6-2016

Similarity Beyond Correlation: Symmetry-Based Approach

Ildar Batyrshin
Instituto Politénico Nacional, batyr1@gmail.com

Thongchai Dumrongpokaphan
Chiang Mai University, tcd43@hotmail.com

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

Comments:
Technical Report: UTEP-CS-16-32

Similarity Beyond Correlation: Symmetry-Based Approach

Ildar Batyrshin[1], Thongchai Dumrongpokaphan[2],
Vladik Kreinovich[3], and Olga Kosheleva[3]
[1]Centro de Investigaciń en Computación (CIC)
Instituto Politécnico Nacional (IPN)
México, D.F., batyr1@gmail.com
[2]Department of Mathematics, Faculty of Science
Chiang Mai University, Thailand
tcd43@hotmail.com
[3]University of Texas at El Paso
El Paso TX 79968, USA
vladik@utep.edu, olgak@utep.edu

Abstract

When practitioners analyze the similarity between time series, they often use correlation to gauge this similarity. Sometimes this works, but sometimes, this leads to counter-intuitive results. In this paper, we use natural symmetries – scaling and shift – to explain this discrepancy between correlation and common sense, and then use the same symmetries to come up with more adequate measures of similarity.

1 Correlation, and Why We Need to Go Beyond Correlation: Formulation of the Problem

Practitioners routinely use correlation to detect similarities. When a practitioner is interested in gauging similarity between two sets of related data or between two time series, a natural idea seems to be to look for (sample) *correlation*; see, e.g., [3]:

$$\rho(a,b) = \frac{C_{a,b}}{\sigma_a \cdot \sigma_b},$$

where

$$C_{a,b} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (a_i - \overline{a}) \cdot (b_i - \overline{b}),$$

$$\overline{a} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} a_i, \quad \overline{b} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{b} b_i, \quad \sigma_a \stackrel{\text{def}}{=} \sqrt{V_a}, \quad \sigma_b \stackrel{\text{def}}{=} \sqrt{V_b},$$

$$V_a \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (a_i - \overline{a})^2, \quad V_b \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (b_i - \overline{b})^2.$$

Of course, correlation has its limitations. Practitioners understand that correlation only detects *linear* dependence. In some cases, the dependence is non-linear; in such cases, simple correlation does not work, and more complex methods are needed to detect dependence.

However, in simple cases, when we do not expect nonlinear dependencies, correlation should be a perfect measure of similarity. And often it is. But sometimes, it is not. Let us give an example.

Example of a simple case when correlation is not an adequate measure of similarity. Let us consider a simple case, when we ask people to evaluate several newly released movies on a scale from 0 to 5, and then we compare their evaluations a_i, b_i, ..., of different movies i to gauge how similar their tastes are; see, e.g., [1].

For simplicity, let us assume that for six movies, the first person gave them the following grades:

$$a_1 = 4, \quad a_2 = 5, \quad a_3 = 4, \quad a_4 = 5, \quad a_5 = 4, \quad a_6 = 5,$$

while the second person gave

$$b_1 = 5, \quad b_2 = 4, \quad b_3 = 5, \quad b_4 = 4, \quad b_5 = 5, \quad b_6 = 4.$$

From the common sense viewpoint, these two viewers have similar tastes – they seem to like all the movies very much. This similarity is especially clear if we compare them with the evaluations of a picky third person who does not like any new moves at all:

$$c_1 = 0, \quad c_2 = 1, \quad c_3 = 0, \quad c_4 = 1, \quad c_5 = 0, \quad c_6 = 1.$$

However, if we compute correlations, we will get exactly opposite conclusions:

- between a_i and c_i, there is a perfect correlation $\rho = 1$, while
- between a_i and b_i, there is a perfect *anti*-correlation $\rho = -1$.

In other words:

- a cheerful viewer a and a gloomy viewer c – who, from the commonsense viewpoint, are opposites – have a perfect positive correlation, while
- two cheerful viewers a and b – who, from the commonsense viewpoint, are almost Siamese twins – show perfect negative correlation.

