Knowledge and Information Systems
https://doi.org/10.1007/s10115-025-02436-z

RESEARCH

Numerical dependency analysis (NDA): a new method for estimating the statistical dependence (not correlation) of two variables

Abolfazl Zanghaei[1,2] · **Hassan Doosti**[3] · **Ali Ameri**[1]

Received: 26 January 2024 / Revised: 18 December 2024 / Accepted: 30 March 2025
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2025

Abstract
Dependence and correlation represent distinct statistical concepts. While there are methods to measure linear and nonlinear correlation between two variables, understanding the statistical dependence between them remains a topic of great interest. In this paper, we propose a heuristic, numerical, and algorithmic approach to estimate the dependence coefficient between two variables. With this approach, first, the **X**–**Y** scatter plot is transformed into a functional scatter plot using a procedure called "functionalizing." Next, a novel concept called "successive triangles" is employed to estimate the dependence of **Y** on **X**. The proposed method offers several advantages; it is distribution-free, so it is suitable for both Gaussian and non-Gaussian numerical variables. Moreover, it can be used for both numerical and categorical (nominal) variables. This approach can be employed in other applications such as correlation measurement and also template matching for single-dimensional patterns. The presented method has been validated by both the simulated and clinical data with promising results.

Keywords Association · Copula · Correlation · Dependence · Dependency · Functional · Numerical · Nonlinear · NDA · Relation

Mathematics Subject Classification 62H20 · 62H05 · 62G05

✉ Abolfazl Zanghaei
zanghaei@sbmu.ac.ir

Hassan Doosti
hassan.doosti@mq.edu.au

Ali Ameri
aliameri86@gmail.com

1 Department of Biomedical Engineering and Biophysics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

2 Sinus and Surgical Endoscopic Research Center, Mashhad University of Medical Sciences, Mashhad, Iran

3 School of Mathematical and Physical Sciences, Macquarie University, Sydney, Australia

Published online: 05 May 2025

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

1 Introduction

Correlation and dependence represent distinct concepts [1–4]. While the primary focus of this article is not the difference between these two notions, it is essential to clarify their meanings. Dependence implies causation [5, 6], whereas correlation signifies an association between two variables. Notably, dependence is a more robust concept than correlation.

In most studies, it is crucial to understand how an independent variable impacts a dependent one. Irrelevant or redundant variables can weaken analytical results [7]. For instance, in kidney transplantation studies, various parameters, such as blood group, family relationships between donors and recipients, age, and laboratory measurement, may influence transplantation outcomes [8]. Identifying which variables most strongly predict transplantation outcomes requires a well-designed experimental plan to measure dependence. However, conducting an ideal experiment where all independent variables except the desired one are fixed can be costly and time-consuming. As a practical alternative, researchers often employ quasi-experimental methods like propensity score matching (PSM) to estimate the causal impact of an independent variable, mimicking the effects of a randomized control trial [9].

While some methods focus on feature importance [10–27], assessing dependence effectively can identify important features. Regression models can capture relationships between variables [28–30], but quantifying the strength of a functional relation remains an area of interest. Common correlation coefficients [31–35], such as Pearson's correlation coefficient [36], measure the correlation, but cannot directly assess the dependence [37]. Consequently, researchers have developed various methods to accurately measure dependence over the past few decades [38–61]. These methods include mutual information [40, 62–64], copula-based techniques [30, 65–68], kernel-base methods [48, 69–71], conditional dependence analysis [72], empirical methods [73], and principal curves [74]. Some approaches specifically address nonlinear dependence between continuous variables [33, 74–79], while others exhibit symmetry or asymmetry [4, 53, 67, 80–83]. Simulated data can be useful for assessing these methods [84, 85].

By increasing mixed datasets that include both numerical and categorical variables, it is necessary to present heuristic numerical algorithms beyond the traditional methods [86–88]. Our proposed method stands out because it accommodates both nominal and numerical data. It is asymmetric, allowing it to serve as a tool for measuring correlation as well. For instance, if the dependence of a random variable $\mathbf{X}$ on $\mathbf{Y}$ is high while the dependence of $\mathbf{Y}$ on $\mathbf{X}$ is low, $\mathbf{X}$ and $\mathbf{Y}$ are correlated but not dependent. Additionally, this method facilitates the construction of causal models between variables [88–92].

1.1 Nonlinear dependency analysis

Our proposed method, "*numerical dependency analysis (NDA)*," yields an asymmetry matrix. In this matrix, the elements of column j indicate how much other variables can predict variable j. Unlike common correlation matrices, which exhibit symmetry, NDA introduces asymmetry:

- *Correlation Matrix (Common Case)*:

$$\mathrm{corr}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{corr}(\boldsymbol{Y}, \boldsymbol{X})$$

- *NDA Matrix (Asymmetric Case)*:

$$\mathrm{NDA}(\boldsymbol{X}, \boldsymbol{Y}) \neq \mathrm{NDA}(\boldsymbol{Y}, \boldsymbol{X})$$

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

In simple terms, the NDA algorithm quantifies the randomness of **Y** variations caused by changes in **X**. While NDA cannot precisely measure the dependency of **Y** on **X**, it provides a good estimation of the statistical dependence of **Y** on **X**. Throughout this paper, when we refer to the "measure of dependence," we specifically mean an estimation of dependency measurement.

1.1.1 Preliminaries and methods

In this section, first, a brief definition of dependence is provided, then the main approach to measuring dependence is presented. Finally, the difference between dependence and correlation is described.

1.1.2 Functional relation and dependency

The association or relation of two variables **X** and **Y** can be expressed as a functional or non-functional relation. In an ideal functional relation, the value of **X** uniquely determines the value of **Y**:

$$\boldsymbol{X} \rightarrow \boldsymbol{Y} \tag{1}$$

Figure 1 shows an example of a functional relation (**Y** is completely dependent on **X** [68]), while Fig. 2 represents a non-functional relationship.

In an ideal functional relationship, **X** can predict (or determine) **Y** uniquely. It means that, for each **X** value, there is no more than one value of **Y**. This is the principal key to predictability. But in a non-functional relationship, there can be more than one **Y** value for the same value of **X**. For example, in Fig. 2, for $\mathbf{X} = 0.2$, there are two values of **Y** (points A and B). Due to noise and other modulating factors, most practical scatter plots are non-functional (Figs. 3 and 4).

