AMSE JOURNALS-AMSE IIETA publication-2017- Series: Advances A; Vol. 54; N°4; pp 476-496
Submitted July 2017; Revised Aug. 10, 2017, Accepted Oct.30, 2017

Probabilistic Rank Correlation
- A New Rank and Comparison Based Correlation Coefficient with a Simple, Pragmatic Transitivity Condition

Ruiting Lian*[1], Changle Zhou[1], Ben Goertzel[2]

[1] Fujian Provincial Key Laboratory of Brain-like Intelligent Systems, Department of Cognitive Science, Xiamen University, Xiamen, China
[2]Hanson Robotics Open Cog Foundation709C, 7/F Bio-Informatics Centre 2 Science Park West Avenue Shatin, Hong Kong, China

Abstract

A novel measure of correlation between data sets is proposed based on applying the notion of "probabilistic support" to compare the pairwise comparisons of measurements. Probabilistic Rank Correlation (PRC) is a crisp instantiation of this idea, in the spirit of traditional rank correlations. It is shown that, under broad conditions, Probabilistic Rank Correlations has a strong, elegant transitivity property. The practical application of the PRC is also illustrated.

Keywords

Correlation, Transitivity, Probabilistic support, Probabilistic Rank Correlation, Correlation measure

1. Introduction

When the need arises to assess the correlation of two data series whose distributions are unknown or expected to be importantly non-normal, analysts often eschew the Pearson correlation (Benesty, J. et al, 2009). They and turn to *rank correlation measures*, such as the Spearman correlation (Croux, C., & Dehon, C., 2010)[1] or the Kendall (Abdi, H, 2007; Kendall, M. G., 1938) or Goodman-Kruskal correlations (*Corder, G.W., D.I., 2009*). Recently researchers

[1] which involves applying the standard Pearson correlation to rank values.

have also introduced a number of new Rank Correlation measures, including Lee's measure (*Le, Chap T., 2007*), which is similar to the Kruskal-Wallis test *(McKight, P. E., & Najab, J., 2010).* Also, the measure proposed by Yilmaz et al (*Yilmaz et al, 2008*) that is specialized for information retrieval applications, and a fuzzy version of the Goodman-Kruskal measure (*Gonzalez-Serna, A. et al. 2012*). Here we present another rank correlation coefficient, the Probabilistic Rank Correlation, which is based on applying the concept of probabilistic support to propositions about pairwise rank comparisons *(Popper, K., & Miller, D. W., 1987)* (and is equivalent to the covariance of a particular pair of Boolean-valued random variables, which pertain to pairwise rank comparisons of the original datasets).

The quantity $supp_{AB} = P(B|A) - P(B)$ is known as the probabilistic support that one event A gives to another event B. If $P(B|A) > P(B)$ then we say that A supports B. We introduce the notion of symmetric probabilistic support,

$$ssupp_{AB} = supp_{AB} P(A) = supp_{BA} P(B) = P(B \cap A) - P(A)P(B) \quad (1)$$

and then, given a random variable X mapping into a poset, construct G^X as a proposition holding true for any pair that X ranks in decreasing order. We then define the Probabilistic Rank Correlation of X and Y as

$$\pi_{XY} = ssupp_{G^X G^Y} \quad (2)$$

(A bit of additional subtlety arises in the context of pairs with equal rank, but the essential formula is the above.)

We show that the Probabilistic Rank Correlation possesses strong and elegant transitivity properties, governing when two positive or negative correlations can be chained together to yield a positive correlation. In essence, if one knows X is correlated with Y and Y is correlated with Z, one can say the following about transitivity to correlation of X with Z:

- Transitivity holds if the rankings implied by Z are independent of the rankings implied by X, when conditioned on the rankings implied by Y
- Transitivity also holds if this independence is only approximate, assuming the correlations involved are strong enough

The main theorem presented here (Theorem 7.1) quantifies the balance between strength and independence involved.

The symmetric probabilistic support is equal to the covariance, in the case of Boolean-valued random variables. Since the standard Pearson correlation is also defined in terms of covariance, it might seem we are not doing anything radically different from usual here. However, the normalization factor in the definition of covariance (which is necessary if using non-Boolean, numerical random variables, as in the case of Pearson and Spearman correlation) makes transitivity conditions dramatically more complicated. In the Probabilistic Rank Correlation, we avoid this normalization factor by taking the covariance of appropriately defined Boolean-valued random variables, which results in a correlation measure with an elegant transitivity condition.

We then give some potentially useful extensions to the basic Probabilistic Rank Correlation formulation. First, we note that our results also apply to a modified version of the Probabilistic Rank Correlation, in which attention is restricted to certain pairs only. In the Restricted Probabilistic Rank Correlation, G^X is defined as a proposition-holding true for any pair that X ranks in decreasing order that also fulfills some other given criterion$\mathcal{C}$. For instance, $\mathcal{C}$ could be used to restrict attention only to those pairs (a, b) fulfilling $d(a, b) < k$ for some k and some metric d on X. This may be useful in time series analysis (Quenouille, M. H., 1949; Tsay, R. S., 2005) in order to restrict attention only to pairs occurring nearby in time, as judged by threshold duration k.