This example clearly shows that we need to go beyond correlation to captures the commonsense meaning of similarity.

Second example. Let us have a somewhat less trivial example – based on the saying that when America sneezes, the world catches cold. Let us use a simplified example. Suppose that the US stock market shows periodic oscillations, with relative values

$$a_1 = 1.0, \quad a_2 = 0.9, \quad a_3 = 1.0, \quad a_4 = 0.9.$$

In line with the above saying, the stock market in a small country X shows similar relative changes, but with a much higher amplitude:

$$b_1 = 1.0, \quad b_2 = 0.5, \quad b_3 = 1.0, \quad b_4 = 0.5.$$

Are these sequence similar? Somewhat similar yet, but not exactly the same: while the US stock market has relatively small 10% fluctuations, the stock market of the country X changes by a factor of two.

However, if we use correlation to gauge the similarity, we will see that these two stock markets have a perfect positive correlation $\rho = 1$. This example confirms that we need to go beyond correlation to capture the commonsense meaning of similarity.

What we do in this paper. In this paper, we describe natural alternatives to correlation, alternatives which are in better agreement with common sense.

Our ultimate goal is not only to introduce these two measures to the research community, but also to convince practitioners to use these new measures of similarity. Because of this practice-oriented goal, we tried our best to make our explanations and derivations as detailed as possible.

2 Why Is Correlation Not Always in Perfect Accordance with Common Sense: Natural Symmetries

Compared values come from measurements. To better understand why, for some time series a_i and b_i, there is sometimes such a discrepancy between commonsense meaning of similarity and correlation, let us recall how we get the values a_i and b_i. Usually, we get these values from measurements (see, e.g., [2]) – or, as in the example of evaluating movies, from expert estimates, which can also be considered as measurements, measurements performed by a human being as a measuring instrument.

Natural symmetries related to measurements. In the general measurement process, we transform actual physical quantities into numbers. For example, when we measure time, we transform an actual moment of time into a numerical value.

In general, to perform such a transformation, we need to select:

- a starting point and
- a measuring unit.

For example, if to measure time, we select the birth year of Jesus Christ as the starting point, and a usual calendar year as a measuring unit, we get the usual date in years. If instead we select the moment 2000.0 and use seconds as units, then we get astronomical time.

Similarly, we can measure temperature in the Fahrenheit (F) scale or in the Celsius (C) scale; these two scales have:

- different starting points: $0°\text{C} = 32°\text{F}$, and
- different units: a difference of 1 degree C is equal to the difference of 1.8 degrees Fahrenheit.

If we change a measuring unit to a new one which is u times smaller, then all numerical values get multiplied by this factor u: the same quantity that had the value x in the original units has the value $x' = u \cdot x$ in the new unit. For example, if we replace meters with centimeters, with $u = 100$, then a height of $x = 2$ m becomes $x' = 100 \cdot 2 = 100$ cm in the new units.

Similarly, if we change from the original the starting point to a new starting point which is s units earlier, then the original numerical value x is replaced by a new value $x' = x + s$.

In general, if we change both the measuring unit and the starting point, we get new units $x' = u \cdot x + s$.

In such general cases, correlation is a good description of similarity. For quantities for which we can arbitrarily select the measuring unit and the starting point, the same time series which is described by the numerical values x_i can be described by values $x'_i = u \cdot x_i + s$ in a different scale.

If we have a perfect correlation $\rho = 1$ between the two time series a_i and b_i, this means that after an appropriate linear transformation, we have $b_i = u \cdot a_i + s$. In other words, if we select an appropriate measuring unit and an appropriate starting point for measuring a, then the values $a'_i = u \cdot a_i + s$ of the quantity a as described in the new units will be identical to the values of the quantity b. And equality is, of course, a perfect case of what we intuitively understand by similarity.

This is why in many cases, correlation is indeed a perfect measure of similarity.

