Determination-based correlation coefficient

Andrii Berko[1,†], Vladyslav Alieksieiev[1,†], Vladyslav Holdovanskyi[1,*,†]

[1] *Lviv Polytechnic National University, 12 Stepana Bandery Street, Lviv, 79014, Ukraine*

Abstract

We introduce a novel approach to defining the correlation coefficient based on local dispersion, which reflects the degree of non-linear relationship between two random variables. Utilizing the theory of random processes, we adapted computations of random function characteristics to address the correlation problem. Throughout our research, we identified and implemented a new correlation coefficient that allows analyzing non-linear relationships between random variables. Our results confirmed that this new approach is effective for processing large volumes of data. Additionally, we explored the relationship between our new correlation coefficient and the popular machine learning model evaluation metric, R2. This simplifies the interpretation of correlation analysis results and makes it more informative for analysis. We also analyzed the satisfaction of Renyi's properties and other useful for practice properties of our new coefficient to understand its behavior under different conditions. We compared it with other approaches and demonstrated the advantages and disadvantages of our method. To demonstrate the effectiveness of our coefficient, we utilized both simulated and real data. Our results on both types of data show the potential of our method and its possible application in practical tasks.

Keywords

correlation analysis, non-linear dependence measure, coefficient of determination

1. Introduction

Correlation plays a key role in the fields of statistics, data analysis, and machine learning. It is a fundamental tool for measuring and establishing relationships between variables. In doing so, it provides valuable information about patterns, the type of relationship, its strength, potential predictive capabilities of future models, and more. Among other metrics and statistics, correlation stands out as a widely used and versatile indicator. It is applied in research across various fields—from finance and economics to psychology and epidemiology[1-10].

The history of correlation analysis spans over 100 years—starting with Pearson's correlation coefficient and other methods from the late 19th to the early 20th century and continuing to the present day. Even today, articles continue to emerge about new methods

of defining correlation or improving existing ones [11-19]. Currently, there are dozens of different coefficients: classical coefficients, coefficients of maximum correlation, coefficients based on joint cumulative distribution functions, coefficients based on information theory, coefficients based on copulas, coefficients based on distance, and so on.

The reason why there are so many correlation coefficients is evident—they all have their advantages and disadvantages. A coefficient that may be applicable in one scenario, with certain objectives, may be less effective or even unsuitable for use in other conditions and for other purposes.

The reasons why preference is given to certain coefficients over others can be quite diverse-lack of independence test or presence of such a test but inability to indicate the strength of the relationship; inability of the coefficient to establish a non-linear relationship or complexity of the algorithm, making its application on large datasets impractical, and so on.

Thus, there is still room in science for new correlation coefficients, and we aim to present our developments and research. The article discusses our journey to this coefficient. Additionally, we will examine and compare the methods of other researchers who had similar ideas and/or implementations. We will show the connection of our coefficient with the coefficient of determination and the Pearson coefficient. We will discuss the properties, advantages, and disadvantages of our coefficient and compare it with others. We will also demonstrate its performance on simulated and real data.

2. Related Works

Modifying the template — including but not limited to: adjusting margins, typeface sizes, line spacing, paragraph and list definitions — is not allowed. The development of a new correlation coefficient based on random processes and the coefficient of determination is situated within the context of extensive scientific research on correlation coefficients and their applications. This section aims to provide an expanded reflection on the relevant literature, which serves as the foundation for our research and the innovative development of the proposed coefficient.

Considering the regression problem as a random function, I utilized the theory of random processes to analyze the characteristics of this function. Foundational texts[20] on random processes aided me in unraveling the nature of this random function and understanding its properties. Their high-quality information provided the groundwork for my research, allowing me to apply mathematical concepts of random processes to study regression models. Such an approach enabled me to develop a new coefficient that takes into account all the significant characteristics of this random function to provide a more precise and adequate model for analyzing dependencies between variables.

The works of Fisher and Pearson[21, 22], particularly their seminal works, have immense significance for contemporary research on correlation and statistical analysis. In their works, they introduced the concept of correlation for the first time and developed mathematical methods for its measurement. They demonstrated the importance of establishing robust measures of correlation for adequately studying relationships between different variables in datasets. The works of Fisher and Pearson have played a pivotal role

in the development of statistical theory and methodology, and their influence can still be felt in modern research.

Rényi's article[23] opens up new perspectives for understanding the essence of correlation and defining its key characteristics. It provides a systematic approach to analyzing the properties of correlation coefficients and identifies seven fundamental criteria that effective measures of correlation should satisfy. These properties described by Rényi are crucial for developing new methods of correlation analysis and evaluating their effectiveness.

The RDC article[24], which investigates the correlation coefficient, provides important context for comparison with other correlation coefficients. In particular, this work establishes that the properties of marginal invariance, i.e., remaining unchanged when scaling or shifting data, are extremely important for the correlation coefficient. This is an important characteristic that ensures the stability of analysis results regardless of the chosen unit of measurement.

Furthermore, important aspects of the work include the use of vector input data and compliance with Rényi's criteria. The use of vector input data allows for considering the multidimensionality of input data, which can be useful in analyzing complex systems. Meeting Rényi's criteria underscores the importance of defined criteria for the effectiveness and robustness of the correlation coefficient.

Additionally, the article explores aspects of algorithm complexity and efficiency, which are key factors in assessing its effectiveness. The ability to capture nonlinear dependencies between variables is also an important aspect that reflects the potential of the correlation coefficient in detecting complex relationships in data. Thus, this article demonstrates the characteristics by which we can compare correlation coefficients.

Moreover, the comparative analysis of different correlation coefficients, as described in [24, 25], provided us with important context for critically evaluating the strengths and weaknesses of existing methodologies.

Contemporary research continues to actively seek correlation coefficients that adequately reflect complex relationships in diverse datasets. Many articles illustrate this search[11-19], exploring new coefficients and their applications in various fields. These studies reflect the ongoing evolution of correlation analysis and the need to develop informative methodologies to meet the demands of modern data analysis.

Moreover, it is worth noting that there is a large body of scientific literature that utilizes various correlation coefficients and conducts correlation analysis[1-10]. This indicates the widespread use of these methods across various fields of science and practical applications. Knowledge and skills in this area are highly valuable and consequently in high demand in both scientific and professional environments.

3. Methods and Materials

In this section, we will delve into the mathematical aspect of our study, focusing on the analysis of regression models in the context of random processes. We will consider the regression problem as a random function, which will open up new perspectives for us in understanding this phenomenon. Following this, we will examine the process of finding the

grid for approximating the random function, which is a key stage in our research. Finally, we will analyze the approach to deriving the correlation coefficient, which will help us uncover the relationship between different variables in the regression model.

3.1. The use of random process theory for regression and correlation tasks.

We were interested in the properties of random function characteristics. Although random functions fall within the realm of random processes rather than regression or correlation tasks, there is an apparent similarity. The advantage of random function characteristics, such as mean and variance, is also evident [20]. When considering them as characteristics of a random variable, we see only scalar quantities characterizing the distribution of the random function. However, the properties of random function characteristics are much more interesting. They transform random features in functions into scalar form while preserving non-random dependencies. Clearly, such properties of these characteristics would be interesting for regression and correlation tasks. Therefore, we decided to explore the possibilities of applying them to these tasks.

The main obstacle is the inability to find functions of these same characteristics. Theoretical formulas work only in theory. In practice, they are found by conducting n experiments, with a series of sections corresponding to specific moments in time (typically considering problems where the argument is time) – $t_1, t_2 \dots t_m$ [20]. And already at this stage, there arises a discrepancy in tasks. When it comes to random processes, especially when we consider processes related to time, it is easy to conduct experiments by measuring the value of the random function at certain moments in time. However, when we talk about a more general regression or correlation task, where we cannot control the sample, such measurements are impossible.

For example, let`s take the iris dataset created by Fisher. Suppose we have two tasks: 1) we are trying to understand the dependence of flower petal length on other indicators (regression task); 2) we are trying to understand the dependence of petal length on time (random process task). In the second case, we can grow n flowers and measure, for example, the length of petals daily. In the first case, we cannot fix certain values of any of the features. For almost any typical regression task, this is impossible. For example, we can take any of the classical datasets: Boston Housing (to find the dependence of house price on its characteristics), Wine (contains results of chemical analysis of wines), Titanic (contains some data of Titanic passengers and whether they survived), Diamonds (to find the dependence of diamond price on its characteristics), Diabetes (to find the dependence between human indicators and whether they have diabetes). Trying to fix diamonds, people's health, house characteristics, etc., is either physically impossible or unethical. Moreover, it contradicts the fact that the sample must be random. Therefore, we need a different method.

Before we continue to search for a way to construct characteristic functions, we would like to pause for a moment and still consider the typical regression task as a random function. In this article, we will consider a simple case with one dependent and one independent variable, as well as one other variable that will correspond to noise or randomness.

Let x be the known independent random variable, z be the unknown independent random variable, y be the dependent variable, which is also the target feature. There is a functional dependence between y and x characterized by the function - $\varphi(x)$. The feature z also affects y, but we do not know what this feature is, or perhaps it is a combination of several features and cannot be measured. We will consider this dependence as:

$$y = \varphi(x) + z, \tag{1}$$

Now, everything looks quite typical, but it is from this moment that differences arise in regression and random process tasks. Random processes suggest that each function is a realization of a random variable z, and accordingly, for measurements of functions $x_1, x_2, \dots x_m$, it is the same. However, in regression tasks, a realization is one value of x, and the experiment measurement is not measurements of the function at m points x, but one pair (x_i, y_i), for each of which z will be different. Such a dependence (1) can be considered as a random function, although for regression tasks, there is a problem with measurements to estimate characteristics.