These ideas may be extended beyond rank correlation per se, to take into account the magnitude of difference between two elements, rather than just the rank ordering. When correlation is known to be transitive this can accelerate certain computational algorithms (*Mahmood, A., & Khan, S., 2010*), it may be valuable to explore the utility of the probabilistic rank correlation in this context. The notion of Probabilistic Rank Correlation and the simple examples given herein may also be useful in an educational context, either as part of a discussion of the transitivity and other properties of correlation measures (e.g. *Castro Sotos et al, 2007*). Alqo, it is part of a more general treatment of uncertainty and its role in quantifying relationships (e.g. *Li, Wei., 2016; Liu, D. et al., 2017*).

2. Transitivity of Correlation

It is a typical error of naive statistics students to assume that positive Pearson correlation is transitive: that if A and B are positively correlated, and B and C are positively correlated, so as A and C. However, simple counterexamples abound for instance, *tigerhood* is positively correlated with *cathood*, and *lionhood* is positively correlated with *cathood*, but *tigerhood* and *lionhood* are anti-correlated.

For a mathematical example, let X and Y be i.i.d. $N(0,1)$ random variables and $\rho \in [0,1]$ be a constant. Let

$$V = \rho X + \sqrt{1-\rho^2}Y \tag{3}$$

Then the variance of V is easily seen to be 1, so that

$$cor(V,X) = \frac{E(VX)}{\sqrt{Var(V)Var(X)}} = \rho \tag{4}$$

as X and Y are i.i.d by construction. Similarly, $cor(V,Y) = \sqrt{1-\rho^2}$. So we see that X and V have nonzero correlation, and V and Y have nonzero correlation, but the correlation of X and Y is zero. Transitivity of non-zero correlation is violated, and in a way that does not depend on the specific value of the constantρ.

The intransitivity of correlation is not a fluke of any particular formalization; it is a fundamental qualitative property of the correlation concept (*Castro Sotos et al, 2007*). However, oftentimes correlation *is* transitive. One would like to have a correlation measure that is transitive as often as possible, consistent with capturing the intuitive notion of correlation; and one would like to have a simple, elegant condition identifying a broad range of cases where transitivity holds.

Probabilistic support, like correlation, is not generally transitive: If A supports B and B supports C, this does not necessarily mean that A supports C. However, there are known conditions under which probabilistic support *is* transitive. In the main work of this paper, we

derive a broader set of conditions than those previously described; and we then show that Probabilistic Rank Correlation inherits an elegant, pragmatic transitivity property from the transitivity properties of probabilistic support.

The contrast between Probabilistic Rank Correlation and standard Pearson correlation, regarding the simplicity of known transitivity conditions, is striking. It is known (*Langford, E., et al. 2001)* that Pearson correlation is transitive if $\rho_{AB}^2 + \rho_{BC}^2 > 1$, but this is a very severe condition. It is a sufficient but not necessary condition, and there are many other cases where Pearson correlation does turn out to be transitive, but there is no conceptually simple relationship characterizing these cases. To get a little more insight into the situation, suppose there are two series of n values, $x = \{x_i\}$ and $y = \{y_i\}$ where $i = 1,2,\ldots,n$. If you replace $x = \{x_i\}$ with $x'_i = \{x'_i\} = \{x_i - mean(x)\}$, and treat $y = y_i$ correspondingly, then the Pearson correlation of x and y is the cosine of the angle between x' and y' (the mean-centered versions of x and y). If you then consider 3 vectors v, w, z, then: intuitively, the worst scenario for $cos(v,z)$ comes about when the projections of v, w, z are in a geodesic on the unit sphere. In this case, if the angle between v and $w = a$, and the angle between w and $z = b$, then $cos(v,z) = cos(a + b)$ and a Taylor expansion yields

$$cos(v,z) \geq \frac{1}{2}(cos(v,w) + cos(w,z))^2 - 1 \qquad (5)$$

Which is meaningful (and can be refined further), but certainly lacks the conceptual simplicity of the results we present here for the Probabilistic Rank Correlation. In harmony with the above comments on covariance, we suggest that the culprit, underlying the complexity of these transitivity related calculations, is ultimately the fact that the x'_i are real-valued rather than Boolean.

The Spearman correlation simply applies the Pearson correlation to ranks, and hence doesn't change the situation fundamentally.

The Kendall τ correlation takes a different approach, somewhat related to the Probabilistic Rank Correlation, but appears not to solve the "transitivity condition" problem in a particularly elegant way either. We will explore this point in detail in section 5.

3. Probabilistic Rank Correlation

Let $(\Omega, \mathcal{F}, P)$ be a probability space (here Ω is a sample space, $\mathcal{F}$ is a σ-algebra on this sample space, and P is a probability measure), and (Π, Σ) is a measurable space (here Σ is a σ-algebra over Π). Let X and Y denote two random variables so that $X: \Omega \to \Pi$, $Y: \Omega \to \Pi$. Let Π possess the partial ordering $<$.

Most of the results about Probabilistic Rank Correlation presented here will assume $<$ is a total ordering, yet the Probabilistic Rank Correlation is also meaningful over posets lacking total ordering, and we plan to explore this aspect in future work.

