AUSTRALASIAN JOURNAL OF COMBINATORICS
Volume **57** (2013), Pages 157–187

Right-invariant generalized metrics applied to rank correlation coefficients*

Dongyang Cheng

3928 Milton Street
Houston, Texas 77005
U.S.A.
dcheng@college.harvard.edu

Petros Hadjicostas

School of Mathematics, Statistics and Operations Research
Victoria University of Wellington
Wellington 6012
New Zealand
petros.hadjicostas@msor.vuw.ac.nz

Abstract

In this paper we investigate properties of rank correlation coefficients that can be derived from right-invariant generalized metrics on the symmetric group. We prove some new inequalities between a number of generalized metrics, and we characterize the sample sizes for which several (right-invariant) rank correlation coefficients can equal zero. Using the Hausdorff generalized metric, we show how to construct circular rank correlation coefficients from regular rank correlation coefficients. In addition, we show how generalized triangle inequalities satisfied by generalized metrics on the symmetric group can be used to create new partial rank correlation coefficients (that measure the association between two variables controlling for the effect of a third one).

1 Introduction

Correlation coefficients range from -1 to 1 and they give a numerical summary of the relationship between two numerical variables. In this paper we concentrate on rank correlation coefficients that are derived from right-invariant metrics, pseudo-metrics,

* Certain parts of this paper were completed while the first author was taking part at the Clark Scholar Program at Texas Tech University (in Lubbock, Texas, USA) in Summer 2010, while the second author served as the faculty mentor.

or semi-metrics on the symmetric group, which we collectively call "generalized metrics", and whose general theory was developed by Diaconis and Graham [8] in 1977. While reviewing and expanding on this theory, we define new right-invariant generalized metrics and their corresponding rank correlation coefficients and prove new inequalities between a number of such generalized metrics.

Of particular interest in this paper is Daniels' [5] rank correlation coefficient that can be used to measure the relationship between two angular variables (and in general between two circular variables). We indicate that this circular rank correlation coefficient (introduced in 1950) was re-invented several decades later by different authors doing independent work in different countries! Inspired by this coefficient and the work of Critchlow [4] on the analysis of partially ranked data, we give a general theory for the creation of right-invariant circular generalized metrics (and their corresponding circular rank correlation coefficients) from any right-invariant generalized metric.

Finally, inspired by the triangle inequality satisfied by metrics on the symmetric group, we give probably the first ever interpretation of the corresponding inequality of the induced rank correlation coefficients: we show how a partial rank correlation coefficient between two variables controlling for the effect of a third variable can be created based on this inequality. (Most of our discussion on this topic is limited to right-invariant metrics and their corresponding rank correlation coefficients that are symmetric with respect to complements.)

We start our paper with a motivation and a review of the theory of right-invariant generalized metrics on the symmetric group.

Given two numerical variables x and y (such as height and weight), one can take measurements on n individuals or objects:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n). \tag{1}$$

One can measure the linear relationship between the x and y variables using the Pearson product moment correlation coefficient

$$r_n(x, y) = \frac{\sum_{i=1}^{n}(x_i - \mathrm{av}(x))(y_i - \mathrm{av}(y))}{\sqrt{\sum_{i=1}^{n}(x_i - \mathrm{av}(x))^2}\sqrt{\sum_{i=1}^{n}(y_i - \mathrm{av}(y))^2}}, \tag{2}$$

where $\mathrm{av}(x)$ and $\mathrm{av}(y)$ are the sample means of the observed values of the x and y variables, respectively. The correlation coefficient r always satisfies the inequality $-1 \leq r \leq 1$. If $r = 1$, there is a perfect positive linear relationship between the observed values of x and y, while if $r = -1$, there is a perfect negative linear relationship between the observed values of x and y.

In case the data (1) are not on an interval scale, some scientists prefer to replace the data by the corresponding ranks:

$$(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n).$$

Here $a = (a_1, a_2, \ldots, a_n)$ is a permutation of the numbers $1, 2, \ldots, n$, and a_i is the rank of x_i: the smallest x value gets rank 1, the second smallest x value gets rank 2, and so on. Also $b = (b_1, b_2, \ldots, b_n)$ are the ranks for $(y_1, y_2, \ldots, y_n)$.

If some of the x values have ties, then some ranks must be averaged. In this paper, we are assuming there are no ties among the x values and no ties among the y values[1].

The Pearson moment product correlation coefficient between the ranks of x and the ranks of y is called the Spearman rank correlation coefficient:

$$r_{S,n}(a,b) = \frac{\sum_{i=1}^{n}(a_i - \mathrm{av}(a))(b_i - \mathrm{av}(b))}{\sqrt{\sum_{i=1}^{n}(a_i - \mathrm{av}(a))^2}\sqrt{\sum_{i=1}^{n}(b_i - \mathrm{av}(b))^2}}. \tag{3}$$

One can show that (for $n > 1$)

$$r_{S,n}(a,b) = 1 - \frac{6\sum_{i=1}^{n}(a_i - b_i)^2}{n(n-1)(n+1)}. \tag{4}$$

The numerator of the fraction of the expression above,

$$SQ_{1,n}(a,b) := \sum_{i=1}^{n}(a_i - b_i)^2, \tag{5}$$

is a measure[2] of "distance" between the permutations a and b. It can be shown that

$$0 \leq SQ_{1,n}(a,b) \leq \frac{n(n-1)(n+1)}{3}.$$

The first inequality holds as equality if and only if $a = b$, while the second inequality holds as equality if and only if

$$a = (n, n-1, \ldots, 2, 1) \circ b,$$

where $\circ$ denotes the composition of permutations f and g.

The set of all permutations, S_n, of the numbers $1, 2, \ldots, n$, endowed with the composition $\circ$ of permutations, has the structure of a group and is called the *symmetric group* on n elements with identity element $e_n := (1, 2, \ldots, n)$. Instead of using $SQ_{1,n}$ to measure the distance between a and b, one can use other "distance" functions, which we call "generalized metrics" on S_n in this paper.

A *generalized metric sequence*[3] is a list[4] $(d_n \,|\, n \in \mathbb{N}^*)$ of functions $d_n : S_n \times S_n \to [0, \infty)$ that satisfy the following properties:

1. For all $n \in \mathbb{N}^*$ and $a, b \in S_n$ we have $d_n(a,b) \geq 0$.

[1] Unfortunately, the methods for tied ranks work well only for the Spearman pseudo-metric $SQ_{1,n}$ (that appears in the formula for the Spearman rank correlation coefficient) and for few other generalized metrics–see later for a precise definition. For an arbitrary generalized metric sequence $(d_n|n \in \mathbb{N}^*)$, one has to use the general approach in Chapter IV of Critchlow [4].

[2] The subscript "1" in $SQ_{1,n}$ (and later in the section, in $D_{1,n}$) can be understood by reading the beginning of Section 4 in this paper, where two new pseudo-metrics, $SQ_{2,n}$ and $D_{2,n}$, are introduced.

[3] The phrase "generalized metric" has a different meaning in Section 3.4 (p. 75) in [6].

[4] We let $\mathbb{N} := \{0, 1, 2, \ldots\}$ be the set of non-negative integers and $\mathbb{N}^*$ be the set of positive integers.

2. For all $n \in \mathbb{N}^*$ and $a, b \in S_n$ we have $d_n(a, b) = d_n(b, a)$.

3. There is $C > 0$ such that for all $n \in \mathbb{N}^*$ and $a, b, c \in S_n$ we have

$$d_n(a, b) \leq C[d_n(a, c) + d_n(c, b)].$$

A *pseudo-metric sequence*[5] is a generalized metric sequence $(d_n \,|\, n \in \mathbb{N}^*)$ that satisfies the following property:

4. For all $n \in \mathbb{N}^*$ and $a, b \in S_n$,

$$d_n(a, b) = 0 \Leftrightarrow a = b.$$

Pseudo-metric sequences were introduced by Estivill-Castro [12, 13].

A *semi-metric sequence* is a generalized metric sequence $(d_n \,|\, n \in \mathbb{N}^*)$ that satisfies Property 3 above with $C = 1$, i.e., it is one that satisfies the triangle inequality. Some of the semi-metrics we will examine in this paper are appropriate for measuring the "distance" between the "ranks" of observations on two angular variables.

A *metric sequence* is a generalized metric sequence $(d_n \,|\, n \in \mathbb{N}^*)$ that is both a pseudo-metric and a semi-metric sequence, i.e., it satisfies Properties 1–4 with 3 satisfied for $C = 1$.

A generalized metric sequence $(d_n \,|\, n \in \mathbb{N}^*)$ is called *right-invariant* if it satisfies the property below:

5. For all $n \in \mathbb{N}^*$ and $a, b, c \in S_n$ we have $d_n(a \circ c, b \circ c) = d_n(a, b)$.

The last property is called *right invariance* and it implies

$$d_n(a, b) = d_n(a \circ b^{-1}, e_n) = d(e_n, b \circ a^{-1})$$

and $d_n(e_n, a) = d_n(e_n, a^{-1})$ for all $a, b \in S_n$. Here a^{-1} and b^{-1} are the inverses of a and b, respectively, in the symmetric group S_n (endowed with the operation of composition). For abbreviation, we use $d_n(a)$ for $d_n(e_n, a)$. One can think of $d_n(a)$ as a *measure of disarray* in the permutation a. Not all plausible measures of disarray in S_n, however, correspond to metrics (for more details, see Estivill-Castro [12, 13] and Estivill-Castro et al. [14]). Right-invariant metrics were introduced by Diaconis and Graham [8]. The reader may also consult [6, Chapter 11].

If right invariance is known to hold for a function $d_n : S_n \times S_n \to [0, \infty)$ (that satisfies Properties 1 and 2), to prove Property 3, it is enough to prove that there is $C > 0$ such that for all $n \in \mathbb{N}^*$ and $x, y \in S_n$ we have

$$d_n(x \circ y) \leq C[d_n(x) + d_n(y)].$$

The generalized metric sequence $(SQ_{1,n} \,|\, n \in \mathbb{N}^*)$, where $SQ_{1,n}$ is defined by Equation (5), is an example of a right-invariant pseudo-metric because it satisfies Property 3 with $C = 2$. On the other hand, the generalized metric sequence

[5] Sometimes the term *quasi-metric* is used in place of *pseudo-metric*.

$(\sqrt{SQ_{1,n}}\,|\,n \in \mathbb{N}^*)$ is an example of a right-invariant metric sequence because it satisfies Property 3 with $C = 1$.

According to Diaconis and Graham [8] and Estivill-Castro [12, 13], other examples of right-invariant metrics on S_n are

(i) $D_{1,n}(a, b) = \sum_{i=1}^{n} |a_i - b_i|$;

(ii) $EX_n(a, b)$ = the minimum number of transpositions (exchanges) required to bring the list $(a_1, \ldots, a_n)$ into the list $(b_1, \ldots, b_n)$;

(iii) $I_n(a, b)$ = the minimum number of pairwise adjacent transpositions required to bring $((a^{-1})_1, \ldots, (a^{-1})_n)$ into the order $((b^{-1})_1, \ldots, (b^{-1})_n)$;

(iv) $H_n(a, b)$ = number of $i \in \{1, 2, \ldots, n\}$ such that $a_i \neq b_i$.

By a result due to Cayley [3], $EX_n(a) = EX_n(a, e_n)$ is equal to n minus the number of cycles in a. By a "pairwise adjacent transposition" we mean a transposition of the form (a_i, a_{i+1}). Note that $I_n(a) = I_n(a, e_n)$ is the number of inversions in the permutation a. An inversion in a is a pair of integers (i, j) with $1 \leq i < j \leq n$ and $a_i > a_j$. The quantity $H_n(a, b)$ is known as the *Hamming distance* between permutations a and b and it was introduced by Hamming [25, pp. 154-155] in 1950. Many of the measures of disorder induced by the above generalized metrics (through $d_n(a) := d_n(a, e_n)$ for $a \in S_n$) have been used by Hadjicostas and Lakshmanan [21, 22, 23] for the analysis of sorting algorithms with erroneous comparisons.