We propose to bypass this problem by discretizing x. We can divide our pairs of points (x_i, y_i) by x, obtaining intervals, and then calculate characteristics within them. Although it is evident that the accuracy of such a method will depend on the division and the intervals themselves, it will allow for different estimates.

Let`s introduce the concept of a grid:

$$\omega_h = \left\{ M_{x \in [x_i; x_{i+1}]}(y), x_i = a + ih, i = \overline{0, \dots N}, h = \frac{b-a}{N} \right\}, \tag{2}$$

This grid defines an equidistant partitioning of y based on x. Although this grid roughly approximates the function of the mathematical expectation and, in general, the function itself, as we will show later. To illustrate this, let's consider a quantity called the average value of the function [26].

$$avf(y) = \frac{1}{b-a} \int_a^b f(x)dx,$$

where a and b are the limits of the interval on which we are looking for the mean value. This quantity is an analog of the simple mean for a continuous function. This formula directly follows from the simple mean:

$$\frac{1}{n} \sum_{i=1}^n y_i = \frac{(b-a)}{n(b-a)} \sum_{i=1}^n y_i = \frac{1}{b-a} \sum_{i=1}^n y_i \ \Delta x.$$

The obtained sum as $n \to \infty$ can be considered as an interval:

$$\lim_{n \to \infty} \frac{1}{b-a} \sum_{i=1}^n y_i \Delta x = \frac{1}{b-a} \int_a^b \varphi(x)dx$$

Avf is bounded below and above. This follows from the mean value theorem for integrals: if f(x) is a continuous function on [a,b], then there exists a c such that:

$$f(c) = \frac{1}{b-a} \int_a^b f(x)dx,$$

and from here follows the following:

$$\inf_{x \in [a,b]} f(x) < f(c) < sup_{x \in [a,b]} f(x).$$

Therefore, our function is bounded both above and below, and the accuracy of the estimate by simple averaging over the interval depends on the number of points, the size of the interval, and the rate of change of the function.

It is evident that for (1), simple averaging yields the following estimate:

$$M_{x\in[a,b]} = avf_{x\in[a,b]}\big(\varphi(x)\big) + M(z).$$

Next, we will demonstrate how well this model approximates the random function. We generated 10,000 points and set y as a certain function of x plus the addition of noise alpha (see Figure 1).
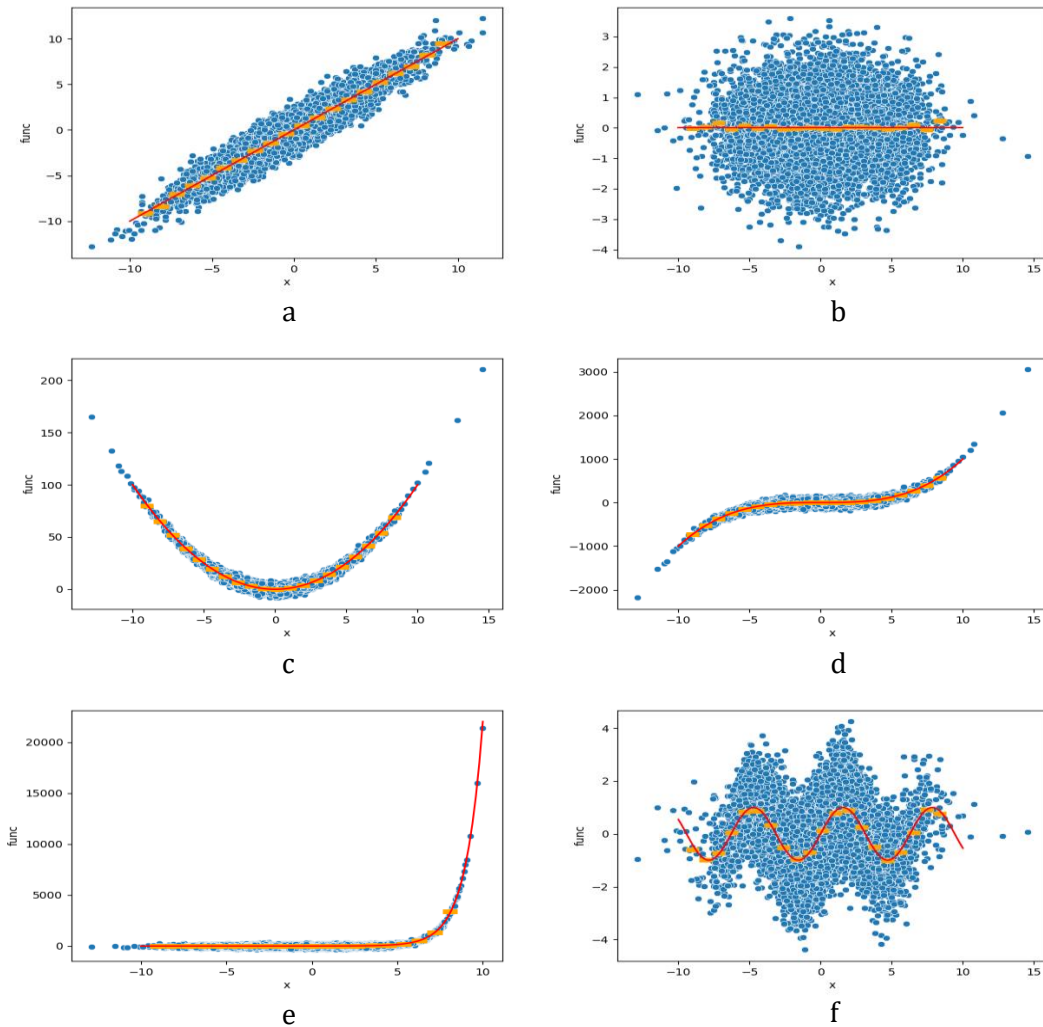
Figure 1: Demonstration of a grid for random functions: a)linear dependency, b) no dependency, c) quadratic, d) cubic, e) exponential, f) sinusoidal. Blue dots represent generated points, orange denotes grid intervals, red indicates the real function.

From the graphs, it can be seen that while the edges of the interval deviate significantly from the true values, the centers of the intervals approximate the function excellently.

Moreover, if instead of intervals (which our grid currently represents), we display a line plot, the approximation will be even more apparent(see Figure 2).
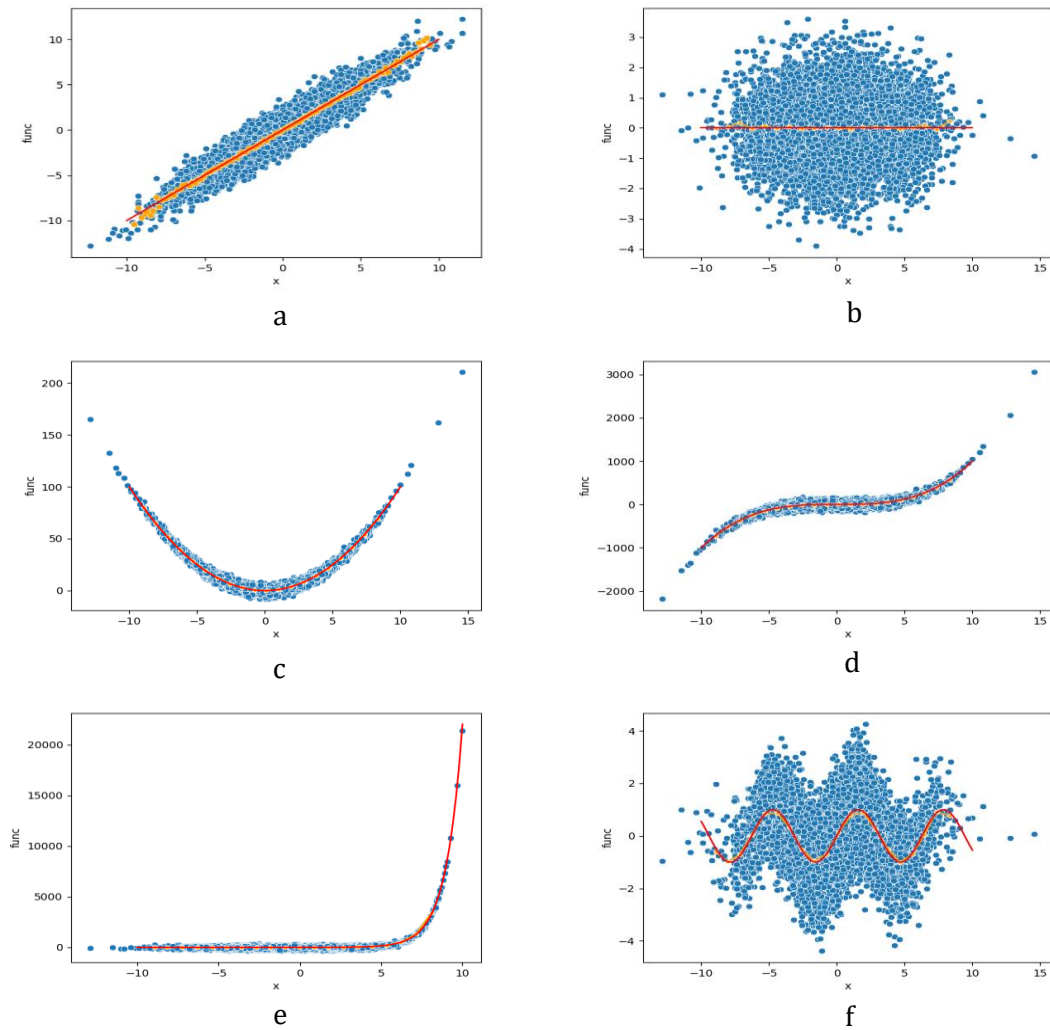
Figure 2. Demonstration of a grid via a line plot for random functions: a) linear dependency, b) no dependency, c) quadratic, d) cubic, e) exponential, f) sinusoidal. Blue dots represent generated points, orange denotes interval centers and lines between them, red indicates the real function.