Next, let for instance $\mathcal{G}^X$ denote a random variable $\mathcal{G}^X: \Omega \times \Omega \to \{0,1\}$, defined via

$$\mathcal{G}^X(\omega, \omega_2) = 1 \Leftrightarrow X(\omega) > X(\omega_2) \tag{6}$$

Similarly, we define L^X via

$$\mathcal{L}^X(\omega, \omega_2) = 1 \Leftrightarrow X(\omega) \leq X(\omega_2) \tag{7}$$

Note that iff $<$ is a total ordering, we have $\mathcal{G}^X = \neg\mathcal{L}^X$ as a general identity (a fact we will use frequently in the following). If $<$ is not total then there may be cases where that $\mathcal{G}^X(\omega, \omega_2) = \mathcal{L}^X(\omega, \omega_2) = 0$ because ω and ω_2 are not comparable according to $<$. It will never be the case, though, that $\mathcal{G}^X(\omega, \omega_2) = \mathcal{L}^X(\omega, \omega_2) = 1$.

Let that $G^X = \{(\omega, \omega_2) \in \Omega \times \Omega : \mathcal{G}^X(\omega, \omega_2) = 1\}$, and define L^X similarly. We define:

DEFINITION 1. The **Probabilistic Rank Correlation** between X and Y is given by

$$\pi_{XY} = ssupp_{G^Y G^X} \tag{8}$$

Next, suppose we have two series of n values drawn from the space Σ, $x = \{x_i\}$ and $y = \{y_i\}$ where $i = 1,2,\ldots,n$. We then let e.g. $G^x = \{(x_i,x_j): x_i > x_j\}$, and define

DEFINITION 2. The **Probabilistic Rank Correlation** between x and y is given by

$$\pi_{xy} = ssupp_{G^y G^x} \tag{9}$$

(Where the probabilities involving $P(G^x)$ and $P(G^y)$, inside the definition of $ssupp$, are taken relative to the universal set defined as the set of pairs (i,j) for $i,j \in \{1,\ldots,n\}$) If the $x = \{x_i\}$ and $y = \{y_i\}$ are interpreted as samples drawn from random variables X and Y, then one can ask the usual probability-theoretic questions about the convergence of Probabilistic Rank Correlation based on samples to Probabilistic Rank Correlation based on the underlying distributions of the random variables. However, we will not concern ourselves with such issues here, focusing instead on the properties of Probabilistic Rank as applied to finite series of values.

Note that, if there are no equally ranked pairs in the dataset x, then $P(G^x) = \frac{1}{2}$, because half of the total set of pairs (x_i,x_j) will be increasing and the other half will be decreasing. Where p_x is the probability that a randomly chosen pair (x_i,x_j) has $x_i = x_j$, the general formula is $P(G^x) = \frac{1-p_x}{2}$, since we have arbitrarily assigned the equally ranked pairs to L^x rather than G^x. We will revisit this issue in Section 8 below, suggesting a more complex approach that is more practical in cases where there are many equally ranked pairs.

3.1 Algorithm for Calculating Probabilistic Rank Correlation between Time Series

We now describe a practical methodology for calculating the Probabilistic Rank Correlation between two time series. This algorithm does not exhaust the general theory presented above, but merely constitutes an initial practical realization.

Assume one has two data series, written as $x = \{x_i\}$, $y = \{y_i\}$, with $i = 1 \ldots n$. Assume a certain *window size* k. Then, we will calculate the Probabilistic Rank Correlation, restricted so as to compare only values lying within the same window. Let $\mathcal{P}_x^{[k]}$ denote the set of pairs (x_i,x_j) so

that $|i-j| \leq k$; and define $\mathcal{P}_y^{[k]}$ similarly. Then, we will calculate G^x and G^y based on these restricted sets of pairs, thus deriving a rank-based measure of whether the local movements of x and y correlate.

The practical procedure here is simply:

1) Start with $g_x = g_y = g_{xy} = l_x = l_y = l_{xy} = 0$
2) Iterate through all pairs (i,j) with $|i-j| \leq k$ and $i \geq j$
3) For each such pair (i,j)
   a) increment g_x by 1 if $x_i > x_j$
   b) increment l_x by 1 if $x_i < x_j$
   c) increment g_y by 1 if $y_i > y_j$
   d) increment l_y by 1 if $y_i < y_j$
   e) increment g_{xy} by 1 if both $x_i > x_j$ and $y_i > y_j$
   f) increment l_{xy} by 1 if both $x_i < x_j$ and $y_i < y_j$
4) Then the Probabilistic Rank Correlation is calculated as the difference

$$\frac{g_{xy}+l_{xy}-g_x g_y - l_x l_y}{2} \tag{10}$$

This procedure deals with equally-ranked cases via the simple heuristic of averaging the results obtained via the two methods of pushing them into the "greater than" category, and pushing them into the "less than" category. Of course, in the case of no equally ranked cases we have

$$\frac{g_{xy}+l_{xy}-g_x g_y - l_x l_y}{2} = g_{xy} - g_x g_y = l_{xy} - l_x l_y \tag{11}$$

