Revista Colombiana de Estadística
Special Issue: Colombian and Brazil Graduate Programs
December 2025, volume 48, issue 3, pp. 509–528
http://doi.org/10.15446/rce.v48n3.123662

Exploring and Comparing Pairwise Nonlinear Association Measures for Continuous Variables

Exploración y comparación de medidas de asociación no lineal por pares para variables continuas

Alisson L. Brito[1,a], Fernanda De Bastiani[1,b], Mikis D. Stasinopoulos[2,c], Robert A. Rigby[2,d], Roberto F. Manghi[1,e], Thomas Kneib[3,f]

[1]Department of Statistics, Federal University of Pernambuco, Recife, Brazil

[2]School of Computing and Mathematical Sciences , University of Greenwich, London, United Kingdom

[3]Campus Institute Data Science, Georg-August-Universität Göttingen, Göttingen, Germany

Abstract

There are many linear and nonlinear measures of association between two continuous pairwise variables. They are used to indicate the strength of the relationship between the two variables. The question thus arises as to which of these measures should be used to explore relationships between two variables in general. The identification of linear and/or nonlinear relationship between two variables can help to avoid problems within a regression framework. The objective of this paper is to examine alternative measures of association that could be employed as a replacement or in conjunction with, standard linear correlation coefficients. The results lead us to conclude that the maximum correlation measure is particularly useful, and capable of detecting linear and nonlinear associations between two continuous variables, while also being relatively computationally efficient. It can be utilized in exploratory analysis and in a modern regression framework.

***Keywords*:** Correlation coefficient; Maximum correlation; Permutation test.

[a]Ph.D. Candidate. E-mail: alisson.limab@ufpe.br
[b]Ph.D. E-mail: fernanda.bastiani@ufpe.br
[c]Ph.D. E-mail: d.stasinopoulos@gre.ac.uk
[d]Ph.D. E-mail: r.a.rigby@gre.ac.uk
[e]Ph.D. E-mail: roberto.manghi@ufpe.br
[f]Ph.D. E-mail: tkneib@uni-goettingen.de

Resumen

Existen numerosas medidas de asociación, tanto lineales como no lineales, entre pares de variables continuas. Estas medidas se utilizan para indicar la fuerza de la relación entre las dos variables. Surge entonces la pregunta de cuál de estas medidas debería emplearse para explorar, en general, las relaciones entre dos variables. La identificación de relaciones lineales o no lineales puede ayudar a evitar problemas en modelos de regresión. El objetivo de este trabajo es examinar medidas alternativas de asociación que podrían utilizarse como reemplazo o en conjunto con los coeficientes de correlación lineal estándar. Los resultados nos llevan a concluir que la medida de correlación máxima es particularmente útil y capaz de detectar asociaciones no lineales entre variables continuas, además de ser relativamente eficiente desde el punto de vista computacional. Puede emplearse tanto en análisis exploratorios como en un marco de regresión moderno.

***Palabras clave*:** Coeficiente de correlación; Correlación máxima; Prueba de permutación.

1. Introduction

In the process of analyzing the data, a researcher often finds it necessary to determine the nature of the relationship between pairwise variables. A typical example, within a general regression framework, such as generalized additive models (GAM), or generalized additive models for location, scale, and shape (GAMLSS) is when the explanatory variables are highly associated. The relationship between covariates can be nonlinear, making the model susceptible to the presence of concurvity, a term introduced by Buja et al. (1989), that can be understood as a nonparametric extension of the concept of multicollinearity and whose diagnostics can be challenging in some cases. That is, high correlation between variables can affect both fitting and interpretation of a model. Detecting the association between variables can help to a better understanding on the underlying data structures before fitting regression models, to better choose of appropriate model forms such as a nonlinear or a transformation terms, and improve the interpretation and inference by accurately representing the true relationships among variables. Scatterplots are a good tool for identifying possible problems, however, they are not convenient when there is a large number of explanatory variables in the data. It is common in practice to use Pearson's correlation coefficient as a way of identifying associations, yet its limitations are well documented in situations where the relationship between variables is not necessarily linear. A variety of measures exists in the literature for determining the degree of association between pairwise variables, including linear and nonlinear relationships. There are also numerous potential forms of nonlinear association between variables, which make it increasingly valuable to gain a deeper understanding of the available measures and to determine which ones is most suitable for preventing potential problems in further analysis.

Conceptual, linear, or nonlinear associations could exist either at a population level or at the data collection level. In the former case, the problem is systemic, while in the second, could be the way of collecting data. In both cases high

association between explanatory variables leads to instability in the fitting process of the model, which in turn affects the interpretation of the model itself. Note that researchers also use high correlations in explanatory variables as a way to eliminate variables from the regression equation, but we would not like to recommend this because one could easily eliminate variables with good interpretation from the model while keeping variables which make no sense.

The purpose of this study is to compare different approaches for describing relationships between two continuous variables, with the goal of providing insights that can help readers apply this knowledge within a regression framework, thereby preventing, or at least recognizing potential problems in subsequent regression analysis. Santos et al. (2013) have already highlighted the importance of identifying the dependence between expression signals in the context of molecular biology. In this paper, we consider some of the measures considered by them, among others. In our article the emphasis is on choosing a *single* measure which could be used taking into account how useful they are in different Scenarios and how easy they are to compute. We use permutation tests to verify if the linear/nonlinear associations are statistically significant and specially we calculate the computational cost to obtain the results for each of the association measures.

Section 2 describes the different nonlinear correlation approaches used in this study. Besides the well known Pearson (r), Spearman (r_s) and Kendall's (τ) correlation coefficients, we considered continuous analysis of variance (r_{CAN}), predictive power score (r_{pps}), canonical correlation (r_C), maximal information coefficient (r_{MIC}), correlation distance (r_D), dynamic partition (r_{dp}) and maximum correlation (r_M). Section 3 describes the permutation test to determine the significance of the association measures' values. Section 4 shows some results considering an artificial data set and the Monte Carlo procedure. In Section 5 an application of the well known Boston data is presented. More information and details about the association measures are given in the Appendix.

2. Association Measures

The relationship between two or more variables can be i) *pairwise*, i.e. between two variables, ii) *full*, i.e. between one of the variables and the rest, and iii) pairwise *partial* i.e. between two variables correcting for the contribution of the rest. This paper concentrates on pairwise nonlinear associations and as with the linear correlation coefficients, here we are looking for measures which could flash up whether a nonlinear relationship exists between two continuous variables.

