

Chapter 2

The Weighted Rank Correlation Coefficient r_W

Abstract Spearman's rank correlation coefficient is not entirely suitable to measure the correlation between two rankings in some applications because it treats all ranks equally. In 2001, we have proposed a weighted rank measure of correlation that weights the distance between two ranks using a linear function of those ranks, giving more importance to higher ranks than lower ones. In this chapter, we analyze its distribution and provide a table of critical values to test whether a given value of the coefficient is significantly different from zero. We also summarize a number of applications for which the new measure is more suitable than Spearman's.

2.1 Introduction

In this chapter we analyze the statistical distribution of the weighted rank correlation coefficient r_W [61, 63, 67, 93] and provide a table of critical values to test whether a given value of the coefficient is significantly different from zero. We also summarize a number of applications for which the new measure is more suitable than Spearman's.

Rank correlation coefficients such as Spearman's [57, Chap. 9], [94] can be used to quantify the similarity between two rankings. However, Spearman's coefficient treats all ranks equally and is, therefore, not entirely suitable for applications such as the one described in the previous chapter, where different weights need to be given to different ranks.

In this chapter, we describe a measure of correlation—adapted from Spearman's rank correlation coefficient—that weighs ranks proportionally to how high they are.¹ This problem has already been considered by Blest in 2000 [10]. His measure, however, is not a symmetric function of the two vectors of ranks. The measure r_W that we have proposed in 2001 does not have this problem. In the next section, we describe rank correlation and provide an example to illustrate the need for weighted measures. In Sect. 2.3, we describe previous approaches to this problem, identifying their drawbacks, which lead us to the measure r_W described here. We also give some

¹We assume that the higher rank is 1, and corresponds to the “best” element in the ranking.

insight into its interpretation. In Sects. 2.4 and 2.5, we analyze the distribution of the measure proposed and provide some illustrative examples. In Sect. 2.6, we discuss the applicability of the r_W measure, identify a few of its potential applications and describe some of its limitations. Conclusions are given in Sect. 2.7. The proofs of results used and the table of critical values of the r_W measure are given in the Appendix.

2.2 Rank Correlation

One interesting fact about rank correlation is that, contrary to other correlation methods, it can be used not only on numerical data but on any data that can be ranked. An example of the use of such methods is the analysis of sales data where the aim is to assess whether there is correlation between marketing activities (i.e., visits to clients) and the number of sales [57, Chap. 9].

Rank correlation can be applied, for instance, to the problem of evaluating rankings of documents generated by search engines, which was introduced in the previous chapter. As shown there, Spearman's coefficient r_S treats all ranks equally and in that situation it should not, as the top ranks are clearly more important. Thus, we need a measure of similarity between rankings that takes rank importance into account. An alternative measure to r_S is Kendall's concordance coefficient [57, Chap. 9]. This coefficient is equivalent to counting the minimum number of transpositions required to transform one ranking into the other. The most striking difference between Spearman's and Kendall's coefficients is that the differences are squared in the former but not in the latter. Therefore, Spearman's coefficient is more affected by larger differences while, on the other hand, Kendall's is more affected by smaller ones.

In 2001 and 2005 [61, 63, 93], we have introduced and analyzed a weighted rank correlation coefficient, r_W , that weighs the distance between two ranks using a linear function of those ranks, giving more importance to higher ranks than lower ones. This measure will be described next. We will also analyze the statistical distribution of r_W in the case of independence between the two vectors of ranks and also for the general case; that is, the case where we make no assumption of independence between the two vectors of ranks. To do so, we will use the same notation and analogous arguments of those used by Ruymgaart, Shorack and Van Zwet (1972) in the proof of their Theorem 2.1 (see [83]). We show that r_W has a normal limit distribution. A table of critical values for r_W will be provided in the Appendix in order to test whether a given value of the coefficient is significantly different from zero, and a number of applications for this new measure will also be given.

2.3 Weighted Rank Measure of Correlation

Here, we describe the construction of the weighted rank measure of correlation r_W and then use the example introduced in the previous chapter to illustrate the advantage of the new measure. As before let us denote by $\mathbf{R} = (R_1, R_2, \dots, R_n)$ and $\mathbf{Q} = (Q_1, Q_2, \dots, Q_n)$ two vectors of ranks obtained on a sample of size n .

The calculation of the distance between two ranks in Spearman's coefficient, $r_S = 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2}{n^3 - n}$, is given by:

$$D_i^2 = (R_i - Q_i)^2$$

which does not take rank importance into account. In 2000, an alternative was proposed by Soares et al. [92]: $((R_i - Q_i)/R_i)^2$. This function has several shortcomings. First, the ranking \mathbf{Q} that will obtain the largest distance from \mathbf{R} is not the inverted ranking, i.e., $Q_i = n - R_i + 1$, which is rather unintuitive. Second, the function is not symmetric, which means that the distance between series \mathbf{R} and series \mathbf{Q} can be different from the distance between \mathbf{Q} and \mathbf{R} . Also in 2000, an adaptation of Kendall's concordance coefficient has been proposed by Blest [10], which also addresses the same issue. However, as in the distance function proposed by Soares et al., Blest's measure is not a symmetric function of the two vectors of ranks.

Here, we use the following alternative distance measure proposed in [61, 63, 93]:

$$\begin{aligned} W_i^2 &= (R_i - Q_i)^2 ((n - R_i + 1) + (n - Q_i + 1)) \\ &= D_i^2 (2n + 2 - R_i - Q_i) \end{aligned}$$

The first term of this product is D_i^2 , exactly as in Spearman's method and represents the distance between R_i and Q_i . The second term represents both the importance of R_i and also the importance of Q_i .