If $(d_n | n \in \mathbb{N}^*)$ is a generalized metric sequence of functions $d_n : S_n \times S_n \to [0, \infty)$ that are not necessarily right-invariant on the symmetric group, then one can create a corresponding right-invariant generalized metric sequence $(d_n^{ri} | n \in \mathbb{N}^*)$ as follows:

$$d_n^{ri}(a, b) = M \sum_{c \in S_n} d_n(a \circ c, b \circ c) \quad (a, b \in S_n), \tag{6}$$

where M is an arbitrary positive constant, e.g., $M = 1/n!$. It can be easily proven that if $(d_n | n \in \mathbb{N}^*)$ is a generalized metric sequence, then so is $(d_n^{ri} | n \in \mathbb{N}^*)$ with the same constant C in Property 3. In addition, if $(d_n | n \in \mathbb{N}^*)$ is a pseudo-metric (resp., semi-metric, metric) then $(d_n^{ri} | n \in \mathbb{N}^*)$ is also a pseudo-metric (resp., semi-metric, metric).

Finally, we mention that there are useful generalized metric sequences $(d_n | n \in \mathbb{N}^*)$ that are neither right-invariant nor left-invariant (see (38)), such as those considered by Block et al. [2]. Those generalized metrics are *invariant under inverse transformations*, i.e., they satisfy $d_n(a, b) = d_n(a^{-1}, b^{-1})$ for all $n \in \mathbb{N}^*$ and $a, b \in S_n$. Such generalized metrics are useful for "dependence orderings on bivariate empirical distributions," but will not be examined in this paper.

The organization of the paper is as follows. In Section 2, as suggested by Diaconis and Graham [8], we use right-invariant metrics (or pseudo-metrics or semi-metrics) to define rank correlation coefficients between two permutations in the symmetric group. Part of the section is devoted to the discussion and history of Daniels' rank correlation coefficient introduced by him in 1950, re-discovered independently by

Guilbaud [20] in 1980, and re-discovered again by Fisher and Lee [16] in 1982 (even though the two authors were not acquainted with his work). Daniels and Guilbaud's identity relating their rank correlation coefficient to the Spearman's and Kendall's correlations–see (10)–was rediscovered independently by Shieh [36], who although acquainted with the work of Fisher and Lee [16], was unaware of the work of Daniels [5] and Guilbaud [20] (and the more general work of Monjardet [32, Section 6] on that subject[6]). To the best of our knowledge, this is the first time anyone has pointed that all of these works are connected together. Daniels' coefficient serves as the inspiration of a number of original results and useful methodology proposals in this paper (e.g., see the descriptions of Sections 3, 6, 7 and 8 below).

In Section 3, we formulate a precise theorem of when two angular variables have Daniels-Gilbaud[7] rank correlation coefficient equal to one. (This theorem can be easily modified to deal with the case when the coefficient equals minus one.) This simple, but clear result gives a useful interpetation of this coefficient that does not depend on the probabilistic framework of Fisher and Lee [16]. In Section 4 we introduce two new right-invariant pseudo-metrics, $D_{2,n}$ and $SQ_{2,n}$: the first one is a variation of Spearman's footrule metric $D_{1,n}$ and the other is a variation of Spearman's pseudo-metric $SQ_{1,n}$. (We do prove, however, that $SQ_{2,n}$ is a multiple of $SQ_{1,n}$, see (17).) We then proceed to prove that the first pseudo-metric is greater than or equal to the Spearman pseudo-metric $SQ_{1,n}$ and less than or equal to $n-1$ times the Spearman footrule metric $D_{1,n}$ (thus refining an inequality that appears in Diaconis and Graham [8]).

In Section 5, we review the results of Marshal [30] about the sample sizes for which Spearman's correlation can be zero and the sample sizes for which Kendall's correlation can be zero, and then we characterize the sample sizes for which other rank correlations can be zero. For example, in Theorem 3 we prove that for each sample size $n \geq 3$, Daniels and Guilbaud's rank correlation coefficient can be zero if and only if n is not of the form $n = 4m + 3$, where $m \in \mathbb{N}$.

In Section 6 we prove a new inequality between three of the semi-metrics discussed in this paper ($D_{2,n}$, I_n and the Daniels-Guilbaud semi-metric) and show how it can be transformed into an inequality between three rank correlation coefficients–see (25).

In Section 7 we give a motivation and a definition for generalized metrics that measure the "distance" between the ranks of two circular variables (such as angles). We also note that the Daniels-Guilbaud semi-metric (and the corresponding rank correlation coefficient) is indeed circular because of a formula due to Shieh [36]. This formula expresses the Daniels-Guilbaud rank correlation coefficient between two sets of ranks (permutations in S_n) as an "average" of all the Kendall's rank correlations between circular permutations of the two sets of ranks.

In Section 8 we use Critchlow's [4] method (for partially ranked data) to create circular rank generalized metrics (and circular rank correlation coefficients) from any (usual) generalized metric. This method uses the Hausdorff (generalized) "distance"

[6]Unfortunately, Monjardet [32] refers to the wrong paper by Daniels when mentioning his inequality (that can be derived from his equality (10)). He does refer to the right paper in later works [33, 34].

[7]This name was suggested by Deza and Deza [6, p. 212].

between two right cosets in the quotient space of the symmetric group with the cyclic group generated by $g_n := (2, 3, \ldots, n, 1)$. In the case of bi-invariant generalized metrics, we note that this method can be simplified (due to a result by Diaconis and Graham that appears in Critchlow [4].) (A bi-invariant generalized metric is one that is both left- and right-invariant.) For the case of the Hamming metric, we show how to obtain the corresponding circular semi-metric and (using a result due to Chris Monico [31]) we explain how to define the corresponding circular Hamming rank correlation coefficient.

In Section 9 we show how the triangle inequality for metrics on the symmetric group can be transformed into an inequality for the corresponding rank correlation coefficients, and then use the latter to define partial rank correlation coefficients of two sets of ranks controlling for a third one. More precisely, we define partial correlation coefficients for generalized metric sequences that are symmetric with respect to complements and satisfy Property 3 with $C = 1$. (In Remark 4, we generalize this by showing how to define such coefficients when the constant C in Property 3 can take any (fixed) positive value.)

The usual computational formula for the partial correlation coefficient between two numerical variables x and y controlling for a third variable z—see Equation (46)—can be used to define not only a Spearman partial rank correlation coefficient (that corresponds to the pseudo-metric $SQ_{1,n}$), but also a Spearman-induced partial rank correlation coefficients that arises from any semi-metric that is symmetric with respect to complements–see Equations (48) and (52). Besides the fact that this was done in the past for the Spearman rank correlation coefficient, we note that it has also traditionally been done on the Kendall (tau) correlation coefficient as well (see, for example, [19, Chapter 5]). We observe in Remark 3 that the absolute value of our partial rank correlation coefficient is greater than or equal to the corresponding Spearman-induced rank correlation coefficient (for each semi-metric that is symmetric with respect to complements).

The paper concludes with Section 10, where we discuss some possible future research topics.

2 Rank correlation coefficients

Given a generalized metric sequence $(d_n \,|\, n \in \mathbb{N}^*)$ of functions $d_n : S_n \times S_n \to [0, \infty)$, Diaconis and Graham [8] defined the corresponding *sequence of rank correlation coefficients* $(r_{d,n} \,|\, n \in \mathbb{N}^*)$ consisting of functions

$$r_{d,n} : S_n \times S_n \to [-1, 1]$$

satisfying

$$r_{d,n}(a, b) = 1 - \frac{2d_n(a, b)}{\max(d_n)}, \tag{7}$$

where

$$\max(d_n) := \max\{d_n(\pi, \sigma) | \pi, \sigma \in S_n\}.$$

The above definition makes sense only for those $n \in \mathbb{N}^*$ such that $\max(d_n) > 0$. Note that $r_{d,n}(a,b) = 1$ if and only if $d_n(a,b) = 0$; and $r_{d,n}(a,b) = -1$ if and only if $d_n(a,b) = \max(d_n)$.

The most widely used rank correlation coefficient is the one introduced by Spearman that was discussed in Section 1; see Equations (3) and (4). Another one is Kendall's (tau) rank correlation coefficient [27]

$$r_{K,n}(a,b) = 1 - \frac{2I_n(a,b)}{\max I_n} = 1 - \frac{4I_n(a,b)}{n(n-1)}, \tag{8}$$

which can be defined for all integers $n > 1$.[8] Finally, one can use Spearman's footrule rank correlation coefficient defined by

$$r_{F,n}(a,b) := 1 - \frac{2D_{1,n}(a,b)}{\max D_{1,n}} = 1 - \frac{2D_{1,n}(a,b)}{\lfloor n^2/2 \rfloor}. \tag{9}$$

Diaconis and Graham [8] pointed out that Kendall [27, p. 32] used the wrong denominator in the definition of Spearman's footrule.

The Hamming rank correlation coefficient can be similarly defined:

$$r_{H,n}(a,b) := 1 - \frac{2H_n(a,b)}{\max H_n} = 1 - \frac{2H_n(a,b)}{n}.$$

Note that $r_{H,n}(a,b) = 1$ if and only if $a = b$, and $r_{H,n}(a,b) = -1$ if and only if $a = c \circ b$ for some *derangement* $c \in S_n$. A derangement c is a permutation in S_n with no fixed points, and there are exactly $\left\lfloor \frac{n!}{e} + \frac{1}{2} \right\rfloor$ of them in S_n; see, for example, Hassani [26]. (Here $\lfloor x \rfloor$ is the greatest integer less than or equal to $x \in \mathbb{R}$.)

Daniels [5] introduced another correlation coefficient, $r_{U,n} : S_n \times S_n \to [-1,1]$, which satisfies the equality

$$3n\, r_{K,n}(a,b) - 2(n+1) r_{S,n}(a,b) = (n-2) r_{U,n}(a,b). \tag{10}$$

To define $r_{U,n}(a,b)$, we first define the function $OT_n : S_n \times S_n \to \mathbb{R}$ as follows[9]: we let $OT_n(a,b)$ be the number of triplets of integers (i,j,k) such that $1 \le i < j < k \le n$ and such that there is an odd permutation $(\alpha,\beta,\gamma) \in S_3$ with

$$\mathrm{Perm}(a_i,a_j,a_k) = \mathrm{Perm}(b_i,b_j,b_k) \circ (\alpha,\beta,\gamma).$$

Here $\mathrm{Perm}(a_i,a_j,a_k)$ is the (ordered) list of ranks of the numbers a_i, a_j, a_k; for example,
$\mathrm{Perm}(10,1,33) = (2,1,3)$. Daniels' correlation coefficient can then be defined as follows:

$$r_{U,n}(a,b) := 1 - \frac{2\, OT_n(a,b)}{\max OT_n} = 1 - \frac{12\, OT_n(a,b)}{n(n-1)(n-2)}, \tag{11}$$

[8] For the traditional rank correlation coefficients, such as the ones due to Spearman and Kendall, we use the more traditional notation, rather than the one introduced in this paper. For example, for Kendall's tau we write $r_{K,n}$ rather than $r_{I,n}$.

[9] The notation OT stands for "odd triplets."

which can be defined only for $n > 2$.

The quantity $OT_n(a,b)$ can also be written as

$$OT_n(a,b) = \#\{(i,j,k) \,|\, 1 \le i < j < k \le n \text{ and } \\ \mathrm{Perm}[(b_i,b_j,b_k)]^{-1} \circ \mathrm{Perm}[(a_i,a_j,a_k)] \text{ is odd}\}.$$

For all $a, b \in S_n$, it can be proven that $0 \le OT_n(a,b) \le \binom{n}{3}$; $OT_n(a,a) = 0$; and $OT_n(a,b) = OT_n(b,a)$. On the other hand, $OT_n(a,b) = 0$ if and only if $a \circ b^{-1}$ is a cyclic permutation of e_n, i.e., there is $m \in \{1,2,\dots,n\}$ such that

$$a = (m+1, m+2, \dots, n, 1, 2, \dots, m) \circ b.$$

(If $m = n$, the numbers $m+1, m+2, \dots, n$ do not exist in the above equality, in which case $a = e_n \circ b = b$.) This means that $OT_n : S_n \times S_n \to [0,\infty)$ is not a metric on S_n, but a *semi-metric* because it does *not* satisfy Property 4 of the definition of a metric.