The standard Pearson's correlation coefficient (Galton, 1888; Pearson, 1920), r, is a summary statistics, detecting linear relationships between two continuous variables by providing the *strength* and *direction* of the relationship. For example, r, takes values between -1 to 1, with negative values showing a negative linear association while positive values showing a positive one. Values close to -1 or 1 show a strong linear association, while those close to zero a weak association. The Spearman's rank correlation coefficient r_s, (Spearman, 1904), and Kendall's τ (Kendall, 1938) do not assume linearity, but are measuring the strength and direction of monotonic relationships.

Nonlinear relationships may exhibit, both positive and negative aspects at different location; so, the direction (negative or positive) is not well defined globally. Therefore, the nonlinear association measures should show strength but not direction. A value close to zero shows a poor nonlinear relationship, a value close to 1 shows a strong relationship. Like its linear counterpart, the nonlinear coefficients do not, on their own, support hypothesis testing. To test the hypothesis that the nonlinear coefficients are zero, the *permutation test* is used based on Efron & Tibshirani (1993), Section 3.

Mathematically, linear relationships are easy to define, but unfortunately there exist a lot of different types of nonlinear associations and the question then becomes; 'which statistical measure is more appropriate to capture the most common nonlinear associations in the data'. This paper tries to answer this question by systematically examining the performance of different measures proposed in the literature. In what follows, we use capital letters (e.g., X) to denote random variables, and lowercase letters (e.g., x) to denote their observed values. The different measures of nonlinear association considered in this paper can be classified into the following categories:

- methods applying *nonparametric smoothing* functions, see below,
- a method based on *information* measures, r_{MIC} (see Appendix A.5) and
- a method based on *distance* measures, (the correlation distance, r_D, see Appendix A.6).

Here, we consider in more detail the methods that apply nonparametric smoothing functions. A nonparametric smoothing function can be fitted to the covariates X_1 and X_2 to model their pairwise nonlinear association. The nonlinear correlation between them can then be defined as the correlation coefficient between the fitted values for X_2 and X_1, denoted as $r_{1,2}$. However, this relationship is asymmetric, since fitting X_1 against X_2 would yield a completely different smooth function and, consequently, a different nonlinear correlation coefficient $r_{2,1}$. A simple way to address this asymmetry is to take the maximum of the two coefficients as the final measure of nonlinear association between X_1 and X_2, i.e., $\max(r_{1,2}, r_{2,1})$. We apply this approach in the calculation of:

i) the predictive power score, r_{pps}, a regression tree based correlation coefficient, see Appendix A.3 and

ii) the dynamic partition correlation, r_{dp}, see Appendix A.7.

The *maximum correlation*, popularized by Breiman & Friedman (1985) needs a smoothing technique to estimate the transformation functions g and f, (see Appendix A.1). The function `ace()` of the R package `acepack` (Spector et al., 2025) uses a `loess` smoother. For our analysis, we use our own function `ACE()`, in the R package `gamlss.prepdata` (Stasinopoulos et al., 2025), which uses either a `loess` or a P-spline smoother (Eilers & Marx, 1996). While the maximum

correlation r_M is not symmetric, X against Y is different from Y against X, we have found that the difference is small to justify intervention.

The canonical correlation r_C is obtained by approximating both X and Y with their B-spline bases, and then computing the canonical correlation between the resulting linear subspaces. That is, the maximum correlation between their two linear manifolds. The CANOVA correlation, r_{CAN}, approximates the relationship between Y and X locally, see Appendix A.2.

3. Permutation Test

Although the measures presented above are useful for measuring the strength of association between two continuous variables, they are not sufficient to test if this association is significant or not. To do this we will need the distribution of the association under the null hypothesis. Fisher (1915) have shown that for the Pearson's correlation coefficient r there is a simple transformation, $z = \operatorname{arctanh}(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ which makes the distribution of Z approximately normal, but this assumes that (X, Y) has a bivariate normal distribution, which is not the case in general. Furthermore for a general measure of correlation the exact distribution under the null hypothesis H_0, (that there is no association between the two random variables X and Y) is not known. Therefore, we cannot confirm if we have a 'high' observed value, in the sense of rejecting or not the null hypothesis, H_0. A practical solution, but computationally intensive, of assessing the significant of any association measure between variables is to use the *permutation test*. The permutation test is an idea introduced by R. A. Fisher in the 1930's (Fisher, 1935). The permutation test is *distributional free* in the sense that it does not require an assumed distribution for X and Y. The concepts behind the permutation tests are well described in Efron & Tibshirani (1993, page 200). In our case, in order to assess the significance of a general measure of association under the null hypothesis of independence (H_0) the Algorithm 1, given in Section 4, is used.

4. Simulation Study Using Artificial Data

To gain further insight into the behavior of the association measures used in this paper, a simulation study was conducted using artificial data to explore the impact of varying functional relationships between two continuous variables. Bivariate samples were generated with varying sample sizes 50, 80, and 100. The type of functional relationship between the samples is illustrated in Figure 1 and described mathematically in Table 1. Scenario 1 in Figure 1 represents a linear association between the two variables. Scenario 2 represents an independent pattern. From Scenario 3 to Scenario 8, a more complex pattern was adopted, see also Table 1.

Algorithm 1 Performing a permutation test to determine the significance of an association measure, say $r_{\hat{\theta}}$

(i) Construct a permuted data set under H_0 by keeping the observations of $X = (x_1, x_2, \ldots, x_n)$ fixed and randomly permuting the corresponding values of $Y = (y_1, y_2, \ldots, y_n)$;

(ii) Compute the chosen measure of association (i.e., r_θ) on the permuted data set;

(iii) Repeat steps (i) and (ii) a large number of B times to obtain the empirical null distribution of the statistic r_θ to compared it with the actual value obtained from the original X and Y, $r_{\hat{\theta}}$.

(iv) Approximate the p-level, by

$$p = \#(r_\theta > r_{\hat{\theta}})/B,$$

for a one tail test, where $\#(r_\theta > r_{\hat{\theta}})$ stand for the number of times the simulated r_θ is greater to the observed $r_{\hat{\theta}}$. Note that for the standard correlation coefficients, r, r_s and τ the test should be a two tail test giving $p = \#(|r_\theta| > |r_{\hat{\theta}}|)/B$.