4. Algebraic Properties of Symmetric and Asymmetric Probabilistic Support

Now we begin some mathematical work, aimed at elucidating the transitivity properties of the correlation measures introduced above. As a preliminary, we first present some elementary algebraic facts relating the symmetric and asymmetric probabilistic support. These facts will be used in the following section to explore the transitivity of Probabilistic Rank Correlation and Probabilistic Comparison Correlation. Firstly, a single step of algebra shows that

$$ssupp_{AB} = supp_{AB}P(A) = supp_{BA}P(A) \tag{12}$$

from which we see that

$$ssupp_{AB} > 0 \;\Leftrightarrow\; supp_{AB} > 0 \;\Leftrightarrow\; supp_{BA} > 0 \tag{13}$$

It is also clear from the definition of ssupp that

$$ssupp_{AB} = ssupp_{BA} \tag{14}$$

and it is easy to see that

$$ssupp_{AB} = ssupp_{\neg A\neg B} \tag{15}$$

Since

$$\begin{aligned} ssupp_{\neg A\neg B} = \\ P(\neg B \cap \neg A) - P(\neg A)P(\neg B) = \\ 1 - P(A) - P(B) + P(A \cap B) - (1 - P(A))(1 - P(B)) = \\ 1 - P(A) - P(B) + P(A \cap B) - (1 - P(A) - P(B) + P(A)P(B)) = \\ P(A \cap B) - P(A)P(B) = \\ ssupp_{AB} \end{aligned} \tag{16}$$

Putting the above together, we find that the following are all equivalent criteria:

$$\begin{aligned} ssupp_{AB} &> 0 \\ ssupp_{\neg A\neg B} &> 0 \\ supp_{AB} &> 0 \\ supp_{BA} &> 0 \\ supp_{\neg A\neg B} &> 0 \\ supp_{\neg B\neg A} &> 0 \end{aligned} \tag{17}$$

We will also have use for the following observation:

LEMMA 1.

$$P(C|B) - P(C|\neg B) = \frac{P(C|B) - P(C)}{1 - P(B)} \tag{18}$$

Proof of ***LEMMA 1*** is attached in the Appendix.

5. Comparison with Kendall Correlation

It is interesting to compare the Probabilistic Rank and Kendall τ correlations via formulating the latter in the mathematical framework we have introduced here. For simplicity, in this section we will restrict ourselves to the case where $>$ is a total order. We will also follow the previous section and lump cases where $\mathrm{x_i} = \mathrm{x_j}$ into L^x; this is sometimes though not usually done in the context of the Kendall correlation, but it doesn't affect the analysis fundamentally and simplifies the formulas a bit.

The Kendall correlation, like the Probabilistic Rank Correlation, is based on the sets of ranked pairs G^x and L^x. It is typically described as "the number of pairs on whose ranking x and y agree, minus the number of pairs on whose ranking x and y disagree." In the formalism we're using here, this means

$$\tau_{xy} = P(\neg(G^x \,\Delta\, G^y)) - P(G^x \,\Delta\, G^y) \tag{19}$$

From this definition, it follows that

$$\tau_{xy} > 0 \;\Leftrightarrow\; P(G^x \cap G^y) > P(G^x) + P(G^y) - \tfrac{1}{4} \tag{20}$$

$$\begin{aligned} \tau_{xy} = P(\neg G^x \,\Delta\, G^y) - P(G^x \,\Delta\, G^y) = \\ (1 - P(G^x \,\Delta\, G^y)) - P(G^x \,\Delta\, G^y) = \\ 1 - 2P(G^x \,\Delta\, G^y) \end{aligned} \tag{21}$$

So, we have

$$\tau_{xy} > 0 \;\Leftrightarrow\; P(G^x \,\Delta\, G^y) < \tfrac{1}{2} \tag{22}$$

and it thus suffices to show that, in general,

$$P(A \,\Delta\, B) < \tfrac{1}{2} \;\Leftrightarrow\; P(A \cap B) > \tfrac{P(A)+P(B)}{2} - \tfrac{1}{4} \tag{23}$$

This follows via

$$\begin{aligned} P(A \,\Delta\, B) &< \tfrac{1}{2} \\ P(A \cup B) - P(A \cap B) &< \tfrac{1}{2} \\ P(A) + P(B) - 2P(A \cap B) &< \tfrac{1}{2} \\ P(A \cap B) &> \tfrac{P(A)+P(B)}{2} - \tfrac{1}{4} \end{aligned} \tag{24}$$

Compared to the relevant condition for the Probabilistic Rank Correlation, the Kendall positive correlation condition is more awkward to work with mathematically, and lacks an equally natural connection with probability theory.

It's easy to check that

$$P(A)P(B) < \tfrac{P(A)+P(B)}{2} - \tfrac{1}{4} \tag{25}$$

occurs iff one of $P(A), P(B)$ is $> \frac{1}{2}$ and the other is not. So for values of $P(A) = P(G^x), P(B) = P(G^y)$ falling into this category, there will be some cases where τ_{xy} and π_{xy} will give different answers regarding whether or not x and y are positively correlated.

6. A Simple Sufficient Condition for Transitivity of the Probabilistic Rank Correlation

As with any other sensible correlation measure, the Probabilistic Rank Correlation will not always be transitive. However, there exist simple and forgiving sufficient conditions under which transitivity will hold.

In this section, we give a sufficient condition for transitivity of Probabilistic Rank Correlation, which is overly severe and not likely to be applicable in realistic situations. Then in the following section we extend this theorem into a more forgiving one, according to which the requirements for transitivity are easier to fulfill.

