

A relative-rank measure for the rank transformation

Abid Hussain^{1,*}, Steve Drekic², and Salman A. Cheema³

¹Department of Statistics, Government College Khayaban-e-sir syed, Rawalpindi, Pakistan.

²Department of Statistics & Actuarial Science, University of Waterloo, Ontario, Canada.

³Department of Applied Sciences, National Textile University, Faisalabad, Pakistan.

Abstract

This research introduces a new ranking scheme capable of ordering competing units in a more continual manner. The classic ranks are functionalized by incorporating data range and quantile coverage simultaneously in the proposed formula. This dual use of additional information is found to enhance the sensitivity of the devised scheme towards offering more competent weights of competing observations. Some key mathematical properties of the proposed approach are highlighted. In particular, the capability of the devised approach in offering unambiguous preferences is established. Finally, the performance of the proposed measure is investigated under a set of diverse parametric settings involving different stochastic formulations, sample sizes, and degrees of correlation.

Keywords: Ranks; magnitudes; relative-rank; correlation.

1 Introduction

The development of novel ranking schemes and analysis of ranked data have gained overwhelming popularity over the years (Erdi, 2019; Acuna-Soto et al., 2021). A variety of interesting applications have been studied in this regard, such as (to name a few) the formation of social hierarchies and associated systematic inequalities (Turchin et al., 2018; Posfai and D’Souza, 2018), health care systems and preferred preemptive pathways (Wagner and Fell, 2001), consumer preferences and market research (Mantegna et al., 1995), DNA sequence analytics (Adlakha and Sharma, 2020), political choices and blocks (Gormley and Murphy, 2008), and content analysis (Cocho et al., 2015). More recently, Iniguez et al. (2022) considered an engaging study targeting the dynamics of rank trajectories with respect to time while allowing the flux of entries across rank boundaries.

The ubiquity (and consequently the utility) of ranks is rooted in their capability of competently reducing the complexity of multidimensional systems into simple measurable entities of interest. The ranks (and ranked data) usually consist of $N \geq 1$ permutations, say $\delta_1, \delta_2, \dots, \delta_N$, on a series of, say $n \geq 1$ items or objects, indexed by $i \in \{1, 2, \dots, n\}$. The analytic challenge in handling such series remains the absenteeism of any vector space structure on the set of all permutations (see, for example, Collas and Irrozki (2021)). It is to be noted that if ties are not permitted, the ranked data carries the deterministic quality. However, the equidistant nature of ranks are sometimes reported as a shortcoming of the ordering mechanism, see, for example, Pinto da Costa and Soares (2005).

In statistics, rank transformation has been commonly used for many years. However, rankings have the major disadvantage of not revealing the magnitude of differences between measurements. Differences between rankings are not usually equal. As an example, three runners may be ranked first, second, and third based on how they complete a marathon. But this does not imply that the amount of time between completions by numbers 1 and 2, as well as between numbers 2 and 3, is the same. The first-place runner may finish five minutes ahead of the second-place runner, who may then finish eight minutes ahead of the

third-place runner. We cannot determine the magnitude of the difference between ranked measurements if we only have the ranks available.

This article devises a new ranking scheme while providing insight into the mathematical foundations of the measure. The newly proposed set of ranking statistics is advocated based on its sensitivity towards data alterations. This sensitivity is achieved by incorporating range and quantile coverage into the formula. In particular, we show that our set of ranking statistics are translation-invariant and offer an independent connection with the statistics of the rank-based distribution-free tests. We conclude the paper with two thorough numerical studies aimed at showing how our proposed approach compares to some of the more popular rank-based correlation measures found in the literature.

2 Proposed relative-rank measure

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a vector of random observations with continuous joint cumulative distribution function (cdf) given by

$$P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = \prod_{i=1}^n F(x_i),$$

where $X_1, X_2, \dots, X_n$ each have the same marginal cdf $F(\cdot)$. Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ represent the corresponding sequence of order statistics, satisfying $X_{(1)} < X_{(2)} < \dots < X_{(n)}$. Then, a new set of ranking statistics, which we refer to as the *relative-rank statistics* and denote by $\{RR_i\}_{i=1}^n$, is formulated as follows:

(1) Define $RR_1 = 1$.

(2) Define

$$Y_i = i - 1 + \frac{(n-1)(X_{(i)} - X_{(i-1)})}{X_{(n)} - X_{(1)}}, \quad i = 2, 3, \dots, n. \quad (1)$$

(3) For $i = 2, 3, \dots, n$, define RR_i to be the $(i-1)^{th}$ smallest value of $\{Y_2, Y_3, \dots, Y_n\}$, so that

$$RR_1 < RR_2 < \dots < RR_n.$$

We now proceed to demonstrate some mathematical characteristics of the proposed scheme.