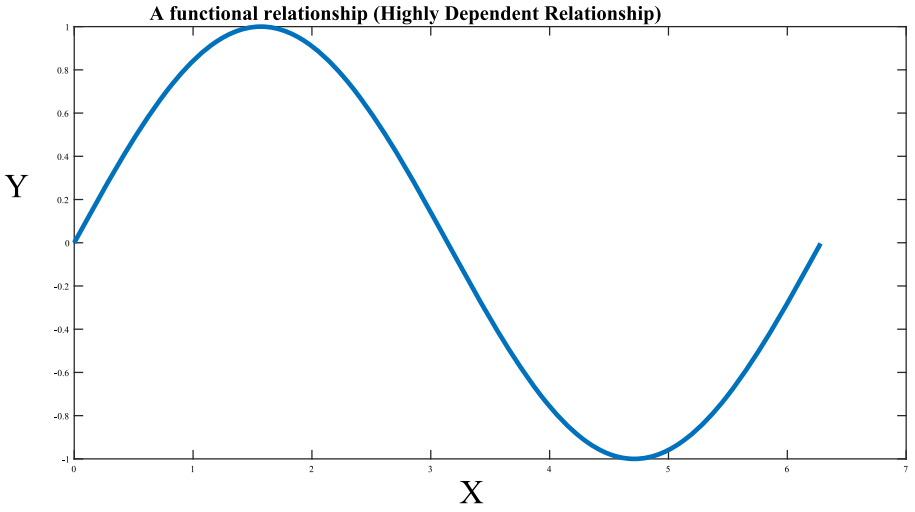

Fig. 1 A functional relationship

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

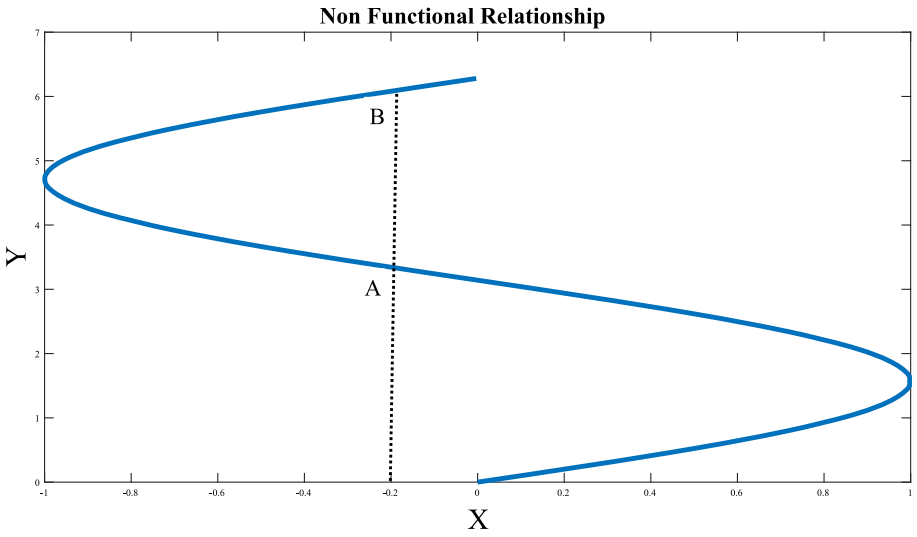

Fig. 2 A non-functional relationship, e.g., for $X = 0.2$, there are two values of Y (**A** and **B**)

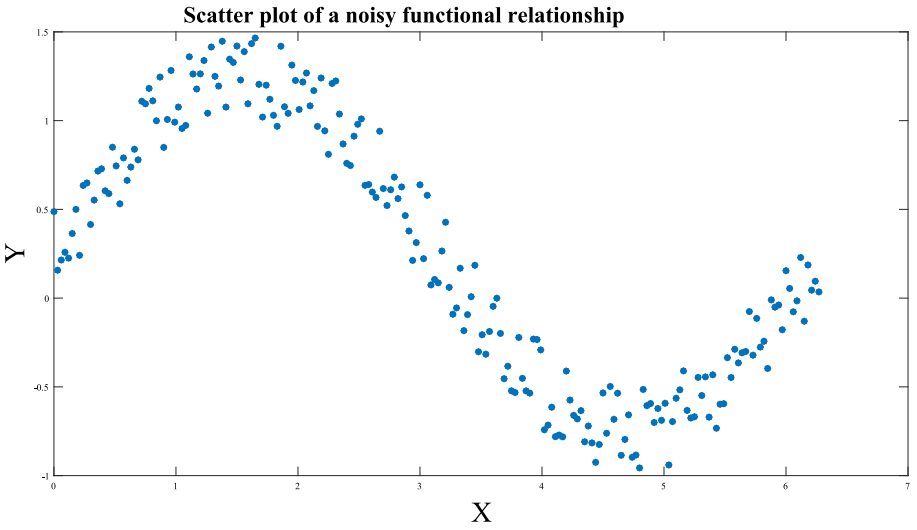

Fig. 3 Scatter plot of an ideal functional relationship that is changed to a nonfunctional due to noise or other modulating factors

Figure 4 shows that **X** cannot predict **Y** properly, but **Y** can predict **X** strongly. In other words, in Fig. 4, **X** is dependent on **Y** strongly. This procedure leads us to the concept of dependency based on a scatter plot.

If **X** and **Y** are mathematically independent, then we have:

$$P(\boldsymbol{Y}|\boldsymbol{X}) = P(\boldsymbol{Y}) \tag{2}$$

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

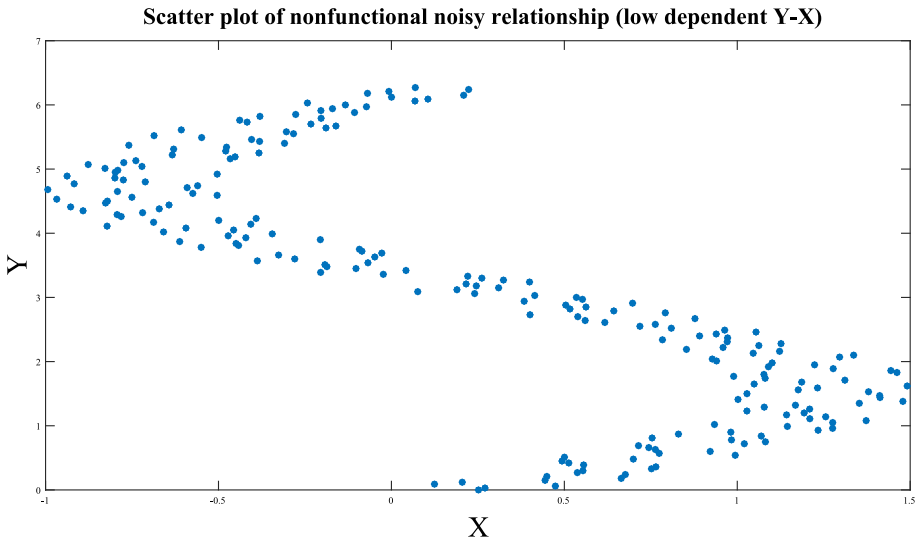

Fig. 4 Scatter plot of a non-functional, noisy relationship

where P is the probability distribution function and $P(\boldsymbol{Y}|\boldsymbol{X})$ is the (discrete) conditional probability distribution of $\mathbf{Y}$, given the nonzero probability $\mathbf{X}$.

Note that in this study, $\mathbf{X}$ and $\mathbf{Y}$ are not general random variables in a statistical sense. They are *sequences of observations* of a (random) process. For example, they may be samples of the height and weight of a population.

If $\mathbf{Y}$ and $\mathbf{X}$ are independent sequences, then we have:

$$|P(\boldsymbol{Y}) - P(\boldsymbol{Y}|\boldsymbol{X})| = 0 \tag{3}$$

1.1.3 Difference between dependence and correlation

Despite dependency, *correlation* only shows the proximity of scatter plot points to the interpolation curve. In other words, correlation indicates how small the variance is around the interpolation curve (principal curve). The interpolation curve of Figs. 4 and 5 is shown in Fig. 3, and the interpolation curve of Fig. 3 is shown in Fig. 1. The correlation between $\mathbf{X}$ and $\mathbf{Y}$ in Fig. 5 is greater than Fig. 4, but $\mathbf{Y}$ has a low dependence on $\mathbf{X}$.

If $\mathbf{X}$ can properly predict $\mathbf{Y}$, then there is both dependence and correlation between $\mathbf{Y}$ and $\mathbf{X}$. Pearson, Spearman, and Kendall coefficients can be used for linear, monotonical, and ranked correlations, respectively.