Not all quantities allow an arbitrary selection of measuring unit and starting point. The problem is that some quantities only allow some of the above symmetries – or none at all.

For example:

- while (as we have mentioned earlier) we can select different *units* for the distance between the points,

- we cannot select an arbitrary *starting point*: there is a natural starting point 0 that corresponds to the distance between the two identical points.

In this case, the fact that we can, e.g., obtain two series of distances a_i and b_i from one another by a shift does not make them similar: since this shift no longer has an intuitive sense.

This was exactly the case of two stock markets: for any price, 0 is a natural starting point, so:

- while scalings $x \to u \cdot x$ make sense,
- shifts $x \to x + s$ change the situation – often drastically.

For movie evaluations, the results are even less flexible: here, both the measuring unit and the starting point are fixed: any transformation will change the meaning. For the same physical distance, we can have two different values, e.g., 100 miles and 160 km, but for evaluations on a scale from 0 to 5, different numbers simply mean different evaluations.

In such cases, correlation – which is based on detecting a general linear dependence – is clearly not an adequate measure of similarity.

So how should we gauge similarity in such cases? Up to now, we showed that natural symmetries explain why correlation is not always a perfect measure of similarity. Let us now use natural symmetries to come up with measures of similarity which are adequate in such situations.

3 Starting Point: Case When No Scaling Is Possible

Description of the case. Let us start with the case when both measuring unit and starting point are fixed, all measurement results are absolute, and no scaling is possible – as in the case of viewers evaluating movies.

Natural idea. In this case, how can we gauge similarity between the two times series $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$? The closer the two tuples, the more similar are these tuples. Thus, a natural measure of dissimilarity is simply the distance $d(a, b)$ between these two tuples:

$$d(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}. \tag{1}$$

From the computational viewpoint, this idea can be slightly improved. The above formula (1) is reasonable. However, our goal is not just to come up with a reasonable idea, but, ideally, to come with an idea to be applied in practice. From the practical viewpoint, the simpler the computations, the easier it is to apply the corresponding idea.

From this viewpoint, the above expression is not perfect; namely, in addition to:

- subtractions $b_i - a_i$ (which are easy to perform even by hand),
- multiplications $(b_i - a_i) \cdot (b_i - a_i)$ (which are also relatively easy to perform), and
- additions to compute the sum $(b_1 - a_1)^2 + (b_2 - a_2)^2 + \ldots$ (also easy),

we also need to compute the square root – which is not easy to perform by hand.

Good news is that the main purpose of gauging similarity is not so much to come up with some "absolute" number describing similarity, but rather to be able to compare the degree of similarity between different pairs (a_i, b_i). For example, in prediction, we can say that if a new situation a is sufficient similar to one of the past situations b – i.e., if the degree of similarity between them exceeds a certain threshold – then it is reasonable to predict that the situation a will come up with the same changes as were observed in the situation b in the past.

From this viewpoint, it does not matter that much how we assign numerical values to different degrees of similarity. We can change the numerical values of these degrees – as long as we preserve the order between them. In particular, when we square all the distances, then clearly larger distances become larger squares, and vice versa. Thus, instead of the original distances (1), we can as well consider their squares

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

Conclusion: in this case, distance is a natural measure of similarity. Summarizing, we can say that in situations when no scaling is possible – like in the case of movie evaluations – a reasonable idea is to use, as a reasonable measure of similarity,

- *not* the correlation (as practitioners are sometimes tempted to), but
- the *distance* (or squared distance) between the two series.

Comment. It should be emphasized once again, that, in contrast to correlation – which attempts to describe *similarity* – distance describes *dissimilarity*:

- the larger correlation, the more similar the two time series, but
- the larger the distance, the less similar are the two time series.

4 Somewhat Surprisingly, When All Scalings Are Allowed, We Come Back to Correlation

Description of the case. To see how good is the distance as the measure of similarity, let us apply this idea to the generic case, when all scalings are applicable. In other words, we consider the case when numerical values of both quantities a_i and b_i are defined only modulo general linear transformations $a \to u \cdot a + s$ and $b \to u' \cdot b + s'$, for any $u > 0$, $u' > 0$, s, and s'.