Let us consider now again the example introduced in the previous chapter concerning the ranking of ten documents according to the relevance of each document to a particular subject. Recall that \mathbf{R} represents the "true" ranking of the ten documents, provided by an expert, and \mathbf{Q} and \mathbf{Z} represent two rankings of the same ten documents provided, for instance, by two competing search engines. The sum of distances between \mathbf{R} and \mathbf{Q} , using this expression ($\sum_{i=1}^n W_i^2$), is 278 and the sum of distances between \mathbf{R} and \mathbf{Z} is 436 (Table 2.1). This means that the distance between \mathbf{R} and \mathbf{Z} is larger, a conclusion that is consistent with the intuitive analysis of the usefulness of suggested rankings in the previous section. In fact, given that ranking \mathbf{R} is more similar to ranking \mathbf{Q} in the most important ranks, the first ones, we expect the difference between rankings \mathbf{R} and \mathbf{Q} to be smaller, which is the case. As for Spearman's coefficient both distances, between rankings \mathbf{R} and \mathbf{Q} and rankings \mathbf{R} and \mathbf{Z} are the same, which is rather unintuitive.

Now, in order to construct a correlation coefficient based on this new distance, we will follow a common strategy, which consists in looking for an affine function of

Table 2.1 Application of the new distance measure to the example of rankings of documents

Document	\mathbf{R}	\mathbf{Q}			\mathbf{Z}		
i	R_i	Q_i	D_i^2	W_i^2	Z_i	D_i^2	W_i^2
D1	1	2	1	19	3	4	72
D2	2	1	1	19	5	9	135
D3	3	4	1	15	2	1	17
D4	4	6	4	48	1	9	153
D5	5	5	0	0	6	1	11
D6	6	3	9	117	4	4	48
D7	7	8	1	7	7	0	0
D8	8	9	1	5	8	0	0
D9	9	10	1	3	9	0	0
D10	10	7	9	45	10	0	0
Sum				278			436

the distance between the two rankings, an expression of the form $A + B \sum_1^n W_i^2$. The idea is to find which two constants A and B make this expression to take values in $[-1, 1]$, as is usual with correlation coefficients; 1 when they are the same ($R_i = Q_i$) and -1 when the rankings are inverted ($R_i = n + 1 - Q_i$). In the first case, we have that $\sum_1^n W_i^2 = 0$ and so A must be 1. In the second case, as is shown in Appendix A.1, the maximum value of the weighted distance $\sum_1^n W_i^2$ between two rankings is $(n^4 + n^3 - n^2 - n)/3$ and is obtained when the rankings are inverted, that is, $Q_i = n + 1 - R_i$. Using this expression for the maximum value of the weighted distance, we obtain therefore that $A + B \cdot (n^4 + n^3 - n^2 - n)/3 = -1$. Using these two conditions for the two constants A and B , the weighted rank measure of correlation, $r_W = A + B \sum_1^n W_i^2$, becomes, after some simplifications:

$$r_W = 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2 ((n - R_i + 1) + (n - Q_i + 1))}{n^4 + n^3 - n^2 - n} \quad (2.1)$$

In the example of Table 2.1 concerning the ranking of documents, the r_W value for \mathbf{R} and \mathbf{Q} is 0.8468. As expected, this is higher than the r_W value for \mathbf{R} and \mathbf{Z} , which is 0.7598.

As stated earlier, rank correlation coefficients measure the monotonicity between two series of matched values. We have just described the weighted r_W measure of correlation, adapted from Spearman's r_S , that takes rank importance into account, unlike the latter. That is, correlation will be more affected by points that are ranked higher in the series than others. Next we analyze the statistical distribution of this measure, starting with the case of independence between the two vectors of ranks. We also analyze the differences between Spearman's r_S and the weighted measure r_W .

2.4 Properties of the Distribution of r_W Under the Null Hypothesis of Independence

In this section, we will study the distribution of r_W under the null hypothesis of independence between the two vectors of rankings. We start by briefly describing some results from linear rank statistics, which will be used next to determine the expected value and the variance of r_W . We provide evidence to suggest that, under the null hypothesis, the standardized value of r_W follows the Gaussian distribution; that is,

$$(r_W - E(r_W))/\sqrt{\text{var}(r_W)} \stackrel{d}{\approx} N(0, 1)$$

As above, let us denote by $\mathbf{R} = (R_1, \dots, R_n)$ the first vector of ranks and by $\mathbf{Q} = (Q_1, \dots, Q_n)$ the second vector of ranks. That is, \mathbf{R} and \mathbf{Q} assume only values in the set \mathcal{R} of all the $n!$ permutations of the integers $(1, \dots, n)$. Under the null hypothesis,

$$H_0 : \mathbf{R} \text{ and } \mathbf{Q} \text{ are independent,}$$

the two rank vectors are both uniformly distributed over \mathcal{R} . This implies that the distribution of

$$r_W = 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2 (2(n+1) - R_i - Q_i)}{n(n^3 + n^2 - n - 1)}$$

is the same as the distribution of

$$1 - \frac{6 \sum_{i=1}^n (i - R_i^*)^2 (2(n+1) - i - R_i^*)}{n(n^3 + n^2 - n - 1)}$$

where $\mathbf{R}^* = (R_1^*, \dots, R_n^*)$ is a random vector taken uniformly from the set \mathcal{R} .

2.4.1 Linear Rank Statistics

A statistic of the form

$$S = \sum_{i=1}^n c(i) a(R_i^*) \tag{2.2}$$

is called a linear rank statistic [78, Chap. 8]. The constants $a(1), \dots, a(n)$ are called the scores and $c(1), \dots, c(n)$ the regression constants. In [78, Chap. 8] it is shown that, under H_0 ,

- (i) $\Pr(R_i^* = r) = \frac{1}{n}, r = 1, \dots, n$
- (ii) if $i \neq j$ then $\Pr(R_i^* = r, R_j^* = s) = \begin{cases} \frac{1}{n(n-1)} & r \neq s = 1, \dots, n; \\ 0 & \text{otherwise} \end{cases} \quad (2.3)$
- (iii) $E(S) = n\bar{c}\bar{a}$
- (iv) $\text{var}(S) = (n-1)s_c^2 s_a^2$,

where \bar{a} and \bar{c} represent the average values of the scores and regression constants, respectively. Similarly, s_a^2 and s_c^2 represent their variances.

We will now find the two first moments of the weighted rank correlation coefficient r_W under the hypothesis of independence between the two vectors of ranks. In particular, its expected value has the desirable property of being equal to zero under independence, which is a common result for correlation coefficients.