1.1.4 Method's overview

In the real world, we deal with scatter plots instead of continuous random variables. So, in this section, dependence measurement is investigated based on a scatter plot. First the principal idea is explained, and then the proposed algorithm is described.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

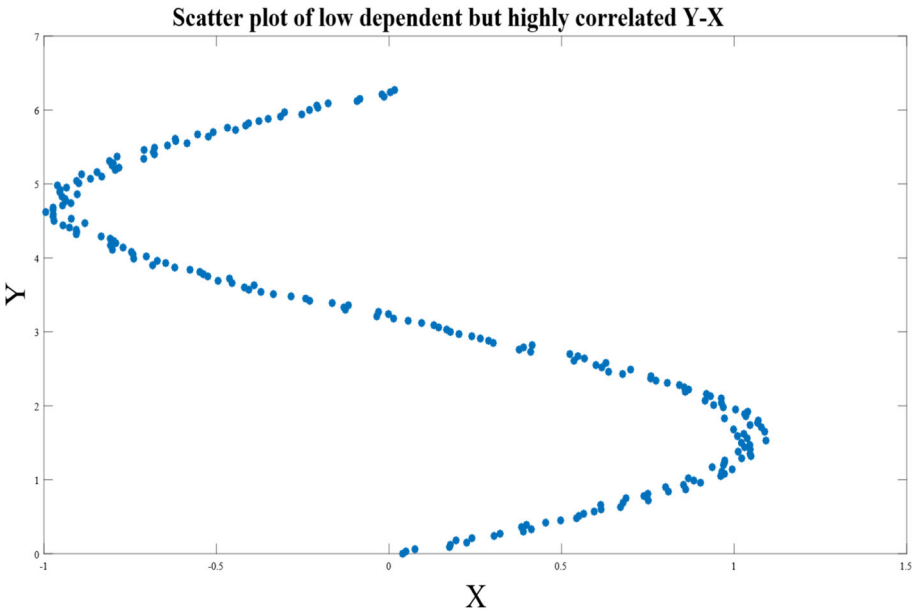

Fig. 5 Scatter plot of low dependent but highly correlated Y–X

2 Main idea and principal assumption

Regarding Eq. 3, the term $|P(\boldsymbol{Y}) - P(\boldsymbol{Y}|\boldsymbol{X})|$ can be defined as a measure of independence. This definition is our *principal assumption*:

$$\text{Independecy} \overset{\text{def}}{=} |P(\boldsymbol{Y}) - P(\boldsymbol{Y}|\boldsymbol{X})|$$

Simultaneously, a clear definition of independence can be presented by the following equation:

$$\text{Dependency} = 1 - \text{Independency}$$

Or:

$$\text{Dependecy} = 1 - |P(\boldsymbol{Y}) - P(\boldsymbol{Y}|\boldsymbol{X})| \tag{4}$$

Basically, to calculate dependency or independency over a scatter plot, a complex combination of Bayes rules and chain rules should be used, but in this paper, an intuitive measure is presented by a new method, named *area of successive triangles (AST)*.

Originally, $\begin{pmatrix} \boldsymbol{X} = x \\ \boldsymbol{Y} = y \end{pmatrix}$ pairs on a scatter plot do not necessarily indicate a functional relation. But if the random variable $\boldsymbol{Y}$ is dependent on $\boldsymbol{X}$, then there exists a regression function $f(\boldsymbol{X})$ that relates $\boldsymbol{Y}$ to $\boldsymbol{X}$. The aim of this article is not to find the specific regression function $f(x)$, but to estimate the probability of the existence of such a function [30]. In other words, the presented method in this paper measures the dependence of $\boldsymbol{Y}$ on $\boldsymbol{X}$.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

If we expand Eq. 4 for each point, after averaging all points, we have:

$$\text{Dependency} = 1 - \frac{\sum_{j=1}^{n}\sum_{i=1}^{n}\left|P(\boldsymbol{Y} = y_j) - P(\boldsymbol{Y} = y_j|\boldsymbol{X} = x_i)\right|}{n^2} \tag{5}$$

Suppose $D(i,\ j)$ as:

$$D(i,\ j) = \left|P(\boldsymbol{Y} = y_j) - P(\boldsymbol{Y} = y_j|\boldsymbol{X} = x_i)\right|$$

Then we have:

$$\text{Dependency} = 1 - \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} D(i,\ j)}{n^2} \tag{6}$$

Suppose there are just 3 points on a scatter plot, and suppose $A = (\mathrm{x}_1,\ y_1)$, $B = (\mathrm{x}_2,\ y_2)$ and $C = (\mathrm{x}_3,\ y_3)$ are three adjacent points; then, for $1 \le i \le 3$ and $1 \le i \le 3$, we can define $D(i,\ j)$ as the strength of dependency of point y_j to the point x_i as follows:

$$D(i,\ j) = \left|P(\boldsymbol{Y} = y_j) - P(\boldsymbol{Y} = j|\boldsymbol{X} = x_{\boldsymbol{i}})\right|$$

Generally, three points on the $\boldsymbol{X} - \boldsymbol{Y}$ plane can form a triangle. If the regression or interpolating function $f(x)$ exists and $f(x)$ is smooth enough, then the simplest relation between three points is a linear relation. It means, three points are located on a straight line, and the area formed by these three points is zero. A smaller area of the triangle indicates greater dependency. On the other hand, if $f(x)$ does not exist, or there is a significant noise on $f(x)$, then the area of the triangle formed by these three points is considerable and also greater than zero. In reality, due to noise, measuring errors, and the nature of $f(x)$, the three adjacent points are not distributed on a line.

In cases where a scatter plot (such as Fig. 6) consists of more than three points, a prior knowledge about the $\boldsymbol{X}$ and $\boldsymbol{Y}$ probability distribution, as well as their joint probability distribution is required.

The dependency due to the first three points or first triangle (D_1) is maximum when points A, B, and C are on a straight line; consequently, we have:

$$D_1 \propto 1 - A_{\triangle ABC} \tag{7}$$

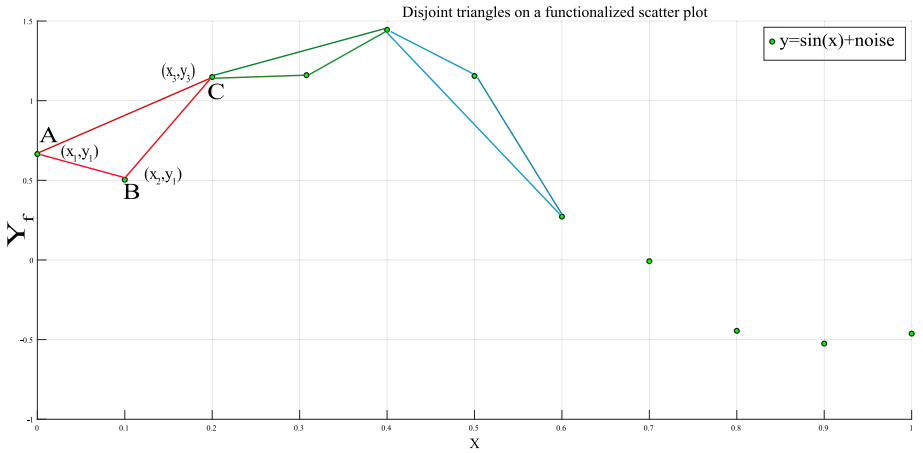

Fig. 6 Disjoint triangles on a functionalized scatter plot

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

where $A_{\Delta ABC}$ is the area of triangle ΔABC.

Considering other triangles, the proportional relation of Eq. 7 can be converted to the following equation:

$$\text{Dependency Approximation} = 1 - \lambda \sum_{i=1}^{N_t} A_{\Delta_i}$$

where A_{Δ_i} is the area of triangle i, λ is a constant, and N_t is the number of triangles.

Although in this method, we cannot have an exact value for λ, as shown in Eq. 8, we can normalize $\lambda \sum_{i=1}^{N_t} A_{\Delta_i}$ term by a similar value (term) of a random sequence. So, with a suitable approximation λ is crossed out from the numerator and denominator. The probability distribution of this random sequence is investigated in the AST of a random dataset section.