Proposition 1 *The set of statistics $\{RR_i\}_{i=1}^n$ generate a simple arithmetic series on $[1, n]$ for a set of distinct observations (i.e., no ties are permitted).*

Proof: The result can be established by showing that $\sum_{i=1}^n RR_i = \frac{n(n+1)}{2}$. Note that

$$\begin{aligned} \sum_{i=1}^n RR_i &= 1 + \sum_{i=2}^n Y_i \\ &= 1 + \sum_{i=2}^n \left(i - 1 + \frac{(n-1)(X_{(i)} - X_{(i-1)})}{X_{(n)} - X_{(1)}} \right) \\ &= \sum_{i=1}^n i - (n-1) + \frac{n-1}{X_{(n)} - X_{(1)}} \sum_{i=2}^n (X_{(i)} - X_{(i-1)}). \end{aligned} \quad (2)$$

The sum of all adjacent differences of a sequence is equal to the range of the sequence, namely $X_{(n)} - X_{(1)}$. Therefore, equation (2) clearly simplifies to yield $\sum_{i=1}^n RR_i = \sum_{i=1}^n i$, which is known to be equal to $\frac{n(n+1)}{2}$.

Proposition 2 *The set of statistics $\{RR_i\}_{i=1}^n$ are translation-invariant.*

Proof: Let us consider the inclusion of a constant c in every observed value, so that $\mathbf{X}^{(c)} = (X_1 + c, X_2 + c, \dots, X_n + c)$ now denotes the vector of random observations. Let $X_{(1)}^{(c)}, X_{(2)}^{(c)}, \dots, X_{(n)}^{(c)}$ represent the corresponding sequence of order statistics. Furthermore, let $\{Y_i^{(c)}\}_{i=2}^n$ be the corresponding set of random variables defined by equation (1). Note that for $i \geq 2$:

$$\begin{aligned} Y_i^{(c)} &= i - 1 + \frac{(n-1) \left(X_{(i)}^{(c)} - X_{(i-1)}^{(c)} \right)}{X_{(n)}^{(c)} - X_{(1)}^{(c)}} \\ &= i - 1 + \frac{(n-1) \left(X_{(i)} + c - (X_{(i-1)} + c) \right)}{X_{(n)} + c - (X_{(1)} + c)} \\ &= i - 1 + \frac{(n-1) \left(X_{(i)} - X_{(i-1)} \right)}{X_{(n)} - X_{(1)}} \\ &= Y_i. \end{aligned}$$

Therefore, if $\{RR_i^{(c)}\}_{i=1}^n$ denotes the corresponding set of relative-rank statistics, it follows that $RR_i^{(c)} = RR_i$.

Proposition 3 *The set of statistics $\{RR_i\}_{i=1}^n$ independently provide the statistics for the rank-based distribution-free tests.*

Proof: Proposition 1 establishes that $\sum_{i=1}^n RR_i = \frac{n(n+1)}{2}$, which is the result of the test statistics for the rank-based tests. Furthermore, once we know $\{RR_i\}_{i=1}^n$, we require the same critical values that are associated with the classic ranking scheme for a specific test.

3 Performance evaluation

The comparative performance of our devised procedure is demonstrated by the launch of well-directed and rigorous simulation-based experiments. The objectives are attained by enumerating the capability of the proposed scheme in retaining the degree of correlation and significance testing of correlation prevalent among two random variables. We consider various parametric settings for these purposes.

Our first study addresses the correlation comparison. In particular, three stochastic formations (i.e., the bivariate normal distribution, bivariate uniform distribution, and bivariate double-exponential (or Laplace) distribution) are considered to meet the goal of comparative assessment. The recovery of the correlation structure is investigated with respect to values of the correlation coefficient parameter $\rho = 0, 0.25, 0.50$, and 0.75 . In addition, each setting is studied for varying sample sizes of $n = 10, 15, 20$, and 25 . The comparative evaluation is conducted with respect to the three most commonly-used and prominent correlation measures, namely the Pearson correlation coefficient, Spearman rank correlation coefficient, and Kendall rank correlation coefficient. Table 1 specifies the pre-defined parametric settings (including the correlation coefficient values under consideration) and presents the estimated values of the correlation coefficients, along with associated standard errors for the four estimating schemes. As highlights, we witness the following observations:

- Regardless of the sample size and parent distributional structure, our proposed ranking scheme produced estimates with comparable and controlled standard errors. This observation underscores the ability of the new ranking scheme in using available information as competently as the existing techniques.