The key point we make in this section is that the transitivity properties of Probabilistic Rank Correlation follow from the transitivity y properties of probabilistic support. Define

DEFINITON 3. B **Screens off** A with respect to C if $P(C|A \cap B) = P(C|B)$ and $P(C|A \cap \neg B) = P(C|\neg B)$

Basically, what is meant by "B screens off A with respect to C" is that C is independent of A, when conditioned on B. Shogenji *(Shojenji, Tomoji., 2003)* has shown using elementary calculations that under this condition, probabilistic support is transitive (a simple result that is closely related to the ideas underlying Bayes networks).

Using this result of Shogenji's, it is direct to conclude that

THEOREM 6.1. Suppose we have three data series, written as $x = \{x_i\}$, $y = \{y_i\}$, $z = \{z_i\}$ where $i = 1,2,\ldots,n$ and the values are drawn from some totally ordered space Σ. Then if

1) $\pi_{xy} > 0$
2) $\pi_{yz} > 0$
3) G^y screens off G^x with respect to G^z,

it follows that

$\pi_{xz} > 0$

PROOF. Recall that $\pi_{xy} = ssupp_{G^x G^y}$.

But, in the text above, it was shown that $ssupp_{xy} > 0 \Leftrightarrow supp_{xy} > 0$.

Thus, the result follows from Shogenji's result regarding the transitivity of probabilistic support. In fact, the calculations Shogenji presents in his paper show something a little, stronger, namely that if

1) $\pi_{xy}\pi_{yz} > 0$
2) G^y screens off G^x with respect to G^z,

Then $\pi_{xz} > 0$

This covers the situation: two negative correlations chain together to yield a positive correlation. We will cover this case below in Theorem 7.1 whose main business, however, is generalizing Theorem 6.1 to handle the case where the screening is only approximate.

7. A Less Stringent Transitivity Condition

Now we present our main theorem, an extension of the above result that is sufficiently strong to be useful in practical situations. In extending the result, we not only make the condition less stringent, but also generalize it to the case where one has two negative correlations and wishes to combine them transitively to get a positive correlation. We define

DEFINITION 4. B **screens off** A with respect to C **with error** ϵ if $|P(C|A \cap B) - P(C|B)| < \epsilon$ and $|P(C|A \cap \neg B) - P(C|\neg B)| < \epsilon$

This is a kind of screening that is far more likely to be observed in actual datasets, as opposed to the ideal, complete screening that Shogenji considers.

This allows us to state our main theorem:

THEOREM 7.1. Suppose we have three data series, written as $x = \{x_i\}, y = \{y_i\}, z = \{z_i\}$ where $i = 1,2,\ldots,n$ and the values are drawn from some totally ordered space Σ. Let $\epsilon > 0$. Let p_y denote the probability of equally-ranked cases for y. Then if

1) $\pi_{xy} * \pi_{yz} > \epsilon(1 - p_y)$

2) G^y screens off G^x with respect to G^z, with error $< \epsilon$

It follows that

$$\pi_{xz} > 0$$

This theorem follows easily from

LEMMA 2. Suppose we have three subsets A, B and C of the set Ω, and $\epsilon > 0$. Then if

1) $supp_{AB} * supp_{BC} > 2\epsilon$ (1-P(B))

2) B screens off A with respect to C, with error $< \epsilon$

It follows that

$$supp_{AC} > 0$$

Proofs of ***THEOREM 7.1*** *and* ***LEMMA 2*** are attached in the Appendix. Finally, we note that:

COROLLARY 1. The conclusions of Theorem 7.1 also hold for Restricted Probabilistic Rank Correlation, in which G^X is defined as a proposition holding true for any pair that X ranks

in decreasing order, that also fulfills some other given criterion $\mathcal{C}$.) The probability p_y must of course be interpreted as the probability of equally ranked pairs under the assumption of criterion $\mathcal{C}$.)

The corollary follows via the same proof as Theorem 7.1. This proof doesn't actually care about the particulars of how G^X is defined, since Lemma 2 holds in any case. The conclusions of Section 8 regarding equally ranked values also apply unproblematically to the Restricted Probabilistic Rank Correlation.

8. Handling Equally Ranked Values

In the above we simply lumped the case $x_i = x_j$ into L^x. From a mathematical perspective this may seem unproblematic since the odds of drawing two samples precisely equal to each other are very low for any reasonably large space Σ. In practical data analysis, though, the situation arises reasonably often. So from the perspective of applying π to real datasets we need to explicitly address this case.

Let G_1^x, L_1^x denote the modified sets obtained by putting the equally-ranked pairs into G^x instead of L^x. We then have

$$P(G_1^x \cap G_1^y) - P(G_1^x)P(G_1^y) = P(G^x \cap G^y) - P(G^x)P(G^y) + (p_{xy} - p_x p_y + p_x P(G^x) + p_y P(G^y)) \quad (26)$$

Where

1) as above, p_x (p_y) is the probability that a randomly selected element of G^x (G^y) is an equally-ranked pair

2) p_{xy} is the probability that a randomly selected element of $G^x \cap G^y$ is an equally-ranked pair

Note that these probabilities will generally be quite small; if a proportion p of data items are equally ranked, then the number of equally ranked pairs will be of order p^2. However, if a large

percentage of the data items are equally ranked, then the extra terms in the above expression for $P(G_1^x \cap G_1^y) - P(G_1^x)P(G_1^y)$ may become significant.