Theorem 1 *Under the hypothesis of independence between two vectors of ranks,*

$$E(r_W) = 0 \text{ and } \text{var}(r_W) = \frac{31n^2 + 60n + 26}{30(n^3 + n^2 - n - 1)}$$

The proof is given in Appendix A.3.

2.4.2 Exact and Asymptotic Distribution of r_W Under the Null Hypothesis of Independence

We will now investigate the asymptotic distribution of $(r_W - E(r_W))/\sqrt{\text{var}(r_W)}$, under the null hypothesis of independence between the two vectors of ranks. Let

$$S_n^{(11)} = \sum_{i=1}^n iR_i^*, \quad S_n^{(12)} = \sum_{i=1}^n iR_i^{*2}, \quad S_n^{(21)} = \sum_{i=1}^n i^2 R_i^*$$

Then, as shown in the Appendix,

$$r_W = \frac{1}{n(n^3 + n^2 - n - 1)} \left(24(n+1) \left(S_n^{(11)} - E(S_n^{(11)}) \right) - 6 \left(S_n^{(12)} - E(S_n^{(12)}) \right) - 6 \left(S_n^{(21)} - E(S_n^{(21)}) \right) \right)$$

To standardize r_W , we divide by the square root of its variance ($E(r_W) = 0$ under the null hypothesis). We start by defining the three constants,

$$a_n^{(11)} = \frac{2\sqrt{30}n(n+1)^2\sqrt{n-1}}{n\sqrt{31n^2 + 60n + 26}\sqrt{n^3 + n^2 - n - 1}},$$

$$a_n^{(12)} = \frac{-6\sqrt{30}n(n+1)\frac{\sqrt{16n^3 + 14n^2 - 19n - 11}}{2160}}{n\sqrt{31n^2 + 60n + 26}\sqrt{n^3 + n^2 - n - 1}}$$

$$a_n^{(21)} = \frac{-6\sqrt{30}n(n+1)\frac{\sqrt{16n^3+14n^2-19n-11}}{2160}}{n\sqrt{31n^2+60n+26}\sqrt{n^3+n^2-n-1}}$$

Then,

$$\frac{r_W}{\sqrt{\text{var}(r_W)}} = a_n^{(11)} \frac{S_n^{(11)} - \mu_n^{(11)}}{\sigma_n^{(11)}} + a_n^{(12)} \frac{S_n^{(12)} - \mu_n^{(12)}}{\sigma_n^{(12)}} + a_n^{(21)} \frac{S_n^{(21)} - \mu_n^{(21)}}{\sigma_n^{(21)}}$$

where

$$\mu_n^{(k\ell)} = E(S_n^{(k\ell)}) \text{ and } \sigma_n^{(k\ell)} = \sqrt{\text{var}(S_n^{(k\ell)})}$$

In [78, Chap. 8] it is shown that as $n \rightarrow \infty$ the following statistic converges in distribution to the Gaussian:

$$\frac{S_n^{(k\ell)} - \mu_n^{(k\ell)}}{\sigma_n^{(k\ell)}} \xrightarrow{d} N(0, 1)$$

On the other hand, as $n \rightarrow \infty$,

$$a_n^{(11)} \rightarrow a^{(11)} = 2\sqrt{\frac{30}{31}}, \quad a_n^{(12)} \rightarrow a^{(12)} = -\frac{1}{90}\sqrt{\frac{30}{31}} \quad \text{and} \quad a_n^{(21)} \rightarrow a^{(21)} = a^{(12)}$$

Therefore [8, p. 288],

$$a_n^{(k\ell)} \frac{S_n^{(k\ell)} - \mu_n^{(k\ell)}}{\sigma_n^{(k\ell)}} \xrightarrow{d} a^{(k\ell)} Z$$

where Z stands for the standard normal distribution. So, the standardized r_W is the sum of three statistics that are asymptotically normal. However, these three statistics are not independent and so we cannot conclude directly that their sum is asymptotically normal.

In order to verify the asymptotic distribution of r_W , we have started by computing some theoretical and empirical distributions in the next section, and compared it with the normal curve.

2.4.3 Simulations

We have calculated the exact distribution of r_W for n up to 14. Due to computational limitations, for larger values of n , we estimated the distribution based on a random sample of one million permutations. For $n = 14$, we observe that there is a small difference between the exact and estimated values for the most important quantiles (Table 2.2). Note that we have decided not to interpolate the critical values because it is a discrete distribution. Instead, we used a common strategy of finding the quantiles

Table 2.2 Difference between the exact and estimated quantiles for $n = 14$

Quantile (%)	0.5	1	2.5	5	95	97.5	99	99.5
Difference	0.0052	0.0042	0.0021	0.0000	0.0010	0.0010	0.0010	0.0010

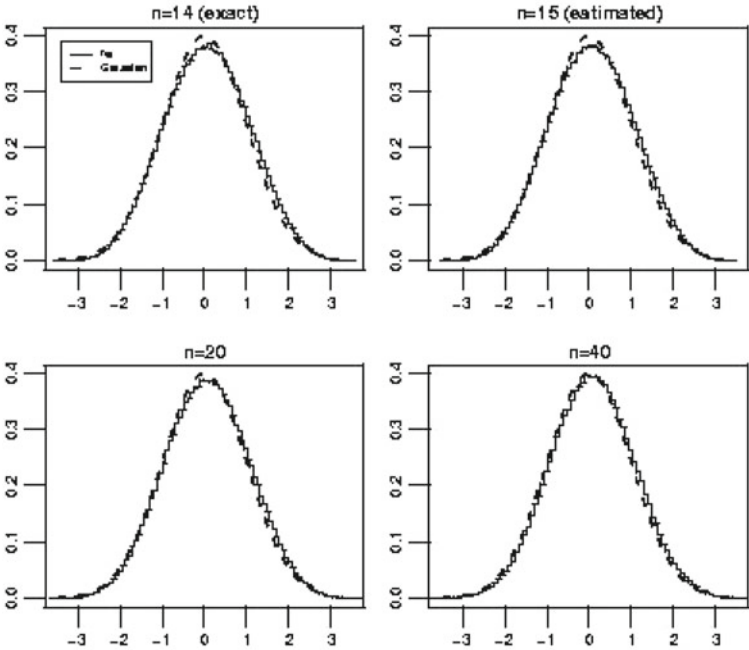


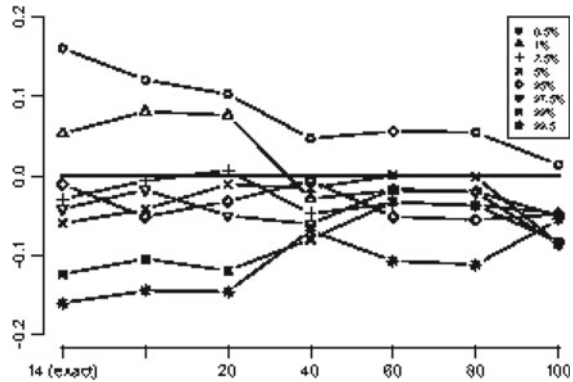
Fig. 2.1 Exact distribution for $n = 14$ and estimated distribution for $n = 15, 20$ and 40 , together with the Standard Normal curve