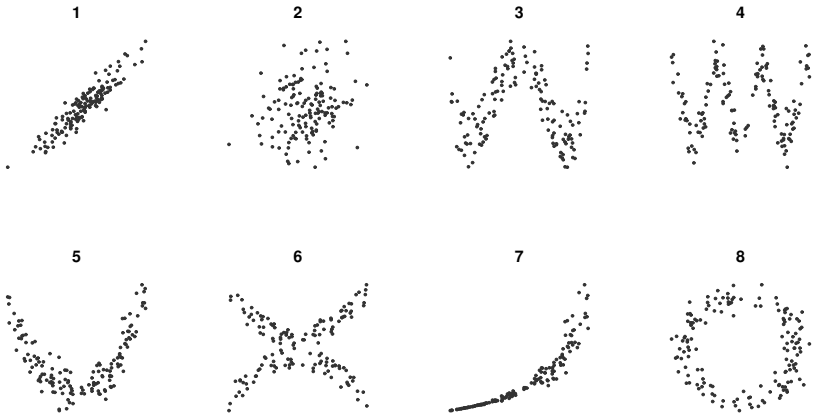

FIGURE 1: Simulated data showing different possibilities of association between two continuous variables.

TABLE 1: Mathematical representation of how the Scenarios in Figure 1 were simulated.

Scenario	Equation
1	$\mathbf{X} \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mathbf{0}, \mathbf{0})^\top$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$
2	$\mathbf{X} \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mathbf{0}, \mathbf{0})^\top$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
3	$X \sim U(-1,\ 1)$ and $Y = 4(X^2 - 1/2)^2 + 1/3X$
4	$X \sim U(-5,\ 9/2\pi)$ and $Y = 30\sin(X) + W$, where $W \sim \mathcal{N}(0,\ 10^2)$
5	$X \sim U(-1,\ 1)$ and $Y = 2X^2 + W$, where $W \sim \mathcal{N}(0,\ 25^2)$
6	$X \sim U(-1,\ 1)$ and $Y = (X^2 + W_1) \cdot W_2$, where $W_1 \sim U(0,\ 1/2)$ and $W_2 \sim U_d(-1,\ 1)$
7	$X \sim U(0,\ 1)$ and $Y = 2 + 1/5X + 1/5X^2 + 3X^4 + e^X + W$, where $W \sim \mathcal{N}(0,\ X^6)$
8	$X = \sin(V\pi) + W$ and $Y = \cos(V\pi) + W$, where $V \sim U(-1,\ 1)$, and $W \sim \mathcal{N}(0,\ (1/8)^2)$

where $U(\cdot)$ and $U_d(\cdot)$ denote the continuous and discrete uniform distributions, respectively.

Each Scenario was replicated 500 times and 10 different association measurements were calculated for all the replicas. The permutation test was employed to ascertain the success rate of rejecting the null hypothesis of non-association between the two variables. For each of the 500 replicas, the value of the descriptive level of the permutation test was approximated by using repermutations with $B = 1000$. In summary, 8 Scenarios with samples of 50, 80, and 100 were considered, with 10 measures of association employed. For each Scenario, $B = 1000$ permutations were used to obtain the permutation test results for each replica.

4.1. Results Obtained Using the Permutation Test

The results obtained from the simulations, showing the percentage of rejection of the null hypothesis H_0 by the permutation test are presented in Tables 2 to 9, at a descriptive level of 5%. This percentage was calculated as $\gamma = r_{H_0}/R$, where r_{H_0} represents the number of rejections for the permutation test through the replicates and R is the number of replicates, i.e, the number of times that the permutation test was performed.

Seven of the Scenarios were generated under the alternative hypothesis, that is, that there is an association between the pairs. The exception is Scenario 2. We feel that the percentage of rejection of the null hypothesis is a good approximation for the power of the test defined as rejection of H_0 when it is false. For Scenario 2, since there is independence between the two variables, the percentage rejection of H_0 provides information on the type I error (the rejection of H_0 when it is true).

Among the 8 Scenarios considered, Scenarios 1 (strong linear association) and 7 (strong nonlinear association) showed similar results, with practically all the measures reaching a rejection percentage of H_0 of 100%, including the Pearson, Spearman and Kendall coefficient measures. The exception is the canonical correlation which did not obtain such precise results, especially for sample sizes 50 and 80. It is also important to note that, despite Scenario 7 showing a nonlinear pattern, the Pearson, Spearman and Kendall measures rejected the null hypothesis 100% of the time.

Table 2 shows the results for Scenario 1, where almost all the measures performed well, with a percentage of rejection of the null hypothesis approximately equal to 100%.

TABLE 2: Percentage of rejection of the null hypothesis (H_0) in the permutation test for the measures of association in Scenario 1 via Monte Carlo simulation.

Measure/ Scenario	r		r_s		τ		r_{CAN}		r_{pps}	
	n	γ	n	γ	n	γ	n	γ	n	γ
1	50	100%	50	100%	50	100%	50	100%	50	100%
	80	100%	80	100%	80	100%	80	100%	80	100%
	100	100%	100	100%	100	100%	100	100%	100	100%
Measure/ Scenario	r_D		r_{dp}		r_M		r_C		r_{MIC}	
	n	γ	n	γ	n	γ	n	γ	n	γ
1	50	100%	50	100%	50	100%	50	64.0%	50	100%
	80	100%	80	100%	80	100%	80	84.6%	80	100%
	100	100%	100	100%	100	100%	100	90.4%	100	100%

For Scenario 2, Table 3, (random pattern), it is possible to see a problem, since the vast majority of the measures had a rejection percentage above 5%, i.e. the number of times the permutation test makes a type I error is greater than 5 out of 100. Particularly noteworthy is the r_{pps} (Predictive Power Score) measure, which showed percentages around 50%, i.e. for this measure, the decision to reject H_0 is wrong approximately half the time in this Scenario. The measures with the best results in this Scenario are r_{MIC} (maximal information coefficient) and r_M (maximum correlation), which had a rejection percentage around 5% in all sample sizes.

The Pearson's r, Spearman's r_s and Kendall's τ measures failed to detect an association between the variables for the other relationship patterns (Scenarios 3 to 6 and 8).

TABLE 3: Percentage of rejection of the null hypothesis (H_0) in the permutation test for the measures of association in Scenario 2 via Monte Carlo simulation.