From these graphs, it can be observed that if the points - the centers of the intervals - are connected, these lines almost overlay the curves of the real function, indicating the potential of the function.

3.2. Derivation of the correlation coefficient

In the previous section, we concluded that our grid (2) effectively approximates the random function. However, the aim of this article is different: based on this grid, we aim to construct the correlation coefficient.

Before defining the measure characterizing the strength of the relationship, it is necessary to establish the concept of absence of correlation.

Definition. We consider the absence of correlation of y from x when $\varphi(x) = const$, or in other words, when x does not influence y. Then (1) takes the form:

$$y = const + z = z,$$

We equated const + z to z because in the case of adding a constant to a random variable, only its mean changes, while the variance and distribution remain unchanged, making it the same random variable.

From here, we initially formed the following quantity for comparing the relationship:

$$\frac{1}{k}\sum_{i=0}^{k}\left[M_{x_i}(y) - M(y)\right]^2,$$

where $M_{x_i}(y)$ is the local mean, and $M(y)$ is the overall mean. This indicator shows the presence and strength of the relationship. If there is no relationship, then by definition all local means tend towards the overall mean. If a relationship exists, then the local means significantly differ from the overall mean. However, this indicator has a significant drawback - we could not normalize it to the interval [0, 1]. We tried to normalize it by variance, the range of the x and y samples, the interval size, and so on, but none of the normalization attempts were successful.

Upon analyzing the previous quantity, we noticed a relationship of this indicator divided by the variance with the machine learning metric - R^2. The relationship was close to linear. And we decided to change the approach by using local unaveraged variances, resulting in the following coefficient:

$$\eta^2 = 1 - \frac{\sum_{i=0}^{n}[y_i - M_x(y_i)]^2}{\sum_{i=0}^{n}[y_i - M(y)]^2}, \tag{3}$$

where $M_x(y_i)$ is the value of the grid for y_i, and $M(y)$ is the overall mean. Thus, the numerator essentially represents the local unaveraged variance, while the denominator is the unaveraged overall variance. This fraction tends towards one when there is no relationship, as the local mean tends towards the overall mean. Also, this fraction tends towards zero when there is a strong relationship because the stronger the relationship between y and x, the less other noise, and therefore, the smaller the local variance. The unit is needed to invert the fraction for the traditional representation of the relationship's presence: 0 means no relationship, 1 means y is fully described by x.

It is also worth noting that the resemblance of the coefficient of determination allows transferring its interpretation to this correlation coefficient. Before we move on to the interpretation and properties of our coefficient, we must recall those similar to it.

3.3. Analysis of Similar Correlation Coefficients

In searching for literature for the article, we came across a similar coefficient that had already been described. Fisher and Pearson [21, 22, 23] mentioned the correlation ratio in their works. This coefficient is defined as the sum of the unaveraged variances in individual categories relative to the total variance. It is obvious that our coefficient differs only by the inversion through unity.

In defense of our approach, it can be noted that the correlation ratio was primarily used for between-class comparison of numerical characteristics, or in other words, for tasks of variance analysis. However, Fisher provides an example of using such a coefficient in a problem similar to regression [21], with the difference that the independent feature is temperature, and its values were discretized into intervals with a step of 0.5 degrees. However, we did not find the application of the correlation ratio for analyzing the relationship between two continuous features.

At the same time, the connection with this coefficient, in our opinion, only strengthens ours because although we used a different approach to arrive at it, and its purpose was slightly different, we can use the achievements made before us. Such achievements will be mentioned further in this article.

Also, in another article, we noticed that the author, using a different idea and approach, arrives at a rank correlation coefficient [25]. It is a rank coefficient, but the author indicates that with an infinite amount of data, their coefficient tends to be very similar to ours. The fact that other scientists arrive at the same or similar coefficients may indicate that the idea of such a coefficient is indeed justified. Moreover, our and their achievements can complement each other since each uses a different approach and consequently obtains different estimates, advantages, and disadvantages.

3.4. Criteria for the presence of a relationship

It's very useful if the presence of a relationship by the correlation coefficient can be confirmed through a statistical criterion. And here we are far from the first, so we decided to turn to Fisher's book [21]. However, Fisher himself wrote in his book that it is not so easy to do this for the correlation ratio, although the task is to assess whether the variances within classes differ significantly from the overall one. He describes the following ways to statistically evaluate the significance of the relationship:

- Analysis through the z-table, but this criterion is not inclined to normality if the number of arrays does not tend to infinity;
- For large samples - $N\eta^2$ - distributed by chi-square.
- Blakeman's Criterion

So there are methods for testing the statistical significance of the coefficient, although they are not ideal, they are available. We did not conduct a more in-depth analysis in the literature, but we are confident that more works on this topic can be found in the search.

Moreover, other articles can be found describing other methods of assessing the coefficient. For example, the next article aims to find a confidence interval for the correlation ratio [27]. In another article, the author tries to reduce the effect of small grouping, thereby improving the estimation of the coefficient itself [28].

3.5. Interpretation

This coefficient can currently have at least two interpretations: the first one is associated with the coefficient of determination, and the second one is associated with the correlation ratio.

The coefficient of determination is interpreted as the proportion of explained variance by the model. That is, if the coefficient of determination is 0.7, then the regression model explains, and accordingly takes into account in its forecasts, 70% of the variance of y. Thus, such a model can be considered quite good.

Therefore, considering that our coefficient essentially follows the formula of the coefficient of determination, we can interpret it as the proportion of explained variance of y through x, that is, the proportion of the variance of y explained by x.

The interpretation of the correlation ratio is different, namely - how the variances within the classes differ from the overall variance and from each other. In our case, we are considering a discretized independent variable, and accordingly our coefficient can be interpreted as: how much the variance within the intervals differs from the overall variance, and therefore how close or far the relationship between x and y is from no correlation.

3.6. Possible Implementations and Modifications

Given that our grid and coefficient are based on simple partitioning and simple averages, it raises a natural question: can these parameters be modified to obtain new algorithms?

One of the first modifications we can consider is partitioning by quantiles. The difference with this partitioning lies in the uniform number of observations in each interval, rather than a constant step size of the grid. Formally, it can be presented as follows:

$$\omega_h = \left\{ M_{x \in [x_{hi}; \, x_{h(i+1)}]}(y), \ x_i \ = \ Q_i(x), \ i = \overline{0, \, \dots N-1}, \ h \ = \ \frac{1}{N} \right\},$$

where $Q_i(x)$ represents the quantile for x_i, and N is the number of observations in the interval.

Another relatively simple modification could be replacing the mean in the grid with the median. The idea is that we cannot guarantee either the normality or symmetry of the distribution law of the random part. If it is skewed, the mean may show a worse or even inadequate result. Formally, it can be presented as follows:

$$\omega_h = \left\{ \overline{M}_{x \in [x_{hi}; \, x_{h(i+1)}]}(y), \ x_i \ = \ a + ih, \ i = \overline{0, \, \dots N}, \ h \ = \ \frac{b-a}{N} \right\}.$$

Another modification could be replacing the mean with a weighted average distance. The logic is as follows: in the original version, we cannot be sure that the mean M(y) calculated over the interval [xi; xi+1] corresponds to the mean M(y) on this interval. Introducing a weighted average distance, where the mean is computed over y and weighted by x, we can obtain an average that approximates f(c), where c is the point relative to which we calculate the weights. Formally, it can be presented as follows:

$$\omega_h = \{ M_{x \in [x_i; \, x_{i+1}]}(y), \ M_{x \in [x_i; \, x_{i+1}]}(y) = \frac{\sum_{i=1}^{n} \omega_i \, y_i}{\sum_{i=1}^{n} \omega_i},$$

$$\omega_i = \frac{1}{\sum_{j=1, \, j \neq i}^{n} d_{i,j}}, \ x_i \ = \ a + ih, \ i = \overline{0, \, \dots N}, \ h \ = \ \frac{b-a}{N} \}$$

Another modification of the partitioning could be the introduction of overlapping intervals, i.e., intervals that have some overlap with neighboring intervals. Although such an extension may seem counterproductive for the initial algorithm, as expanding the interval can shift f(c), which is approximated by the mean, it could work for the implementation with the weighted mean distance. In this case, less weight is given to the points at the ends of the intervals and more to the center. This implementation is planned as a way to deal with a small amount of data. To do this, we introduce the concept of a window w, which specifies the width of the subinterval relative to the step between nodes. Thus, if the step between nodes x_i and x_{i+1} is h, then x_i is the center of the interval $[x_i - \frac{wh}{2}; \ x_i + \frac{wh}{2}]$. It should be noted that we only consider the case when $w > 1$, because for w=1, we get partitioning without overlapping intervals, and for $w < 1$, we get intervals that do not cover all points. Obviously, b-a-wh is the free width for nodes, so based on this, we can find the interval as $h = \frac{b-a}{n+w}$, where n is the number of nodes. From here, we can define the grid as:

$$\omega_h = \left\{ M_{x \in \left[x_i - \frac{w}{2}; \ x_{i+1} + \frac{w}{2}\right]}(y), \ x_i = a + \frac{wh}{2} + ih, \ i = \overline{0, \ ...n}, \ h = \frac{b-a}{n+w} \right\}$$

where n is the number of nodes, and w is the window size. Here, $M_x(y)$ represents any method of calculating the mean.

3.7. Comparison with Other Correlation Coefficients

In this section, we will discuss the advantages and disadvantages of our coefficient compared to others. We based our comparison on the article about RDC[24], as they compared their coefficient with most others.