If $g_n := (2,3,\dots,n,1)$, then $g_n^m = (m+1, m+2, \dots, n, 1, 2, \dots, m)$ for $m = 0, 1, \dots, n-1$, where $g_n^k := g_n \circ g_n^{k-1}$ for $k \ge 1$, $g_n^0 := e_n$ and $g_n^{-k} := (g_n^{-1})^k$ for $k \ge 1$. This means that

$$g_n^{-m} = g_n^{n-m} = (n-m+1, n-m+2, \dots, n, 1, 2, \dots, n-m).$$

It follows that $r_{U,n}(a,b) = 1$ if and only if $OT_n(a,b) = 0$ if and only if there is $m \in \{1,2,\dots,n\}$ such that $a^{-1} = b^{-1} \circ g_n^{n-m}$, i.e., if and only if a^{-1} is a cyclic permutation of b^{-1}.

Daniels [5] mentions that $r_{U,n}(a,b) = -1$ if and only if $OT_n(a,b) = \binom{n}{3}$, if and only if $a \circ b^{-1}$ is a cyclic permutation of $\overline{e}_n := (n, n-1, \dots, 1)$. Since the set of cyclic permutations of $\overline{e}_n$ can be written as

$$\{\overline{e}_n, \overline{e}_n \circ g_n, \overline{e}_n \circ g_n^2, \dots, \overline{e}_n \circ g_n^{n-1}\},$$

we have that $r_{U,n}(a,b) = -1$ if and only if there is $m \in \{0,1,\dots,n-1\}$ such that $a = \overline{e}_n \circ g_n^m \circ b$.

It can be easily shown that Equation (10) implies that

$$OT_n(a,b) = nI_n(a,b) - SQ_{1,n}(a,b). \tag{12}$$

Since I_n and $SQ_{1,n}$ are right-invariant functions on $S_n \times S_n$, so is OT_n (see Property 5 in Section 1).

Because of right invariance,

$$OT_n(a,b) = OT_n(a \circ b^{-1}, e_n) = OT_n(e_n, b \circ a^{-1}).$$

Thus, we may assume without loss of generality that $b = e_n$, and define the measure of disarray

$$OT_n(a) := OT_n(a, e_n) = \#\{(i,j,k) \,|\, 1 \le i < j < k \le n \text{ and } \mathrm{Perm}[a_i,a_j,a_k] \\ \text{is an odd permutation in } S_3\}.$$

Fisher and Lee [16] re-discovered Daniels' rank correlation coefficient without being acquainted with his work. They also gave the formula

$$r_{U,n}(a,b) = \binom{n}{3}^{-1} \sum_{1 \le i < j < k \le n} \delta_{a,b}(i,j,k), \tag{13}$$

where

$$\begin{aligned}\delta_{a,b}(i,j,k) :=& \operatorname{sgn}(a_i - a_j)\operatorname{sgn}(a_j - a_k)\operatorname{sgn}(a_k - a_i)\\ &\times \operatorname{sgn}(b_i - b_j)\operatorname{sgn}(b_j - b_k)\operatorname{sgn}(b_k - b_i),\end{aligned}$$

with $\operatorname{sgn}(x) = 1$ if $x > 0$, -1 if $x < 0$, and 0 if $x = 0$. The equivalence between Daniels' formula and Fisher and Lee's formula was proven, for example, by Shieh [36] and Shieh et al. [37, Appendix 4] even though in neither reference is Daniels' paper cited! For more details (and for an additional formula due to Shieh), see Section 7 of this paper.

Daniels did not emphasize the importance of his correlation coefficient, nor did he state or prove the triangle inequality for OT_n (i.e., Property 3 with $C = 1$). Fisher and Lee [16] do not mention the triangle inequality either. This was proven by by Monjardet in [33, Proposition 2] in a more general framework than the one considered here.

3 Applications of Daniels' rank correlation coefficient

Let θ and ϕ be two angular variables. Fisher and Lee [17] mentioned that a natural way of defining linear relationship between these two variables is to write

$$\theta \equiv \phi + \alpha_0 \quad \bmod (2\pi) \quad \text{for positive association, and} \tag{14}$$

$$\theta \equiv -\phi + \alpha_0 \quad \bmod (2\pi) \quad \text{for negative association,} \tag{15}$$

for some arbitrary angle α_0. They call such a dependence between θ and ϕ as toroidal-linear.

As examples of two angular variables (whose association we want to study) one may consider peak times of successive measurements of blood pressure, converted into angles (Fisher and Lee [16, 17]; Downs [11]). Another example of a pair of angular variables θ and ϕ are wind directions at two different times in a given day at a weather station (Fisher [15, pp. 149-150]). One can then collect data over a period of n days for the same weather station, or collect data for n weather stations on a single day (but at two different times).

If Daniels' correlation coefficient of the ranks of two angular variables θ and ϕ equals 1, the relationship between θ and ϕ is not always described by (14). Similarily, if Daniels' correlation coefficient of the ranks of θ and ϕ is -1, angular variables θ and ϕ are not necessarily related through (15). This, for example, is indirectly mentioned in the discussion preceeding the introduction of Daniels' correlation coefficient in Section 2 of Fisher and Lee [16] (even though the authors were not aware of Daniels' work). In any case, by avoiding the notation and probabilistic framework of Fisher and Lee [16], we formulate below a theorem that gives necessary and sufficient conditions of when the angular variables have $r_U = 1$. A similar theorem can be formulated for the case $r_U = -1$ when we replace "strictly increasing function" by "strictly decreasing function." In the theorem below the angular data are described by the n pairs

$$(\theta_1, \phi_1), (\theta_2, \phi_2), \ldots, (\theta_n, \phi_n),$$

and we assume there are no ties among the θ values and no ties among the ϕ values. Without loss of generality we assume the θ values are listed in increasing order. The proof of the theorem is elementary and thus is omitted.

Theorem 1 *Assume* $0 \le \phi_i, \theta_i < 2\pi$ *for* $i = 1, 2, \ldots, n$ *with* $\theta_i \ne \theta_j$ *and* $\phi_i \ne \phi_j$ *for* $i \ne j$, *and* $\mathrm{Perm}(\theta) = e_n = (1, 2, \ldots n)$. *Then* $r_{U,n}(\mathrm{Perm}(\theta), \mathrm{Perm}(\phi)) = 1$ *if and only if there is an* $m \in \{1, 2, \ldots, n\}$ *and a strictly increasing function* $g : \mathbb{R} \to \mathbb{R}$ *such that* $\theta_i = g(\phi_i)$ *for* $i = m, m+1, \ldots, n$ *and* $2\pi + \theta_i = g(\phi_i)$ *for* $i = 1, 2, \ldots, m-1$.

4 Some new right-invariant pseudo-metrics

In this section we define two new right-invariant pseudo-metrics that are indirectly related to Daniels' semi-metric OT_n defined earlier–see also Section 6. We define the generalized metric sequences $(D_{2,n} | n \in \mathbb{N}^*)$ and $(SQ_{2,n} | n \in \mathbb{N}^*)$ by introducing the functions $D_{2,n} : S_n \times S_n \to \mathbb{R}$ and $SQ_{2,n} : S_n \times S_n \to \mathbb{R}$ through the formulas

$$D_{2,n}(a, b) = \sum_{1 \le i < j \le n} |(a_i - b_i) - (a_j - b_j)| \tag{16}$$

and

$$SQ_{2,n}(a, b) = \sum_{1 \le i < j \le n} |(a_i - b_i) - (a_j - b_j)|^2$$

for all $a, b \in S_n$. The subscript 2 in the notation for $D_{2,n}$ and $SQ_{2,n}$ indicates that the pseudo-metrics are modifications of $D_{1,n}$ and $SQ_{1,n}$, respectively, that involve two indices under the sum (i and j). One can easily show that $D_{2,n}$ is a metric while $SQ_{2,n}$ is a pseudo-metric with $C = 2$ in Property 3 (see Section 1). To show right-invariance for $D_{2,n}$ one has to observe that

$$D_{2,n}(a, b) = \frac{1}{2} \sum_{1 \le i,j \le n} |(a_i - b_i) - (a_j - b_j)|$$

for $a, b \in S_n$. A similar observation can be used to show that $SQ_{2,n}$ is right-invariant.

Unfortunately, the pseudo-metric $SQ_{2,n}$ does not add anything to our collection of pseudo-metrics because

$$SQ_{2,n}(a, b) = n SQ_{1,n}(a, b) \quad (a, b \in S_n). \tag{17}$$

It follows that

$$\max SQ_{2,n} = \frac{n^2(n+1)(n-1)}{3},$$

and thus the non-parametric rank correlation coefficient induced by $SQ_{2,n}$ is the one induced by $SQ_{1,n}$, i.e., the Spearman rank correlation coefficient. To prove (17), note that for $a, b \in S_n$,

$$\begin{aligned}
SQ_{2,n}(a,b) = & \sum_{1\le i<j\le n} (a_i - b_i)^2 + \sum_{1\le i<j\le n} (a_j - b_j)^2 \\
& - 2 \sum_{1\le i<j\le n} (a_i - b_i)(a_j - b_j) \\
= & \sum_{i=1}^{n} (n-i)(a_i - b_i)^2 + \sum_{j=1}^{n} (j-1)(a_j - b_j)^2 \\
& - \sum_{1\le i,j\le n} (a_i - b_i)(a_j - b_j) + \sum_{k=1}^{n} (a_k - b_k)^2 \\
= & \sum_{i=1}^{n} (n - i + (i-1) + 1)(a_i - b_i)^2 - 0 = nSQ_{1,n}(a,b).
\end{aligned}$$

Note that

$$\max D_{2,n} = \frac{n(n-1)(n+1)}{3} = \max SQ_{1,n}, \tag{18}$$

and $D_{2,n}(a,b) = \max D_{2,n}$ if and only if $a = (n, n-1, \ldots, 1) \circ b$. To prove these claims, observe first that

$$D_{2,n}(a,b) \le \sum_{1\le i<j\le n} |a_i - a_j| + \sum_{1\le i<j\le n} |b_i - b_j| = 2 \sum_{1\le i<j\le n} (j - i). \tag{19}$$

The latter sum equals $n(n-1)(n+1)/3$, while equality in (19) holds if and only if

$$(a_i - a_j)(b_i - b_j) \le 0$$

for all $i, j \in \{1, 2, \ldots, n\}$ with $i < j$. The latter condition holds if and only if $a = (n, n-1, \ldots, 1) \circ b$, and our claims have been proven.

The metric $D_{2,n}$ is connected to $SQ_{1,n}$ and $D_{1,n}$ through

$$SQ_{1,n}(a,b) \le D_{2,n}(a,b) \le (n-1)D_{1,n}(a,b) \tag{20}$$

for $a, b \in S_n$. The inequality $SQ_{1,n}(a,b) \le (n-1)D_{1,n}(a,b)$ is proved in Diaconis and Graham [8], and equality here holds if and only if $a = b$ or $a = (n, 2, 3, \ldots, n-1, 1) \circ b$ – see also Hadjicostas and Monico [24]. To prove the right inequality in (20) we use the triangle inequality to get

$$\begin{aligned}
D_{2,n}(a,b) \le & \sum_{1\le i<j\le n} |a_i - b_i| + \sum_{1\le i<j\le n} |a_j - b_j| \\
= & \sum_{i=1}^{n} (n-i)|a_i - b_i| + \sum_{j=1}^{n} (j-1)|a_j - b_j|,
\end{aligned}$$

from which we obtain

$$D_{2,n}(a,b) \leq \sum_{i=1}^{n}(n-i+i-1)|a_i-b_i| = (n-1)D_{1,n}(a,b).$$

To prove the left inequality in (20), note first that

$$\sum_{1\leq i<j\leq n}((a_i-b_i)-(a_j-b_j)) = -2\sum_{i=1}^{n} i(a_i-b_i). \tag{21}$$

When $b = e_n$, equality (21) gives

$$\sum_{1\leq i<j\leq n}((a_i-i)-(a_j-j)) = 2\sum_{i=1}^{n} i^2 - 2\sum_{i=1}^{n} ia_i = SQ_{1,n}(a). \tag{22}$$

Since $x \leq |x|$ for each number x, it follows from the definition of $D_{2,n}(a)$ that $SQ_{1,n}(a) \leq D_{2,n}(a)$.

We finish the section by giving necessary and sufficient conditions for each equality to hold in (20). Due to space limitations we omit the proofs of our claims.