Measure/ Scenario	r		r_s		τ		r_{CAN}		r_{pps}	
	n	γ	n	γ	n	γ	n	γ	n	γ
2	50	5.0%	50	4.4%	50	4.2%	50	6.6%	50	49.4%
	80	7.2%	80	5.6%	80	5.8%	80	6.8%	80	50.0%
	100	4.8%	100	5.2%	100	4.8%	100	5.4%	100	48.0%
Measure/ Scenario	r_D		r_{dp}		r_M		r_C		r_{MIC}	
	n	γ	n	γ	n	γ	n	γ	n	γ
2	50	4.6%	50	6.4%	50	5.2%	50	6.6%	50	4.2%
	80	5.4%	80	4.8%	80	4.4%	80	5.2%	80	4.4%
	100	4.6%	100	3.2%	100	5.0%	100	5.0%	100	5.0%

For Scenario 3, shown in Table 4, only the r_{CAN}, r_{pps}, r_{MIC} and r_M measures showed good results at all sample sizes. The r_D and r_{dp} measures had satisfactory results only at sample sizes 80 and 100. For sample size $n = 50$, both measures had a rejection rate of 93.2%. In Scenario 4 (Table 5), the best results are for the r_{CAN} and r_{MIC} measures, with a rejection percentage of 100% in all sample sizes. In addition, the r_{pps}, r_{dp} and r_M measures also showed good results at sample sizes $n = 80$ and 100, with a rejection percentage above 95%.

TABLE 4: Percentage of rejection of the null hypothesis (H_0) in the permutation test for the measures of association in Scenario 3 via Monte Carlo simulation.

Measure/ Scenario	r		r_s		τ		r_{CAN}		r_{pps}	
	n	γ	n	γ	n	γ	n	γ	n	γ
3	50	4.2%	50	7.6%	50	8.0%	50	100%	50	100%
	80	3.6%	80	7.2%	80	7.6%	80	100%	80	100%
	100	5.0%	100	8.4%	100	8.0%	100	100%	100	100%
Measure/ Scenario	r_D		r_{dp}		r_M		r_C		r_{MIC}	
	n	γ	n	γ	n	γ	n	γ	n	γ
3	50	93.2%	50	93.2%	50	97.2%	50	19.4%	50	100%
	80	100%	80	99.4%	80	99.0%	80	79.2%	80	100%
	100	100%	100	99.4%	100	99.8%	100	97.0%	100	100%

TABLE 5: Percentage of rejection of the null hypothesis (H_0) in the permutation test for the measures of association in Scenario 4 via Monte Carlo simulation.

Measure/ Scenario	r		r_s		τ		r_{CAN}		r_{pps}	
	n	γ	n	γ	n	γ	n	γ	n	γ
4	50	3.6%	50	4.4%	50	4.6%	50	100%	50	86.8%
	80	3.8%	80	5.2%	80	5.0%	80	100%	80	100%
	100	2.8%	100	4.0%	100	4.4%	100	100%	100	100%
Measure/ Scenario	r_D		r_{dp}		r_M		r_C		r_{MIC}	
	n	γ	n	γ	n	γ	n	γ	n	γ
4	50	10.4%	50	93.4%	50	92.0%	50	21.4%	50	100%
	80	12.8%	80	99.8%	80	95.2%	80	82.6%	80	100%
	100	17.0%	100	99.8%	100	97.2%	100	96.6%	100	100%

In Scenario 5 (Table 6), the r_{CAN}, r_{pps} and r_M measures showed the best results. The r_{MIC}, r_D and r_{dp} measures showed greater competitiveness only for sample sizes above 80. For Scenario 6 (Table 7), only the maximum correlation measure performed well in all sample sizes. The r_C and r_{MIC} measures achieved interesting results only for n greater than or equal to 80. The others failed to detect the association between the variables effectively.

TABLE 6: Percentage of rejection of the null hypothesis (H_0) in the permutation test for the measures of association in Scenario 5 via Monte Carlo simulation.

Measure/ Scenario	r		r_s		τ		r_{CAN}		r_{pps}	
	n	γ	n	γ	n	γ	n	γ	n	γ
5	50	12.2%	50	15.2%	50	19.0%	50	98.0%	50	97.2%
	80	12.4%	80	15.0%	80	18.8%	80	99.8%	80	99.4%
	100	12.2%	100	14.6%	100	20.6%	100	100%	100	99.8%
Measure/ Scenario	r_D		r_{dp}		r_M		r_C		r_{MIC}	
	n	γ	n	γ	n	γ	n	γ	n	γ
5	50	94.0%	50	74.2%	50	95.0%	50	45.8%	50	84.6%
	80	100%	80	97.6%	80	98.2%	80	54.8%	80	98.4%
	100	100%	100	99.6%	100	99.6%	100	75.8%	100	100%

TABLE 7: Percentage of rejection of the null hypothesis (H_0) in the permutation test for the measures of association in Scenario 6 via Monte Carlo simulation.

Measure/ Scenario	r		r_s		τ		r_{CAN}		r_{pps}	
	n	γ	n	γ	n	γ	n	γ	n	γ
6	50	16.2%	50	15.4%	50	21.0%	50	17.6%	50	61.0%
	80	14.8%	80	14.2%	80	21.0%	80	15.4%	80	67.8%
	100	14.6%	100	14.4%	100	20.0%	100	15.4%	100	73.0%
Measure/ Scenario	r_D		r_{dp}		r_M		r_C		r_{MIC}	
	n	γ	n	γ	n	γ	n	γ	n	γ
6	50	39.2%	50	5.4%	50	98.4%	50	79.8%	50	44.4%
	80	72.4%	80	5.4%	80	99.4%	80	96.4%	80	97.8%
	100	91.0%	100	6.0%	100	100%	100	99.4%	100	99.8%

We have discussed Scenario 7 (Table 8) earlier in conjunction with Scenario 1.

TABLE 8: Percentage of rejection of the null hypothesis (H_0) in the permutation test for the measures of association in Scenario 7 via Monte Carlo simulation.