of a discrete distribution but with a slight change. With a discrete distribution, all the values in a given interval satisfy the definition of quantile of order p . This might bring a difficulty in choosing the quantile especially for small n . Therefore, for each confidence level α_i , we have multiplied it by the total number of permutations, $n!$. If the result is not an integer, we use the next lower/higher integer for small/large confidence levels, respectively. Finally, we picked the corresponding order statistic.

In Fig. 2.1, we plot the distribution for $n = 14$ and $n = 15$, respectively, the last exact and the first estimated distributions. The graphs indicate that the sample size is adequate. In the same figure, we also plot the estimated distributions for $n = 20$ and 40 , respectively. In all graphs, the values of r_W have been standardized and we plot the normal curve for comparison.

The empirical distribution of r_W does not lie symmetrically about zero. This is because the distribution of the values of our statistic is not symmetric; it is a little skewed. This does not strike us as a problem, since we think there is no reason for a

Fig. 2.2 Difference between the estimated quantiles of r_W and the quantiles of the Standard Normal



measure of correlation to be necessarily symmetric. The measure r_W was constructed so that the first ranks are more important than the other ranks and this influences its symmetry. Unlike Blest's measure, our statistic treats both rankings similarly; that is, $r_W(\mathbf{R}, \mathbf{Q}) = r_W(\mathbf{Q}, \mathbf{R})$. However, this does not mean that the distribution of its values, which lie in the interval $(-1, 1)$, is symmetric. In fact it is not. For instance, for $n = 3$, the values of the statistic are $-1, -0.5, -0.5, 0.375, 0.625$, and 1 ; these values have mean zero, but are not symmetric about zero. For the same reason, the percentiles are also not symmetric about zero (e.g., the 5% and 95% percentiles for $n = 5$, which are given in Table A.1 (Appendix), are not the same in absolute value).

We have calculated the difference between the quantiles for the standardized r_W and the standard normal for a few values of n and observed that the differences are small (Fig. 2.2). For $n > 40$, we have observed that these differences are always smaller than 0.1. This means that the differences between the nonstandardized r_W and the approximation given by the normal distribution is smaller than $0.1\sqrt{\text{var}(r_W)}$. For instance, for $n = 51$, the difference between r_W and the approximation given by the normal curve is smaller than 0.003.

2.4.4 Comparison Between r_W and r_S

In the last subsections, we have presented an adaptation of Spearman's rank correlation coefficient, which assigns more importance to higher ranks. Here we start with a comparison of the weighted measure r_W with Spearman's coefficient r_S to point out the differences and describe the conditions under which the new coefficient should be used.

Despite the similarities between the two measures r_W and r_S , they may yield quite different values when applied to the same pair of series. We illustrate these differences using a few examples.

We start by measuring the correlation between rankings \mathbf{R} and \mathbf{Q} . The former is defined as $\mathbf{R} = (1, 2, \dots, n-1, n)$, where n is the number of elements in the

ranking. Ranking \mathbf{Q} is obtained from \mathbf{R} by swapping the elements $(1, \dots, p)$ and $(q - p + 1, \dots, q)$, after inverting the order in each of them:

$$\mathbf{R} = (\boxed{1, \dots, p}, p + 1, \dots, q - p, \boxed{q - p + 1, \dots, q}, q + 1, \dots, n)$$

$$\mathbf{Q} = (\boxed{q, \dots, q - p + 1}, p + 1, \dots, q - p, \boxed{p, \dots, 1}, q + 1, \dots, n).$$

We plot in Fig. 2.3 the value of $r_W(\mathbf{R}, \mathbf{Q}) - r_S(\mathbf{R}, \mathbf{Q})$ for a few values of n , p , and q . Note that although some of the differences are already quite large, achieving values close to 0.15, it is possible to obtain even larger differences. Furthermore, if both p and q are represented as proportions of n , the differences are independent of the size of the ranking (in the examples the differences decrease with the size of the ranking, n , because we have used values of p which represent smaller proportions).

Having proved that the two rank correlation coefficients, r_W and r_S , can give quite different results, it is of importance now to decide when to use r_W . The new measure should be used instead of Spearman's coefficient in applications for which it is known that the importance of concordance between the series decreases with the ranks. In other words, assuming that $f(i)$ is a function that represents the importance of rank i , r_W should be used rather than r_S if:

$$i < j \Rightarrow f(i) > f(j). \quad (2.4)$$

Note that we assume that 1 is the highest rank and n is the lowest one, where n is the number of elements in the series. Again, let us illustrate with some more examples. We measure the difference between the weighted correlation of a ranking \mathbf{R} and each of two rankings \mathbf{Q} and \mathbf{Z} , i.e., $r_W(\mathbf{R}, \mathbf{Q}) - r_W(\mathbf{R}, \mathbf{Z})$. As before, $\mathbf{R} = (1, 2, \dots, n - 1, n)$, where n is the number of elements. Ranking \mathbf{Q} is obtained from \mathbf{R} by swapping the elements $(1, \dots, p)$ and $(q + 1, \dots, q + p)$:

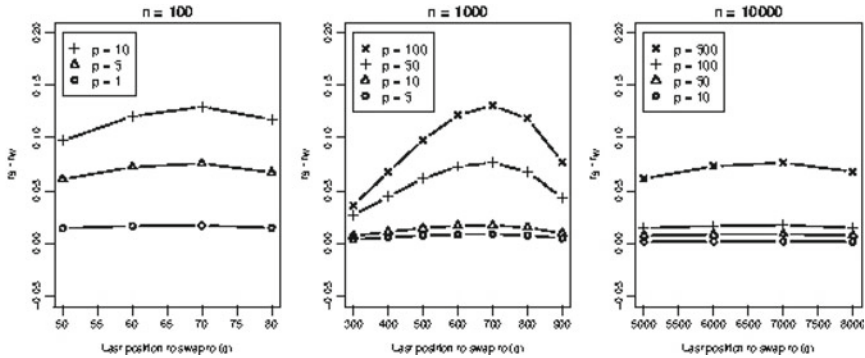


Fig. 2.3 Difference between Spearman's coefficient (r_S) and the new weighted measure of correlation (r_W) on a few illustrative examples

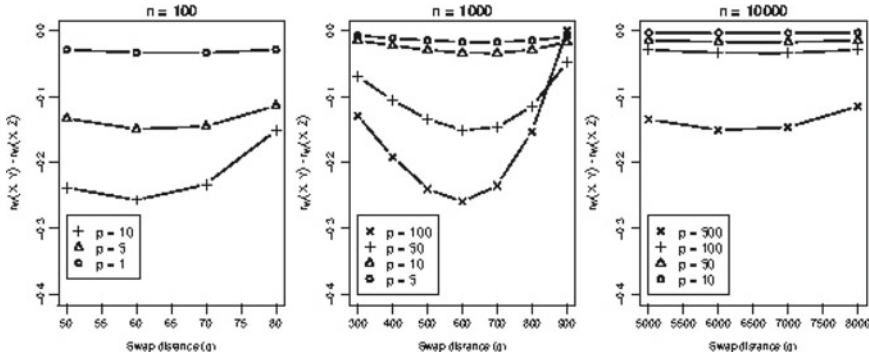


Fig. 2.4 Difference between two weighted rank correlation values, $r_W(\mathbf{R}, \mathbf{Q})$ and $r_W(\mathbf{R}, \mathbf{Z})$, where \mathbf{Q} and \mathbf{Z} are obtained by applying a symmetric procedure to \mathbf{R} , for a few illustrative examples of different parameters of the procedure

$$\mathbf{R} = (\boxed{1, \dots, p}, p+1, \dots, q, \boxed{q+1, \dots, q+p}, q+p+1, \dots, n)$$

$$\mathbf{Q} = (\boxed{q+1, \dots, q+p}, p+1, \dots, q, \boxed{1, \dots, p}, q+p+1, \dots, n).$$

Ranking \mathbf{Z} is obtained from \mathbf{R} using a symmetric procedure to the one used to generate \mathbf{Q} . We swap the elements $(n-p+1, \dots, n)$ and $(n-q-p+1, \dots, n-q)$:

$$\mathbf{R} = (1, \dots, n-q-p, \boxed{n-q-p+1, \dots, n-q}, n-q+1, \dots, n-p, \boxed{n-p+1, \dots, n})$$

$$\mathbf{Z} = (1, \dots, n-q-p, \boxed{n-p+1, \dots, n}, n-q+1, \dots, n-p, \boxed{n-q-p+1, \dots, n-q})$$

According to the assumption above (2.4), the concordance between \mathbf{R} and \mathbf{Z} is clearly much higher than between \mathbf{R} and \mathbf{Q} . However, the value of r_S is the same in both cases. The difference between the values of the weighted measure in the two cases is illustrated in Fig. 2.4 for a few values of n , p , and q .

Note that we have made a very weak assumption, (2.4), namely that the higher the rank, the higher its importance. This enables us to consider a wide range of applications, a few of which are enumerated in Sect. 2.6. However, in some applications it is possible to make stronger assumptions. For instance, it may be known that only the top 1, 5, or 10 alternatives in the ranking will be considered. In these cases, it may be more suitable to use more specific measures that only take those ranks into consideration and ignore all the others. Other weighted measures of correlation will be given in later chapters.

Our main claim is that the weighted measure r_W is more appropriate than traditional rank correlation coefficients for a wide range of applications where higher

ranks are more important than lower ones. Although r_W gives more importance to higher ranks, it still takes the whole ranking into account rather than simply assuming that some ranks matter and others do not. Therefore, it may be used as a general measure of similarity between two rankings. By treating rankings as a whole, generality is gained (i.e., it may be applied to a wide range of ranking problems) at the cost of the ability to capture specificities of individual problems (e.g., only the top-5 ranks are considered). Therefore, we do not claim that it should replace problem-specific measures. We believe that it may be more useful as a complement to those measures, assessing the general concordance between rankings, while other measures may provide a more specific assessment.

2.5 The Asymptotic Distribution of r_W for the General Case

In the last section we have seen that, in the case of independence between two rankings, the weighted measure of correlation r_W seems to converge to the Gaussian distribution, according to the simulations realized. Now we make no independence assumptions; that is, we study the asymptotic distribution of r_W for the general case and we give the formal proof that r_W converges to the normal distribution (see also [67]). This section is rather technical and can be skipped at a first reading. First,

$$\begin{aligned} r_W &= 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2 (2n + 2 - R_i - Q_i)}{n^4 + n^3 - n^2 - n} \\ &= 1 - \frac{6}{n} \sum_{i=1}^n \left(\frac{R_i}{n+1} - \frac{Q_i}{n+1} \right)^2 \left(\frac{2n+2-R_i-Q_i}{n-1} \right) \end{aligned}$$

Therefore, the asymptotic behavior of r_W is the same as the one of $1 - 6W_n$, where

$$W_n = \frac{1}{n} \sum_{i=1}^n \left(\frac{R_i}{n+1} - \frac{Q_i}{n+1} \right)^2 \left(2 - \frac{R_i}{n+1} - \frac{Q_i}{n+1} \right).$$

W_n is a statistic of the type $\frac{1}{n} \sum_{i=1}^n a_n(R_i, Q_i)$, where $a_n(i, j)$ is a real number for $i, j = 1, 2, \dots, n$.