3.7.1. Renyi's Properties

A. Renyi in his work [22] outlined a set of properties that correlation coefficients ρ^* should possess:

1. ρ^* is defined for any pair of non-constant X and Y.
2. $\rho^*(X, Y) = \rho^*(Y, X)$
3. $0 \leq \rho^* \leq 1$
4. $\rho^* = 0$, if there is no relationship between X and Y.
5. For bijective Borel-measurable functions: $f, g: \mathbb{R} \to \mathbb{R}, \rho^*(X, Y) = \rho^*(f(X), g(Y))$.
6. $\rho^* = 1$, if $Y = f(X)$, or $X = f(Y)$.
7. If $(X, Y) \sim \mathcal{N}(\mu, \sigma)$, then $\rho^* = |\rho(X, Y)|$, where ρ is the correlation coefficient.

We'll refer to the source, as A. Renyi himself discussed these properties for the correlation ratio in his work and indicates that it does not satisfy properties 1), 4), and 5). Property 2) was not explicitly mentioned, but it can be easily addressed by taking:

$$\rho^*(X, Y) = \max(\rho^*(X, Y); \ \rho^*(Y, X))$$

Property 7 is not obvious, but Fisher's book mentions that the square root of the correlation ratio corresponds to the correlation coefficient if the relationship is linear. A. Renyi also mentions this.

Property five is partially fulfilled since:

$$\eta^2 = sup_f \, R(f(x), y)$$

However, here we can try to apply the same approach as for maximal correlation. Although this is more akin to a hypothesis, articles attempting similar endeavors can be found[20].

Property 4 is rejected because for multivalued functions, such as a circle equation, our coefficient will give a falsely zero value. However, such cases are unlikely to occur frequently in practice. In the real world, there are no cases where there can be many target values for one independent feature value. Here, there should be a third feature that will precisely determine (for example, for a circle, this could be a categorical feature indicating which semicircle the point belongs to).

Property one is rejected because x and y must have finite variances. However, again, for practical use, it is very doubtful that any real data are prone to infinite variances.

3.7.2. Other properties of the coefficient

For the correlation coefficient, many other coefficients are important, such as marginal invariance, vector inputs, and so on. Let's consider them.

Marginal Invariance. Our coefficient does not depend on the scaling or shifting of x or y. When x changes, the grid changes with it. When y changes, according to the properties of variance, shifting has no effect, and scaling is factored out of the numerator and denominator.

Vector Inputs. Although in this article we do not discuss coefficients for multiple features, it is easy to modify the grid so that the segmentation is not based on a single feature. We have even conducted preliminary experiments that yielded positive results, but we do not discuss them in this article.

Algorithm Complexity. Let's first consider the complexity of computing the variance, as our coefficient is almost entirely composed of them. The complexity of computing the variance is $O(n)$. We have K computations of variances in the numerator and one in the denominator. It is worth noting that the variances in the numerator are local because they are calculated not for n observations, but for ki. Therefore, the sum of variances in the numerator will have a complexity of $O(k_1 + k_2 + ... + k_i) = O(n)$. Accordingly, the total complexity of the algorithm is $O(2n)$, but for large data, the factor of two can be disregarded, so the complexity is $O(n)$.

3.7.3. Comparison with other coefficients

The article on the RDC coefficient[24] compares it with several other coefficients, namely: Pearson, Spearman, Kendall, CCA, KCCA, ACE, MIC, dCor, HSIC, CHSIC. They were compared based on the following characteristics: ability to recognize nonlinear dependencies, vector input, marginal invariants, satisfaction of all Renyi's Properties, coefficient lying in [0, 1] range, and algorithm complexity.

Of all these properties, our coefficient satisfies: marginal invariant, vector inputs (although its effectiveness has not been proven yet), it lies within [0, 1], and it outperforms all previously mentioned algorithms in algorithm complexity, except for Pearson, CCA, and ACE. However, our coefficient does not satisfy all Renyi's properties. Although we have disputed their practical significance for our coefficient earlier.

However, there is an advantage of our coefficient over others, namely its interpretation. While other coefficients typically indicate the strength of the relationship, our coefficient can be interpreted as the proportion of explained variance of y by x.

This is extremely important in regression tasks in machine learning. In practice, you cannot be sure that any actions to improve the model will bring significant results. You can rely on the difference between training and test sample metrics, have experience, or understand algorithms, but there is no exact tool for analyzing how well the model can learn from certain data, just as there are no tools to understand how much it can be improved.

Our coefficient can provide such a tool. Because the proportion of explained variance of y by x directly estimates how much variance the model can explain using the x feature. Similarly, if our coefficient is applied to the residuals of the model, it can show whether a certain feature still has a connection, and if so, our coefficient will quantitatively show how much more the model can be improved.

4. Experiment

In this section, we embark on a journey of experimentation aimed at validating and enhancing our approach. Our experiments are divided into three parts, each focusing on distinct aspects of our research. Implementation can be found on the GitHub repository[29].

The first part of our experiments revolves around assessing the efficiency of the correlation coefficient calculation algorithm. Here, we analyze the computational performance of the algorithm, aiming to optimize its speed and accuracy.

Moving forward, the second part of our experiments is dedicated to the selection of the parameter N. We explore the impact of varying N on the overall performance of our method, seeking to identify the optimal value that maximizes the effectiveness of our approach.

Lastly, in the third part of our experiments, we delve into investigating the relationship between the correlation coefficient, the coefficient of determination, and Pearson's coefficient. Through comprehensive analysis, we aim to elucidate the interplay between these key metrics and gain deeper insights into their significance within the context of our study.

4.1. Comparison of the coefficient efficiency

For practical use of the coefficient, it should scale well for large datasets, as they are most likely to be used with it. We compared our implementation in R with the Pearson coefficient and RDC (as implemented in their article[5]). The results of these experiments were sufficient to give us an idea of how well our coefficient performs on large datasets. For each dataset size, we conducted a total of 250 measurements of algorithm execution time with feature regeneration to eliminate the influence of the sample. The results are presented in Table 1.

Table 1
Time running results in R environment

Coef. / n	1000	10.000	100.000	1.000.000
Person's ρ, avg	2.069187e-05	7.849789e-05	0.0005257692	0.005731739
Person's ρ, std	5.063336e-05	2.112877e-05	0.0001095381	0.0008129202
RDC, avg	0.002380476	0.0243814	0.2436182	3.285504
RDC, std	0.0003120181	0.004626209	0.01469011	0.2986529
η^2, avg	0.0004041605	0.002601263	0.01333642	0.1807116
η^2, std	0.0003125966	0.005066414	0.006118725	0.03149566

As seen from the table, the results of our experiments indicate that our coefficient is slower than the Pearson correlation coefficient but still outperforms RDC. Furthermore, the experiments showed that the coefficient scales well with a large amount of data.

We also conducted experiments in the Python environment, using the pandas and numpy libraries. Our algorithm implementation was compared with the implementation of the Pearson correlation coefficient for the pandas.DataFrame object. Similar to the first experiment, we had 250 iterations of experiments with data generation on each iteration. In this set of experiments, we only compared the Pearson coefficient with ours. The results are presented in Table 2.

Table 2
Time running results in Python environment

| Coef. / n	1.000	10.000	100.000	1.000.000	10.000.000
Person's ρ, avg	0.000073	0.000188	0.001217	0.012289	0.122185
Person's ρ, std	0.000258	0.000392	0.000413	0.000500	0.001821
η^2, avg	0.003401	0.004186	0.012494	0.122029	1.425318
η^2, std	0.000484	0.000392	0.000694	0.007234	0.042040

Based on the experimental results, we can confirm that our coefficient is fast. Although it is slower than the Pearson correlation coefficient, it performs well with a large amount of data. Even when the amount of data reaches 10 million, it is computed in less than 1.5 seconds, indicating that it can be applied to large datasets.

4.2. Selection of Parameter N

Unfortunately, our coefficient has a parameter - N. In this implementation of the coefficient (3), more specifically, the grid on which we compute it (2), N strongly influences the coefficient itself. N is responsible for the average number of observations within the interval and the width of the interval itself.

As we have discussed, reducing the interval should better approximate the random function, but with too small N, a simple average may not provide an adequate estimate. Therefore, in this section, we will conduct a series of experiments to determine the optimal N.