Measure/ Scenario	r		r_s		τ		r_{CAN}		r_{pps}	
	n	γ	n	γ	n	γ	n	γ	n	γ
7	50	100%	50	100%	50	100%	50	100%	50	100%
	80	100%	80	100%	80	100%	80	100%	80	100%
	100	100%	100	100%	100	100%	100	100%	100	100%
Measure/ Scenario	r_D		r_{dp}		r_M		r_C		r_{MIC}	
	n	γ	n	γ	n	γ	n	γ	n	γ
7	50	100%	50	100%	50	100%	50	93.6%	50	100%
	80	100%	80	100%	80	100%	80	99.8%	80	100%
	100	100%	100	100%	100	100%	100	100%	100	100%

In the 8th Scenario presented in Table 9, only the maximum correlation measure achieved good results, with a rejection percentage above 98% for $n = 50$ and 80, and 99.8% for $n = 100$. None of the other measures managed to capture this relationship pattern at any sample size, with the exception of the maximal information coefficient, which proved to be competitive at sample sizes 80 and 100, with a rejection percentage above 97%.

TABLE 9: Percentage of rejection of the null hypothesis (H_0) in the permutation test for the measures of association in Scenario 8 via Monte Carlo simulation.

Measure/ Scenario	r		r_s		τ		r_{CAN}		r_{pps}	
	n	γ	n	γ	n	γ	n	γ	n	γ
8	50	1.2%	50	0.2%	50	0.0%	50	8.4%	50	62.4%
	80	0.6%	80	0.2%	80	0.0%	80	6.6%	80	70.4%
	100	0.6%	100	0.2%	100	0.0%	100	10.0%	100	74.2%
Measure/ Scenario	r_D		r_{dp}		r_M		r_C		r_{MIC}	
	n	γ	n	γ	n	γ	n	γ	n	γ
8	50	9.8%	50	7.6%	50	98.8%	50	18.6%	50	41.4%
	80	18.2%	80	5.4%	80	98.8%	80	51.4%	80	97.6%
	100	30.0%	100	4.4%	100	99.8%	100	78.8%	100	100%

Table 10 shows the running time (in hours) of the Monte Carlo simulation (500 replicates) to calculate the percentage of rejection of the null hypothesis for all the measures in the 8 Scenarios. The measures that required the least execution time were Pearson, Spearman, Kendall and maximum correlation, in that order. The longest times were recorded for Predictive Power Score with 73.78h and dynamic partitioning with 171.27h.

TABLE 10: Execution time of the Monte Carlo simulation to calculate the percentage of rejection of the null hypothesis (H_0) for the measurements in the 8 Scenarios.

Measure	**Time (hours)**
Pearson's correlation (r)	0.36
Spearman's correlation (r_s)	0.42
Kendall's correlation (τ)	0.54
Continuous analysis of variance (r_{CAN})	3.76
Predictive power score (r_{pps})	73.78
Canonical correlation (r_C)	1.14
Maximal information coefficient (r_{MIC})	1.93
Correlation distance (r_D)	0.68
Dynamic partition (r_{dp})	171.27
Maximum correlation (r_M)	0.56

The highlight here is the maximum correlation measure, r_M, which performed best in all 8 Scenarios and had a very low execution time, similar to the time recorded for the linear correlation measures. The maximal information coefficient, r_{MIC}, also proved to be very competitive, with results similar to those obtained by the r_M measure. However, although the execution time recorded for this measure was not too high when compared to the Predictive Power Score and dynamic partitioning measures, it was still considerably longer than the time required by the maximum correlation measure.

5. Application to the Boston Data

In this Section we present results of nonlinear correlation analysis applied to variables from the data set used in the study by Harrison & Rubinfeld (1978). That study examined Housing data from the city of Boston during the 1970s. The data set, commonly known as the Boston Housing Data, is widely recognized in the fields of machine learning and statistical modeling, and has been extensively employed in numerous studies, particularly those related to housing price prediction.

The database comprises 506 neighborhoods within the Boston metropolitan area and includes 14 variables, as indicated in Table 11. Two variables with very few distinct values were eliminated from the data set, `chas` (2 distinct observations) and `rad` (9 dist. obs.). The scatter plots of the remaining variables together with Pearson correlation coefficients are shown in Figure 2. The plot was generated using the `ggpairs()` function from the **GGally** `R` package (Schloerke et al., 2025).

It is evident from Figure 2 that the variables exhibit a disparate pattern of associations. In certain instances, the nature of the relationship remains ambiguous, particularly with regard to whether it is linear or nonlinear.

TABLE 11: Description of the variables present in the Boston Housing data set.

Name	Description
`crim`	Crime rate per capita by city.
`zn`	Proportion of residential land zoned for lots over 25 000 square feet.
`indus`	Proportion of non-retail business acres per city.
`chas`	Charles River dummy variable (1 if the tract limits the river; 0 otherwise).
`nox`	Concentration of nitric oxides (parts per 10 million).
`rm`	Average number of rooms per household.
`age`	Proportion of owner-occupied units built before 1940.
`dis`	Weighted distances to five Boston job centers.
`rad`	Accessibility index for radial roads.
`tax`	Full value property tax for U$10 000.
`ptratio`	Student-teacher ratio per city.
`block`	$1000(B_k - 0,63)^2$ where B_k is the proportion of black people per city.
`lstat`	Percentage of lower status population.
`medv`	Average value of owner-occupied homes in U$1 000.00.

The maximum correlation coefficients of all remaining variables of the Boston Housing Data are shown in Figure 3. The figure was generated with the function `data_mcor()` of the R package **gamlss.prepdata** (Stasinopoulos et al., 2025).