(i) For $n = 1$ we obviously have $D_{2,n}(a,b) = (n-1)D_{1,n}(a,b) = 0$ for $a = b = (1)$. Let $n \geq 2$ and consider the set

$$A_n := \{c \in S_n |\, c_i \neq i \text{ for at most two } i \in \{1,\dots,n\}\}.$$

If $c \in A_n$ and $c \neq e_n$, then there are i and j with $1 \leq i < j \leq n$ such that $a_i = j$ and $a_j = i$. This means that $\#A_n = \binom{n}{2} + 1$. One can show that $D_{2,n}(a,b) = (n-1)D_{1,n}(a,b)$ if and only if there is $c \in A_n$ such that $a = c \circ b$.

(ii) It follows from (22), and the fact that $x = |x|$ if and only if $x \geq 0$, that (in the case $b = e_n$) we have $SQ_{1,n}(a) = D_{2,n}(a)$ if and only if

$$a_1 - 1 \geq a_2 - 2 \geq \ldots \geq a_n - n.$$

We can actually characterize all $a \in S_n$ that satisfy $SQ_{1,n}(a) = D_{2,n}(a)$ (or more generally, all $a, b \in S_n$ that satisfy $SQ_{1,n}(a,b) = D_{2,n}(a,b)$).

Define recursively the sequence of sets $(B_n : n \in \mathbb{N}^*)$ as follows. For $n = 1$, let $B_1 := \{(1)\}$. For $n > 1$ define $B_{n1} \subseteq S_n$ by attaching to each element c of B_{n-1} the number n at the beginning of the list to create (n, c) in S_n. Also, define $B_{n2} \subseteq S_n$ by inserting to each element f of B_{n-1} the number n after $n-1$ and before the next integer in f. Thus, if

$$f = (\dots, f_{i-1}, n-1, f_{i+1}, \dots) \in B_{n-1},$$

we create $(\dots, f_{i-1}, n-1, n, f_{i+1}, \dots) \in B_{n2}$. Finally, define $B_n := B_{n1} \cup B_{n2}$.

It is clear that $\#B_n = 2^{n-1}$. Note also that $\overline{e}_n \in B_n$, and for $n > 1$, we have in particular that $\overline{e}_n \in B_{n1}$. One can show by induction that $SQ_{1,n}(a,b) = D_{2,n}(a,b)$ if and only if there is $c \in B_n$ such that $a = c \circ b$.

5 When does a rank correlation coefficient equal to zero?

If $d_n : S_n \times S_n \to \mathbb{R}$ is the n^{th} member of the generalized metric sequence $(d_\nu | \nu \in \mathbb{N}^*)$ and $r_{d,n} : S_n \times S_n \to [-1, 1]$ is the corresponding rank correlation coefficient given by Equation (7), then $r_{d,n}(a, b) = 0$ if and only if

$$d_n(a, b) = \frac{\max(d_n)}{2}.$$

Marshall [30] proved that if $n > 3$, Spearman's rank correlation coefficient, $r_{S,n}$, can be zero if and only if n is not of the form $n = 4m + 2$, where m is a positive integer. He also proved that Kendall's rank correlation coefficient, $r_{K,n}$, can be zero if and only if n is either of the form $n = 4m$ or $n = 4m + 1$, where m is a positive integer. For example, for the Hamming rank correlation coefficient, $r_{H,n}$, it is clear that it can be zero if and only if n is an even positive integer.

Knuth [28, p. 74, Ex. 104] calls a permutation $a \in S_n$ *well-balanced* if

$$\sum_{k=1}^{n} k a_k = \sum_{k=1}^{n} (n + 1 - k) a_k.$$

It turns out a permutation $a \in S_n$ is well-balanced if and and only if $r_{S,n}(a, e_n) = 0$ (i.e. the Spearman's rank correlation coefficient of a with the identity e_n is zero).

Marshall [30] proved a result that shows for which integers n there are two permutations for which the Spearman's footrule rank correlation coefficient can be zero, but he used the wrong denominator for the correlation coefficient: he used the one suggested in Section 2.20 in Kendall [27]. In Equation (7), instead of dividing by the maximum of $D_{1,n}$ over $S_n \times S_n$, Kendall [24] divides by $2(n^2 - 1)/3$. (In another suggestion, he proposes dividing by $n^2/2$, which does not equal the maximum of $D_{1,n}$ when n is odd.) Below we prove a similar result using Equation (9), which is the one suggested by Diaconis and Graham [8].

Theorem 2 *For each integer $n > 1$, Spearman's footrule rank correlation coefficient $r_{F,n}$ can be zero if and only if n is not of the form $4m+2$ where $m \in \mathbb{N} = \{0, 1, 2, \ldots\}$.*

Proof. Notice that $r_{F,n}(a, b) = 0$ if and only if

$$D_{1,n}(a, b) = \frac{\left\lfloor \frac{n^2}{2} \right\rfloor}{2}.$$

If n is even and $r_{F,n}(a, b) = 0$, it can be easily shown that n is a multiple of 4 (since $D_{1,n}$ is always even). To prove the converse (when n is even), we show by induction that if $n = 4l$ ($l \in \mathbb{N}^*$), rankings a and b in S_n can be constructed so that $r_{F,n}(a, b) = 0$. For $n = 4$, consider the two rankings $\alpha = (1, 2, 3, 4)$ and $\beta = (2, 1, 4, 3)$, for which $D_{1,n}(\alpha, \beta) = 4$ and $r_{F,n}(\alpha, \beta) = 0$.

Assume now that $n = 4l$ is a multiple of 4 greater than or equal to 8 and that we can find a and b in $S_{n-4} = S_{4l-4}$ so that $r_{F,n-4}(a, b)$ is zero. Without loss of generality we may assume $b = e_{n-4}$. Consider

$$\pi = (1, 2, \ldots, 4l - 1, 4l) \quad \text{and} \quad \sigma = (4l - 1, 1, \epsilon_1, \epsilon_2, \ldots, \epsilon_{4l-4}, 2, 4l),$$

where $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{4l-4}) = (a_1 + 2, a_2 + 2, \ldots, a_{4l-4} + 2)$, which is a rearrangement of $(3, 4, \ldots, 4l - 2)$. We have $\sum_{i=1}^{n-4} |\epsilon_i - (i + 2)| = 4l^2 - 8l + 4$ (by the induction hypothesis), so $D_{1,n}(\pi, \sigma) = 4l^2$, which implies $r_{F,n}(\pi, \sigma) = 0$.

Next we show by induction that if $n = 2k + 1$ ($k \in \mathbb{N}^*$), rankings a and b in S_n can be constructed so that $r_{F,n}(a, b) = 0$. For $n = 3$, consider the two rankings $a = (1, 2, 3)$ and $b = (2, 1, 3)$, for which $D_{1,n}(a, b) = 2$ and $r_{F,n}(a, b) = 0$. For $n = 5$, consider the two rankings $a = (1, 2, 3, 4, 5)$ and $b = (4, 1, 3, 2, 5)$, for which $D_{1,n}(a, b) = 6$ and $r_{F,n}(a, b) = 0$.

Assume now that $n = 2k + 1 \geq 5$ and that we may find a and b in $S_{n-4} = S_{2k-3}$ so that $r_{F,n-4}(a, b)$ is zero. Again assume $b = e_{n-4}$. Consider $\pi = e_{2k+1}$ and

$$\sigma = (2k, 2, \epsilon_1, \epsilon_2, \ldots, \epsilon_{2k-3}, 1, 2k + 1),$$

where $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{2k-3}) = (a_1 + 2, a_2 + 2, \ldots, a_{2k-3} + 2)$, which is a rearrangement of $(3, 4, \ldots, 2k - 1)$. We have $\sum_{i=1}^{n-4} |\epsilon_i - (i + 2)| = k^2 - 3k + 2$ (by the induction hypothesis), so $D_{1,n}(\pi, \sigma) = k^2 + k$, which implies $r_{F,n}(\pi, \sigma) = 0$. This completes the proof of the theorem. □

Next we investigate for which $n > 2$ there are at least two permutations for which Daniels' correlation coefficient $r_{U,n}$ can be zero.

Theorem 3 *For each integer $n \geq 3$, Daniels' correlation coefficient $r_{U,n}$ can be zero if and only if n is not of the form $n = 4m + 3$, where $m \in \mathbb{N}$.*

Proof. If $r_{U,n}(a, b) = 0$ for some $a, b \in S_n$, then $n(n - 1)(n - 2)/12$ is an integer, and thus $n \not\equiv 3 \pmod 4$. Conversely, assume first that $n \equiv 2 \pmod 4$ (where $n \geq 6$), and let $b = e_n$ and

$$a = \left(1, 2, 3, \ldots, \frac{n}{2}, n, n - 1, n - 2, \ldots, \frac{n + 2}{2}\right).$$

We have

$$I_n(a, b) = \frac{\frac{n}{2}\left(\frac{n}{2} - 1\right)}{2} \quad \text{and} \quad SQ_n(a, b) = \sum_{i=0}^{\frac{n}{2}-1} \left(\frac{n}{2} - 2i - 1\right)^2 = \frac{n^3 - 4n}{24}.$$

Using (12), we obtain $OT_n(a, b) = n(n - 1)(n - 2)/12$, which implies $r_{U,n}(a, b) = 0$.

Now we proceed with the cases $n \equiv 0 \pmod 4$ and $n \equiv 1 \pmod 4$ (where $n \in \mathbb{N}^*$ with $n \geq 4$). We will show by induction on n that there are two permutations whose Kendall and Daniels rank correlations are both zero. For the base case, consider the two rankings $(1, 2, 3, 4)$ and $(2, 4, 1, 3)$ for the first case, for which $r_{K,n} = r_{S,n} = r_{U,n} = 0$; while for the second case consider $(1, 2, 3, 4, 5)$ and $(2, 5, 3, 1, 4)$, for which $r_{K,n} = r_{S,n} = r_{U,n} = 0$.

Assume now that n is an integer greater than or equal to 8 and that it is possible to find a and b in S_{n-4} so that $r_{K,n-4}(a, b) = r_{U,n-4}(a, b) = 0$. Assume again that $b = e_{n-4}$. Consider $\pi = e_n$ and

$$\sigma = (n - 1, 1, \epsilon_1, \epsilon_2, \ldots, \epsilon_{n-4}, n, 2),$$

where $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{n-4}) = (a_1 + 2, a_2 + 2, \ldots, a_{n-4} + 2)$, which is a rearrangement of $(3, 4, \ldots, n-2)$. Since $r_{U,n-4}(a,b) = r_{K,n-4}(a,b) = 0$, we get from (10) that $6SQ_{1,n-4}(a,b) = (n-4)(n-3)(n-5)$. In addition, the increment in the sum of squares contributed by the integers $n-1, 1, n, 2$ is $2n^2 - 8n + 10$, which implies $6SQ_{1,n}(\pi,\sigma) = (n-1)n(n+1)$, i.e., $r_{S,n}(\pi,\sigma) = 0$. Furthermore, we have $r_{K,n}(\pi,\sigma) = 0$ because $I_{n-4}(a,b) = \frac{(n-4)(n-5)}{4}$ (by the induction hypothesis), and the increment in the number of inversions due to the numbers $n-1, 1, n, 2$ is $2n-5$. By (10), $r_{U,n}(\pi,\sigma) = 0$. This completes the induction. □

In the next theorem we investigate for which integers $n \geq 2$ the *modified Spearman's footrule* (*rank correlation coefficient*), defined by

$$r_{MF,n}(a,b) := 1 - \frac{2D_{2,n}(a,b)}{\max D_{2,n}} = 1 - \frac{6D_{2,n}(a,b)}{n(n-1)(n+1)}, \tag{23}$$

equals zero. (Because of (18) and (20), we have $r_{MF,n}(a,b) \leq r_{S,n}(a,b)$, i.e., the modified Spearman's footrule is always less than or equal to the Spearman rank correlation coefficient.)

Theorem 4 *For each integer $n \geq 2$, the modified Spearman's footrule $r_{MF,n}$ can be zero if and only if n is not of the form $4m+2$ where $m \in \mathbb{N}$.*

Proof. If $r_{MF,n}(a,b) = 0$, then $D_{2,n}(a,b) = n(n-1)(n+1)/6$. Because of (21), the integer $D_{2,n}(a,b)$ must be even. Thus n cannot be of the form $4m+2$ where $m \in \{0, 1, 2, \ldots, \}$.