If we define $J(s, t) = (s - t)^2(2 - s - t)$, $0 \leq s, t \leq 1$, then $J(s, t)$ is a limit of the score function,

$$J_n(s, t) = a_n(i, j) = J\left(\frac{i}{n+1}, \frac{j}{n+1}\right), \quad (2.5)$$

for i and j such that $\frac{i-1}{n} < s \leq \frac{i}{n}$ and $\frac{j-1}{n} < t \leq \frac{j}{n}$. Hence, W_n can be written as (see [7]),

$$W_n = \int \int J_n(F_n, G_n) dH_n, \quad (2.6)$$

where F_n and G_n are the empirical marginal distribution functions of F and G , respectively; H_n is the bivariate empirical distribution function of H . Now, let us define the population moment $\mu = \int \int J(F, G) dH$. By analogy to r_W , we define the population weighted rank correlation coefficient between two variables X and Y to be,

$$\begin{aligned} \rho_W(X, Y) &= 1 - 6\mu \\ &= 1 - 6 \int \int (F(x) - G(y))^2 (2 - F(x) - G(y)) dH(x, y), \end{aligned}$$

or, by using copulas [58]

$$\rho_W(X, Y) = 1 - 6 \int_{[0,1]^2} (u - v)^2 (2 - u - v) dc(u, v),$$

where the copula $c(u, v) = P(F(X) \leq u, G(Y) \leq v)$, $0 \leq u, v \leq 1$.

Next, we present the conclusion that r_W is asymptotically Gaussian distributed.

Theorem 2.1 r_W is an asymptotic normal and consistent (ANC) estimator of ρ_W

Proof We want to prove that r_W is an asymptotic normal and consistent (ANC) estimator of ρ_W ; first,

$$\sqrt{n}(r_W - \rho_W) = -6\sqrt{n}(W_n - \mu) = -6\sqrt{n} \left[\int \int J_n(F_n, G_n) dH_n - \mu \right].$$

We start by considering the empirical processes $U_n(F) = \sqrt{n}(F_n - F)$, $V_n(G) = \sqrt{n}(G_n - G)$, $U_n^*(F) = \sqrt{n}(F_n^* - F)$, $V_n^*(G) = \sqrt{n}(G_n^* - G)$, where $F_n^* = \left[\frac{n}{n+1} F_n \right]$ and $G_n^* = \left[\frac{n}{n+1} G_n \right]$. Let now $\bar{\Delta}_n = [X_{1n}, X_{nn}] \times [Y_{1n}, Y_{nn}]$ where X_{in} and Y_{in} denote the i th order statistics and $B_{0n}^* = \sqrt{n} \int \int [J_n(F_n, G_n) - J(F_n^*, G_n^*)] dH_n$.

We will now prove that $J_n(F_n, G_n) = J(F_n^*, G_n^*)$ and so $B_{0n}^* = 0$ for all n . In fact, the function F_n , for instance, is a step function and so there is always an $i \in \{0, 1, \dots, n\}$ such that $F_n = \frac{i}{n}$; similarly for G_n . This means that by (2.5) $J_n(F_n, G_n) = J\left(\frac{i}{n+1}, \frac{j}{n+1}\right)$ for some i and j . Now, by the definition above, $\frac{i}{n+1} = F_n^*$ and $\frac{j}{n+1} = G_n^*$. So, $B_{0n}^* = 0$ for all n .

Because $B_{0n}^* = 0$ for all n , then an assumption similar to 2.3(b) in [83] (see Appendix A.5) is satisfied, that is, $B_{0n}^* \rightarrow_p 0$. We will now use the same argument of these authors, adapting it to our situation because our score function $a_n(i, j)$ is bivariate and the score functions used in [83], $a_n(i)$ and $b_n(i)$ have just one variable

(see Appendix A.5). Nevertheless, the adaption follows from the same steps of their proof. The asymptotic convergence of r_W to the normal distribution may be uniform over a class of distribution functions. However, in this work, we are not interested in proving uniform convergence, but only convergence for a single distribution.

Now we can write,

$$\sqrt{n}(W_n - \mu) = \sum_{i=1}^3 A_{in} + B_{0n}^* + B_{1n}^*,$$

where

$$\begin{aligned} A_{1n} &= \sqrt{n} \int \int J(F, G) d(H_n - H), \quad A_{2n} = \int \int U_n(F) \frac{\partial J}{\partial s}(F, G) dH, \\ A_{3n} &= \int \int V_n(G) \frac{\partial J}{\partial t}(F, G) dH, \quad B_{0n}^* \text{ is defined above; and} \\ B_{1n}^* &= \sqrt{n} \int \int [J(F_n^*, G_n^*) - J(F, G)] dH_n - A_{2n} - A_{3n}. \end{aligned}$$

2.5.1 $\sum_{i=1}^3 A_{in}$ is Asymptotically Normal Distributed

As in [83], we can prove the asymptotic normality of A_{1n} , A_{2n} , and A_{3n} based on the fact that J is a continuous function and its partial derivatives are continuous and bounded on $(0, 1)^2$.

Let us start by noting that $A_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{1in}$ where $A_{1in} = J(F(X_i), G(Y_i)) - \mu$. In fact,

$$A_{1n} = \sqrt{n} \int \int J(F, G) d(H_n - H) = \sqrt{n} \left(\int \int J(F, G) dH_n - \int \int J(F, G) dH \right)$$

Now, as in Eq. 2.6 we get,

$$\begin{aligned} A_{1n} &= \frac{\sqrt{n}}{n} \sum_{i=1}^n (J(F(X_i), G(Y_i)) - \mu) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (J(F(X_i), G(Y_i)) - \mu). \end{aligned}$$