Let`s start with a data size of 10,000. We generated data - an independent feature and noise (both following a normal distribution). We compared our coefficient with the coefficient of determination, as our coefficient is its estimate and we can calculate it accurately because we know its function. For each experiment, we varied the noise from zero to complete noise to account for random noise. We also conducted 50 experiments in each case(see Figure 3).
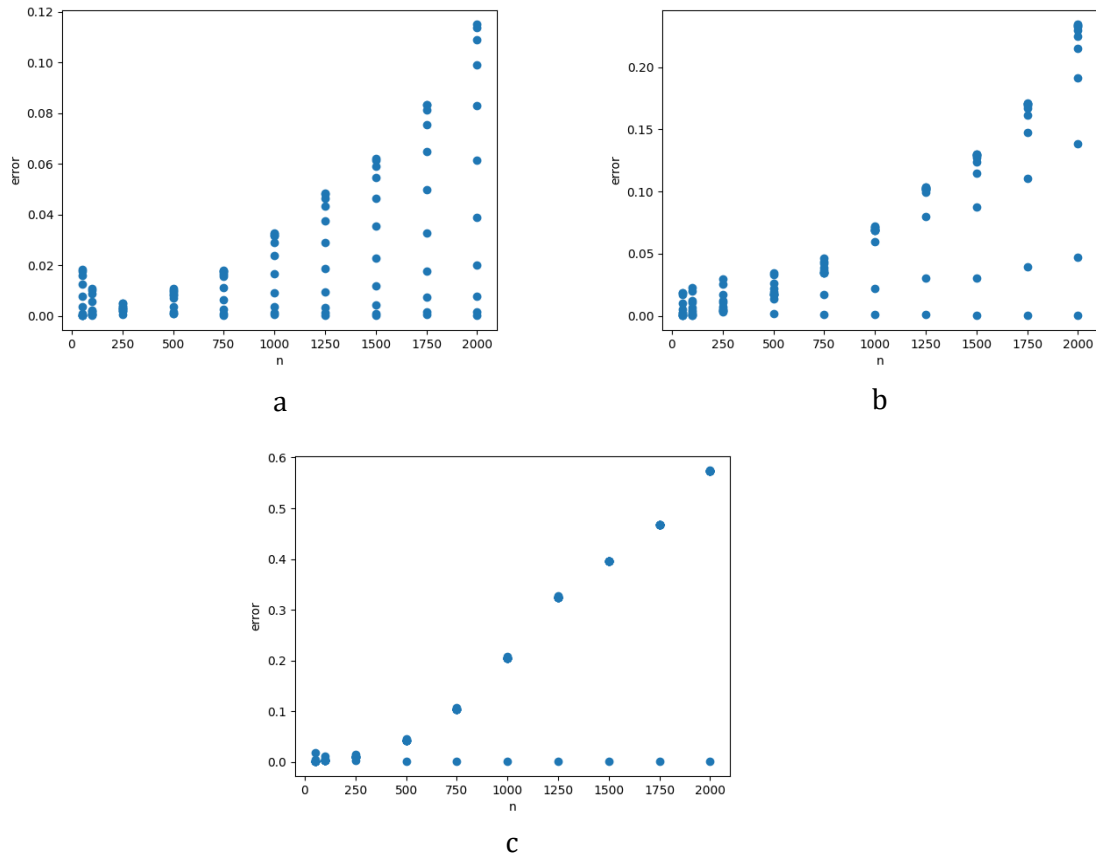

Figure 3: Error from N. a) linear, b) quadratic, c) exponential dependencies.

From the graphs, several conclusions can be drawn:

1. As discussed earlier, the error strongly depends on the rate of change of the function - linear dependence does not exceed an error of 0.12, quadratic exceeded 0.2, and exponential reached 0.6;
2. The optimal N for a data size of 10,000 is 250 and does not depend on the function, as reducing N may not necessarily result in a smaller error.
3. With N = 250 and a data size of at least 10,000, the error will not exceed 0.05.

Further verification may not be necessary, but we decided to perform it. We conducted a similar experiment for a data size of 100,000, for the quadratic dependency(see Figure 4).
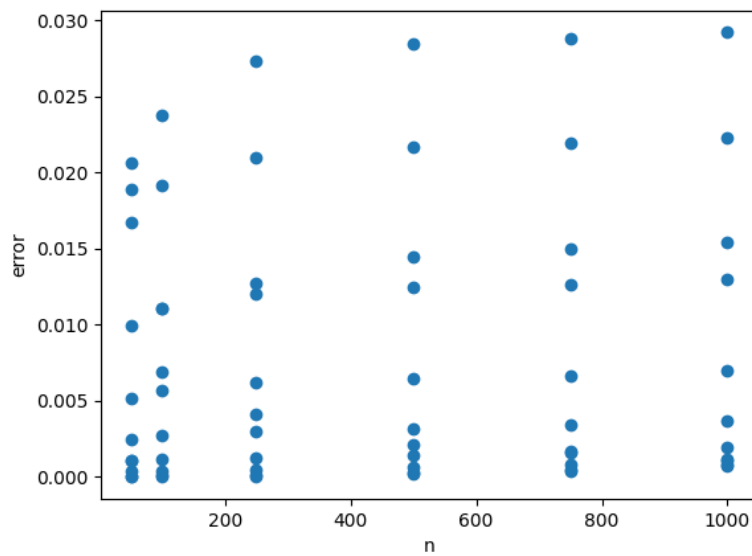
Figure 4: Coefficient error for the case of 100,000.

We see that for any N, the error does not exceed 0.03. Next, we conducted a similar experiment but for small datasets (from 100 to 5000). First, let`s consider the linear dependency. The following graphs display the number of data points in the synthetic sample on the x-axis, and the error of our coefficient (the absolute difference between our coefficient and the coefficient of determination) for different N on the y-axis(see Figure 5).
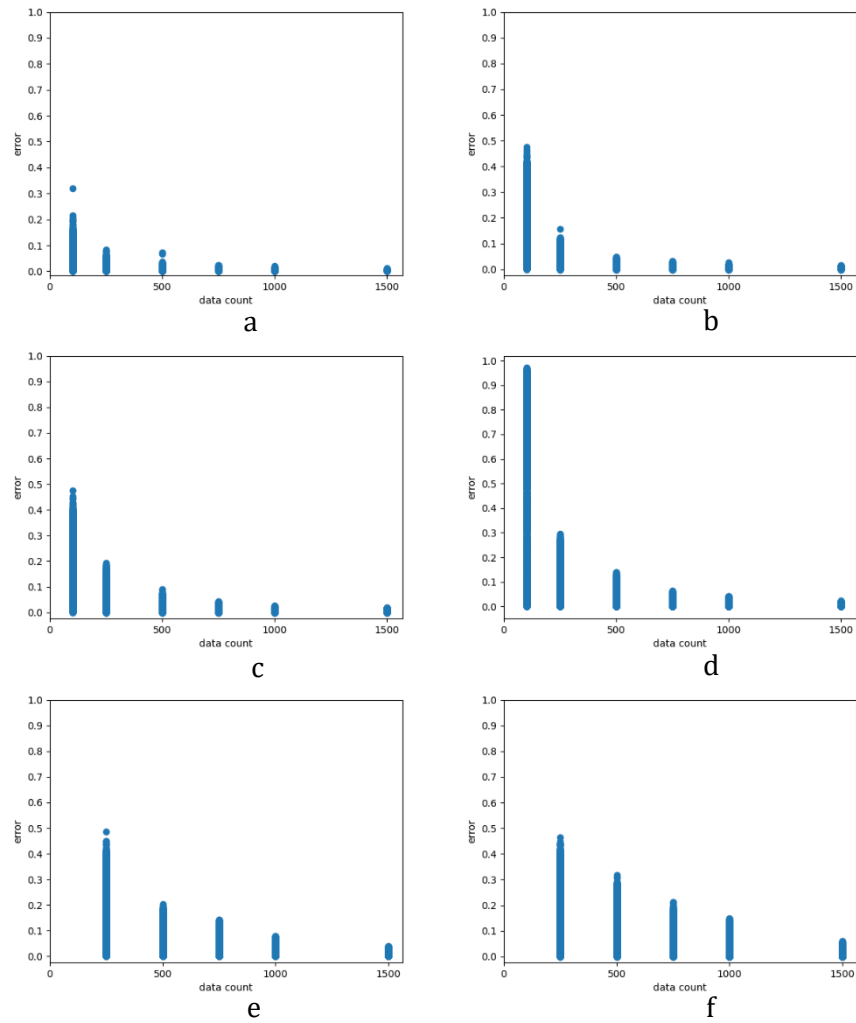
Figure 5: Error from the number of data points for linear dependency. N = a) 50, b) 75, c) 100, d) 150, e) 200, f) 250.

As a result of this experiment, it is evident that the accuracy of the method increases with the increase in the number of data points. Furthermore, starting from 750-1000 data points, the error ceases to change significantly. Interestingly, for small samples (relative to previous experiments), the optimal N is no longer 250 (although only cases up to 200 are shown in the graphs, but beyond this point, the error only increases.

At the same time, empirically we have established that the maximum error approximately follows the following law:

$$\max\left(|r^2 - \eta^2|\right) = \frac{N}{K} - 0.1,$$

where N is the grid parameter and K is the amount of data. And the obvious fact that as N decreases and K increases, the error decreases. However, N cannot be reduced infinitely, as for very small N, a simple average will be a terrible estimate. Along with this, the graphs

show that for linear dependency, N=30 is sufficient. And although for a sample of 100 observations, the coefficient has too high an error, for 250, the maximum error does not exceed 0.1, which is still rough, but it can already be applied in practice.