As shown in Fig. 6, the area between disjoint triangles is ignored. To overcome this issue, we use *Area of successive triangles* (AST) (Fig. 8), instead of area of disjoint triangles. In the following sections, AST is explained in detail.

3 Idea implementation

The proposed method includes the following two different sub-algorithms:

1 To measure dependency, a *functionalization* method was developed, which transforms the input scatter plot to a functional scatter plot.
2 The AST or "Area of Successive Triangles" method was developed to determine the randomness of a functional scatter plot.

To implement the above two different sub-methods, the following algorithm was designed. First, suppose $i = 0$.

A. *Functionalize* $\mathbf{Y}$ with respect to $\mathbf{X}$. Then replace $\boldsymbol{X} \leftarrow X_{\text{functionalized}}$ and

$$\boldsymbol{Y} \leftarrow \boldsymbol{Y}_{\text{functionalized}}$$

$$\boldsymbol{i} = \boldsymbol{i} + 1$$

The functionalization method is explained in the section "Functionalizing of a scatter plot."

B. Normalize $\mathbf{X}$ and $\mathbf{Y}$ between 0 and 1.
C. Measure the variations of $\mathbf{Y}$ caused by $\mathbf{X}$ using AST method and get the $AST_X(i)$.
D. Generate a random variable $\boldsymbol{X}_{\text{rnd}}$ with the same distribution as $\mathbf{X}$ to obtain $\boldsymbol{X}_{\text{rnd}} - Y$ scatter plot.
E. Measure the variations of $\mathbf{Y}$ caused by $\boldsymbol{X}_{\text{rnd}}$ by AST method and obtain $\text{AST}_{X_{\text{rnd}}}(i)$.
F. To reduce the functionalization side effects, permute (x_k, y_k) pairs to derive a new rearrangement of $\mathbf{X}$ and $\mathbf{Y}$ pairs (see example 1 for more details), where x_k is an element of X, and y_k is an element of Y. Obtaining these permutations is only necessary in cases where there are multiple y_k for a single x_k.
G. Repeat previous steps n times, to obtain other instances or values of AST_X and $\text{AST}_{X_{\text{rnd}}}$, where n is an arbitrary integer, representing the number of instances or runs. Larger n gives better results. This means that the algorithm is dependent on n. Notice that, here n is not the number of triangles.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

H. Replace AST_X and $AST_{X_{\text{rnd}}}$ by their average values:

$$\overline{\text{AST}}_X = \left(\sum_{i=1}^{i=n} \text{AST}_X(i)\right)/n$$

and:

$$\overline{\text{AST}}_{X_{\text{rnd}}} = \left(\sum_{i=1}^{i=n} \text{AST}_{X_{rnd}}(i)\right)/n$$

I. Divide AST_X by $\text{AST}_{X_{\text{rnd}}}$ to obtain a measure of normalized independence of **Y** from **X:**

$$\text{Normalized Independency} = \frac{\overline{\text{AST}}_X}{\overline{\text{AST}}_{X_{\text{rnd}}}} \tag{8}$$

The ratio of $\frac{\overline{\text{AST}}_X}{\overline{\text{AST}}_{X_{\text{rnd}}}}$ is a measure of both randomness and independence. A small ratio indicates low randomness. The term $\frac{1}{\overline{\text{AST}}_{X_{\text{rnd}}}}$ can be regarded as the constant of equation (λ) in the previous section.

J. *Numerical Dependency Analysis* (***NDA***) dependency coefficient can be calculated by the following equation.

$$\text{NDA} = 1 - \text{Normalized Independency}$$

Or:

$$\text{NDA} = 1 - \frac{\overline{\text{AST}}_X}{\overline{\text{AST}}_{X_{\text{rnd}}}}$$

NDA is asymmetric, i.e., $\text{NDA}_{XY} \neq \text{NDA}_{YX}$. NDA value is between − 1 and + 1. $NDA < 0$ shows no dependence, while values greater than 0 are indicative of dependency. Large NDA values indicate high dependence and also high correlation.

3.1 Functionalizing of a scatter plot

As mentioned earlier, in the real world, data are usually in the form of scatter plots, such as Figs. 3 and 4. It means that, for one **X** value there may be more than one **Y** value, which contradicts the concept of functionality. To convert (rearrange) a non-functional (***X***, ***Y***) scatter plot to a functional form, a method called *functionalizing by fixed intervals* was used. In the following subsections, functionalization procedure will be examined.

3.2 The functionalizing algorithm

In this section, the concept of functionalization and functional relations is further explained. In a functional scatter plot, there is only one y_i corresponding to each x_i. If the interval or step size between x_i and x_{i+1} is regarded as st_i, then we have:

$$\text{st}_i = x_{i+1} - x_i$$

In general, it is not necessary that the distances st_i to be equal. But in the functionalization method with fixed distances, after functionalization, these distances are considered the same.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Ranked (ordered) and categorical data usually have fixed intervals in the X axis, but scale data may have variable intervals.

To convert a non-functional scatter plot to a functional one, first a new arrangement of x_i should be generated. In this new arrangement there is just one y_i for each x_i. So, the aim of functionalization is to determine all x_i. During the functionalizing procedure, the corresponding values of x_i, *i.e.*, y_i remain unchanged, so the histogram of $\boldsymbol{Y}$ does not change. But, the histogram of $\mathbf{X}$ may change slightly. The following algorithm shows the functionalizing procedure.

1. Sort (x_i, y_i) pairs in ascending order based on the $\mathbf{X}$ axis. Suppose that the first point is $(x_1.y_1)$, and other points are $(x_2, y_2) \ldots (x_n, y_n)$. where n is the number of $\mathbf{X}$. After sorting we have:

$$x_1 \leq x_2 \ldots x_{n-1} \leq x_n$$

2. Calculate the step size between x_i and x_{n+1} by the following equation:

$$\text{st} = \frac{x_n - x_1}{n-1}$$

3. Suppose $i = 0$.
4. Increase the i value by one, $(i = i + 1)$.
5. Suppose X' is the horizontal axis of functionalized scatter plot and $x_i\prime$ is a member of X'. Calculate the $x_i\prime$ from the following equation:

$$x_i' = \text{st}(i-1) + x_1$$

6. If there is only one y_i value for x_i, keep this point i.e., (x_i, y_i) as a member of the functional scatter plot, and then go to step 4. Otherwise, follow the next steps.
7. Consider n_i as the number of (x_i, y_j) pairs at x_i, where $(1 \leq j \leq n_i)$. Redistribute all (x_i, y_j) pairs in the interval x_i' to $x_i' + st * n_i$ uniformly by the step size st in the same order as the original.
8. Repeat steps 4–7 for the next points (up to $i = n$).

Functionalization does not change $\mathbf{Y}$ values, only the $\mathbf{X}$ values. The minimum step size of $\mathbf{X}$ should be physically meaningful. For example, variations of less than 1 kg in the patient's weight can be neglected in many studies; therefore, the resolution of the step size should be selected larger than 1 kg. The following example shows the functionalization process.

Example 1 Assume numeric sequences as $\boldsymbol{X} = [1, 2, 2, 3]$ and $\boldsymbol{Y} = [1, 1, 2, 3]$ (Fig. 7).