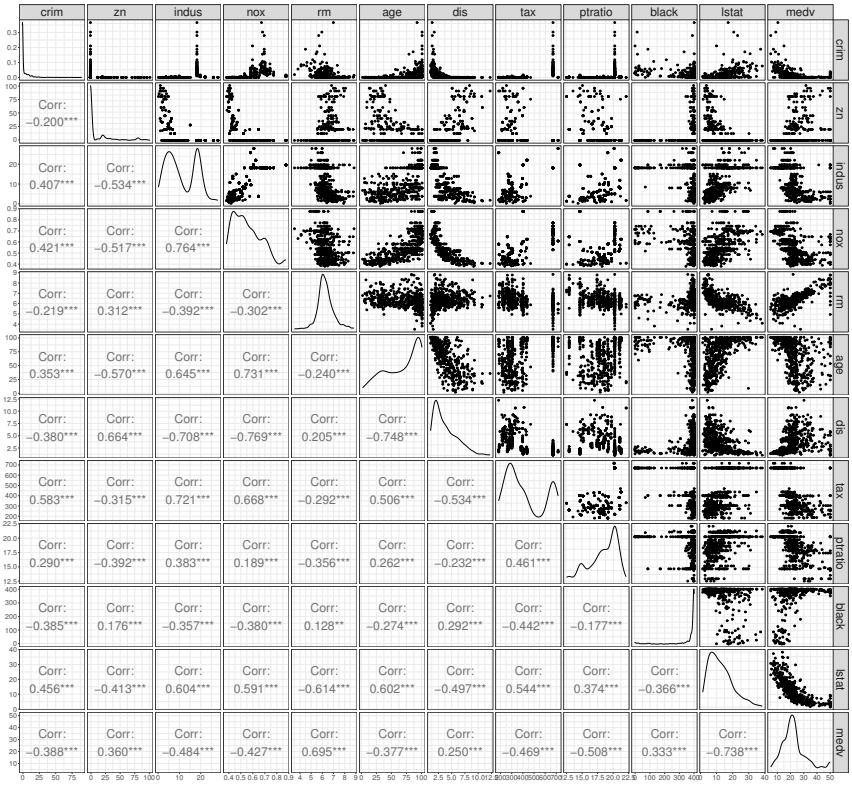

FIGURE 2: Pairwise scatter plots of the continuous variables in the Boston dataset with corresponding Pearson correlation coefficients.

FIGURE 3: Maximum correlation matrix of the continuous variables in Boston data set.

There is an obviously difference between the linear, Pearson's correlation coefficients, and the nonlinear maximum correlations. For example, the entrance for `tax` against `crim` has the very high value of 0.947 for the maximum correlation and the moderate 0.583 value for the linear correlation. Note that for the Boston Housing Data, having only 12 variables we can afford to look also at the scatter plots, something almost impossible for bigger data sets.

6. Conclusion

Pearson's, Spearman's, and Kendall's correlation coefficients are indeed very good of detecting linear relationship. However, they are less effective when examining nonlinear configuration.

The maximum correlation r_M and the maximal information coefficient r_{MIC} demonstrated the best performance across all patterns, particularly in samples of size $n > 80$. While we have not simulate bigger samples (due to computational cost) we believe that will be suitable for larger samples. In addition, this two measures exhibited the lowest percentages in the assessment of Scenario 2 (false positive). Furthermore, the maximum correlation measure r_M has the advantage

of requiring less computational time than the maximal information coefficient to run especially using the permutation test. The dynamic partitioning r_{dp} measure demonstrated notable competitiveness for $n = 80$ and 100, although it exhibited suboptimal performance in Scenarios 6 and 8 within the simulation study. Moreover, this measure entail a significant computational burden.

It should be noted that this is not an exhaustive list of all possible measures of association. And it is crucial to recognize that the existence of an association between two continuous variables does not necessarily imply causation. In a future research, it would be extend to categorical variables and also to check whether correlations based on copula concepts could improve our understanding.

Acknowledge

We acknowledge the partial financial support from Capes, CNPq (projects 306561/2020-4, 404872/2023-9 and 302413/2022-7), as well as FACEPE – *Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco* (grant number IBPG-0378-1.02/21), which supports Brito's ongoing doctoral research, of which this study is a part, through a PhD scholarship.

[Received: September 2025 — Accepted: November 2025]

References

Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G. & Furlanello, C. (2013), 'Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers', *Bioinformatics* **29**(3), 407–408.

Breiman, L. & Friedman, J. H. (1985), 'Estimating optimal transformations for multiple regression and correlation', *Journal of the American statistical Association* **80**(391), 580–598.

Buja, A. (1990), 'Remarks on functional canonical variates, alternating least squares methods and ace', *The Annals of Statistics* pp. 1032–1069.

Buja, A., Hastie, T. & Tibshirani, R. (1989), 'Linear smoothers and additive models', *The Annals of Statistics* pp. 453–510.

Edelmann, D., Fokianos, K. & Pitsillou, M. (2019), 'An updated literature review of distance correlation and its applications to time series', *International Statistical Review* **87**(2), 237–262.

Efron, B. & Tibshirani, R. J. (1993), *An introduction to the bootstrap*, CRC press.

Eilers, P. H. & Marx, B. D. (1996), 'Flexible smoothing with b-splines and penalties', *Statistical science* **11**(2), 89–121.

Fisher, R. A. (1915), 'Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population', *Biometrika* **10**(4), 507–521.

Fisher, R. A. (1935), *The Design of Experiments*, Oliver and Boyd, Edinburgh.

Fung, W.-K., Zhu, Z.-Y., Wei, B.-C. & He, X. (2002), 'Influence diagnostics and outlier tests for semiparametric mixed models', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**(3), 565–579.

Galton, F. (1888), 'Co-relations and their measurement, chiefly from anthropometric data', *Proceedings of the Royal Society of London* **45**, 135–145.

Gebelein, H. (1941), 'Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung', *ZAMM – Journal of Applied Mathematics and Mechanics* **21**(6), 364–379.

Harold, H. (1936), 'Relations between two sets of variates', *Biometrika* **28**, 321–377.

Harrell, Frank E., J. (2015), *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd edn, Springer.

Harrison, Jr, D. & Rubinfeld, D. L. (1978), 'Hedonic housing prices and the demand for clean air', *Journal of environmental economics and management* **5**(1), 81–102.

He, G., Müller, H.-G. & Wang, J.-L. (2004), 'Methods of canonical analysis for functional data', *Journal of Statistical Planning and Inference* **122**(1-2), 141–159.

He, X. & Shen, L. (1997), 'Linear regression after splin transformation', *Biometrika* **84**(2), 474–481.

Kendall, M. G. (1938), 'A new measure of rank correlation', *Biometrika* **30**, 81–89.

Pearson, K. (1920), 'Notes on the history of correlation', *Biometrika* **13**, 25–45.

Ramsay, J. O. (1988), 'Monotone regression splines in action', *Statistical science* pp. 425–441.

Ranjan, C. & Najari, V. (2020), 'Package nlcor: Compute nonlinear correlations', *ResearchGate* .

Rényi, A. (1959), 'On measures of dependence', *Acta mathematica hungarica* **10**(3-4), 441–451.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. & Sabeti, P. C. (2011), 'Detecting novel associations in large data sets', *Science* **334**(6062), 1518–1524.

Rizzo, M. L. & Székely, G. J. (2024), *energy: E-statistics: Multivariate inference via the energy of data*. R package version 1.7-12. https://CRAN.R-project.org/package=energy

Santos, S. S., Takahashi, D. Y., Nakata, A. & Fujita, A. (2013), 'A comparative study of statistical methods used to identify dependencies between gene expression signals', *Briefings in Bioinformatics* **15**(6), 906–918.