Conversely, assume $n = 2k-1$, where $k \in \mathbb{N}^* - \{1\}$. Let $b = e_n$ and

$$a = (1, 3, \ldots, 2k-1, 2, 4, \ldots, 2k-2).$$

It follows that

$$(a_1 - b_1, a_2 - b_2, \ldots, a_n - b_n) = (0, 1, 2, \ldots, k-1, -(k-1), -(k-2), \ldots, -1).$$

Then

$$D_{2,n}(a,b) = \sum_{0 \leq i < j \leq k-1} (j-i) + \sum_{1 \leq i < j \leq k-1} (j-i) + \sum_{i=0}^{k-1} \sum_{j=1}^{k-1} (i+k-j).$$

Expanding the sums either by hand or using a symbolic computation package, we obtain $2D_{2,n}(a,b) = \max D_{2,n}$, which proves that $r_{MF,n}(a,b) = 0$.

Finally assume $n = 4m$ ($m \in \mathbb{N}^*$), and let $\sigma = e_n$ and

$$\pi = (1, 3, \ldots, 4m-1, 2, 4, \ldots, 2m-2, 4m, 2m+2, 2m+4, \ldots, 4m-2, 2m).$$

In other words, π consists of the odd integers in order followed by the even integers in order with the exception that $4m$ switches position with $2m$. Then

$$\begin{aligned}(\pi_1 - \sigma_1, \ldots, \pi_n - \sigma_n) =& (0, 1, \ldots, 2m-1, -(2m-1), \ldots, \\ & -(m+1), m, -(m-1), \ldots, -1, -2m).\end{aligned}$$

Using a symbolic computation package, a tedious but straightforward calculation (similar to the one above for the case $n = 2k-1$) yields $2D_{2,n}(\pi,\sigma) = \max D_{2,n}$, which implies $r_{MF,n}(\pi,\sigma) = 0$. This completes the proof of the theorem. □

6 A new inequality between three generalized metrics

In this section we prove a new inequality between the three generalized metrics $D_{2,n}$, I_n and OT_n for $n > 2$:

$$|D_{2,n}(a,b) - nI_n(a,b)| \leq 3\,OT_n(a,b) \quad (a,b \in S_n). \tag{24}$$

Equality holds if and only if a^{-1} is a cyclic permutation of b^{-1} (i.e., if and only if $OT_n(a,b) = 0$).

In terms of the corresponding rank correlation coefficients, after suppressing the dependence on a and b, when $n > 2$, inequality (24) becomes

$$|2(n+1)(1 - r_{MF,n}) - 3n(1 - r_{K,n})| \leq 3(n-2)(1 - r_{U,n}),$$

where the three coefficients are defined by Equations (8), (11) and (23). Equivalently, the above inequality can be written as

$$(n-2)(3r_{U,n} - 2) \leq 3nr_{K,n} - 2(n+1)r_{MF,n} \leq (n-2)(4 - 3r_{U,n}). \tag{25}$$

Fix $a \in S_n$ and without loss of generality assume $b = e_n$. To prove (24), for each pair of integers (i,j) with $1 \leq i < j \leq n$, we define:

$$\begin{aligned}
\mu_{ij} &:= \#\{k \in \mathbb{N} \cap [1,n] \,|\, i < j < k \text{ and } \mathrm{Perm}[a_i, a_j, a_k] \text{ is odd}\};\\
m_{ij} &:= \#\{k \in \mathbb{N} \cap [1,n] \,|\, i < k < j \text{ and } \mathrm{Perm}[a_i, a_k, a_j] \text{ is odd}\};\\
\lambda_{ij} &:= \#\{k \in \mathbb{N} \cap [1,n] \,|\, k < i < j \text{ and } \mathrm{Perm}[a_k, a_i, a_j] \text{ is odd}\}.
\end{aligned}$$

Recall from Section 2 that $\mathrm{Perm}[x,y,z]$ is the (ordered) list of ranks of the numbers x, y, z, i.e., $\mathrm{Perm}[x,y,z] \in S_3$. Note that

$$\sum_{1 \leq i < j \leq n} (\mu_{ij} + m_{ij} + \lambda_{ij}) = 3\,OT_n(a). \tag{26}$$

For the numbers x and y let $\delta(x,y) = 1$ if $x > y$, and 0 otherwise. We observe that

$$I_n(a) = \sum_{1 \leq i < j \leq n} \delta(a_i, a_j). \tag{27}$$

We claim (and prove later in the section) that for each pair of integers (i,j) with $1 \leq i < j \leq n$ we have

$$(a_i - a_j) - (i - j) = n\delta(a_i, a_j) + m_{ij} - \mu_{ij} - \lambda_{ij}. \tag{28}$$

Since $0 \leq \mu_{ij} + m_{ij} + \lambda_{ij} \leq n - 2$, the latter equality implies

$$n\delta(a_i,a_j) - (m_{ij} + \mu_{ij} + \lambda_{ij}) \leq |(a_i - a_j) - (i-j)| \leq n\delta(a_i,a_j) + (m_{ij} + \mu_{ij} + \lambda_{ij}). \tag{29}$$

Using (26)–(29) and the definition of $D_{2,n}(a)$ we get

$$nI_n(a) - 3\,OT_n(a) \leq D_{2,n}(a) \leq nI_n(a) + 3\,OT_n(a), \tag{30}$$

which entails inequality (24).

Proof of equality (28). (i) First assume $a_i < a_j$, in which case $\delta(a_i, a_j) = 0$. Since

$$\lambda_{ij} = \#\{k \in \mathbb{N} \cap [1, n] \,|\, k < i < j \text{ and } a_i < a_k < a_j\}$$

and

$$\mu_{ij} = \#\{k \in \mathbb{N} \cap [1, n] \,|\, i < j < k \text{ and } a_i < a_k < a_j\},$$

we have

$$(i - 1 - \lambda_{ij}) + (n - j - \mu_{ij}) = \#\{k \in \mathbb{N} \cap [1, n] \,|\, \text{not-}(i \le k \le j) \text{ and } \\ \text{not-}(a_i < a_k < a_j)\}.$$

Since in addition

$$m_{ij} = \#\{k \in \mathbb{N} \cap [1, n] \,|\, i < k < j \text{ and not-}(a_i < a_k < a_j)\}$$

and there are $(a_i - 1) + (n - a_j)$ integers $k \in [1, n] - \{i, j\}$ such that a_k is not between a_i and a_j, we have

$$(i - 1 - \lambda_{ij}) + (n - j - \mu_{ij}) + m_{ij} = (a_i - 1) + (n - a_j).$$

The last equality implies (28) for the case $a_i < a_j$.

(ii) In the case $a_i > a_j$ we have $\delta(a_i, a_j) = 1$ and a similar reasoning as in case (i) gives

$$(i - 1 - \lambda_{ij}) + (n - j - \mu_{ij}) + m_{ij} = a_i - a_j - 1,$$

the number of integers $k \in [1, n] - \{i, j\}$ such that a_k is between a_i and a_j. By simplifying the last equality, we obtain (28). □

Remark 1 Summing both sides of (28) over all integers (i, j) with $1 \le i < j \le n$ and using (22) and (27) we obtain

$$-\sum_{1 \le i < j \le n} (m_{ij} - \mu_{ij} - \lambda_{ij}) = nI_n(a) - SQ_{1,n}(a). \tag{31}$$

Subtracting (31) from (26) we get

$$2 \sum_{1 \le i < j \le n} m_{ij} = 3\, OT_n(a) - (nI_n(a) - SQ_{1,n}(a)).$$

On the other hand, it follows from the definition of $OT_n(a)$ that

$$OT_n(a) = \sum_{1 \le i < j \le n} m_{ij}.$$

If follows easily from the last two equalities that

$$OT_n(a) = nI_n(a) - SQ_{1,n}(a),$$

which gives another proof of (12) provided of course one has a proof of equality (28)!

Remark 2 Equality (12) implies $2SQ_{1,n}(a) \leq 2nI_n(a)$. Adding this one to inequality $SQ_{1,n}(a) \leq D_{2,n}(a)$ from Section 4 gives us

$$3SQ_{1,n}(a) \leq 2nI_n(a) + D_{2,n}(a).$$

Replacing $SQ_{1,n}(a)$ with $nI_n(a) - OT_n(a)$ in the above inequality, we obtain after some simple algebra

$$nI_n(a) - 3OT_n(a) \leq D_{2,n}(a),$$

which is the left inequality in (30).

7 Right-invariant generalized metrics for circular data

In this section we discuss right-invariant generalized metrics for angular (and in general, circular) data consisting of n pairs of angles

$$(\theta_1, \phi_1), (\theta_2, \phi_2), \ldots, (\theta_n, \phi_n). \tag{32}$$

As before, we assume that there are no ties among the θ values and no ties among the ϕ values. Finally, we assume that all the angles are in the interval $[0, 2\pi)$. To motivate our definition below about circular generalized metrics, we first prove the following result.

Lemma 1 *Let $n \in \mathbb{N}^* - \{1\}$ and assume the list $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ consists of n distinct numbers such that $0 \leq \phi_i < 2\pi$ for $i = 1, \ldots, n$. If $0 < c < 2\pi$, and $\boldsymbol{\phi'} = (\phi'_1, \ldots, \phi'_n)$ with $\phi'_i = \phi_i + c \pmod{2\pi}$ for $i = 1, \ldots, n$, then there is an integer $m \in \{0, 1, \ldots, n-1\}$ such that*

$$\mathrm{Perm}(\boldsymbol{\phi'}) = g_n^m \circ \mathrm{Perm}(\boldsymbol{\phi}), \tag{33}$$

where $g_n := (2, 3, \ldots, n, 1)$.

Proof. Let $a := [\mathrm{Perm}(\boldsymbol{\phi})]^{-1} = (a_1, \ldots, a_n) \in S_n$ and $b := [\mathrm{Perm}(\boldsymbol{\phi'})]^{-1} = (b_1, \ldots, b_n) \in S_n$ be the inverses of the lists of ranks of vectors $\boldsymbol{\phi}$ and $\boldsymbol{\phi'}$, respectively. Then the lists of ordered statistics for $\boldsymbol{\phi}$ and $\boldsymbol{\phi'}$ are $(\phi_{a_1}, \ldots, \phi_{a_n})$ and $(\phi'_{b_1}, \ldots, \phi'_{b_n})$, respectively. We have

$$c \leq \phi_{a_1} + c < \ldots < \phi_{a_n} + c < 2\pi + c < 4\pi.$$

If $c < 2\pi \leq \phi_{a_1} + c$, then $\phi'_i = (\phi_i + c) \mod 2\pi = \phi_i + c - 2\pi$ for $i = 1, \ldots, n$, in which case $\mathrm{Perm}(\boldsymbol{\phi}) = \mathrm{Perm}(\boldsymbol{\phi'})$, that is, equality (33) holds trivially with $m = 0$. Similarly, if $\phi_{a_n} + c \leq 2\pi < 2\pi + c$, we have again $\mathrm{Perm}(\boldsymbol{\phi}) = \mathrm{Perm}(\boldsymbol{\phi'})$, and equality (33) holds again trivially with $m = 0$.

