity, still seem very favorable to the index δ, as it has broadly been shown in this work.

References

[1] Blest, D.C., (2000): Rank Correlation - An alternative measure. *Australian and New Zeland, Journal of Statistics*, **42**, 101-111.

[2] Fechner, T.G. (1897): *Kollectionmasslehre*. Leipzig: Lipps Ed..

[3] Gini, C. (1914): Di una misura della dissomiglianza tra due gruppi di quantitá e delle sue applicazioni allo studio delle relazioni statistiche, Atti del Reg.Istit.Veneto delle Scien.Lett. ed Arti, 1914-15.

[4] Kendall, M.G. (1962): *Rank Correlation Methods*. London: Griffin.

[5] Lauro, N. (1977): Considerazioni sulla metrica degli indici di cograduazione, Giornate di lavoro AIRO 1977, Parma.

[6] Mango, A. (1997): *Rank Correlation Coefficients: A New Approach*. Computing Science and Statistics. Computational Statistics and Data Analysis on the Eve of the 21^{st} Century. Proceeding of the Second World Congress of the IASC, **29**, 471-476.

[7] Spearman, C. (1904): The proof and measurement of association between two things. *Am.J.Psych.*, **15**, 88.

[8] Tarsitano, A. (2005): Weighted rank correlation and hierarchical clustering, Classification and Data Analysis. In S. Zani and A. Cerioli (Eds): *Book of Short Papers*, MUP, Parma, 517-520.