Sarmanov, O. (1962), 'Maximum correlation coefficient (nonsymmetric case)', *Selected translations in mathematical statistics and probability* **2**, 207–210.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A. & Crowley, J. (2025), *GGally: Extension to 'ggplot2'*. R package version 2.4.0. https://CRAN.R-project.org/package=GGally

Spearman, C. (1904), ''General intelligence', objectively determined and measured', *The American Journal of Psychology* **15**, 201–292.

Spector, P., Friedman, J., Tibshirani, R., Lumley, T., Garbett, S., Baron, J., Klar, B. & Chasalow, S. (2025), *acepack: ACE and AVAS for Selecting Multiple Regression Transformations*. R package version 1.6.3. https://CRAN.R-project.org/package=acepack

Stasinopoulos, M. D., Rigby, R. & De Bastiani, F. (2025), *gamlss.prepdata: Prepering Data for Fitting a Generalized Additive Model for Location Scale and Shape*. R package version 0.1.19. https://www.gamlss.com/

Székely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007), 'Measuring and testing dependence by correlation of distances', *The Annals of Statistics* **35**(6), 2769–2794.

Van der Laken, P. (2021), *ppsr: Predictive Power Score*. R package version 0.0.2. https://CRAN.R-project.org/package=ppsr

Wang, G., Lin, N. & Zhang, B. (2012), 'Functional linear regression after spline transformation', *Computational Statistics & Data Analysis* **56**(3), 587–601.

Wang, T. & Zhu, L. (2018), 'Flexible dimension reduction in regression', *Statistica Sinica* pp. 1009–1029.

Wang, Y., Li, Y., Cao, H., Xiong, M., Shugart, Y. Y. & Jin, L. (2015), 'Efficient test for nonlinear dependence of two continuous variables', *BMC bioinformatics* **16**, 1–8.

Yi, L. (2025), *canova: CANOVA: Efficient test for nonlinear dependence of two continuous variables*. R package version 0.1.0. https://github.com/liyistat/canova

Yu, Y. (2008), 'On the maximal correlation coefficient', *Statistics & Probability Letters* **78**(9), 1072–1075.

Appendix A. Nonlinear Measures of Association

Appendix A.1. Maximum Correlation

The maximum correlation is defined as

$$r_M(X,Y) = r(g^*, f^*) = \max_{g,f} \; r\Big[g(Y), f(X)\Big], \tag{1}$$

where r is the Pearson's correlation coefficient and r_M is the maximum correlation between X and Y. That is, the maximum correlation is the simple correlation between $g(Y)$ and $f(X)$ where the functions g and f are unknown transformations of the original random variables Y and X, respectively. This measure was introduced by Gebelein (1941) and popularized by Breiman & Friedman (1985). It is capable of assessing the degree of both linear or nonlinear association between two variables. The maximum correlation has been the subject of studies over the years. These include Rényi (1959), Sarmanov (1962), Buja (1990), Yu (2008), among others. Note that $0 \leq r_M(X,Y) \leq 1$ and the maximum correlation value between X and Y is equal to zero only in cases where X and Y are independent. For the bivariate case, the functions $f()$ and $g()$ are obtained by minimizing $E\{[g(Y) - f(X)]^2\}/Var[g(Y))]$ which yields the so-called *optimal transformations* (Breiman & Friedman, 1985). This method can be applied not only when both variables are continuous, but also when one or both are ordered continuous variables, ordered categorical variables.

The maximum correlation measure can be computed in R using the package `acepack` (Spector et al., 2025). For the analysis of this paper we used our one version function `ACE()` in the package `gamlss.prepdata` (Stasinopoulos et al., 2025).

Appendix A.2. Continuous Analysis of Variance

Continuous analysis of variance (CANOVA), defined by Wang et al. (2015), is based on the same idea as ANOVA. However, instead of studying the variability of Y over all values of X, it study the variability within a given neighborhood constructed on the basis of the numerical values of X. The method is based on the hypothesis that similar values of X lead to similar values of Y, i.e. Y can be understood as a function of X, i.e. $\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$ and f is a non-constant smoothing function. Based on this idea and given two random variables X and Y with x_i and y_i the i-th respective values for X and Y, the sum of squares of the inner neighbourhood is defined as

$$W = \sum_{i,j} (y_i - y_j)^2, \quad j < i, \quad |rank(x_i) - rank(x_j)| < K, \tag{2}$$

where K is a constant integer value. The neighbourhood structure is defined by $|rank(x_i) - rank(x_j)|$. When X and Y show some association, the value of the W statistic tends to be small. The CANOVA results for the association between the two variables X and Y can be obtained using the `canova` function from the R package `canova` (Yi, 2025).

Appendix A.3. Predictive Power Score (PPS)

The *Predictive Power Score (PPS)* was introduced by Van der Laken (2021). This measure is commonly calculated based on the results of the evaluation metrics of a decision tree or a GLM model. An interesting feature of this measure is that it makes it possible to assess the degree of association between variables when one of them is categorical.

When the study of the relationship involves only quantitative variables, the r_{pps} value is obtained based on the values of the root mean square error - RMSE (in the case of the GLM) or the mean absolute error - MAE (in the case of the decision tree). Thus, if we consider the method based on fitting a GLM, the r_{pps} value can be obtained by

$$r_{pps} = 1 - \frac{\text{RMSE}_{model}}{\text{RMSE}_{null}},$$

where the null model is the one that do not have explanatory variables. It is important to emphasize that there are not many studies in the literature involving this measure and its evaluation more often involves practical results. The analysis in this paper was performed using the function `score()` of the package `ppsr` (Van der Laken, 2021).

Appendix A.4. Canonical Correlation

Canonical correlation analysis, proposed by Harold (1936), is a technique in multivariate analysis seeking to find the best linear combinations of two sets of continuous variables, $\mathbf{X}$'s and $\mathbf{Y}$'s. That is, the method looks for linear combinations of the two variables say $\mathbf{u} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{v} = \mathbf{Y}\boldsymbol{\alpha}$ with the maximal correlation. Let $\mathbf{B}_X$ be the beta spline base for variable $\mathbf{X}$, i.e. $\mathbf{X} = \mathbf{B}_X\boldsymbol{\gamma}_X$ and $\mathbf{B}_Y$ be the beta spline base for variable $\mathbf{Y}$ i.e. $\mathbf{Y} = \mathbf{B}_Y\boldsymbol{\gamma}_Y$. The canonical correlation $r_C(\mathbf{X}, \mathbf{Y})$ of $\mathbf{X}$ and $\mathbf{Y}$ is the canonical correlation between the linar subspaces $\mathbf{B_x}$ and $\mathbf{B_y}$. The method only measure linear relationships between set of variables. Ramsay (1988), He & Shen (1997), Fung et al. (2002), He et al. (2004), Wang et al. (2012) and Wang & Zhu (2018), use transformations of the variables i.e. splines, before applying the canonical correlation technique.