Assume $\phi_{a_k} + c \leq 2\pi < \phi_{a_{k+1}} + c$ for some $k \in \{1, \ldots, n-1\}$. Then

$$\phi_{a_1} + c - 2\pi < \ldots < \phi_{a_k} + c - 2\pi \leq 0 < \phi_{a_{k+1}} + c - 2\pi < \ldots < \phi_{a_n} + c - 2\pi < c,$$

and thus

$$\phi'_{a_1} - 2\pi < \ldots < \phi'_{a_k} - 2\pi < 0 < \phi'_{a_{k+1}} < \ldots < \phi'_{a_n} < c.$$

Since $\phi_{a_n} + c - 2\pi < \phi_{a_1} + c$, that is, $\phi'_{a_n} < \phi'_{a_1}$, we obtain

$$0 < \phi'_{a_{k+1}} < \ldots < \phi'_{a_n} < \phi'_{a_1} < \ldots < \phi'_{a_k} < 2\pi,$$

which entails

$$b = (b_1, \ldots, b_n) = (a_{k+1}, \ldots, a_n, a_1, \ldots, a_k) = a \circ g_n^k.$$

This implies $b^{-1} = g_n^{-k} \circ a^{-1}$, i.e., $\text{Perm}(\boldsymbol{\phi'}) = g_n^{n-k} \circ \text{Perm}(\boldsymbol{\phi})$. This means that equality (33) holds with $m = n - k \in \{1, \ldots, n-1\}$. □

The previous lemma tells us that if we rotate the origin clockwise by c radians, then the ranks of the new values of the angles in ϕ are connected to the ranks of the angles of the original angles through Equation (33). If without loss of generality, we assume the ranks of the angles $(\theta_1, \ldots, \theta_n)$ are $(1, 2, \ldots, n) = e_n$, then we want any right-invariant generalized metric sequences $(d_n | n \in \mathbb{N}^*)$ applied to the ranks of the data in (32) to satisfy

$$d_n(\text{Perm}(\boldsymbol{\phi}), e_n) = d_n(\text{Perm}(\boldsymbol{\phi'}), e_n) = d_n(g_n^m \circ \text{Perm}(\boldsymbol{\phi}), e_n)$$

for all $n \in \mathbb{N}^* - \{1\}$ and all $m \in \{0, 1 \ldots, n-1\}$. This leads us to the following definition:

Definition 1 A right-invariant generalized metric sequence $(d_n | n \in \mathbb{N}^*)$ of functions $d_n : S_n \times S_n \to [0, \infty)$ is called **circular** if

$$d_n(a, b) = d_n(g_n^m \circ a, g_n^s \circ b) \tag{34}$$

for all $n \in \mathbb{N}^* - \{1\}$, $m, s \in \{0, 1 \ldots, n-1\}$, and $a, b \in S_n$ (where $g_n = (2, 3, \ldots, n, 1)$).

It is clear that a right-invariant generalized metric sequence $(d_n | n \in \mathbb{N}^*)$ is circular if and only if $d_n(g_n^m \circ a) = d_n(a)$ for all $n \in \mathbb{N}^* - \{1\}$, $m \in \{0, 1 \ldots, n-1\}$ and $a \in S_n$. Such a sequence obviously does not satisfy Property 4 (in Section 1), so a circular generalized metric sequence cannot be a pseudo-metric sequence.

The Daniels-Guilbaud semi-metric sequence $(OT_n | n \in \mathbb{N}^*)$ is indeed circular. This follows directly from the following formula proved in Shieh [36] and Shieh et al. [37]:

$$r_{U,n}(a, b) = \frac{3}{n(n-2)} \sum_{m=0}^{n-1} \sum_{s=0}^{n-1} r_{K,n}(g_n^m \circ a, g_n^s \circ b).$$

In terms of the corresponding semi-metrics, this formula becomes:

$$OT_n(a, b) = \frac{1}{n} \sum_{m=0}^{n-1} \sum_{s=0}^{n-1} I_n(g_n^m \circ a, g_n^s \circ b) - \frac{(n-1)n(n+1)}{6}.$$

8 Creation of circular generalized metrics

In this section we show how Critchlow's [4] theory can be used so that a right-invariant generalized metric sequence $(d_n|n \in \mathbb{N}^*)$ can be transformed into a circular one. Even though Critchlow's [4] theory was applied for analyzing partially ranked data, it can be easily modified for our purpose.

For each $n \in \mathbb{N}^*$, consider the cyclic group

$$C_n := \{g_n^m \,|\, m \in \{0, 1 \ldots, n-1\} \text{ and } g_n := (2, 3, \ldots, n, 1)\},$$

which is a subgroup of the symmetric group S_n of order n. For each permutation $a \in S_n$ consider the right coset

$$C_n a := \{g_n^m \circ a \,|\; m \in \{0, 1 \ldots, n-1\} \text{ and } g_n := (2, 3, \ldots, n, 1)\},$$

and the coset (quotient) space S_n/C_n:

$$S_n/C_n := \{C_n a \,|\, a \in S_n\}.$$

Essentially, S_n/C_n partitions S_n into $(n-1)!$ equivalent classes, each of size n. The idea is to create (through a certain process) a right-invariant generalized metric sequence $(\breve{d}_n|n \in \mathbb{N}^*)$ on S_n/C_n from the given right-invariant generalized metric sequence $(d_n|n \in \mathbb{N}^*)$ on S_n. We can then define a circular right-invariant generalized metric sequence $(\mathring{d}_n|n \in \mathbb{N}^*)$ on S_n by

$$\mathring{d}_n(a, b) := \breve{d}_n(C_n a, C_n b) \tag{35}$$

for $n \in \mathbb{N}^*$ and $a, b \in S_n$. Such a generalized metric does satisfy Equation (34) in Definition 1. One can then define the corresponding circular rank correlation coefficient as follows:

$$\mathring{r}_{d,n}(a, b) := 1 - \frac{2\mathring{d}_n(a, b)}{\max(\mathring{d}_n)} \quad (a, b \in S_n).$$

By right invariance, the denominator in the above formula equals

$$\max(\mathring{d}_n) = \max\{\mathring{d}_n(a, b) \,|\, a, b \in S_n\} = \max\{\breve{d}_n(C_n a, C_n e_n) \,|\, a \in S_n\}.$$

The discrete maximization above, however, is a difficult combinatorial problem even for circular generalized metrics $\mathring{d}_n$ induced by quite simple generalized metrics d_n.

But the main question is how to define the right-invariant generalized metric sequence $(\breve{d}_n|n \in \mathbb{N}^*)$ on S_n/C_n from $(d_n|n \in \mathbb{N}^*)$. The most often used and simplest method of achieving that is through the Hausdorff generalized metric as modified by Dieudonné [10, pp. 53-54] and introduced to statisticians by Critchlow [4, pp. 14-26]:

$$\breve{d}_n(C_n a, C_n b) := \max\left\{ \max_{c \in C_n a} \min_{f \in C_n b} d_n(c, f), \max_{f \in C_n b} \min_{c \in C_n a} d_n(c, f) \right\} \tag{36}$$

for $a, b \in S_n$ (or equivalently, for cosets $C_n a, C_n b \in S_n / C_n$). Observe that, for each $c \in C_n a$, the quantity $\min_{f \in C_n b} d_n(c, f)$ is the distance of permutation c to the coset $C_n b$. Then $\max_{c \in C_n a} \left(\min_{f \in C_n b} d_n(c, f) \right)$ is the worst (largest) of these distances to the coset $C_n b$. Similarly, $\max_{f \in C_n b} \left(\min_{c \in C_n a} d_n(c, f) \right)$ is the largest distance to the coset $C_n a$ from permutations in $C_n b$. The Hausdorff distance between cosets $C_n a$ and $C_n b$ is the largest of these two maximum distances.

It can be proven that the induced sequence of Hausdorff generalized metrics $(\breve{d}_n \,|\, n \in \mathbb{N}^*)$ satisfies Property 3 with the same positive constant C as the sequence $(d_n \,|\, n \in \mathbb{N}^*)$.

The calculation of the Hausdorff generalized "distance" through Equation (36) is a formidable task and (in this paper) we shall only concentrate on right-invariant generalized metrics that are also *left-invariant* because, in this case, as proven in a lemma in Critchlow [4, pp. 21–24],

$$\breve{d}_n(C_n a, C_n b) = \min\{d_n(c, f) \,|\, c \in C_n a,\ f \in C_n b\} = \min_{c \in C_n a} d_n(c, b). \tag{37}$$

Critchlow attributes this result and its proof to Diaconis and Graham [9]. See also [7, Chapter 6D]. We remind the reader that $d_n : S_n \times S_n \to [0, \infty)$ is left-invariant if and only if

$$d_n(c \circ a, c \circ b) = d_n(a, b) \tag{38}$$

for all $a, b, c \in S_n$.

Imitating Equation (6) from Section 1, if $(d_n | n \in \mathbb{N}^*)$ is a right-invariant generalized metric sequence of functions $d_n : S_n \times S_n \to [0, \infty)$ that are not necessarily left-invariant on the symmetric group, then one can create a corresponding bi-invariant generalized metric sequence $(d_n^{bi} | n \in \mathbb{N}^*)$ as follows:

$$d_n^{bi}(a, b) = M \sum_{c \in S_n} d_n(c \circ a, c \circ b) \quad (a, b \in S_n), \tag{39}$$

where M is an arbitrary positive constant, e.g., $M = 1/n!$. Equations (6) and (39), however, involve a lot of computational cost, so we would stick with the two natural bi-invariant (generalized) metrics in this paper: H_n and EX_n.

Using Equations (35) and (37), we can define the Hamming circular semi-metric as follows:

$$\mathring{H}_n(a, b) = \breve{H}_n(C_n a, C_n b) = \min_{c \in C_n a} H_n(c, b) \quad (a, b \in S_n).$$

For example, if $a = (4, 3, 2, 1)$ and $b = (3, 2, 1, 4)$, then

$$C_4 a = \{(4, 3, 2, 1), (1, 4, 3, 2), (2, 1, 4, 3), (3, 2, 1, 4)\}$$

and $\mathring{H}_4(a, b) = \min\{4, 4, 4, 0\} = 0$. On the other hand, if $b = (3, 2, 4, 1)$, then $\mathring{H}_4(a, b) = \min\{3, 4, 3, 2\} = 2$.

Because of the bi-invariance of H_n, the Hamming circular semi-metric $\mathring{H}_n(a,b)$ can be calculated using left cosets as well:

$$\mathring{H}_n(a,b) = \hat{H}_n(aC_n, bC_n) = \min_{c\in aC_n} H_n(c,b) \quad (a,b \in S_n),$$

where $\hat{H} : (C_n\backslash S_n) \times (C_n\backslash S_n) \to [0,\infty)$ can be defined via a modification of (36) on the set[10] $C_n\backslash S_n$ of left cosets

$$aC_n := \{a \circ g_n^m \mid m \in \{0,1\ldots,n-1\} \text{ and } g_n := (2,3,\ldots,n,1)\} \tag{40}$$

$$= \{(a_1,a_2,\ldots,a_n),(a_2,a_3,\ldots,a_n,a_1),\ldots,(a_n,a_1,\ldots,a_{n-1})\}. \tag{41}$$

Since a_1 appears in all possible positions in the members of aC_n, there is $c \in aC_n$ such that $c_{a_1} = a_1$. In such a case, $\mathring{H}_n(c,e_n) \neq n$, and thus

$$0 \leq \max \mathring{H}_n \leq n-1.$$

We are now ready to define the Hamming circular rank correlation coefficient:

$$\mathring{r}_{H,n}(a,b) := 1 - \frac{2\min_{c\in C_n a} H_n(c,b)}{\max(\mathring{H}_n)} = 1 - \frac{2\min_{c\in C_n a} H_n(c,b)}{n - \frac{1}{2}(3+(-1)^n)}.$$

The denominator in the above expression is $n-1$ when n is odd and $n-2$ when n is even (see below). For the first example above (with $a = (4,3,2,1)$ and $b = (3,2,1,4)$) we have $\mathring{r}_{H,4}(a,b) = 1$, while for the second example (with the same a but $b = (3,2,4,1)$) we have $\mathring{r}_{H,4}(a,b) = -1$.

Since $\mathring{H}_n(\overline{e}_n, e_n) = n-1$ when n is odd, it is obvious that (in this case) $\max \mathring{H}_n = n-1$. The case n being even is more complicated. To prove that (in such a case) $\max \mathring{H}_n = n-2$, we need the following result due to Chris Monico [31]:

Lemma 2 *Assume $n = 2k$ with $k \in \mathbb{N}^*$ and let $a = (a_1,\ldots,a_{2k}) \in S_{2k}$. If $g_{2k} = (2,3,\ldots,2k,1)$, at least one of the permutations in the left coset aC_{2k}, given by* (40) *or* (41)*, has at least two fixed points.*

Proof. Consider the $2k$ integers $\alpha_i := a_i - i \,(\mathrm{mod}\ 2k)$, $i = 1,2,\ldots,2k$, with $0 \leq \alpha_i < 2k$. If these integers were distinct, then

$$\{\alpha_1,\ldots,\alpha_{2k}\} = \{0,1,\ldots,2k-1\}$$

and

$$0 = \sum_{i=1}^{2k}(a_i - i) = \sum_{i=1}^{2k}\alpha_i = \sum_{j=0}^{2k-1} j = 2k^2 - k = k \,(\mathrm{mod}\ 2k),$$

a contradiction because $k \neq 0 \,(\mathrm{mod}\ 2k)$. Therefore there are two integers l_1 and l_2 such that $1 \leq l_1 < l_2 \leq 2k$ and $a_{l_1} - l_1 = a_{l_2} - l_2 \,(\mathrm{mod}\ 2k)$. Then

$$a_{l_1} = l_1 + r + 2ks \quad \text{and} \quad a_{l_2} = l_2 + r + 2ks$$

[10] To agree with Critchlow [4, p. 15], we are forced to use the opposite notation from the one used by Dieudonné [10, pp. 52-55].

for some $r, s \in \mathbb{Z}$ with $r \in \{0, 1, \ldots, 2k-1\}$. Since $l_1 < l_2$, we have

$$1 \leq a_{l_1} = l_1 + r + 2ks < a_{l_2} = l_2 + r + 2ks \leq 2k,$$

from which we obtain

$$1 - (2k-1) - (2k-1) \leq 1 - l_1 - r \leq 2ks \leq 2k - l_2 - r \leq 2k - 2 - 0.$$

This entails

$$-2 + \frac{3}{2k} \leq s \leq 1 - \frac{1}{k}.$$

It follows that $s = -1$ or $s = 0$.