The random variables A_{1in} are independent and identically distributed (i.i.d.) with mean zero. If we choose $\delta = \frac{1}{4}$, $D = p_0 = q_0 = 2$, $r(u) = \frac{1}{u(1-u)}$ then we have an assumption similar to Assumption 2.1 in the statement of Theorem 2.1 in [83] (See Appendix A), that is, $J(F, G) \leq D(r(F))^a (r(G))^b$ with $a = \frac{\delta - \frac{1}{2}}{p_0} = -\frac{1}{8}$ and $b = \frac{\delta - \frac{1}{2}}{q_0} = -\frac{1}{8}$; $\frac{\partial J}{\partial s}(F, G) \leq D(r(F))^{a+1} (r(G))^b$ with $a = \frac{\delta - \frac{1}{2}}{p_1} = -\frac{1}{8}$ and

$$b = \frac{\delta - \frac{1}{2}}{q_1} = -\frac{1}{8} \text{ and } \frac{\partial J}{\partial t}(F, G) \leq D(r(F))^b(r(G))^{a+1} \text{ with } a = \frac{\delta - \frac{1}{2}}{p_2} = -\frac{1}{8} \text{ and } b = \frac{\delta - \frac{1}{2}}{q_2} = -\frac{1}{8}.$$

Taking this assumption into account and by application of Hölder's inequality,

$$\int \int |\phi(F)\psi(G)| dH \leq \left[\int |\phi|^{p_0} dI \right]^{\frac{1}{p_0}} \left[\int |\psi|^{q_0} dI \right]^{\frac{1}{q_0}},$$

$$\forall p_0 > 0, q_0 > 0 : \frac{1}{p_0} + \frac{1}{q_0} = 1;$$

where ϕ and ψ are functions on $(0, 1)$, dI denotes Lebesgue measure restricted to the unit interval, we note that A_{1in} has a finite absolute moment of order $2 + \delta_0$ for some $\delta_0 > 0$ (see Appendix A.6).

Let us consider now A_{2n} . As $U_n(F) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (I(X_i \leq x) - F)$ we can write $A_{2n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{2in}$, where $A_{2in} = \int \int (I(X_i \leq x) - F) \frac{\partial J}{\partial s}(F, G) dH$ are i.i.d with mean zero. If we choose $\delta = \frac{1}{4}$, $D = p_1 = q_1 = 2$, $r(u) = \frac{1}{u(1-u)}$ then an assumption similar to 2.1 in [83] is satisfied. Again, by applying Hölder's inequality and similarly to A_{1in} , it follows that A_{2in} has a finite absolute moment of order $2 + \delta_1$ for some $\delta_1 > 0$.

Let us consider now A_{3n} . As $V_n(G) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (I(Y_i \leq y) - G)$ we can write $A_{3n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{3in}$ where $A_{3in} = \int \int (I(Y_i \leq y) - G) \frac{\partial J}{\partial t}(F, G) dH$ are i.i.d with mean zero. If we choose $\delta = \frac{1}{4}$, $D = p_2 = q_2 = 2$, $r(u) = \frac{1}{u(1-u)}$ then an assumption similar to assumption 2.1 in [83], is satisfied. By application of Hölder's inequality and similarly to A_{1in} , it follows that A_{3in} has a finite absolute moment of order $2 + \delta_2$ for some $\delta_2 > 0$.

From the above conclusions: $A_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{1in}$ where A_{1in} are i.i.d. with mean zero; $A_{2n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{2in}$ where A_{2in} are i.i.d with mean zero; $A_{3n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{3in}$ where A_{3in} are i.i.d with mean zero and because A_{1in} , A_{2in} , A_{3in} have a finite absolute moment of order larger than 2, we get $\sum_{i=1}^3 A_{in} \rightarrow_d N(0, \sigma^2)$ as $n \rightarrow \infty$. The expression for the variance corresponds to Eq. 3.10 in [83] and is given by

$$\sigma^2 = Var \left[J(F(X), G(Y)) + \int \int (I(X \leq x) - F) \frac{\partial J}{\partial s}(F(x), G(y)) dH(x, y) + \int \int (I(Y \leq y) - G) \frac{\partial J}{\partial t}(F(x), G(y)) dH(x, y) \right].$$

2.5.2 B_{1n}^* is Asymptotically Negligible

We have already seen that an assumption similar to 2.3(b) in [83] is satisfied. If we consider the mean value theorem (see [84]),

$$\sqrt{n}J(F_n^*, G_n^*) = \sqrt{n}J(F, G) + U_n^*(F) \frac{\partial J}{\partial s}(\phi_n^*, \psi_n^*) + V_n^*(G) \frac{\partial J}{\partial t}(\phi_n^*, \psi_n^*)$$

for all (x, y) in $\bar{\Delta}_n$ with $\phi_n^* = F + \alpha_3(F_n^* - F)$ and $\psi_n^* = G + \alpha_4(G_n^* - G)$, where α_3 and α_4 are numbers between 0 and 1, then B_{1n}^* can be decomposed as a sum of seven terms ($\sum_{i=1}^5 B_{\gamma_{in}}^* + B_{6n}^* + C_n^*$) which are all asymptotically negligible by the same arguments used in Sects. 5 and 6 of Ruymgaart et al. [83].

2.5.3 r_W is Asymptotically Normal Distributed

We have thus that $\sqrt{n}(W_n - \mu) \rightarrow N(0, \sigma^2)$ in distribution and it is immediate that r_W is an asymptotic normal and consistent (ANC) estimator of ρ_W : $\sqrt{n}(r_W - \rho_W) \rightarrow N(0, 36\sigma^2)$.

2.6 Examples of Application of r_W

The motivation underlying this work applies to a broad range of applications involving rankings of alternatives representing the preferences stated by humans or recommendations provided by decision support systems.

A general application is the evaluation of methods to predict rankings. The evaluation of these methods consists of comparing the ranking \mathbf{R} of a set of n objects generated by a ranking prediction method for a given situation with the target ranking \mathbf{Q} of the same set of objects on the same situation. A few examples of ranking prediction applications are recommendation of data analysis tools, stock trading support, information retrieval, recommender systems and user preferences, which will be discussed in more detail next.

The recommendation of data analysis tools is an important problem in knowledge discovery in databases (KDD) or data mining. Due to its interactive and iterative nature, an important part of the KDD process is often spent trying different preprocessing methods (e.g., discretization of numeric attributes) and learning algorithms (e.g., decision trees or support vector machines), and tuning their parameters [12].