- Regardless of the parent distributional structure, our proposed rank-based measure quantifies the extent of correlation more closely than the other two existing rank-based measures. This fact becomes more obvious as the sample size is increased.
- The performance of the non-rank-based Pearson correlation coefficient is observed to be dominating our devised measure. This result is explainable due to the distributional assumptions exploited by the Pearson correlation coefficient. However, it is to be noted that these assumptions are usually considered limitations of the measure. In fact, our proposed method does not involve such assumptions for operational conduct. Moreover, even if the distributional structure is available, some distributions are known not to carry a defined variance-covariance structure fundamental for the calculation of the Pearson correlation coefficient.
- It appears, based on the findings of this numerical study, that our newly proposed method outperforms existing rank-based measures while enumerating the underlying extent of linear dependencies.

Our second simulation-based study compares the new ranking procedure with the ordinary method used to test for the significance of correlation. Paired data values from the bivariate normal distribution with correlations of 0, 0.25, 0.50, and 0.75 were generated for selected sample sizes of $n = 8, 12, 16, 20, 25$, and 30. The resulting values were then ranked using three methods (i.e., Spearman, Kendall, and Relative). The Pearson product-moment correlation values were also computed between the two sets of initially generated values. Significance tests of correlation were performed using the test statistic

$$t = \frac{r\sqrt{n-2}}{1-r^2},$$

which has a Student's t distribution with $n - 2$ degrees of freedom (see, for example, [Zimmerman \(1994\)](#)). The increments in correlation from 0 to 0.75 made it possible to examine both type-I and type-II errors. All tests were directional and performed at the 5% level of significance. There were 10,000 replications of the sampling procedure and subsequent significance tests for each sample size and each degree of correlation.

The results of the simulations are presented in Table 2. For various degrees of correlation between the two random variables and different sample sizes, Table 2 shows the probabilities of rejecting the null hypothesis of independence. According to the table columns, we performed t -tests on the initial scores using Pearson's product-moment approach and the three competing rank-based procedures (i.e., Spearman's ρ , Kendall's τ , and our relative rank approach).

The probability of type-I errors (first row in each panel) remained close to the 5% level of significance in all four procedures. There was a noticeable difference between the tests based on Pearson's product-moment approach and those connected with the rank-based procedures. However, evidence from Table 2 does suggest that the relative ranking method is more accurate than Spearman's ρ and Kendall's τ tests. As a result, in terms of type-I and type-II performance, the relative ranking method appears to be an acceptable alternative.

4 Summary

A new ranking scheme is proposed and several of its mathematical attributes are demonstrated in this article. It is established that the devised scheme uses the observed data more competently to launch the estimated hierarchies among the competing entities when interest lies in the assessment of the extent of the linear relationship exhibited among two variables. The comparative performance enumeration of the proposed scheme with respect to the three most commonly-used existing measures reveals performance hierarchies of the form $\rho(Pearson) > \rho(Relative) > \rho(Spearman) > \rho(Kendall)$. In the next phase of research, it will be interesting to investigate the behavior of the devised approach in the extension of ordinal log-linear models by replacing the equidistant ranks with our proposed scheme.

Disclosure statement

All authors of this article declare that there is no conflict of interest regarding the publication of this article.

Data availability statement

No new data is associated with this manuscript.

References

- Acuna-Soto, C., Liern, V., & Perez-Gladish, B. (2021). Normalization in TOPSIS-based approaches with data of different nature: application to the ranking of mathematical videos. *Annals of Operations Research*, 296, 541-569.
- Adlakha, K., & Sharma, S. (2020). Brand positioning using multidimensional scaling technique: an application to herbal healthcare brands in indian market. *Vision: The Journal of Business Perspective*, 24(3), 345-355.
- Cocho, G., Flores, J., Gershenson, C., Pineda, C., & Sánchez, S. (2015). Rank diversity of languages: generic behavior in computational linguistics. *PLoS ONE*, 10, e0121898.
- Collas, F., & Irurozki, E. (2021). Concentric mixtures of Mallows models for top- k rankings: sampling and identifiability. In International Conference on Machine Learning (pp. 2079-2088). PMLR.
- Erdi, P. (2019). *Ranking: The Unwritten Rules of The Social Game We All Play*. Oxford University Press.
- Gormley, I. C., & Murphy, T.B. (2008). Exploring voting blocs within the Irish electorate: a mixture modeling approach. *Journal of the American Statistical Association*, 103, 1014-1027.
- Iniguez, G., Pineda, C., Gershenson, C., & Barabási, A. L. (2022). Dynamics of ranking. *Nature Communications*, 13, 1646.
- Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C. K., Simons, M., & Stanley, H. E. (1995). Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Physical Review E*, 52(3), 2939.
- Pinto da Costa, J., & Soares, C. (2005). A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics*, 47(4), 515-529.
- Pósfai, M., & D'Souza, R. M. (2018). Talent and experience shape competitive social hierarchies. *Physical Review E*, 98(2), 020302(R).
- Turchin, P., Currie, T. E., Whitehouse, H., & Spencer, C. (2018). Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. *Proceedings of the National Academy of Sciences*, 115(2), E144-E151.
- Wagner, A., & Fell, D. A. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478), 1803-1810.
- Zimmerman, D. W. (1994). A note on modified rank correlation. *Journal of Educational and Behavioral Statistics*, 19(4), 357-362.