As seen in Fig. 7, the scatter plot has 4 points. So, according to the second step of the functionalizing algorithm, we have:

$$\text{st} = \frac{(3-1)}{4-1} = \frac{2}{3}$$

For the first point on $\mathbf{X}$-axis i.e.,$x_1 = 1$ there is only one point, i.e., $y_0 = 1$. It means the first point of functionalized scatter is:

$$x_1' = \frac{2}{3}(1-1) + 1 = 1$$
$$(x_1', y_1') = (1, 1)$$

But, for $x_2 = 2$ there are two points (2, 1) and (2, 2), thus we can distribute points (2, 1) and (2, 2) with two permutations:

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

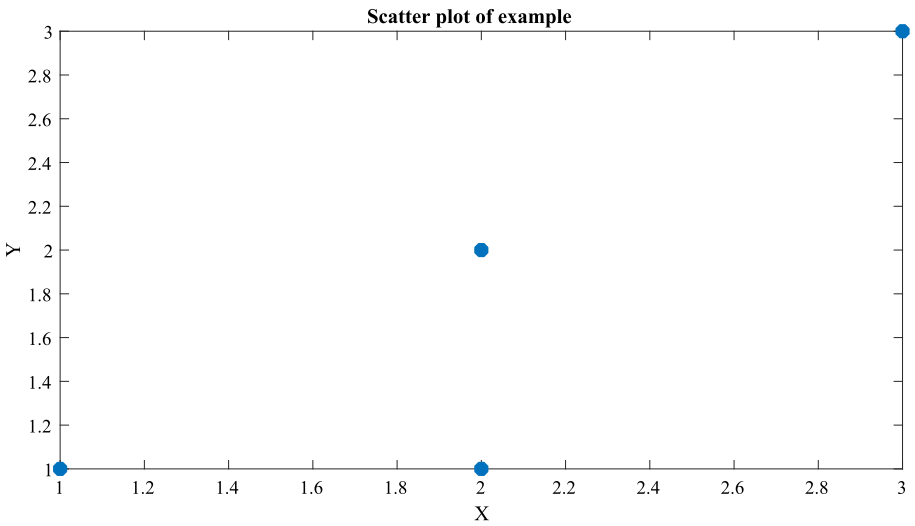

Fig. 7 Data of example 1

Permutation 1

$$x_2' = \frac{2}{3}(2-1)+1 = \frac{5}{3}$$
$$(x_2',\, y_2') = \left(\frac{5}{3},\, 2\right)$$
$$x_3' = \frac{2}{3}(3-1)+1 = \frac{7}{3}$$
$$(x_3',\, y_3') = \left(\frac{7}{3},\, 1\right)$$

Permutation 2
The second permutation is like first the permutation, and only the place of y_i are changed.

$$x_2' = \frac{2}{3}(2-1)+1 = \frac{5}{3}$$
$$(x_2',\, y_2') = \left(\frac{5}{3},\, 1\right)$$
$$x_3' = \frac{2}{3}(3-1)+1 = \frac{7}{3}$$
$$(x_3',\, y_3') = \left(\frac{7}{3},\, 2\right)$$

The last point at **X**-axis is $x_2 = (3,\, 3)$ which is a single point. So, the last point in the functionalized scatter plot is:

$$x_4' = \frac{2}{3}(4-1)+1 = 3$$
$$(x_3',\, y_3') = (3,\, 3)$$

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

The result of the functionalization algorithm can be either of two following permutations:

$$\boldsymbol{X}' = \left[1, \frac{5}{3}, \frac{7}{3}, 3\right] \text{ and } \boldsymbol{Y}' = [1, 2, 1, 3]$$
$$\boldsymbol{X}' = \left[1, \frac{5}{3}, \frac{7}{3}, 3\right] \text{ and } \boldsymbol{Y}' = [1, 1, 2, 3]$$

In scatter plots with more points, there may be more permutations.

4 Data normalization

In order to correctly compare the results of different data, data should be normalized between zero and one, so that:

$$0 \leq \boldsymbol{X} \leq 1 \text{ and } 0 \leq \boldsymbol{Y} \leq 1$$

Area of square between $0 \leq \boldsymbol{X} \leq 1$ and $0 \leq \boldsymbol{Y} \leq 1$ is one. As explained in the future sections, this value represents the probability of one.

4.1 Area of successive triangles (AST)

To measure the randomness of $\boldsymbol{Y}'$ variation due to the changes of $\boldsymbol{X}'$, the new concept of successive triangles was used in the $\boldsymbol{X}' - \boldsymbol{Y}'$ plane. Remember that $\boldsymbol{X}'$ data were sorted in the functionalization step, i.e., for any three subsequent $x_i{}'$ in $\boldsymbol{X}'$ we have:

$$x'_{i-1} < x'_i < x'_{i+1} \quad x'_{i-1}, x'_i, x'_{i+1} \in \boldsymbol{X}'$$

We draw triangles on the functionalized $\boldsymbol{X}' - \boldsymbol{Y}'$ scatter plot, so that each triangle consists of three successive points. Two successive triangles overlap at two points. Triangles show the magnitude of variation of $\mathbf{Y}$ due to (small) $\mathbf{X}$ variation.

It should be noted that, as shown in Fig. 6, if the triangles are considered separately, the area between the disjoint triangles is ignored in the calculations. To overcome this problem, instead of the area of disjoint triangles, the Area of Successive Triangles (AST) is used (Fig. 8). Figure 8 shows a functional curve, where points A, B, C, and D are four successive points on this curve. Although this figure does not show a scatter plot, which has been mentioned so far, the curve is the limit state of a scatter plot. To measure the randomness of variations of $\mathbf{Y}$ due to changes of $\mathbf{X}$, successive triangles were plotted using the following steps:

1- Start from the first point on the curve (A). It is the first point of the first triangle. The next two points are the other vertices of this triangle (ΔABC).
2- Continue with the next triangle, which starts with the second point (B) of the first triangle (ΔABC).
3- If there is an intersection between the first and second triangles, there will be a third triangle (ΔBEC).
4- Repeat the above steps for other points.
5- The total area of all these successive triangles is a measure of curve randomness or dependency. We name it the *Area of Successive Triangles*, or AST.

Example 2 Calculate the AST of Fig. 8.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

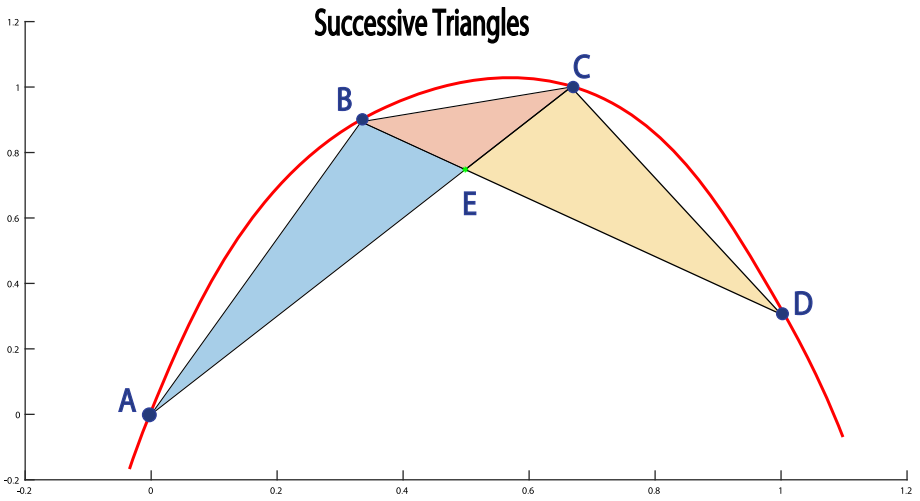

Fig. 8 Successive triangles on a curvature

In this example, the AST of Fig. 8 is:

$$AST = A_{\Delta ABE} + A_{\Delta BEC} + A_{\Delta ECD}$$

where $A_{\Delta ABE}$ denotes the area of triangle ΔABE and so on. A low AST indicates low randomness and high dependency.

Example 3 Figure 9 depicts successive triangles of $\begin{pmatrix} X \\ \sin(X) \end{pmatrix}$, and Fig. 10 shows successive triangles of $\begin{pmatrix} \sin(Y) \\ Y \end{pmatrix}$.