A simple approach to handling equally ranked values is to calculate two different correlations, i.e. π_{xy} as defined above, and

$$\pi_{xy:1} = ssupp(G_1^x, G_1^y) \tag{27}$$

One could then look at a correlation value as consisting of an interval

$$(min(\pi_{xy}, \pi_{xy:1}), max(\pi_{xy}, \pi_{xy:1})) \tag{28}$$

or as a heuristic one could look at the midpoint of this interval. Note that the theorems derived above apply equally well to $\pi_{xy:1}$ as to π_{xy}. So, if we let

$$\pi_{xy:min} = min(\pi_{xy}, \pi_{xy:1}) \tag{29}$$

We then have:

THEOREM 8.1. Suppose we have three data series, written as $x = \{x_i\}, y = \{y_i\}, z = \{z_i\}$ where $i = 1,2,\ldots,n$ and the values are drawn from some totally ordered space Σ. Let $\epsilon > 0$ Then if

1) $\pi_{xy:min} * \pi_{yz:min} > 2\epsilon(1 - p_y)$; or else (a weaker condition) both

a) $\pi_{xy} * \pi_{yz} > 2\epsilon(1 - p_y)$

b) $\pi_{xy:1} * \pi_{yz:1} > 2\epsilon(1 - p_y)$

2) G^y screens off G^x with respect to G^z, with error $< \epsilon$

3) G_1^y screens off G_1^x with respect to G_1^z, with error $< \epsilon$

it follows that

$\pi_{xz:min} > 0$

More experimentation on real-world datasets may lead to different ways of coping with datasets having a large number of equally-ranked pairs. We note that Kendall correlation also deals with equally-ranked pairs in a manner that is not fully conceptually satisfactory (typically, a pair that is equally-ranked in dataset x and unequally-ranked in dataset y is ignored and counted as neither a "mismatch" or a "match", but there is some variation in how this is handled).

9. Practical Examples

In this section, we illustrate the practical application of the PRC, in the context of examples from the literature illustrating non-transitivity of Pearson correlation. In one of these, the PRC is transitive; in the other, it is intransitive, but less severely so than the Pearson correlation.

Table 1 compares Pearson correlation and PRC on data from (*Langford, E., et al. 2001)* regarding triples, home runs and base hits among 2000 New York Yankees with at least 300 at-bats. Here, according to Pearson correlation, triples and base hits are positively correlated, and base hits and home runs are positively correlated, but triples and home runs are negatively correlated. PRC on the other hand shows all pairs as positively correlated, with the correlation between base hits and home runs being very weak but still positive.

Conceptually, what is happening here is that there are some types of players for whom getting a lot of base hits correlated with getting a lot of triples, and other types of players for whom getting a lot of base hits correlated with getting a lot of home runs. But few players for whom getting a lot of triples is correlated with getting a lot of home runs. The "triples versus home runs" correlation aggregates statistics over multiple player types in a way that masks interesting distinctions that appear if one looks only at players who get a lot of triples, or only at players who get a lot of home runs.

Table 1. Baseball Data: Here PRC is transitive while Pearson is not

	Triple vs Base Hits	Base Hits vs Home Run	Triple vs Home Run
Pearson	0.52	0.29	-0.09
PRC	0.36	0.39	0.03

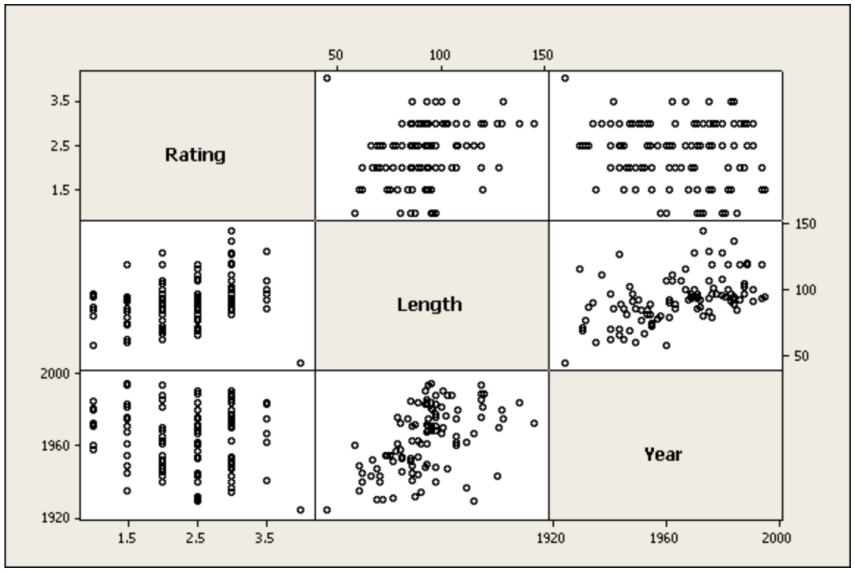

Figure 1. Scatterplot matrix of Rating, Length and Release Year of films

Figure 1 shows the scatterplot matrix of length, rating and release year of films, regarding to the film data from *(Moore, Thomas L., 2006).* Table 2 compares Pearson correlation and PRC on the data that shows in Figure 2. The Pearson correlation shows a strong intransitivity: length and rating are positively correlated, and so are length and release year, but rating and release year are negatively correlated. The PRC shows this same intransitivity, but much less strongly.