However, the accuracy of our grid depends on how quickly the function changes, so we propose considering cases with quadratic and exponential functions. The first case represents a function that changes quite rapidly, while the second one represents a function that changes very quickly(see Figure 6).
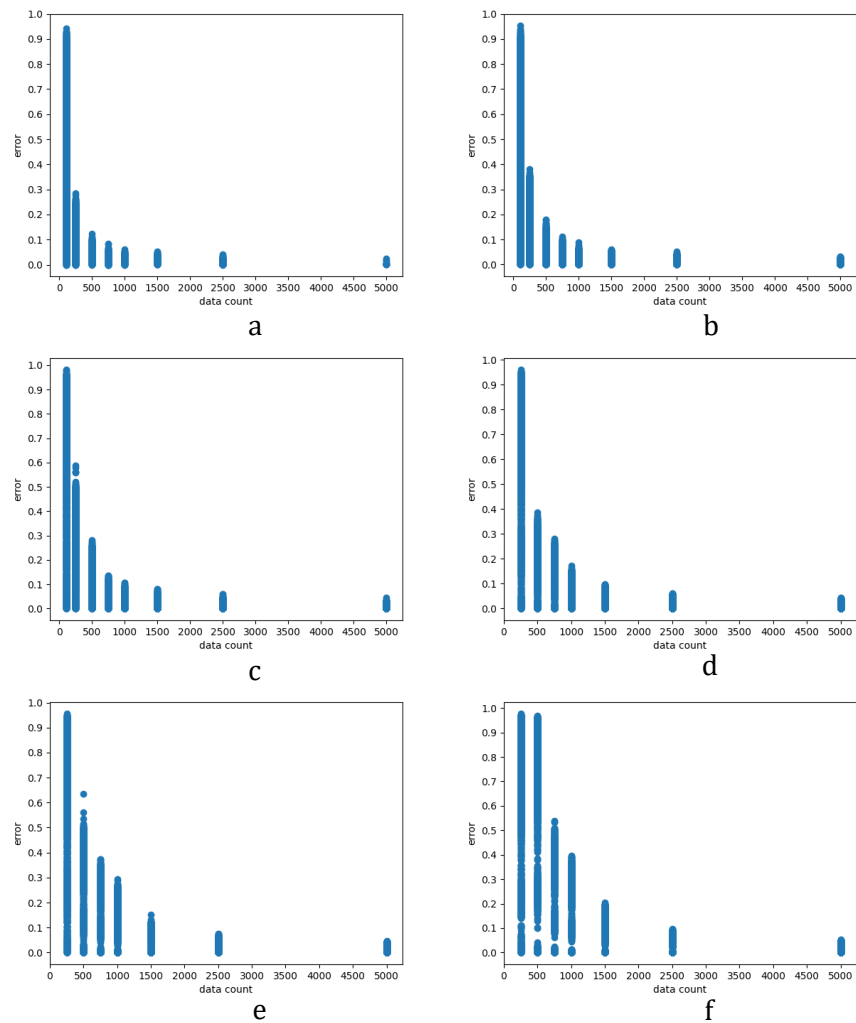
Figure 6: Error from the number of data points for quadratic dependency. N = a) 50, b) 75, c) 100, d) 150, e) 200, f) 250.

From these graphs, even for rapidly growing functions, the error stabilizes around 1500-2500. It is also evident that the empirical law of decreasing error with the amount of data will be approximately the same, with the difference being that for the quadratic function, the constant will no longer be -0.1 but +0.1. Let's move on to the exponential function(see Figure 7).
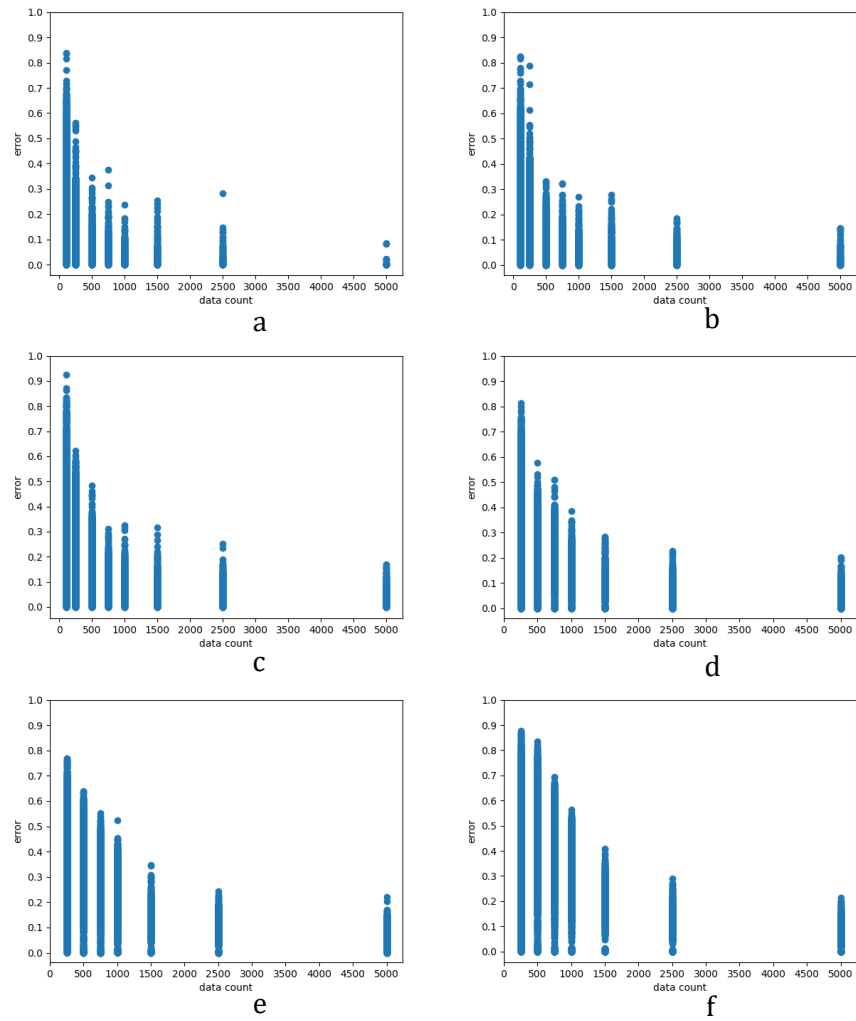
Figure 7: Error from the number of data points for exponential dependency. N = a) 50, b) 75, c) 100, d) 150, e) 200, d) 250.

From the graphs, it can be seen that the maximum error has increased even further. Moreover, the estimation, even with a data quantity of 5000, is very rough for almost all N values. Here lies the same regularity – the error is proportional to N and inversely proportional to the amount of data.

Therefore, in this section, we examined the absolute error of our coefficient compared to the real coefficient of determination. We observed that on average, for a data quantity of 10,000 or more, we can use N = 250, because with smaller N values, the result may be unstable and less reliable. In such cases, the error does not exceed 0.05-0.1. For data quantities up to 5,000, the best value was N = 50. Furthermore, it was found that the error is distributed proportionally to the N parameter and inversely proportional to the data quantity. With smaller data, the error did not exceed 0.1.

It became evident that the rate of change of the function plays an extremely significant role. We could observe how the transition from linear to quadratic, and from quadratic to exponential dependencies affects the error, which needs to be taken into account in research.

4.3. The relationship between the coefficient and r^2 and ρ

One of the methods for estimating the correlation coefficient is to compare it with other coefficients while varying the noise for different bijective relationships. We decided to conduct this experiment slightly differently: we compared our coefficient with the actual value of the determination coefficient and compared the square root of our coefficient with the Pearson coefficient. The first comparison was made due to the explicit similarity between η^2 and r^2. The second comparison was done because we have often encountered references in the literature that in the case of linear dependence, $\sqrt{\eta^2} = \rho$ [3, 4]. We decided to verify this relationship as follows: we calculated the Pearson coefficient not between x and y, but between f(x) and y, where f(x) is the specified relationship function.

On the graph, each point corresponds to normally distributed noise (with a mean of zero, and the variable changing the standard deviation of the noise from 0 to 3), the independent feature x was generated from a normal distribution with a mean of 0, and a standard deviation of 2. We constructed a loop for noise values from 0 to 1 with a step of 0.01 (100 points in total), on each of which we regenerated points to exclude the sampling factor. The graphs depict our coefficient, its square root, the determination coefficient, and the Pearson coefficient. The latter was calculated not between x and y, but between f(x) and y (Figures 9-14).
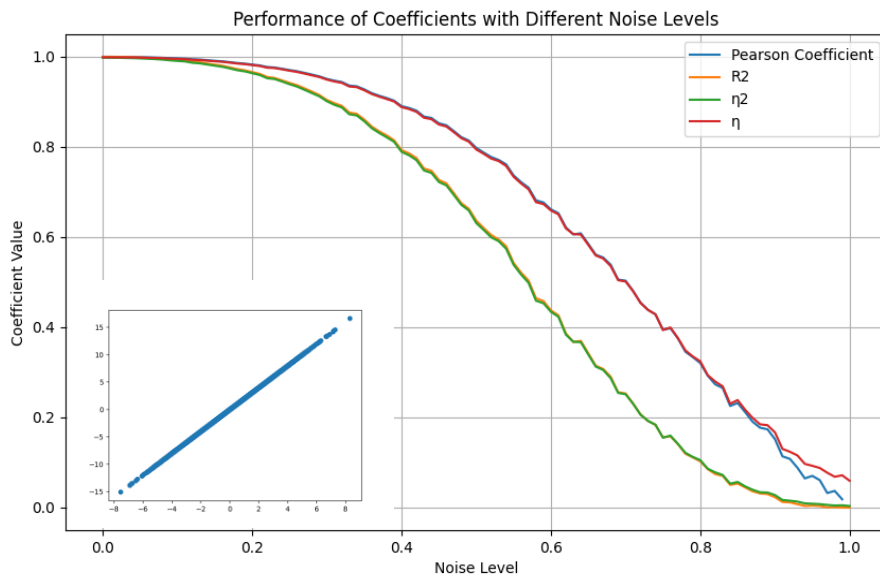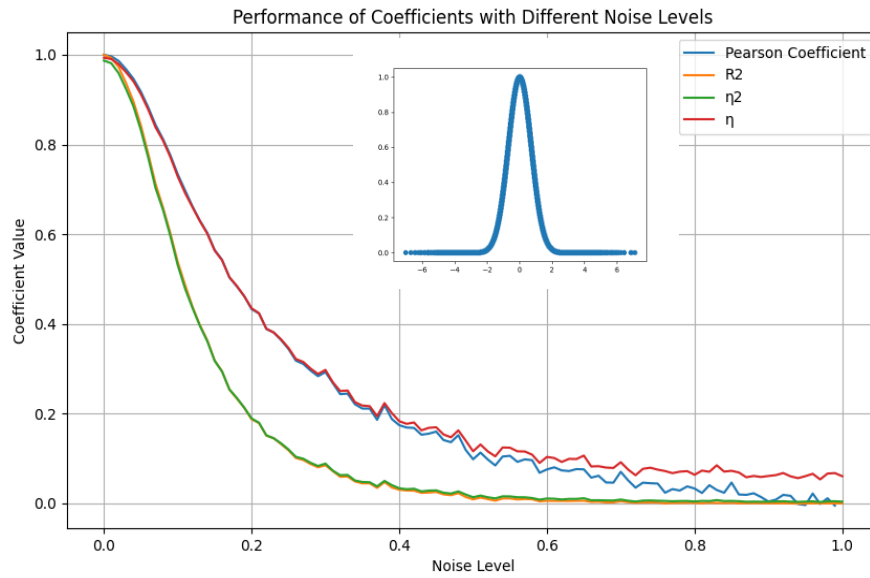