4.2 AST of a random dataset

AST itself is not a good measure of independence or randomness, due to its bias. But why is there a bias in AST? This bias is because **X** and **Y** can be of different data types, i.e., categorical, ordinal, or numeric with different lengths. Our simulations show that the AST of random data depends on three factors: *data type, data length,* and *data distribution*. In the Simulation Results section, the answer of mentioned question is explained in detail through simulated data. What happens if **Y** is completely independent of **X**?

To answer this question, a random or pseudorandom $\boldsymbol{X}_{\text{rand}}$ variable was generated, so that its size was equal to the size of **X**. Subsequently, a new scatter plot of $\boldsymbol{X}_{\text{rand}} - Y$ is obtained, and the AST is calculated over the $\boldsymbol{X}_{\text{rand}} - Y$ scatter plot. $\boldsymbol{X}_{\text{rand}} - Y$ scatter plot indicates that there is no relation between $\boldsymbol{X}_{\text{rand}}$ and **Y**, making AST to be a relatively large value.

Regarding the type of **Y** data, the AST of $\boldsymbol{X}_{\text{rand}} - Y$ scatter plot has a typical value or a bias value. Simulations show that the AST bias is highly independent of the size and type of **X**, but AST is strongly dependent on **Y** levels. In other words, the number of **Y** levels imposes a bias on AST. For example, if **Y** has only two levels, e.g., positive or negative, the

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

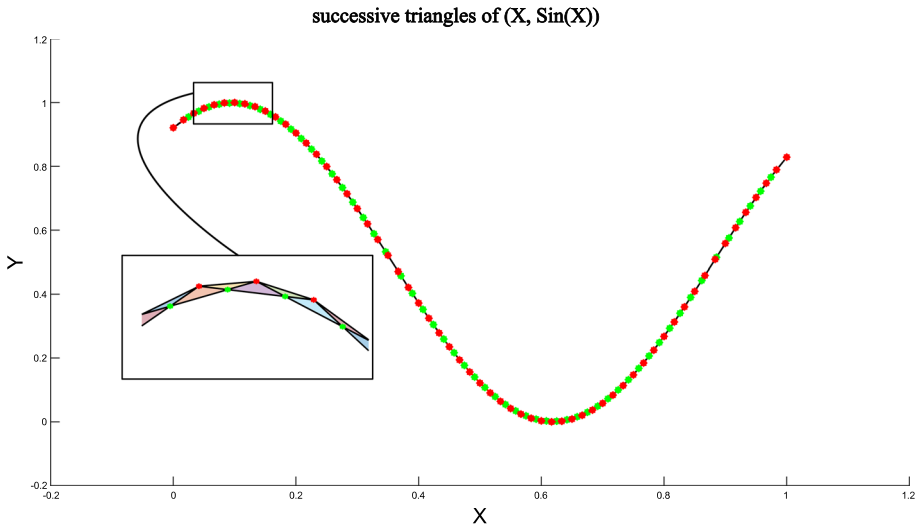

Fig. 9 Successive triangles of sin(X) and a magnified section show successive triangles. This figure shows a small Area of Successive Triangles (AST) for a functional relationship

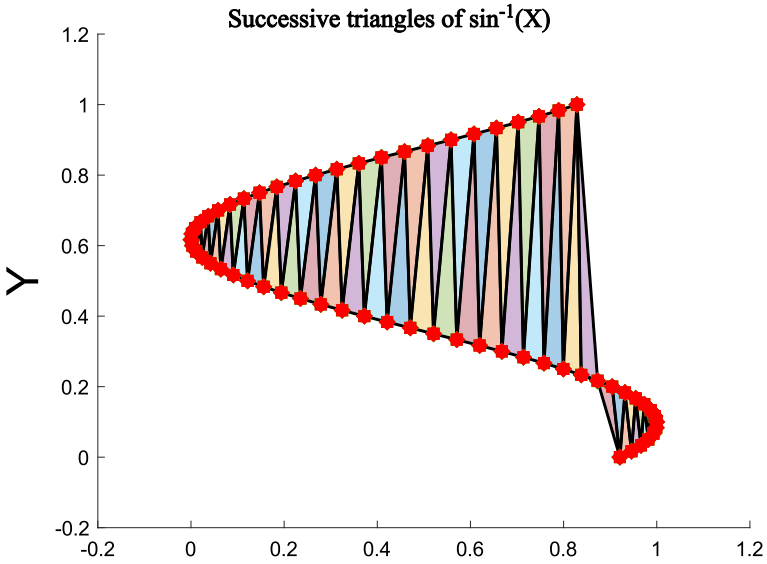

Fig. 10 Successive triangles of sin(Y). This figure shows a large AST for a non-functional relationship

AST is naturally large, but if **Y** is numeric, e.g., integers 1 to 100, the AST is naturally small. Figure 11 shows that regarding $\boldsymbol{X}_{\text{rand}} - Y$ scatter plot, increasing the number of **Y** levels leads to a decrease in AST. Clearly, the number of **Y** levels is independent of sample size or **Y** length.

Another parameter that affects the value of AST is **X** distribution. For example, a uniform distribution of $\boldsymbol{X}_{\text{rnd}}$ has less AST than a normal distribution. Therefore, we select $\boldsymbol{X}_{rnd}$ with

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

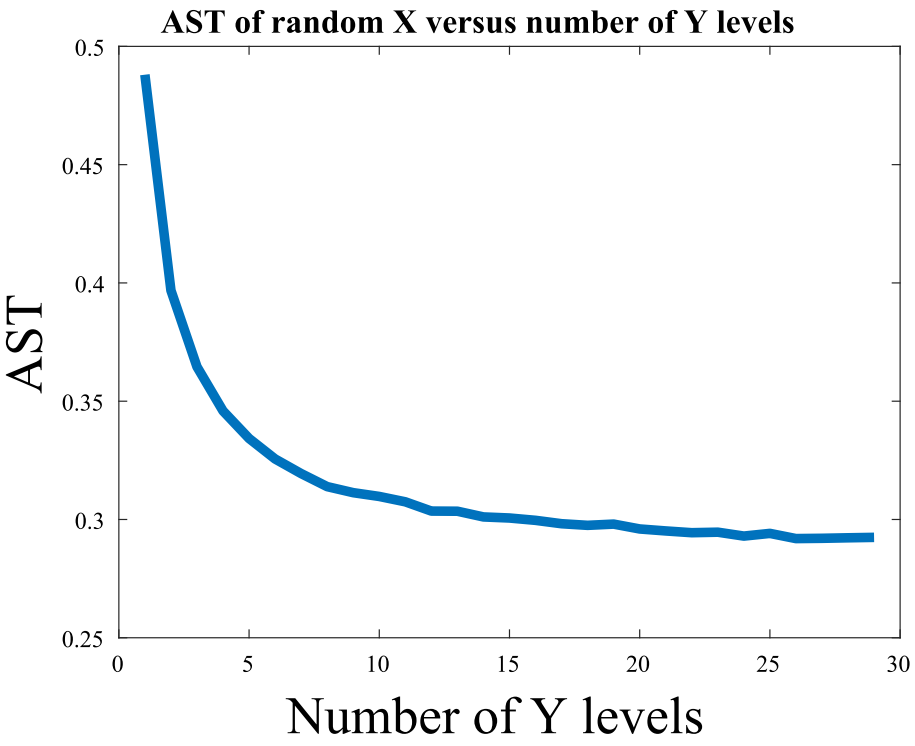

Fig. 11 AST of $X_{\text{rand}} - Y$ versus the number of **Y** level. **X** values are integers between 1 and 100

the same distribution (identically distributed) as **X**. So, in order to compare different AST(s), in step I of the *Idea implementation* section, we should normalize AST of **X** by AST of $\boldsymbol{X}_{\text{rnd}}$ (or $AST_{X_{\text{rnd}}}$).