Table 2. Film Data: Here PRC is less intransitive than Pearson

	Length vs Rating	Length vs Year	Rating vs Year
Pearson	0.32	0.51	-0.15
PRC	0.26	0.37	-0.06

Table 3 shows a synthetic example; here we see that the Pearson correlation gives a highly intransitive result (high correlation A vs B and A vs C, but very low correlation B vs C). The PRC gives results similar to what would be found via probability multiplication (A vs B and B vs C are correlated around .48, and A vs C is correlated around .19).

Table 3. Synthetic Data: Here PRC is transitive while Pearson is not

	A vs B	A vs C	B vs C
Pearson	0.68	0.68	0.12
PRC	0.48	0.48	0.19

These examples illustrate that the theoretical advantages of the PRC regarding transitivity, actually translate into less perverse behavior regarding transitivity on real-world datasets. Further exploration on more diverse datasets will also be valuable, of course. [2]

10. Conclusion

We have proposed a novel correlation measure, the Probabilistic Rank Correlation, and shown that it possesses elegant and pragmatic transitivity criteria. The Probabilistic Rank Correlation's close connection with the notion of probabilistic support provides it with a solid conceptual and intuitive basis. It is closely, mathematically and conceptually, related to the Pearson and Spearman correlations, as one might expect; but via calculating the covariance of an appropriately defined set of Boolean random variables, its transitivity criteria avoid the complexities associated with the transitivity of Pearson or Spearman correlations. It shares with the Kendall correlation a foundation in rank comparisons, but seems more intuitive than the Kendall correlation, due to its foundation on covariance. We have also proposed the Probabilistic Comparison Correlation, which extends the ideas underlying Probabilistic Rank Correlation to form a correlation measure taking into account the magnitude of differences between elements, not merely the rank. We suggest that the Probabilistic Rank and Comparison Correlations should be viewed as viable alternatives to the standard correlation measures, in both theoretical and practical contexts. We have shown that the PRC measure gives more sensible results than Pearson correlation on some simple standard test problems. Finally, we conjecture that our novel correlation measures may be especially applicable for measuring correlation between data series with strongly non-normal distributions, e.g. fractal distributions, and/or to cases where the rank or comparative relations between data items are more important than their precise magnitude.

References

1. Abdi H. The Kendall rank correlation coefficient. 2007. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA, pp. 508-510.
2. J. Benesty, J. Chen, Y. Huang, I. Cohen. Pearson correlation coefficient. 2009. *Noise Reduction in Speech Processing*. Springer Berlin Heidelberg, pp. 1-4.
3. Castro Sotos, Ana Elisa, Andreas Vesaliusstraat, Vanhoof Stijn, Onghena Patrick Van Den Noortgate. The non-transitivity of Pearson's correlation coefficient: an educational perspective. 2007. Proc. 56th Session of the Int. Statistical Institute 2007, pp. 22-29.

[2] We note in passing that in some of our commercial work, we have applied PRC extensively to correlations between proprietary financial time series data, with interesting results.

4. G.W. Corder, D.I. Foreman. Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach. 2009. Wiley.
5. C. Croux, C. Dehon. Influence functions of the Spearman and Kendall correlation measures. 2010. *Statistical Methods & Applications*, vol. 19, no. 4, pp. 497-515.
6. Eric Langford, Neil Schwertman, Margret Owens. Is the Property of Being Positively Correlated Transitive? 2001. The American Statistican, vol. 55, no. 4, pp. 322-325.
7. A. Gonzalez-Serna, R. A. McGovern, P. R. Harrigan, F. Vidal, A. F. Y. Poon, S. Ferrando-Martinez, E. Ruiz-Mateos, Correlation of the virological response to short-term maraviroc monotherapy with standard and deep-sequencing-based genotypic tropism prediction methods. 2012. Antimicrobial Agents and Chemotherapy, vol. 56, no. 3, pp. 1202-1207.
8. M. G. Kendall. A new measure of rank correlation. 1938. Biometrika, vol. 30, no. 1/2, pp. 81-93.
9. E. Langford, N. Schwartzman, M. Owens, Is the property of being positively correlated transitive? 2001. The American Statistician, vol. 55, pp. 322-325.
10. C. T. Le. A New Rank Test Against Ordered Alternatives in K-Sample Problems. 2007. *Biometrical Journal*, vol. 30, no. 1, pp. 87-92.
11. Wei Li, Reform on Mathematical Modelling Teaching Contents in the Era of Big Data. 2016. AMSE Journals-2016-Series: Advances A; vol. 59, no. 1, pp. 129-144.
12. D. Liu, W. Y. Dong, B. X. Wang, Methods for Supplier Library Construction and Parts Similarity Measurement in Web-Based Parts Library Platform. 2017. AMSE JOURNALS-AMSE IIETA publication-2017-Series: Advances A; vol. 54, no. 1, pp. 87-105.
13. A. Mahmood, S. Khan, Exploiting transitivity of correlation for fast template matching. 2010. *IEEE Transactions on Image Processing*, vol. 19, no. 8, pp. 2190-2200.
14. P. E. McKight, J. Najab. Kruskal-Wallis Test. 2010. *Corsini Encyclopedia of Psychology*.
15. Thomas L. Moore, Paradoxes in film ratings. 2006. Journal of Statistics Education, vol. 14, no. 1): n1.
16. K. Popper, D. W. Miller, Why probabilistic support is not inductive. 1987. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 321, no. 1562, pp. 569-591.
17. M. H. Quenouille. Approximate tests of correlation in time-series. 1949. Journal of the Royal Statistical Society. Series B (Methodological). vol. 11, no. 1, pp. 68-84.
18. Tomoji Shogenji. A Condition for Transitivity in Probabilistic Support. 2003. British Journal for the Philosophy of Science, vol. 54, pp. 613-616.
19. R. S. Tsay. Analysis of financial time series, 2005. vol. 543, John Wiley & Sons.