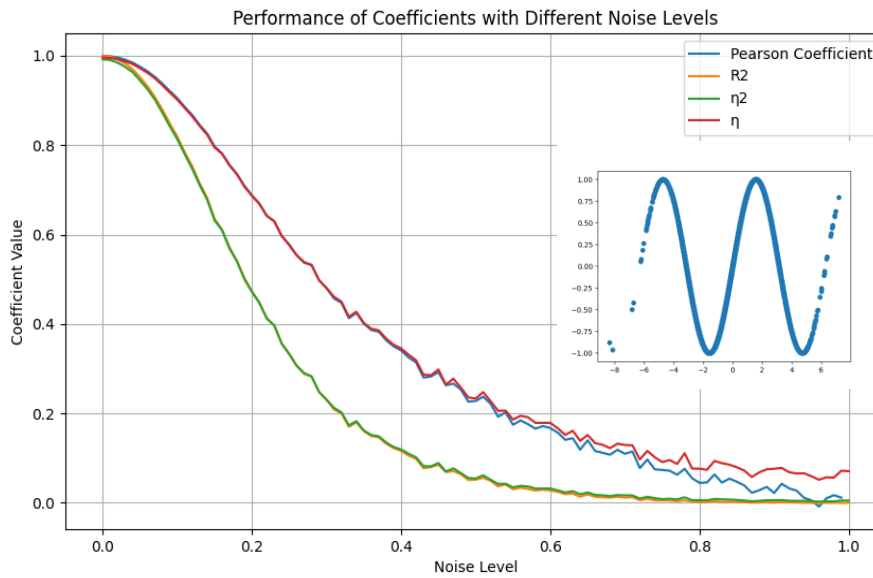**Figure 8**: The relationship between η^2, r^2, and ρ for a linear function.
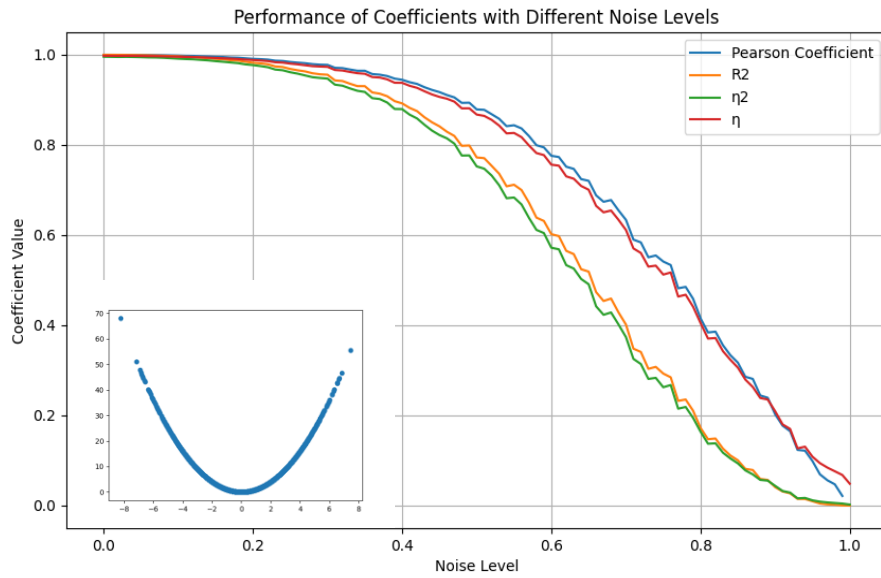
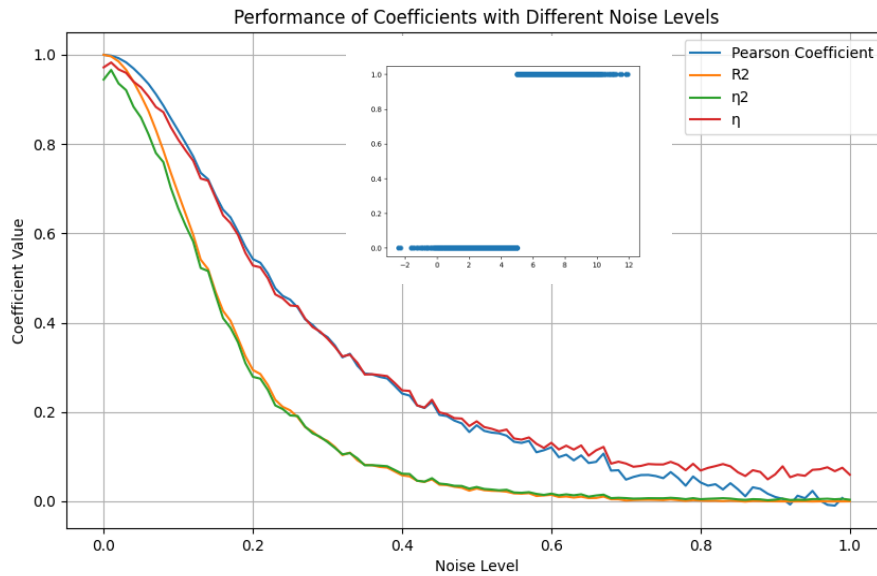**Figure 9:** The relationship between η^2, r^2, and ρ for a quadratic function.
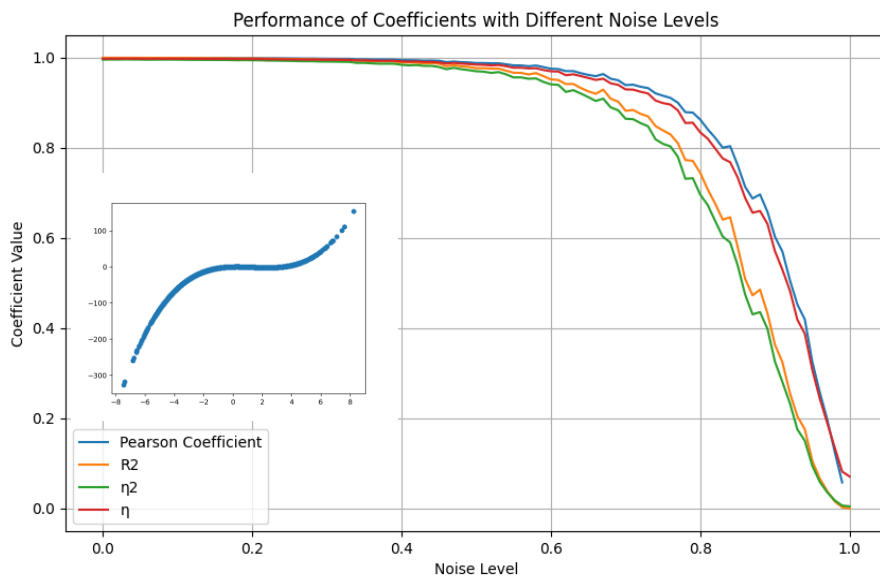
Figure 10: The relationship between η^2, r^2, and ρ for a polynomial function.

Figure 11: The relationship between η^2, r^2, and ρ for a exponent function.

Figure 12: The relationship between η^2, r^2, and ρ for a sinusoidal function.

Figure 13: The relationship between η^2, r^2, and ρ for a step function.

From the graphs, it is evident that our coefficient and its square root approximate the determination coefficient, which is the maximum Pearson correlation coefficient. Therefore, η^2 can be used as r^2, and η as ρ.

5. Results

Overall, we conducted three experiments. The first aimed to demonstrate the algorithm's speed and scalability for large datasets. The second sought to find the optimal value of parameter N for both small and large samples. In the third experiment, we explored the relationship of our coefficient with the coefficient of determination and the Pearson correlation coefficient. Additionally, we highlighted our coefficient's capability to capture nonlinear dependencies.

In the first experiment, we evaluated the speed of our coefficient. We implemented two versions: in Python and R. We compared the first version only to the Pearson correlation coefficient and the second one additionally to RDC (as described in their paper). We decided not to compare it to other coefficients, as RDC had already been compared with most popular coefficients. Since our coefficient was theoretically supposed to be faster, we limited our comparison to this.

The experiment was conducted with varying amounts of data: 1,000, 10,000, 100,000, and 1,000,000 (with an additional 10,000,000 in Python environment) randomly generated pairs of (x_i, y_i) Moreover, for each amount of data, we ran 250 iterations to obtain a better averaged estimate. Points were regenerated at each iteration to eliminate sample bias.

The results of this experiment showed that our coefficient outperforms RDC in terms of speed and accordingly outperforms the other coefficients compared in their paper. On the other hand, our coefficient significantly lagged behind the Pearson correlation coefficient.

Despite this, it is highly scalable and suitable for large datasets. For 10,000,000 data points in the Python environment, the average computation time of the algorithm did not exceed 1.5 seconds, which we consider a remarkable result.

Regarding the second experiment, we endeavored to experimentally determine the parameter N on which the accuracy of our coefficient estimation depends. Given that our coefficient approximates the coefficient of determination, we calculated the error (absolute difference between our coefficient and the coefficient of determination) experimentally since the coefficient of determination is known in our simulated data (the considered function is seen as the model). We split this overall experiment into three smaller ones: for small data (up to 5,000 inclusive), for medium data (10,000 points), and large data (100,000).