If $s = 0$, then $1 \leq a_{l_1} = l_1 + r < a_{l_2} = l_2 + r \leq 2k$ and $1 \leq l_1 < l_2 \leq 2k - r$. Since

$$a \circ g_{2k}^{2k-r} = (a_{2k+1-r}, \ldots, a_{2k}, a_1, \ldots, a_{2k-r}),$$

it follows that the element of $a \circ g_{2k}^{2k-r}$ in position $r + l_1$ is $a_{l_1} = r + l_1$ and the element of $a \circ g_{2k}^{2k-r}$ in position $r + l_2$ is $a_{l_2} = r + l_2$, i.e., $a \circ g_{2k}^{2k-r}$ has at least two fixed points.

If $s = -1$ then $1 \leq a_{l_1} = l_1 + r - 2k < a_{l_2} = l_2 + r - 2k \leq 2k$, and the element of $a \circ g_{2k}^{2k-r}$ in position $l_i + r - 2k$ is $a_{l_i} = l_i + r - 2k$ for $i = 1, 2$. Again, $a \circ g_{2k}^{2k-r}$ has at least two fixed points and the proof of the lemma is complete. □

Lemma 2 implies that, for each $a \in S_{2k}$, there is $c_0 \in aC_{2k}$ such that $H_{2k}(c_0, e_{2k}) \leq 2k - 2$. Thus

$$\mathring{H}_{2k}(a, e_{2k}) = \min_{c \in aC_{2k}} H_{2k}(c, e_{2k}) \leq 2k - 2. \tag{42}$$

To show that there is an $a \in S_n$ so that equality can be achieved in (42), consider the left coset $\overline{e}_{2k} C_{2k}$. Each permutation in this coset that starts with an even number has no fixed points because even integers occupy odd positions and vice versa. On the other hand, each permutation in the coset that starts with an odd integer, say $2\lambda - 1$ (where $\lambda \in \{1, 2, \ldots, k\}$), has exactly two fixed points, in positions λ and $\lambda + k$. This shows that $\mathring{H}_{2k}(\overline{e}_{2k}, e_{2k}) = 2k - 2$, which proves that $\max \mathring{H}_{2k} = 2k - 2$.

Another bi-invariant metric on S_n is EX_n, which induces the following Cayley circular semi-metric:

$$\mathring{EX}_n(a, b) = \breve{EX}_n(aC_n, bC_n) = \min_{c \in aC_n} EX_n(c, b) = n - \max_{c \in aC_n} CY_n(c \circ b^{-1})$$

for $a, b \in S_n$. Here, $CY_n(f)$ is the number of cycles in list $f \in S_n$. In order to be able to define the corresponding Cayley circular rank correlation coefficient, we would need to know a general formula for $\max \breve{EX}_n$ over $(S_n/C_n) \times (S_n/C_n)$. This is a topic for a future paper as is the analysis for other generalized metrics used in this paper (most of which are not bi-invariant).

9 Partial rank correlation coefficients

If $(d_n | n \in \mathbb{N}^*)$ is a generalized metric sequence with $\max d_n > 0$ for each $n \in \mathbb{N}^*$ with $n \geq N$ (for some $N \in \mathbb{N}^*$), Property 3 from Section 1 can be expressed in terms of the corresponding rank correlation coefficient sequence:

$$r_{d,n}(a, c) + r_{d,n}(c, b) \leq 2 - \frac{1}{C} + \frac{r_{d,n}(a, b)}{C} \quad (a, b, c \in S_n). \tag{43}$$

Even though in Section 1 we let $C > 0$, for the rest of the section we will assume that $C \geq 1$. If $C = 1$, $(d_n | n \in \mathbb{N}^*)$ becomes a sequence of metrics and inequality (43) simplifies to

$$r_{d,n}(a,c) + r_{d,n}(c,b) \leq 1 + r_{d,n}(a,b). \tag{44}$$

It is not clear what the meaning of inequalities (43) and (44) is. Even Diaconis [7, pp. 103-104] wonders about the meaning of (44). In this section, we shall attempt to use (44) to define *partial rank correlation coefficients.*

Given three numerical variables x, y, z (such as height, arm length and weight), one can take measurements on n individuals or objects:

$$(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n). \tag{45}$$

One can measure the linear relationship between the x and y controlling for z using the following formula (see [1], [18], and [29]):

$$r_n[(x,y) \bullet z] = \frac{r_n(x,y) - r_n(x,z) r_n(y,z)}{\sqrt{[1 - r_n(x,z)^2][1 - r_n(y,z)^2]}}, \tag{46}$$

where $r_n(w_1, w_2)$ denotes the Pearson product moment correlation coefficient between numerical variables w_1 and w_2, defined by Equation (2) in Section 1.

We call the quantity $r_n[(x,y) \bullet z]$ the *partial*[11] *linear correlation coefficient between variables x and y controlling for z* and originally it is defined as follows. One runs a linear regression between x and z and a linear regression between y and z and then calculates the residuals from each linear regression. The quantity $r_n[(x,y) \bullet z]$ is defined to be the Pearson product moment correlation coefficient between the two sets of residuals. Formula (46) follows from this definition.

If we replace the data (46) with their ranks[12]

$$(a_1, b_1, c_1), (a_2, b_2, c_2), \ldots, (a_n, b_n, c_n), \tag{47}$$

they we may calculate the *Spearman partial rank correlation coefficient* between the two variables controlling for the third one:

$$\tilde{r}_{S,n}[(a,b) \bullet c] = \frac{r_{S,n}(a,b) - r_{S,n}(a,c) r_{S,n}(b,c)}{\sqrt{[1 - r_{S,n}(a,c)^2][1 - r_{S,n}(b,c)^2]}}. \tag{48}$$

Here we use $r_{S,n}(\cdot,\cdot)$ to denote the Spearman rank correlation coefficient given by Equation (3) (or equivalently by (4)).

It is not clear how formulas (46) and (48) can be used to define partial rank correlation coefficient corresponding to other generalized metrics (besides $SQ_{1,n}$ that

[11] It should be noted, however, that this is *not* a conditional linear correlation between x and y given z unless certain conditions hold; see [1] and [29]. This is the reason we do not use the notation $r_n(x, y|z)$.

[12] Here $a = (a_1, \ldots, a_n)$ are the ranks of the x values, $b = (b_1, \ldots, b_n)$ are the ranks for the y values, and $c = (c_1, \ldots, c_n)$ are the ranks for the z values. We assume that there are no ties in the x values, and similarly for the y values and the z values.

gives rise to (48)). Even though one can define a kind of "linear regression" based on ranks using generalized metrics[13] (see [7, pp. 106-107] and [35]), we propose a different method for creating partial rank correlation coefficients induced by metrics (with $C = 1$) that are symmetric with respect to complements.

A right-invariant generalized metric sequence $(d_n | n \in \mathbb{N}^*)$ is called *symmetric with respect to complements* if

$$d_n(a) + d_n(\overline{a}) = \max d_n \tag{49}$$

for all $n \in \mathbb{N}^*$ and $a \in S_n$. Here

$$\overline{a} = \overline{e}_n \circ a = (n + 1 - a_1, n + 1 - a_2, \ldots, n + 1 - a_n)$$

is the *complement or conjugate*[14] of list a. Note that, even for generalized metrics that do not satisfy (49), Property 3 in Section 1 (along with right invariance) implies

$$d_n(a) + d_n(\overline{a}) \geq \frac{d(\overline{e}_n)}{C}.$$

In case $d(\overline{e}_n) = \max d_n$ (which happens, for example, when d_n is $SQ_{1,n}$, I_n or OT_n), we then have

$$d_n(a) + d_n(\overline{a}) \geq \frac{\max d_n}{C}.$$

If $d(\overline{e}_n) = \max d_n$ (for all $n \in \mathbb{N}^*$) and $C = 1$ (e.g., when $d_n = \sqrt{SQ_{1,n}}$), then

$$d_n(a) + d_n(\overline{a}) \geq \max d_n.$$

If $(d_n | n \in \mathbb{N}^*)$ is a generalized metric sequence that is symmetric with respect to complements and $(r_{d,n} | n \geq N)$ is the corresponding induced rank correlation coefficient sequence (for some $N \in \mathbb{N}^*$), then we have

$$r_{d,n}(\overline{a}) = -r_{d,n}(a) \quad (n \geq N,\ a \in S_n). \tag{50}$$

Note that (50) and the right invariance of $r_{d,n}$ imply

$$r_{d,n}(\overline{a}, b) = -r_{d,n}(a, b) \quad \text{and} \quad r_{d,n}(\overline{a}, \overline{b}) = r_{d,n}(a, b) \tag{51}$$

for all $n \geq N$ and $a, b \in S_n$.

Examples of generalized metric sequences $(d_n | n \in \mathbb{N}^*)$ that are symmetric with respect to complements are $(SQ_{1,n} | n \in \mathbb{N}^*)$, $(I_n | n \in \mathbb{N}^*)$ and $(OT_n | n \in \mathbb{N}^*)$. The last two are semi-metrics for they satisfy Property 3 with $C = 1$, and we concentrate for the rest of the section on such examples only. (We do, however, briefly discuss the

[13] If this is possible and reasonable, then one can define $r_n[(x, y) \bullet z]$ from any generalized metric by using the procedure used above for creating the usual $r_n[(x, y) \bullet z]$ and by replacing the usual linear regression with the linear regression based on ranks with "distance" defined through the generalized metric.

[14] See Section 4 in [21], Section 5 in [22], Section 4 in [23] and pp. 11–12 in [27].

general case at the end of the section.) For such cases, for $a, b, c \in S_n$, we define[15] the *partial rank correlation coefficient of a and b controlling for c* by

$$r_{d,n}[(a,b) \bullet c] := \frac{r_{d,n}(a,b) - r_{d,n}(a,c) r_{d,n}(b,c)}{\min \Omega_{d,n}[(a,b) \bullet c]},$$

where

$$\Omega_{d,n}[(a,b) \bullet c] := \{[1 - r_{d,n}(a,c)][1 - r_{d,n}(b,c)],\ [1 + r_{d,n}(a,c)][1 + r_{d,n}(b,c)]\}$$

if $r_{d,n}(a,b) - r_{d,n}(a,c) r_{d,n}(b,c) \leq 0$, and

$$\Omega_{d,n}[(a,b) \bullet c] := \{[1 - r_{d,n}(a,c)][1 + r_{d,n}(b,c)],\ [1 + r_{d,n}(a,c)][1 - r_{d,n}(b,c)]\}$$

if $r_{d,n}(a,b) - r_{d,n}(a,c) r_{d,n}(b,c) \geq 0$.

It is interesting that, in either case, the geometric mean of the two numbers in the set $\Omega_{d,n}[(a,b) \bullet c]$ is equal to the denominator of the formula for the Spearman partial rank correlation coefficient in (48)–see also Remark 3 below. The fact that $-1 \leq r_{d,n}[(a,b) \bullet c] \leq 1$ follows from inequality (44)–valid when $C = 1$–and equalities (51).