Appendix A.5. Maximal Information Coefficient (MIC)

The maximal information coefficient (r_{MIC}) was proposed by Reshef et al. (2011). MIC is part of the wider larger class of maximal information-based nonparametric exploration (MINE) statistics used to identifying and classifying relationships. Theoretically MIC is based on the concepts of entropy and mutual information, ideas which are easily defined when the two variables of interest are discritized. The idea behind it is that if a relationship exists between X and Y then a grid, on their scatterplot, can encapsulate this relationship. Thus, to calculate the MIC of X and Y all grids up to a maximal grid resolution, dependent on the sample size are examined.

Let $D = (x_i, y_i)$ for $i = 1, \ldots, n$ be the pair of observations and $D|_G(x, y)$ the many possible partitions of the data into bins along the X and Y-axis. Note that the number of partitions is a function of the number of observations n i.e. a typically value is $n^{0.6}$. Then the mutual information (entropy) of the discritized set is calculated, $\mathrm{MI}(x, y) = \sum_{i=1}^{n} p(x_i, y_i) \log \frac{p(x_i, y_i)}{p(x_i)p(y_i)}$. Note that the mutual information gives the cost of approximating the joint distribution $p_{X,Y}(x, y)$ using the marginals $p_X(x)$ and $p_Y(y)$. If the approximation is good the mutual information will be small. The mutual information is then scaled by $\log_2 \min(x, y)$. Summarizing the procedure the r_{MIC} is defined as;

$$r_{MIC} = \max_{\text{over grids } D|_G} \frac{\mathrm{MI}(D|_G, x, y)}{\log_2 \min(x, y)}.$$

To calculate r_{MIC} in this paper we use the `R` package `minerva` (Albanese et al., 2013).

Appendix A.6. Correlation Distance

Similarly to the case of correlation (standardization of covariance), the correlation distance comes from the same concept. Thus, Székely et al. (2007) introduced the covariance distance as a new measure of association between two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ of dimension p and q respectively. The definition of the covariance distance function depends on the joint characteristic function between $\boldsymbol{X}$ and $\boldsymbol{Y}$, denoted by $\phi_{\boldsymbol{X},\boldsymbol{Y}}(\mathbf{t}, \mathbf{s}) = \mathbb{E}\big[\exp(i(\mathbf{t}'\boldsymbol{X} + \mathbf{s}'\boldsymbol{Y}))\big]$, where $(\mathbf{t}, \mathbf{s}) \in \mathbb{R}^{p+q}$ and $i^2 = -1$. The corresponding marginal characteristic functions for $\boldsymbol{X}$ and $\boldsymbol{Y}$ are denoted by $\phi_{\boldsymbol{X}}(\mathbf{t}) = \mathbb{E}\big[\exp(i(\mathbf{t}'\boldsymbol{X}))\big]$ and $\phi_{\boldsymbol{Y}}(\mathbf{s}) = \mathbb{E}\big[\exp(i(\mathbf{s}'\boldsymbol{Y}))\big]$, respectively. The covariance distance is then defined as the non-negative square root of a weighted L_2 distance between the joint and marginal characteristic functions between the two random vectors. That is,

$$\mathcal{V}^2(\boldsymbol{X}, \boldsymbol{Y}) = \int_{\mathbb{R}^{p+q}} = \left|\phi_{\boldsymbol{X},\boldsymbol{Y}}(\mathbf{t}, \mathbf{s}) - \phi_{\boldsymbol{X}}(\mathbf{t})\phi_{\boldsymbol{Y}}(\mathbf{s})\right|^2 \omega(\mathbf{t}, \mathbf{s}) \mathrm{d}t \mathrm{d}s, \tag{3}$$

where $\omega(\cdot, \cdot) : \mathbb{R}^{p+q} \to \mathbb{R}$ is the weight function for which the integral exist (Edelmann et al., 2019). Standardizing the equation (3), we get the correlation distance given below.

$$\mathcal{R}^2(\boldsymbol{X}, \boldsymbol{Y}) = \begin{cases} \frac{\mathcal{V}^2(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{\mathcal{V}^2(\boldsymbol{X},\boldsymbol{X})\mathcal{V}^2(\boldsymbol{Y},\boldsymbol{Y})}}, & \mathcal{V}^2(\boldsymbol{X}, \boldsymbol{X})\mathcal{V}^2(\boldsymbol{Y}, \boldsymbol{Y}) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Unlike Pearson's correlation coefficient, which only captures linear relationships between variables and can be zero even in cases where the variables show some kind of association. The correlation distance can measure both linear and nonlinear associations and its value is zero only in cases where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.

To calculate the correlation distance in this paper we use the function `dcor` from the R package `energy` (Rizzo & Székely, 2024).

Appendix A.7. Dynamic Partition

The relationship between the variables X and Y can also be studied using a segmented linear model, which according to Harrell (2015) can be understood as a linear spline function or piecewise linear function. Thinking of the bivariate case, the intuitive idea here is that X and Y can present linear patterns at different scales and/or in different directions. In other words, we can have more than one model (regression line) to represent the relationship between X and Y. However, this relationship can be represented by a single model that considers all the configurations of the relationship between the variables, the mathematical definition of which can be seen in the Equation (5).

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - x_a) I_a, \tag{5}$$

where x_a is the value of X at which there is a discontinuity in the linear relationship between the variables, commonly called a node and I_a is an indicator function, such that

$$I_a = \begin{cases} 0, & \text{if } x_i \leq x_a \\ 1, & \text{if } x_i > x_a. \end{cases}$$

In many situations, however, the relationship between X and Y has more than one point of discontinuity. In these cases, the (5) model can easily be extended to the case where we have x_k, $k = 1, \ldots, K$ knots.

To calculate the dynamic partition correlation measure in this paper we use the R package `nlcor` (Ranjan & Najari, 2020).