Starting point is distance: reminder. When the units and the starting points are fixed, we get the usual distance – or, to be more precise, squared distance $d^2(a,b) = \sum\limits_{i=1}^{n} (a_i - b_i)^2$.

How to take care of possible re-scalings of the quantity a. If this distance is small, this means that the time series a_i and b_i are similar. However, when this distance is large, this does not necessarily mean that the time series a_i and b_i are not similar – maybe we chose a wrong unit and/or a wrong starting point for measuring a, and the distance will be much smaller if we use a different unit and/or a different starting point. From this viewpoint, instead of considering the distance $d(a,b)$ between the original numerical values a_i and b_i, it makes more sense to consider the distance between b_i and re-scaled values $u \cdot a_i + s$ – and consider the smallest possible value of this distance as a measure of this dissimilarity:

$$D_g(a,b) = \min_{u,s} d^2(u \cdot a + s, b) = \min_{u,s} \sum_{i=1}^{n} (b_i - (u \cdot a_i + s))^2. \tag{3}$$

How to take care of re-scalings of the quantity b. The formula (3) takes care of re-scaling the values a_i, but it may change if we re-scale the values b_i. At first glance, it may seem that it we can solve this problem by also taking the minimum also over all possible re-scalings of b as well. However, this will not work: e.g., if we choose a very large measuring unit for measuring b, then the numerical values of b_i can become very small – and thus, the value (3) can also become arbitrarily small, and the minimum will always be 0.

To make the formula (3) scale-invariant, it is reasonable:

- instead of considering the *absolute* size of the discrepancies $b_i - (u \cdot a_i + s)$ between the values b_i and the values $u \cdot a_i + s$ predicted by a_i,
- to consider *relative* size, relative to the size of how the values b_i themselves are different from 0,

i.e., the value

$$D'_g(a,b) = \frac{\min\limits_{u,s} \sum\limits_{i=1}^{n} (b_i - (u \cdot a_i + s))^2}{\sum\limits_{i=1}^{n} b_i^2} = \min_{u,s} \frac{\sum\limits_{i=1}^{n} (b_i - (u \cdot a_i + s))^2}{\sum\limits_{i=1}^{n} b_i^2}. \tag{4}$$

Need to take care of possible shifts in b. The formula (4) take care of scaling $b_i \to u' \cdot b_i$ – which, as one can check, does not change the value (4), but it still does change with the shift $b_i \to b_i + s'$.

Again, at first glance, it may seem reasonable to consider all possible shifts of b and to take the minimum – but then, after shifting b by a large amount, we do not change the numerator but we can make the denominator arbitrarily large. Thus, the result will be a meaningless 0.

Good news is that instead of taking the *minimum* over all possible shifts, we can get a meaningful result if we take the *maximum* overall all possible shifts. Thus, we arrive at the following definition.

Definition 1. *For every two tuples $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$, we define a measure of dissimilarity as*

$$d_g(a,b) = \max_{u',s'} \min_{u,s} \frac{\sum\limits_{i=1}^{n} ((u' \cdot b_i + s') - (u \cdot a_i + s))^2}{\sum\limits_{i=1}^{n} (u' \cdot b_i + s')^2}. \tag{5}$$

Discussion. From this expression, it is not even clear whether this expression in symmetric in terms of a and b, i.e., whether $d_g(a,b) = d_g(b,a)$. This is indeed true, and it is easy to see once we realize that $d_g(a,b)$ is directly related to the usual sample correlation:

Proposition 1. $d_g(a,b) = 1 - \rho^2(a,b)$**.**

Proof. To compute the expression (5), let us first compute the minimum

$$D'_g(a,b') \stackrel{\text{def}}{=} \min_{u,s} \frac{\sum\limits_{i=1}^{n} (b'_i - (u \cdot a_i + s))^2}{\sum\limits_{i=1}^{n} (b'_i)^2}, \tag{6}$$

where we denoted $b'_i \stackrel{\text{def}}{=} u' \cdot b_i + s'$. The denominator $\sum\limits_{i=1}^{n} (b'_i)^2$ does not depend on u and s, so minimizing the ratio (6) is equivalent to minimizing its denominator

$$J_g \stackrel{\text{def}}{=} \sum_{i=1}^{n} (b'_i - (u \cdot a_i + s))^2. \tag{7}$$

To compute the minimum of the expression (7), we differentiate J_g with respect to u and s and equate the derivatives to 0.

Differentiating with respect to s, we get

$$u \cdot \sum_{i=1}^{n} a_i + s \cdot n - \sum_{i=1}^{n} b'_i = 0. \tag{8}$$

Dividing both sides of this equation by n, we get a simpler formula

$$u \cdot \overline{a} + s = \overline{b'}, \tag{9}$$

where, in line with the notations from Section 1, $\overline{b'}$ means the arithmetic average of the values b'_i.

Similarly, differentiation with respect to u leads to

$$u \cdot \sum_{i=1}^{n} a_i^2 + s \cdot \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} b'_i \cdot a_i = 0, \tag{10}$$

i.e., to

$$u \cdot \overline{a^2} + s \cdot \overline{a} = \overline{a \cdot b'}, \tag{11}$$

where we denoted

$$\overline{a^2} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} a_i^2 \text{ and } \overline{a \cdot b'} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} a_i \cdot b'_i.$$

We now have two equations (9) and (11) to find two unknowns u and s. To find u, we can eliminate s; this can be done if we multiply both sides of the equation (9) by $\overline{a}$ and subtract the result from the equation (11), then we get

$$u \cdot \left(\overline{a^2} - (\overline{a})^2 \right) = \overline{a \cdot b'} - \overline{a} \cdot \overline{b}'. \tag{12}$$

We can easily check that $\overline{a^2} - (\overline{a})^2 = V_a$ and $\overline{a \cdot b'} - \overline{a} \cdot \overline{b}' = C_{a,b'}$, thus

$$u = \frac{C_{a,b'}}{V_a}. \tag{13}$$

Now, from (9), we conclude that

$$u \cdot a_i + s - b'_i = u \cdot a_i + s - b'_i - (u \cdot \overline{a} + s - \overline{b'}) = u \cdot (a_i - \overline{a}) - (b'_i - \overline{b'}). \tag{14}$$

Thus,

$$(u \cdot a_i + s - b'_i)^2 = u^2 \cdot (a_i - \overline{a})^2 - 2u \cdot (a_i - \overline{a}) \cdot (b'_i - \overline{b'}) + (b'_i - \overline{b'})^2, \tag{15}$$

and therefore,

$$\sum_{i=1}^{n} (u \cdot a_i + s - b'_i)^2 = u^2 \cdot \sum_{i=1}^{n} (a_i - \overline{a})^2 - 2u \cdot \sum_{i=1}^{n} (a_i - \overline{a}) \cdot (b'_i - \overline{b'}) + \sum_{i=1}^{n} (b'_i - \overline{b'})^2. \tag{16}$$

Dividing both sides by n, we conclude that

$$\frac{J_g}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} (u \cdot a_i + s - b'_i)^2 = u^2 \cdot V_a - 2u \cdot C_{a,b'} + V_{b'}, \tag{17}$$

where

$$V_{b'} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (b'_i - \overline{b'})^2.$$

Substituting the expression (13) into this formula, we conclude that

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (u \cdot a_i + s - b'_i)^2 = \frac{C_{a,b'}}{V_a} - 2\frac{C_{a,b'}}{V_a} + V_{b'} = V_{b'} - \frac{C_{a,b'}}{V_a} =$$

$$V_{b'} \cdot \left(1 - \frac{C_{a,b'}}{V_a \cdot V_{b'}}\right) = V_{b'} \cdot (1 - \rho^2(a, b')). \tag{18}$$