It is clear that the partial rank correlation coefficient $r_{d,n}[(\cdot,\cdot) \bullet \cdot] : (S_n \times S_n) \times S_n \to [-1, 1]$ (as defined above) is right-invariant, i.e.,

$$r_{d,n}[(a \circ f, b \circ f) \bullet c \circ f] = r_{d,n}[(a,b) \bullet c]$$

for all $n \geq N$ and $a, b, c, f \in S_n$ (for which both sides of the above equality make sense). We can also easily prove that

$$r_{d,n}[(\overline{a}, b) \bullet c] = -r_{d,n}[(a,b) \bullet c], \quad r_{d,n}[(\overline{a}, \overline{b}) \bullet c] = r_{d,n}[(a,b) \bullet c],$$
$$\text{and} \quad r_{d,n}[(a,b) \bullet \overline{c}] = r_{d,n}[(a,b) \bullet c].$$

Remark 3 Suppose that given a metric sequence[16] $(d_n | n \in N^*)$, we define the *Spearman-induced partial rank correlation coefficient of a and b controlling for c* as follows:

$$R_{d,n}[(a,b) \bullet c] := \frac{r_{d,n}(a,b) - r_{d,n}(a,c) r_{d,n}(b,c)}{\sqrt{[1 - r_{d,n}(a,c)^2][1 - r_{d,n}(b,c)^2]}} \tag{52}$$

for all $n \in \mathbb{N}^*$ and $a, b, c \in S_n$ such that $0 < d_n(f,g) < \max d_n$ for $(f,g) \in \{(a,b), (a,c), (b,c)\}$. In other words, $R_{d,n}[(a,b) \bullet c]$ is obtained from formula (48) by replacing $r_{S,n}$ with $r_{d,n}$. Since for $x, y > 0$ we have $\min(x,y) \leq \sqrt{xy}$, we obtain

$$|R_{d,n}[(a,b) \bullet c]| \leq |r_{d,n}[(a,b) \bullet c]| \leq 1,$$

which shows that (indeed) $R_{d,n}[(a,b) \bullet c] \in [-1, 1]$. We also have

$$R_{d,n}[(\overline{a}, b) \bullet c] = -R_{d,n}[(a,b) \bullet c],\ R_{d,n}[(\overline{a}, \overline{b}) \bullet c] = R_{d,n}[(a,b) \bullet c],$$

[15] It is clear that we have to assume that $-1 < r_{d,n}(a,c) < 1$ and $-1 < r_{d,n}(b,c) < 1$ (i.e, $0 < d_n(a,c) < \max d_n$ and $0 < d_n(b,c) < \max d_n$), otherwise the definition is meaningless.

[16] That is, we assume $(d_n | n \in \mathbb{N}^*)$ satisfies Property 3 with $C = 1$.

and $R_{d,n}[(a,b) \bullet \overline{c}] = R_{d,n}[(a,b) \bullet c]$.

Note that, if $a, b, c \in S_n$ are the lists of ranks for n triplets of observations from the numerical variables x, y, z, respectively–see (45) and (47)–and $r_{d,n}$ is the Kendall tau correlation coefficient (i.e., $d_n = I_n$ and $r_{d,n} = r_{K,n}$), then Equation (52) has been used for many decades to define "Kendall's partial tau correlation coefficient for x and y when z is held constant" (see, for example, [19, Chapter 5]).

Remark 4 For generalized metric sequences $(d_n | n \in \mathbb{N}^*)$ that satisfy Property 3 for a general $C \geq 1$ and are symmetric with respect to complements, inequality (43) allows us to define a partial rank correlation coefficient in a way that generalizes the definition of $r_{d,n}[(\cdot,\cdot) \bullet \cdot]$ above (for metric sequences). We let

$$r_{d,n}[(a,b) \bullet c] := \frac{(2C-1)r_{d,n}(a,b) - C^2 r_{d,n}(a,c) r_{d,n}(b,c)}{\min A_{d,n}[(a,b) \bullet c; C]},$$

where[17]

$$\begin{aligned} A_{d,n}[(a,b) \bullet c; C] : \;=\; & \{[(2C-1) - C r_{d,n}(a,c)][(2C-1) - C r_{d,n}(b,c)], \\ & [(2C-1) + C r_{d,n}(a,c)][(2C-1) + C r_{d,n}(b,c)]\} \end{aligned}$$

if $(2C-1)r_{d,n}(a,b) - C^2 r_{d,n}(a,c) r_{d,n}(b,c) \leq 0$, and

$$\begin{aligned} A_{d,n}[(a,b) \bullet c; C] : \;=\; & \{[(2C-1) - C r_{d,n}(a,c)][(2C-1) + C r_{d,n}(b,c)], \\ & [(2C-1) + C r_{d,n}(a,c)][(2C-1) - C r_{d,n}(b,c)]\} \end{aligned}$$

if $(2C-1)r_{d,n}(a,b) - C^2 r_{d,n}(a,c) r_{d,n}(b,c) \geq 0$. Even though inequality (43) and equalities (51) can be used to show that $-1 \leq r_{d,n}[(a,b) \bullet c] \leq 1$, the above definition (for a general $C \geq 1$) is not as elegant as the one for the case $C = 1$ nor does it have any connection to the Spearman-induced partial rank correlation coefficient defined in Remark 3 above, which resembles the one in formula (48)[18]. In addition, if $d_n = SQ_{1,n}$ and $C = 2$, then $r_{d,n}[(\cdot,\cdot) \bullet \cdot]$ (as defined in this remark) clearly does not equal the Spearman partial rank correlation coefficient given in (48)[19].

10 Future research

The research in this paper reveals many open problems that require study and careful examination. For example, if all permutation pairs (a, b) in $S_n \times S_n$ are assumed to be equally likely and the sample size n goes to infinity, what is the asymptotic distribution of the right-invariant generalized metrics (and their corresponding rank

[17] Using a notation introduced earlier in the section, $A_{d,n}[(a,b) \bullet c; C = 1] = \Omega_{d,n}[(a,b) \bullet c]$.

[18] One can generalize $R_{d,n}[(\cdot,\cdot) \bullet \cdot]$ in (52) for generalized metrics $(d_n | n \in \mathbb{N}^*)$ as well that satisfy Property 3 for a general $C \geq 1$ and are symmetric with respect to complements. The resulting formula, however, is as not as elegant as the one for $C = 1$. To avoid confusion, we omit the relevant discussion.

[19] This is the reason we put a tilde over $r_{S,n}$ in (48) so that we can distinguish the usual Spearman partial rank correlation coefficient from the one defined in this remark with $d_n = SQ_{1,n}$ and $C = 2$.

correlation coefficients) studied in this paper? For most of the "old" generalized metrics, the asymptotic theory is well-established–see, for example, [8, 19, 36, 37].

Even if a general asymptotic theory is not feasible, it would be interesting to study the asymptotic properties of the new right-invariant metric $D_{2,n}$ defined by (16) (and those of the corresponding rank correlation coefficient $r_{MF,n}$, defined by (23)).

The creation of circular generalized metrics using Critchlow's theory (via the Hausdorff generalized metric) was studied in Section 8, but a detailed analysis was given only for the Hamming circular rank correlation coefficient. It will be interesting to investigate the properties of other circular rank correlation coefficients generated in the same way.

Finally, the partial rank correlation coefficients introduced in Section 9 require further analysis and how they relate to the ones introduced in Remark 3 (e.g., one may study which one of the two has the smallest asymptotic variance).

References

[1] K. Baba, R. Shibata and M. Sibuya, Partial correlation and conditional correlation as measures of conditional independence, *Aust. N. Z. J. Stat.* 46 (2004), 657–664.

[2] H. W. Block, D. Chhetry, Z. Fang and A. R. Sampson, Metrics on permutations useful for positive dependence, *J. Statist. Plann. Inference* 62 (1997), 219–234.

[3] A. Cayley, A note on the theory of permutations, *Phil. Mag.* 34 (1849), 527–529.

[4] D. E. Critchlow, *Metric methods for analyzing partially ranked data*, Springer-Verlag, Berlin (1985).

[5] H. E. Daniels, Rank correlation and population models (with discussion), *J. Roy. Statist. Soc. Ser. B* 12 (1950), 171–191.

[6] M. M. Deza and E. Deza, *Encyclopedia of distances*, Springer, New York (2009).

[7] P. Diaconis, *Group representations in probability and statistics*, Institute of Mathematical Statistics, Hayward, California (1988).

[8] P. Diaconis and R. L. Graham, Spearman's footrule as a measure of disarray, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 39 (1977), 262–268.

[9] P. Diaconis and R. L. Graham, Personal communication to D. E. Critchlow (1985).

[10] J. Dieudonné, *Treatise on Analysis*, Vol. II. Academic Press, New York (1969).

[11] T. D. Downs, Rotational angular correlations, in: *Biorhythms and human reproduction* (Eds. M. Ferin, F. Halberg, R. Richart and R. Vande Wiele), Chapter 7, pp. 97–104, Wiley, New York, NY (1974).

[12] V. Estivill-Castro, *Sorting and measures of disorder*, Ph.D. thesis, University of Waterloo, Ontario, Canada (1991).

[13] V. Estivill-Castro, Generating nearly sorted sequences—the use of measures of disorder, *Electron. Notes Theor. Comput. Sci.* 91 (2004), 56–95.

[14] V. Estivill-Castro, H. Mannila and D. Wood, Right invariant metrics and measures of presortedness, *Discrete Appl. Math.* 42 (1993), 1–16.

[15] N. I. Fisher, *Statistical analysis of circular data*, Cambridge University Press, New York, NY (1995).

[16] N. I. Fisher and A. J. Lee, Nonparametric measures of angular-angular association, *Biometrika* 69 (1982), 315–321.

[17] N. I. Fisher and A. J. Lee, A correlation coefficient for circular data, *Biometrika* 70 (1983), 327–332.

[18] J. L. Fleiss and J. M. Tanur, A note on the partial correlation coefficient, *Amer. Statist.* 25 (1971), 43–45.

[19] J. D. Gibbons, *Nonparametric measures of association*, Sage Publications, Newbury Park, California, USA (1993).

[20] G. Th. Guilbaud, Relation entre les deux coefficients de corrélation de rangs, *Math. Sci. Hum.* 72 (1980), 45–59.

[21] P. Hadjicostas and K. B. Lakshmanan, Bubble sort with erroneous comparisons, *Australas. J. Combin.* 31 (2005), 85–106.

[22] P. Hadjicostas and K. B. Lakshmanan, Measures of disorder and straight insertion sort with erroneous comparisons, *Ars Combin.* 98 (2011), 259–288.

[23] P. Hadjicostas and K. B. Lakshmanan, Recursive merge sort with erroneous comparisons, *Discrete Appl. Math.* 159 (2011), 1398–1417.

[24] P. Hadjicostas and C. Monico, A re-examination of the Diaconis-Graham inequality, *J. Combin. Math. Combin. Comput.* (to appear).

[25] R. W. Hamming, Error detecting and error correcting codes, *Bell System Tech. J.* 29 (1950), 147–160.

[26] M. Hassani, Derangements and applications, *J. Integer Seq.* 6 (2003), Article 03.1.2, 8 pp.

[27] M. G. Kendall, *Rank correlation methods*, Oxford University Press, New York, NY (1975).

[28] D. E. Knuth, *The art of computer programming: Generating all tuples and permutations*, Volume 4, Fascicle 2, Addison-Wesley, Upper Saddle River, NJ (2005).

[29] A. J. Lawrance, On conditional and partial correlation, *Amer. Statist.* 30 (1976), 146–149.

[30] E. I. Marshall, Conditions for rank correlation to be zero, *Sankhyā Ser. B* 56 (1994), 59–66.

[31] C. Monico, Personal communication (2011).

[32] B. Monjardet, Concordance et consensus d'ordres totaux: les coefficients K et W, *Rev. Statist. Appl.* 33 (1985), 55–87.

[33] B. Monjardet, Concordance between two linear orders: The Spearman and Kendall coefficients revisited, *J. Classification* 14 (1997), 269–295.

[34] B. Monjardet, On the comparison of the Spearman and Kendall metrics between linear orders, *Discrete Math.* 192 (1998), 281–292.

[35] P. K. Sen, Estimates of the regression coefficient based on Kendall's tau, *J. Amer. Statist. Assoc.* 63 (1968), 1379–1389.

[36] S. R. Shieh, *Some extensions of U- and V-statistics*, Ph.D. thesis, University of Wisconsin-Madison, Madison, Wisconsin, USA (1990).

[37] G. S. Shieh, R. A. Johnson and E. W. Frees, Testing independence of bivariate circular data and weighted degenerate U-statistics, *Statist. Sinica* 4 (1994), 729–747.

(Received 29 July 2012; revised 12 July 2013)