Table 1: Correlation coefficients with their standard errors (in brackets) for three well-known bivariate distributions (10,000 simulations).

Density	n	ρ	Pearson	Spearman	Kendall	Relative
Normal	10	0	0.0049 (0.0033)	0.0033 (0.0033)	0.0009 (0.0024)	-0.0033 (0.0033)
		0.25	0.2334 (0.0032)	0.2169 (0.0032)	0.1608 (0.0024)	0.2286 (0.0032)
		0.50	0.4764 (0.0027)	0.4401 (0.0028)	0.3354 (0.0024)	0.4608 (0.0028)
		0.75	0.7294 (0.0017)	0.6803 (0.0020)	0.5381 (0.0019)	0.6985 (0.0019)
	15	0	-0.0013 (0.0027)	0.0035 (0.0027)	0.0010 (0.0019)	-0.0008 (0.0027)
		0.25	0.2379 (0.0025)	0.2251 (0.0026)	0.1633 (0.0019)	0.2291 (0.0025)
		0.50	0.4845 (0.0021)	0.4548 (0.0022)	0.3301 (0.0017)	0.4666 (0.0021)
		0.75	0.7380 (0.0013)	0.6449 (0.0015)	0.5419 (0.0014)	0.7093 (0.0014)
	20	0	0.0010 (0.0023)	-0.0005 (0.0023)	0.0005 (0.0016)	0.0009 (0.0023)
		0.25	0.2430 (0.0022)	0.2317 (0.0021)	0.1584 (0.0016)	0.2340 (0.0022)
		0.50	0.4905 (0.0017)	0.4598 (0.0019)	0.3352 (0.0014)	0.4684 (0.0019)
		0.75	0.7415 (0.0011)	0.7071 (0.0013)	0.5408 (0.0012)	0.7174 (0.0012)
	25	0	0.0014 (0.0020)	0.0001 (0.0020)	0.0007 (0.0014)	0.0000 (0.0020)
		0.25	0.2472 (0.0019)	0.2280 (0.0020)	0.1636 (0.0014)	0.2329 (0.0019)
		0.50	0.4908 (0.0016)	0.4669 (0.0017)	0.3319 (0.0013)	0.4725 (0.0016)
		0.75	0.7419 (0.0009)	0.7111 (0.0011)	0.5405 (0.0010)	0.7186 (0.0011)
Uniform	10	0	-0.0013 (0.0033)	-0.0036 (0.0033)	-0.0054 (0.0025)	-0.0060 (0.0033)
		0.25	0.2459 (0.0034)	0.2278 (0.0034)	0.1925 (0.0027)	0.2349 (0.0034)
		0.50	0.4938 (0.0031)	0.4766 (0.0032)	0.4142 (0.0027)	0.4785 (0.0031)
		0.75	0.7478 (0.0024)	0.7365 (0.0025)	0.6896 (0.0023)	0.7391 (0.0024)
	15	0	0.0008 (0.0027)	-0.0025 (0.0027)	0.0002 (0.0019)	-0.0010 (0.0027)
		0.25	0.2508 (0.0028)	0.2374 (0.0027)	0.1898 (0.0021)	0.2418 (0.0027)
		0.50	0.4987 (0.0025)	0.4824 (0.0025)	0.4157 (0.0021)	0.4831 (0.0025)
		0.75	0.7453 (0.0020)	0.7382 (0.0020)	0.6898 (0.0018)	0.7387 (0.0019)
	20	0	0.0016 (0.0023)	-0.0038 (0.0023)	-0.0020 (0.0016)	0.0011 (0.0023)
		0.25	0.2524 (0.0024)	0.2384 (0.0024)	0.1818 (0.0018)	0.2433 (0.0023)
		0.50	0.4963 (0.0022)	0.4880 (0.0022)	0.4155 (0.0018)	0.4883 (0.0021)
		0.75	0.7490 (0.0017)	0.7391 (0.0017)	0.6870 (0.0016)	0.7395 (0.0017)
	25	0	-0.0001 (0.0021)	0.0017 (0.0020)	0.0024 (0.0014)	0.0019 (0.0020)
		0.25	0.2450 (0.0021)	0.2395 (0.0021)	0.1850 (0.0016)	0.2402 (0.0021)
		0.50	0.4963 (0.0019)	0.4935 (0.0019)	0.4165 (0.0016)	0.4944 (0.0019)
		0.75	0.7502 (0.0015)	0.7426 (0.0015)	0.6880 (0.0014)	0.7433 (0.0015)
Laplace	10	0	0.0004 (0.0041)	-0.0028 (0.0325)	-0.0003 (0.0028)	0.0019 (0.0037)
		0.25	0.2303 (0.0039)	0.2123 (0.0034)	0.1658 (0.0027)	0.2155 (0.0035)
		0.50	0.4671 (0.0034)	0.4334 (0.0030)	0.3368 (0.0025)	0.4391 (0.0031)
		0.75	0.7156 (0.0022)	0.6613 (0.0023)	0.5412 (0.0021)	0.6793 (0.0022)
	15	0	0.0000 (0.0034)	0.0018 (0.0584)	0.0007 (0.0022)	0.0038 (0.0029)
		0.25	0.2276 (0.0032)	0.2192 (0.0028)	0.1591 (0.0021)	0.2201 (0.0028)
		0.50	0.4816 (0.0027)	0.4378 (0.0025)	0.3319 (0.0020)	0.4438 (0.0025)
		0.75	0.7156 (0.0022)	0.6760 (0.0018)	0.5384 (0.0016)	0.6869 (0.0017)
	20	0	0.0028 (0.0029)	0.0021 (0.0878)	0.0011 (0.0019)	0.0022 (0.0026)
		0.25	0.2448 (0.0028)	0.2182 (0.0024)	0.1601 (0.0018)	0.2187 (0.0025)
		0.50	0.4841 (0.0024)	0.4453 (0.0021)	0.3315 (0.0017)	0.4486 (0.0021)
		0.75	0.7339 (0.0015)	0.6864 (0.0015)	0.5399 (0.0014)	0.6884 (0.0015)
	25	0	-0.0029 (0.0027)	0.0452 (0.1203)	0.0003 (0.0016)	-0.0030 (0.0023)
		0.25	0.2435 (0.0025)	0.2210 (0.0022)	0.1603 (0.0016)	0.2277 (0.0022)
		0.50	0.4870 (0.0021)	0.4471 (0.0019)	0.3330 (0.0015)	0.4500 (0.0019)
		0.75	0.7355 (0.0013)	0.6912 (0.0014)	0.5396 (0.0012)	0.6924 (0.0013)