One of the first applications that have been considered and that was indeed one of the motivations to develop weighted correlation coefficients comes from the field of machine learning; more precisely, meta-learning: given a certain number of algorithms to perform a given task, one would like to rank those algorithms from 1, the most adequate, to n , the worst. Then, given a certain method to recommend the algorithms, the weighted correlation coefficient r_W is used to evaluate the method. This is not a usual machine learning problem. Traditionally, the supervised learning approach to problems where each example can be a member of one set of n possible classes is *classification*. That is, a set of prelabelled examples is used to induce a model that selects a single one of those classes as the prediction for a new example. In this approach, a lot of information that can be useful in some situations is lost, because none but the “best” class is kept.

In the problem of selecting the best algorithm for a given task [91], for instance, a classification approach would provide one suggestion of an algorithm. We thus would know that the suggested algorithm is expected to be the best but no information about the other candidates algorithms is given. In this case, a ranking of the alternative algorithms (i.e., classes) provides complete information about the expected relative performance of all candidates and enables a more flexible decision process. The user may simply decide to run the algorithm ranked highest but he or she may also, if enough time or computational resources are available, decide to try the first few alternatives. An expert user might even have reasons to ignore the first recommendation, opting, for instance, to use the recommendation in the second rank.

As our example shows, rankings are particularly important for meta-learning, i.e., algorithm selection using past performance information [44, 56, 91]. Other areas where it may be advantageous to use a ranking approach, rather than the usual supervised classification, are medicine (e.g., diagnosis of an illness or choice of an adequate test or treatment) and control systems (e.g., choice of the correct action to carry out). Two areas where rankings are already widely used are information retrieval [53] and recommender systems [87].

Given that ranking is a learning task different from existing ones, like classification, regression, or clustering, it requires different evaluation procedures. That is in the evaluation process of ranking that the weighted correlation coefficient will be used and indeed this application was one of the driving forces for the development of weighted correlation coefficients. In [61, 93], an evaluation framework has been developed that consists of comparing the ranking suggested by the ranking method, called the *recommended ranking*, with the true ranking, called the *ideal ranking* [11]. The two rankings can be compared by using, for instance, a rank correlation coefficient. Nevertheless, as is obvious in this application, the top ranks are the most important; the user will try one, two, or maybe three of the top recommended ranks, but will probably have no time or resources to try all of the n orderings. Therefore it is very important that the *recommended ranking* is similar to the *ideal ranking* in the top ranks and it is not so important that the two rankings are similar in the last positions. The first idea to compare the two rankings was by means of the Spearman correlation coefficient; nevertheless, as this coefficient gives the same importance

(weight) to all of the ranks, the results were not good. This motivated thus the development of the correlation coefficient r_W described above, that appeared for the first time in [61, 93], and was later developed in [63, 67].

The evaluation of stock trading support systems is also a potential application of the weighted rank measure of correlation. This problem has traditionally been tackled as a regression (i.e., predict the value of an individual stock) or as a supervised classification problem (i.e., predict whether to buy, keep or sell a stock). However, what investors want is, in fact, to have a grading of the stocks in question, such that they can make a decision concerning which ones to invest in [35]. Such a grading can be represented as a ranking. The accuracy of a system that predicts the ranking of a set of stocks could be evaluated by measuring the correlation between the predicted ranking to the true ranking of the stocks. To maximize profit, the stocks ranked higher are more important than the ones with lower ranks. Therefore, weighted measures would be more suitable to evaluate such a system than traditional ones.

Two problems that are usually handled as ranking tasks are information retrieval [4] and recommender systems [14]. Evaluation strategies in these areas, usually handle the uncertainty concerning how many alternatives will actually be tried out by the user, by simulating a number of different Top- N scenarios, i.e., by assuming that the user will select the N higher ranked alternatives for different values of N (e.g., 1, 5, 10, etc.). The corresponding results are either presented to the user (possibly represented in a chart) or summarized into one value. However, although the problems that motivated this work are equally relevant in the evaluation of information retrieval systems [4], correlation-based evaluation is not very common, except in the problem of database selection, where Spearman's coefficient has been used [18]. This is true despite the most commonly used evaluation measures are based on relevance assessment, which is an arguable approach [33]. Surprisingly, rank importance is rarely taken into account [40]. Similar remarks can also be made concerning recommender systems [14].

An example which also involves both rankings representing human preferences and generated by models, is the work of [9]. This work investigates methods to infer user preferences concerning health profiles. Evaluation is performed by comparing the predicted rankings to explicitly stated rankings. The authors have used Spearman's r_S to assess ranking similarity. However, it is common knowledge that when stating their preference as rankings, humans rank the most preferred alternatives, i.e., the ones ranked at the top, more accurately than the others. Therefore, the weighted measure r_W or other would be more appropriate than the traditional one.

Many other examples could be given where weighted correlation makes sense, as for instance in bioinformatics. In Chap. 4 we describe an application of weighted correlation in gene expression data where it is clear that the higher absolute expression values in microarray data are more important. Also, weighted correlation is not only important in the application per se, but also because it allowed the development of a new method of weighted Principal Component Analysis.

2.7 Conclusions About r_W

In this chapter, we describe a new rank measure of correlation r_W . It is applicable to problems where the level of correlation between two series of rankings is affected by the importance of each rank. We compare the new measure with Spearman's rank correlation coefficient and show that the weighted measure is clearly more suitable for such problems.

In r_W , we have used a linear rank weighting function to assign more importance to higher ranks (the first ranks). Although r_W is more suitable than Spearman's r_S for the type of applications we are concerned with, like those just described, the linear function may not be the best one in all of them. In the next chapter, we will analyze whether other weighting functions can be more adequate for specific situations. In information retrieval and recommender systems, for instance, the exponential weight function has been used in problem-specific measures [14, 40].

We have analyzed the new measure's asymptotic distribution and computationally show its tendency to the Gaussian curve. Next, the formal proof has been given. We have first concentrated on the null hypothesis that the two rankings are independent and then we have developed tests to do inference for other values of r_W .

Finally, in the last subsection, we claim that there is a wide range of applications where the weighted correlation coefficient r_W can be used to measure the concordance between two rankings.

Rankings and Preferences

New Results in Weighted Correlation and Weighted
Principal Component Analysis with Applications

Pinto da Costa, J.

2015, X, 91 p. 12 illus., 4 illus. in color., Softcover

ISBN: 978-3-662-48343-5