The results on 10,000 points showed that with N = 250, the error remained below 0.05, which is generally acceptable. Furthermore, increasing N led to less accurate results, while decreasing N resulted in less reliable outcomes.

The results on 100,000 points revealed that N started to play a significantly smaller role, and even with N = 1000, the error did not exceed 0.03.

When it comes to small data, the experiment indicated that having 1,000 data points is sufficient to ensure that the error is less than 0.1, assuming the function grows slower than an exponent. For small data, the error decreases inversely proportional to N, where choosing an overly small N is not advisable, as the simple average estimate in this scenario would be inadequate. Experimentally, we concluded that the optimal N is approximately 50.

In the following experiment, we examined our coefficient's ability to capture nonlinear dependencies excluding multivalued functions, as discussed earlier, as our coefficient cannot work with them. Instead of comparing to other coefficients, we chose to compare it to the coefficient of determination and the Pearson correlation coefficient. Given that our coefficient approximates the former and according to the works of Fisher and Renyi, the square root of our coefficient should approximate the supremum of the Pearson correlation coefficient. We calculated the latter not between x and y, as in that case, we wouldn't gain much information, but between the value of f(x) and y, where f(x) is the specified function. This allowed us to compare the value of our coefficient with the real correlation value and thereby confirm the connection of our coefficient with the Pearson correlation coefficient.

We generated x as normally distributed with a mean of zero and a standard deviation of two. The noise followed a normal distribution with a mean of zero and a standard deviation varying from 0 to 3 (scaled to 0-1 in the graphs). The noise increment was 0.01, and the dataset consisted of 10,000 data points.

From the graphs, it is evident that our coefficient and its square root well approximate the coefficient of determination and the Pearson correlation coefficient, respectively. This experiment demonstrated that we can use our coefficient as a nonlinear analogue to the Pearson correlation coefficient. Moreover, approximating the coefficient of determination offers a new tool for analysis, particularly useful for model construction. The coefficient of determination is a machine learning metric; it requires a model. It does not show data properties; it indicates how well a model can predict this data. With the correlation coefficient based on determination, we can evaluate the proportion of explained variance of the dependent feature through the independent one. Therefore, before model construction,

we can assess how well we can train the model on this data and how well it will predict the target feature.

6. Conclusions

Using the theory of random processes, a grid was derived to approximate the random function of dependence between the independent and target features. Based on this grid and our definition of absence of correlation, a correlation coefficient was derived. Similar coefficients found before us were analyzed, as well as differences in approaches.

The relationship between our coefficient and the determination coefficient was demonstrated, based on which a useful interpretation of our coefficient was shown. The connection of our coefficient with the determination coefficient and the Pearson coefficient was also demonstrated. An optimal parameter N, which affects the accuracy of the method, was selected.

The properties of our coefficient were also investigated. These properties were used to compare our coefficient with others. The drawbacks of our coefficient can be identified as: not satisfying all of the Renyi's properties. The algorithm's speed and scalability on large data were also investigated. The coefficient performed very well.

Its potential for use in machine learning tasks was also discussed. Considering its relationship with the determination coefficient, it becomes possible to assess the potential of the model and the data in general, namely the proportion of variance of the target feature that can be explained using the investigated independent feature.

7. References

[1] Zhang, K. (2019). BET on Independence. J. Amer. Statist. Assoc., 114 no. 528, 1620–1637.
[2] Papageorgiou, S. N. (2022). On correlation coefficients and their interpretation. Journal of Orthodontics, 49(3), 359–361. https://doi.org/10.1177/14653125221076142
[3] Kirişçi, M. (2022). Correlation coefficients of Fermatean fuzzy sets with a medical application. Journal of Mathematical Sciences and Modelling, 5(1), 16-23. https://doi.org/10.33187/jmsm.1039613
[4] Karim, I. N. (2021). CORRELATION ANALYSIS: Are Religious Scavengers More Motivated? Jurnal Mahasiswa Humanis, 1(2), 59. e-ISSN 2774-8863.
[5] Madraximova, M. A., Turdaliyev, S. M., & Abduraxmonov, D. A. (2022). CORRELATION COEFFICIENT AS A MATHEMATICAL SOLUTION OF ECONOMIC ISSUES. INTERNATIONAL JOURNAL OF RESEARCH IN COMMERCE, IT, ENGINEERING AND SOCIAL SCIENCES, 16(06), 72. ISSN: 2349-7793. Retrieved from https://www.gejournal.net/index.php/IJRCIESS
[6] Tosepu, R., Gunawan, J., Effendy, D. S., Ahmad, L. O. A. I., Lestari, H., Bahar, H., & Asfiang, P. (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. Science of The Total Environment, 725, 138436. doi: 10.1016/j.scitotenv.2020.138436
[7] Johnson, J. E., McGuone, D., Xu, M. L., Jane-Wit, D., Mitchell, R. N., Libby, P., & Pober, J. S. (2022). Coronavirus Disease 2019 (COVID-19) Coronary Vascular Thrombosis:

Correlation with Neutrophil but Not Endothelial Activation. The American Journal of Pathology, 192(1), 112–120. https://doi.org/10.1016/j.ajpath.2021.09.004

[8] Noviana, N., & Oktaviani, L. (2022). The Correlation Between College Student Personality Types and English Proficiency Ability at Universitas Teknokrat Indonesia. *Journal of English Language Teaching and Learning (JELTL)*, 3(1), 54–60. E-ISSN: 2723-617X. Available online at: http://jim.teknokrat.ac.id/index.php/english-language-teaching/index

[9] Kothari, V., Vij, S., Sharma, S. K., & Gupta, N. (2021). Correlation of Various Water Quality Parameters and Water Quality Index of Districts of Uttarakhand. Environmental and Sustainability Indicators, 9, 100093. https://doi.org/10.1016/j.indic.2020.100093

[10] Savira, R., Rifai, M., & Wahyunengsih, W. (2022). Correlation between TikTok Use and Teenagers' Self Esteem. Indonesian Journal of Learning Studies, 19. ISSN: 2775-5231.

[11] Ejegwa, P. A., Ajogwu, C. F., & Sarkar, A. (2023). A hybridized correlation coefficient technique and its application in classification process under intuitionistic fuzzy setting. Iranian Journal of Fuzzy Systems, 20(4), 103-120.

[12] Jin, Y.; Hussain, M.; Ullah, K.; Hussain, A. A New Correlation Coefficient Based on T-Spherical Fuzzy Information with Its Applications in Medical Diagnosis and Pattern Recognition. Symmetry 2022, 14, 2317. https://doi.org/ 10.3390/sym14112317

[13] Ali, W., Ali, M., Hussain, I., Ullah, S. S., Alroobaea, R., Hussain, S., Binmahfoudh, A., & Umar, F. (2022). A New Correlation Coefficient for -Spherical Fuzzy Sets and Its Application in Multicriteria Decision-Making and Pattern Recognition. Research Article, Volume 2022, Article ID 4471945. https://doi.org/10.1155/2022/4471945

[14] Bhatia, M., Arora, H. D., & Naithani, A. (2023). Some New Correlation Measures Based on Fermatean Fuzzy Sets using Decision Making Approach in Pattern Analysis and Supplier Selection. International Journal of Mathematical, Engineering and Management Sciences, 8(2), 245-263. https://doi.org/10.33889/IJMEMS.2023.8.2.015

[15] Chamani Mohasses, F., Solouki, M., Ghareyazie, B., Fahmideh, L., & Mohsenpour, M. (2020). Correlation between gene expression levels under drought stress and synonymous codon usage in rice plant by in-silico study. PLOS ONE, 15(8), e0237334. https://doi.org/10.1371/journal.pone.0237334Baak, M., Koopman, R., Snoek, H., & Klous, S. (2019, March 12). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. phik. Retrieved from https://phik.rtfd.io

[16] Yoo, Y., Escobedo, A. R., & Skolfield, J. K. (2020). A New Correlation Coefficient for Comparing and Aggregating Non-strict and Incomplete Rankings. School of Computing, Informatics, and Decision Systems Engineering, Arizona State University. Retrieved from https://www.elsevier.com/open-access/userlicense/1.0/

[17] Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2020). Methods and Algorithms for Correlation Analysis in R. Journal of Open Source Software, 5(52), 2306. DOI: 10.21105/joss.02306.

[18] Deb, N. and Sen, B. (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. Preprint. Available at https://arxiv.org/abs/1909.08733.

[19] Karlin, S., & Taylor, H. M. (1975). A First Course in Stochastic Processes. 2nd Revised edition. Academic Press.

[20] Fisher, R. A. (1937). The Design of Experiments (2nd ed.). Edinburgh: Oliver and Boyd.

[21] Pearson, K. (1915). On the Partial Correlation Ratio.

[22] A. Renyi. On measures of dependence. Acta Mathematica Academiae Scientiarum Hungaricae, 10:441–451, 1959.

[23] Lopez-Paz, D., Hennig, P., & Scholkopf, B. (2013). The Randomized Dependence Coefficient. Max Planck Institute for Intelligent Systems

[24] Chatterjee, S. (2010). New Coefficient of Correlation. Mathematics Subject Classification.

[25] Pilkington, A. (2020). "Average Value of Function". Retrieved from https://www3.nd.edu/~apilking/Math10550/Lectures/Dennis_Post_Break/Ave_Value.pdf

[26] Fleishman, A. I. (1980). Confidence Intervals for Correlation Ratios. Educational and Psychological Measurement.

[27] Kelley, T. L. (1935). "An Unbiased Correlation Ratio Measure". Harvard University. Communicated July 16, 1935.

[28] Holdovanskyi, V. (2024). DeterminationCorrelationCoefficient . https://github.com/BattleToady/DeterminationCorrelationCoefficient.git