Table 2: Probability of rejecting the null hypothesis of independence (bivariate normal distribution).

n	ρ	Pearson	Spearman	Kendall	Relative
8	0	0.0515	0.0474	0.0271	0.0528
	0.25	0.0888	0.0761	0.0548	0.0871
	0.50	0.2422	0.1897	0.1501	0.2219
	0.75	0.6440	0.4910	0.4192	0.5793
12	0	0.0511	0.0483	0.0488	0.0518
	0.25	0.1214	0.1118	0.0986	0.1244
	0.50	0.3939	0.3321	0.3192	0.3581
	0.75	0.8588	0.7634	0.7547	0.8095
16	0	0.0508	0.0505	0.0488	0.0507
	0.25	0.1596	0.1371	0.1214	0.1461
	0.50	0.5300	0.4616	0.4299	0.4902
	0.75	0.9496	0.9039	0.8866	0.9221
20	0	0.0515	0.0504	0.0472	0.0511
	0.25	0.1938	0.1653	0.1605	0.1736
	0.50	0.6470	0.5678	0.5507	0.5961
	0.75	0.9854	0.9601	0.9577	0.9695
25	0	0.0475	0.0517	0.0439	0.0529
	0.25	0.2234	0.2010	0.1891	0.2135
	0.50	0.7434	0.6808	0.6683	0.6930
	0.75	0.9964	0.9874	0.9887	0.9915
30	0	0.0479	0.0500	0.0486	0.0495
	0.25	0.2642	0.2480	0.2336	0.4281
	0.50	0.8215	0.7618	0.7641	0.7916
	0.75	0.9995	0.9973	0.9974	0.9975