Thus, the ratio (6) can be described as

$$D'_g(a, b') = \frac{V_{b'} \cdot (1 - \rho^2(a, b'))}{\overline{(b')^2}}. \tag{19}$$

One can easily check that the correlation does not change under a linear transformation of one of the variables, so $\rho(a, b') = \rho(a, b)$. Thus, the formula (19) takes a simplified form

$$D'_g(a, b') = \frac{V_{b'}}{\overline{(b')^2}} \cdot (1 - \rho^2(a, b)). \tag{20}$$

Here, $V_{b'} = \overline{(b')^2} - \left(\overline{b'}\right)^2 \leq \overline{(b')^2}$ thus the ratio $\dfrac{V_{b'}}{\overline{(b')^2}}$ is always smaller than or equal to 1. The largest possible value 1 of this ratio is attained when $\overline{b'} = 0$ – which we can always achieve by selecting an appropriate shift s' (namely, $s' = -\overline{b}$). In this case, the value $D'_g(a, b')$ is equal to $1 - \rho^2(a, b)$. Thus,

$$d_g(a, b) = \max_{u', s'} D'_g(a, b') = 1 - \rho^2(a, b).$$

The proposition is proven.

Discussion. So, we confirmed that our approach makes sense – and it even leads to a non-statistical explanation of correlation. This enables us to use correlation beyond its usual Gaussian distribution case).

Let us now see what this approach results in in situations when only some of the natural symmetries are meaningful.

5 Case When Only Scaling Makes Sense – But Not Shift

Description of the case. Let us consider the case when a starting point is fixed, but we can choose an arbitrary measuring unit. (This is true, e.g., in the above the case of stock markets.)

In this case, we can have transformations $a_i \to a_i' = u \cdot a_i$ and $b_i \to b_i' = u' \cdot b_i$.

Analysis of this situation. In this case, instead of considering the distance $d(a,b)$ between the original numerical values a_i and b_i, it makes more sense to consider the distance between b_i and re-scaled values $u \cdot a_i$ – and consider the smallest possible value of this distance as a measure of this dissimilarity:

$$D_u(a,b) = \min_u d^2(u \cdot a, b) = \min_u \sum_{i=1}^{n} (b_i - u \cdot a_i)^2. \tag{21}$$

This takes care of re-scaling the values a_i. To take case of re-scalings of the values b_i, we can use the same idea as in the general case, and consider the ratio

$$D_u'(a,b) = \frac{\min\limits_u \sum\limits_{i=1}^{n} (b_i - u \cdot a_i)^2}{\sum\limits_{i=1}^{n} b_i^2} = \min_u \frac{\sum\limits_{i=1}^{n} (b_i - u \cdot a_i)^2}{\sum\limits_{i=1}^{n} b_i^2}. \tag{22}$$

It turns out that this ratio does not change if we re-scale b_i as well – this follows, e.g., from Proposition 2 proven below. So, we arrive at the following definition:

Definition 2. *For every two tuples $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$, we define a measure of dissimilarity as*

$$d_u(a,b) = \min_u \frac{\sum\limits_{i=1}^{n} (b_i - u \cdot a_i)^2}{\sum\limits_{i=1}^{n} (b_i)^2}. \tag{23}$$

Comment. The following result provides an explicit formula for this measure of dissimilarity.

Proposition 2. $d_u = 1 - \dfrac{\left(\overline{a \cdot b}\right)^2}{\overline{a^2} \cdot \overline{b^2}}$.

Comment. In particular, when $\overline{a} = \overline{b} = 0$, we have $\overline{a \cdot b} = C_{a,b}$, $\overline{a^2} = V_a$, $\overline{b^2} = V_b$, and thus, this formula turns into the correlation-related formula $d_u = 1 - \rho^2(a,b)$.