4.3 Robustness and criteria NDA

NDA originally is sensitive to smoothness of variations; therefore, it is important to be the variations of the regression function $f(x)$ less than the variation of noise.

Suppose $f(x)$ is a continues function, then we have:

$$F(\omega) = FT(f(x))$$

where

$FT(f(x))$ is Fourier transform of functionalized $f(x)$, and $||\ \ ||$ denotes the norm operator.

Suppose $\omega_{\max}$ is the maximum meaningful angular frequency that exists in $F(\omega)$. Then NDA has a good result if there are enough samples per maximum $\omega_{\max}$. In other words, there must be enough samples per $T_{\min} = \frac{2\pi}{\omega_{\max}}$. It seems about 10 samples per a period of $T_{\min}$ may be sufficient. It means that the proportion of the population that is sampled should be large enough to represent the regression function $f(x)$.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 1 Interpreting of NDA

From	To	Interpretation (NDA)
0.5	1	Very high dependency
0.3	0.5	High dependency
0.15	0.3	Moderate dependency
0.02	0.15	Low dependency
0	0.02	Very weak dependency
- 1	0	No dependency

For periodic dependencies (functions) such as $\sin(\omega x)$ there must be enough samples in each period to obtain a good NDA result; otherwise, NDA has misleading results, because NDA is basically sensitive to the smoothness of the regression function.

On the other hand, the variations of $f(x)$ should be small enough. As mentioned earlier, due to the sampling effect, measuring error or the effect of other variables, $f(x)$ is noisy. The ratio $\frac{\text{variance of noise on } f(x)}{\text{variance of } f(x)}$ should be small enough in order to obtain a correct NDA.

5 Results

To validate the developed method, first, simulated data were used, i.e., linear, sinusoidal, nominal, and piecewise function simulated data. Then NDA was applied to the clinical data. Finally, as an application, NDA was used as a method for template matching.

5.1 Interpreting NDA

NDA is not a correlation coefficient [93], therefore, the interpretation of NDA differs from common correlation coefficients. The NDA coefficient can be interpreted as shown in Table 1.

As it can be concluded from Table 1, the NDA values of approximately 0.15 indicate moderate dependence. Values 0.3 to 0.5 show high dependence, and values greater than 0.5 indicates strong dependence.

Of course, NDA is asymmetric; for example, in Figs. 9 and 10, we have $NDA_{XY} = 0.99$, $NDA_{YX} = -.38,$, respectively. These two figures are similar, and only the axes of $\mathbf{X}$ and $\mathbf{Y}$ are displaced.

5.2 Simulation results

There are some methods to simulate dependent variables [84, 85]. To assess the proposed method, three types of linear, nonlinear, and nominal simulated variables were generated by adding noise to the functional relations.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

5.3 Linear dependence of simulated data

To validate the proposed method with linear simulated data, the **X** variable was generated with the following equation:

$$\boldsymbol{X} = \{\boldsymbol{t} | 0 \leq \boldsymbol{t} \leq 50, \boldsymbol{t} \in \mathbb{R}\}$$
$$\boldsymbol{Y} = \boldsymbol{X} + A * \left(\hat{U}(0,\ 1) - 0.5\right) \tag{9}$$

For simplicity, $\boldsymbol{X}$ was not defined as a random variable, so it has a strict definition (Eq. 9). In Eq. 9, $\hat{U}(0,\ 1)$ represents the uniformly distributed random in the interval (0, 1). and A is a coefficient that determines the noise level or noise amplitude. Figure 12 shows an example of a linear relation between X and Y, as well as the NDA and correlation coefficients for this scatter plot.

Figure 13 shows changes in the NDA, Pearson, Spearman, and Kendall coefficients as the noise amplitude of Eq. 9 increased. NDA is more sensitive to noise, because noise reduces dependency.

It is worth mentioning that although the NDA coefficient measures dependence and is different from other correlation coefficients, the NDA coefficient can also be used to measure correlation, because variables that are dependent are also correlated.

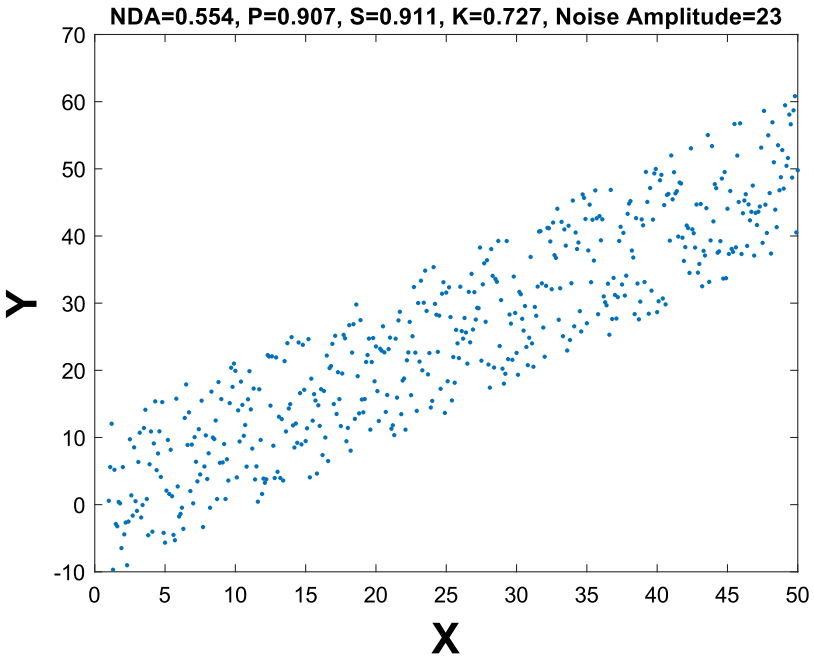

Fig. 12 A noisy linear relation between X and Y variables

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

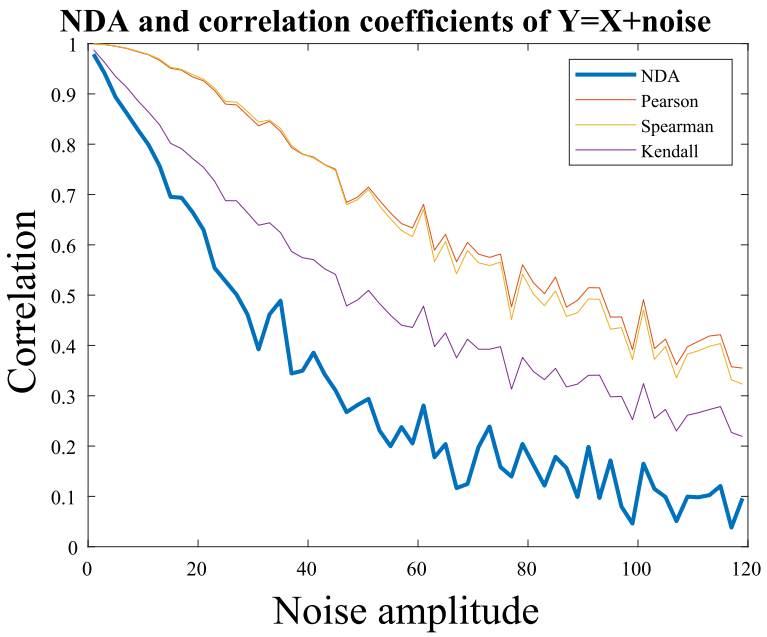

Fig. 13 Correlation of $Y = X$ versus noise level

5.4 Sinusoidal dependence of simulated data

Figure 14 shows a typical sinusoidal scatter plot of:

$$\boldsymbol{X} = \{\boldsymbol{t} | 0 \leq \boldsymbol{t} \leq 50, \boldsymbol{t} \in \mathbb{R}\}$$

$$\boldsymbol{Y} = \sin(\boldsymbol{X}/5) + A * \hat{U}(0, 1)$$

As it can be seen from the title of Fig. 14, Pearson, Kendall, and Spearman's coefficients are very small (near zero), while NDA exhibits significant dependency of 0.22.