20. Emine Yilmaz, Aslam Javed, Robertson Stephen. A new Rank Correlation coefficient for information retrieval. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.

Appendix

PROOF OF LEMMA 1

$$\begin{aligned} P(C|B) - P(C|\neg B) = \\ \frac{P(C\cap B)}{P(B)} - \frac{P(C\cap(\neg B)}{P(\neg B)} = \\ \frac{P(C\cap B)}{P(B)} - \frac{P(C)-P(C\cap(\neg B))}{1-P(B)} = \\ \frac{P(C\cap B)(1-P(B))-(P(C)-P(C\cap B)P(B)}{P(B)(1-P(B))} = \\ \frac{P(C\cap B)-P(C)P(B)}{P(B)(1-P(B))} = \\ \frac{P(C|B)-P(C)}{1-P(B)} \end{aligned} \tag{30}$$

PROOF OF THEOREM 7.1. Let $A = G^x, B = G^y, C = G^z$. We know that

$\pi_{AB} = ssupp_{AB} = supp_{AB}P(A) = supp_{BA}P(B)$, and that $P(B) = \frac{1-p_y}{2}$.

So Lemma 2 yields the condition

$$supp_{AB} * supp_{BC} > 2\epsilon(1 - P(B)) = 2\epsilon\frac{1-p_y}{2} = \epsilon(1 - p_y) \tag{31}$$

PROOF OF LEMMA 2. The proof follows the course of the proof of the theorem in *(Shojenji, Tomoji., 2003)*, but with the addition of the error ϵ, which must be propagated through all the calculations appropriately. We have

$$\begin{aligned} P(C|A) = P(C \cap B|A) + P(C \cap \neg B|A) = \\ P(C|B \cap A)P(B|A) + P(C|\neg B \cap A)P(\neg B|A) = \\ (P(C|B) \pm \epsilon)P(B|A) + (P(C|\neg B) \pm \epsilon)P(\neg B|A). \end{aligned} \tag{32}$$

Where the last step comes from the "screening off" assumption. Meanwhile,

$$P(C) = P(C \cap B) + P(C \cap \neg B) = P(C|B)P(B) + P(C|\neg B)P(\neg B). \tag{33}$$

Let $r \pm \epsilon$ denote the interval $(r - \epsilon, r + \epsilon)$, and interpret equations involving quantities of the form $r \pm \epsilon$ using standard interval arithmetic.

Then, from the above two equations it follows that

$$\begin{aligned} P(C|A) - P(C) = \\ (P(C|B) \pm \epsilon)P(B|A) + (P(C|\neg B) \pm \epsilon)P(\neg B|A) \\ -P(C|B)P(B) - P(C|\neg B)P(\neg B) = \\ P(C|B)(P(B|A) - P(B)) + P(C|\neg B)(P(\neg B|A) - P(\neg B)) \\ \pm 2\epsilon(P(B|A) + P(\neg B|A)) = \\ P(C|B)(P(B|A) - P(B)) + P(C|\neg B)((1 - P(B|A)) - (1 - P(B))) \pm 2\epsilon = \\ P(C|B)(P(B|A) - P(B)) - P(C|\neg B)(P(B|A)) - P(B)) \pm 2\epsilon = \\ (P(C|B) - P(C|\neg B))(P(B|A) - P(B)) \pm 2\epsilon \end{aligned} \tag{34}$$

Using Lemma 1 from earlier, we then have

$$P(C|A) - P(C) = \left(\frac{P(C|B) - P(C)}{1 - P(B)}\right)(P(B|A) - P(B)) \pm 2\epsilon = \frac{((P(C|B) - P(C))(P(B|A) - P(B)) \pm 2\epsilon(1 - P(B))}{1 - P(B)} \tag{35}$$

So, to have

$$P(C|A) - P(C) > 0 \tag{36}$$

it suffices to have

$$(P(C|B) - P(C))(P(B|A) - P(B)) > 2\epsilon(1 - P(B)) \tag{37}$$

This is equivalent to stating that

$$supp_{BC} * supp_{AB} > 2\epsilon(1 - P(B)) \tag{38}$$

and hence completes the proof.