In this case, correlation can be reconstructed as $\rho(a,b) = \sqrt{1 - d_u(a,b)}$. In general, we can therefore view the expression $\sqrt{1 - d_u(a,b)}$ as an analogue of correlation.

In the above example of two stock markets for which $\rho(a,b) = 1$ but common-sense similarity is not perfect, we have

$$\overline{a^2} = \frac{1^2 + 0.9^2 + 1^2 + 0.9^2}{4} = 0.905, \quad \overline{b^2} = \frac{1^2 + 0.5^2 + 1^2 + 0.5^2}{4} = 0.625, \text{ and}$$

$$\overline{a \cdot b} = \frac{1 \cdot 1 + 0.9 \cdot 0.5 + 1 \cdot 1 + 0.9 \cdot 0.5}{4} = 0.725.$$

Thus, here,

$$d_u = 1 - \frac{(0.725)^2}{0.905 \cdot 0.625} \approx 1 - 0.929 = 0.071 > 0.$$

Hence, the above equivalent of correlation $\sqrt{1 - d_u}$ is approximately equal to 0.96, which is smaller than 1 – as desired.

Proof of Proposition 2. We want to minimize the ratio

$$\frac{\sum\limits_{i=1}^{n} (b_i - u \cdot a_i)^2}{\sum\limits_{i=1}^{n} (b_i)^2}$$

with respect to u. The denominator of this ratio does not depend on u, so to find the minimum, it is sufficient to minimize the numerator

$$J_u \stackrel{\text{def}}{=} \sum_{i=1}^{n} (b_i - u \cdot a_i)^2.$$

Differentiating J_u with respect to u and equating the derivative to 0, we conclude that

$$\sum_{i=1}^{n} (u \cdot a_i - b_i) \cdot a_i = u \cdot \sum_{i=1}^{n} a_i^2 - \sum_{i=1}^{n} a_i \cdot b_i = 0.$$

Dividing both sides of this equality by n, we conclude that $u \cdot \overline{a^2} = \overline{a \cdot b}$, i.e., that the optimal u has the form

$$u = \frac{\overline{a \cdot b}}{\overline{a^2}}.$$

For this optimal value u, the numerator J_u takes the form

$$J_u = \sum_{i=1}^{n} (u \cdot a_i - b_i)^2 = u^2 \cdot \sum_{i=1}^{n} a_i^2 - 2u \cdot \sum_{i=1}^{n} a_i \cdot b_i + \sum_{i=1}^{n} b_i^2.$$

Thus,

$$\frac{J_u}{n} = u^2 \cdot \overline{a^2} - 2u \cdot \overline{a \cdot b} + \overline{b^2}.$$

Substituting the above optimal value of u into this expression, we conclude that

$$\frac{J_u}{n} = \frac{\left(\overline{a \cdot b}\right)^2}{\overline{a^2}} - 2\frac{\left(\overline{a \cdot b}\right)^2}{\overline{a^2}} + \overline{b^2} = \overline{b^2} - \frac{\left(\overline{a \cdot b}\right)^2}{\overline{a^2}}.$$

To get the desired value $d_u(a,b)$, we need to divide this expression by $1/n$ of the denominator $\sum\limits_{i=1}^{n} b_i^2$, i.e., by the value $\overline{b^2}$. After this division, we get the desired expression

$$d_u(a,b) = 1 - \frac{\left(\overline{a \cdot b}\right)^2}{\overline{a^2} \cdot \overline{b^2}}.$$

The proposition is proven.

6 Case When Only Shift Makes Sense – But Not Scaling

Description of the case. Let us consider the case when a measuring unit is fixed, but we can choose an arbitrary starting point.

In this case, we can have transformations $a_i \to a_i' = a_i + s$ and $b_i \to b_i' = b_i + s'$.

Analysis of the situation. In this case, instead of considering the distance $d(a,b)$ between the original numerical values a_i and b_i, it makes more sense to consider the distance between b_i and shifted values $a_i + s$ – and consider the smallest possible value of this distance as a measure of this dissimilarity:

$$D_s(a,b) = \min_s d^2(a+s,b) = \min_s \sum_{i=1}^{n} (b_i - (a_i + s))^2. \tag{24}$$

This takes care of shifting the values a_i.

It turns out that this ratio does not change if we shift b_i as well – this follows, e.g., from Proposition 3 proven below. So, we arrive at the following definition:

Definition 3. *For every two tuples $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$, we define a measure of dissimilarity as*

$$D_s(a,b) = \min_s \sum_{i=1}^{n} (b_i - (a_i + s))^2. \tag{25}$$

Comment. The following result provides an explicit formula for this measure of dissimilarity.

Proposition 3. $D_s(a,b) = n \cdot (V_a + V_b - 2C_{a,b})$.

Comment. In the previous two cases, there was a possibility to re-scale b_i. To make the resulting measure of (dis)similarity independent on such re-scaling, we had to divide the squared distance by $\sum\limits_{i=1}^{n} b_i^2$, i.e., consider *relative* discrepancy instead of the absolute one.

In the case when only shifts make physical sense, re-scaling of b_i is not possible, so there is no need for such a division.

What we *can* do is make sure that the value of dissimilarity does not depend on the sample size – in the sense that if we combine two identical samples, the dissimilarity will be the same. Such doubling does not change the sample variances and covariances V_a, V_b, and $C_{a,b}$; thus, after this doubling, the above D_s also doubles. To make it independent on such doubling, we can therefore divide the above expression D_u by the sample size and thus, get a new measure

$$d_u \stackrel{\text{def}}{=} \frac{D_u}{n} = V_a + V_b - 2C_{a,b}.$$

Such a division was not needs in the above two cases – since there, as we can see from Propositions 1 and 2, the division by the sum $\sum\limits_{i=1}^{n} b_i^2$ automatically resulted in doubling-invariance.

Proof of Proposition 3. Differentiating the expression

$$J_s \stackrel{\text{def}}{=} \sum_{i=1}^{n} (b_i - (a_i + s))^2$$

with respect to s and equating the derivative to 0, we conclude that

$$\sum_{i=1}^{n} (a_i + s - b_i) = \sum_{i=1}^{n} a_i + s \cdot n - \sum_{i=1}^{n} b_i = 0.$$

Dividing both sides of this equality by n, we conclude that $\overline{a} + s = \overline{b}$, so

$$s = \overline{b} - \overline{a}.$$

For this value s, we have

$$a_i + s - b_i = a_i + \overline{b} - \overline{a} - b_i = (a_i - \overline{a}) - \left(b_i - \overline{b}\right).$$

Thus, the objective function J_s takes the form

$$J_s = \sum_{i=1}^{n} (a_i + s - b_i)^2 = \sum_{i=1}^{n} \left((a_i - \overline{a}) - \left(b_i - \overline{b}\right)\right)^2 =$$

$$\sum_{i=1}^{n} (a_i - \overline{a})^2 - 2\sum_{i=1}^{n} (a_i - \overline{a}) \cdot \left(b_i - \overline{b}\right) + \sum_{i=1}^{n} \left(b_i - \overline{b}\right)^2.$$

If we divide both sides of this equality by n, we get

$$\frac{J_s}{n} = V_a - 2C_{a,b} + V_b,$$

which implies the desired formula for $D_s(a, b)$.

The proposition is proven.

Acknowledgments

This work is supported by Chiang Mai University, Thailand. It was also supported in part:

- by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721,
- by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation, and
- by a grant Mexico's Instituto Politecnico Nacional.

References

[1] I. Batyrshin, "Fuzzy Logic and Non-Statistical Association Measures", *Proceesings of the 6th World Conference on Soft Computing*, Berkeley, California, May 22–25, 2016.

[2] S. G. Rabinovich, *Measurement Errors and Uncertainty. Theory and Practice*, Springer Verlag, Berlin, 2005.

[3] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.