According to Fig. 15, the Pearson, Spearman, and Kendall coefficients do not change significantly by the increase of the noise level in Fig. 14. But NDA shows a rapid decrease by increasing the noise level.

5.5 Nominal dependence of simulated data

NDA considers nominal data as numerical variables.

$$\boldsymbol{X} = \{\boldsymbol{t} | 0 \leq \boldsymbol{t} \leq 50, \boldsymbol{t} \in \mathbb{R}\}$$

$$\boldsymbol{Y} = \sin(\boldsymbol{X}/5) + A * \hat{U}(0, 1)$$

where $\lfloor \rfloor$ denotes the floor function.

Figure 16 shows a typical four-level nominal data with noise (Note that the data are nominal and not discrete.). The solid lines are guidelines that show the levels of nominal data. As

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

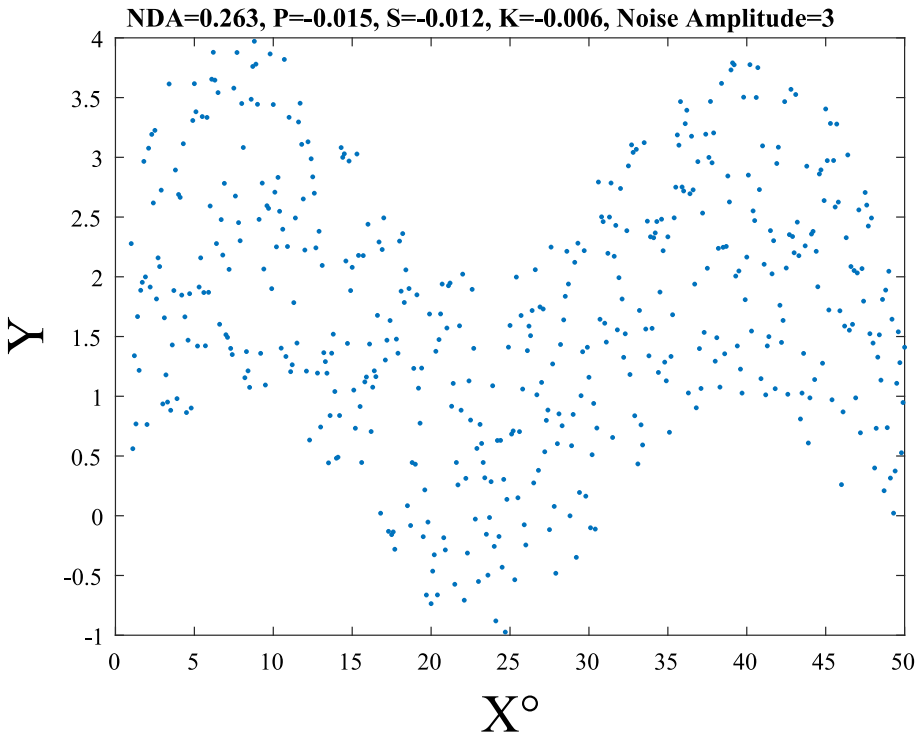

Fig. 14 A typical $Y = \sin(X/5)$ + noise scatter plot. NDA and correlation coefficients are shown in the figure title. The unit of X-axis is in degrees

shown on the top of Fig. 16, the values of Pearson, Kendall, and Spearman coefficients are small, while the value of NDA is a suitable value, i.e., 0.317, according to Table 1.

Figure 17 shows that the NDA coefficient decreases as noise level increases, while the Pearson, Spearman, and Kendall coefficients do not change significantly.

5.6 Nonlinear dependence of simulated data

Figure 18 shows a typical nonlinear scatter plot of:

$$\boldsymbol{X} = \{\ \sin(\boldsymbol{t}/5) | 0 \leq \boldsymbol{t} \leq 50,\ \boldsymbol{t} \in \mathbb{R}\}$$

$$\boldsymbol{Y} = \{\ \boldsymbol{t} + A * \hat{U}(0,\ 1) | 0 \leq \boldsymbol{t} \leq 50,\ \boldsymbol{t} \in \mathbb{R}\}$$

According to Fig. 19 the Pearson, Spearman, Kendall coefficients, and NDA are low.

5.7 Experimental results

5.7.1 In-hospital mortality patients with ST-segment-elevation myocardial infarction

To validate the proposed method by clinical data, a freely available data on ST-Segment-Elevation Myocardial Infarction (STEMI) was used [94]. These data are based on a cohort

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

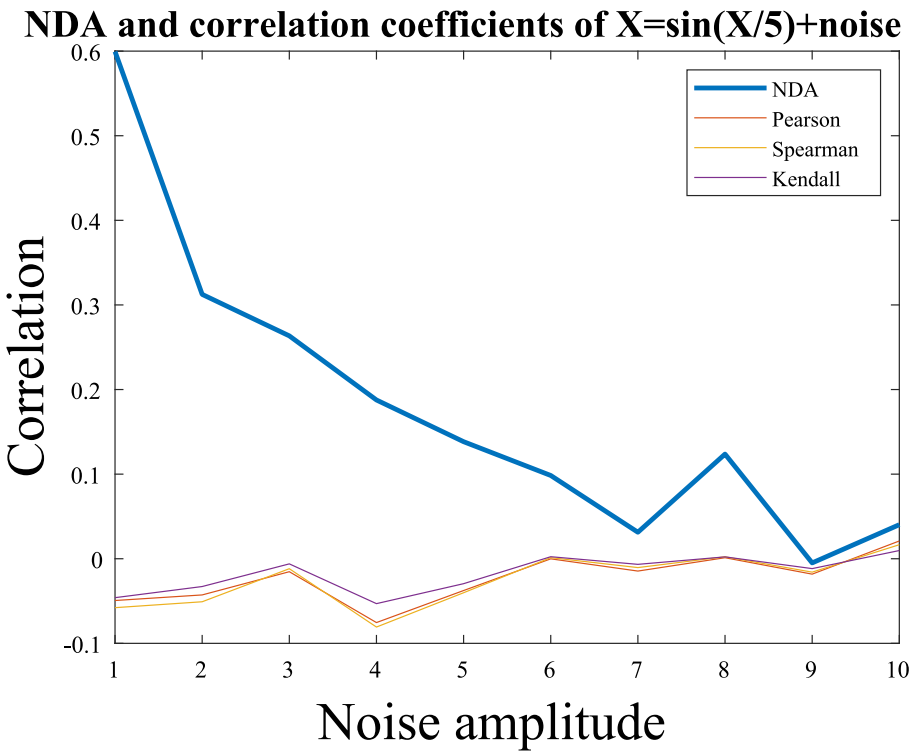

Fig. 15 Correlation of $Y = \sin(X/5)$ + noise versus noise level

study of 2,816 patients with STEMI from Imam Ali Hospital at the Kermanshah University of Medical Sciences, from July 2017 to May 2020. Briefly. There are 55 independent variables, which is a large number for prediction purposes. So, to find important features, and to measure the effect of independent variables on patient outcome or *In-Hospital Mortality*, NDA was used. In addition, the NDA results were compared to Spearman, Kendall, and Pearson coefficients (Table 2). As seen from Table 2, there are some differences between NDA results and correlation methods.

As it can be seen from Table 2, the *worst Killip* class is highly dependent on In-Hospital Mortality. The Killip classification system itself is a valid tool that classifies the mortality risk of patients with acute myocardial, so it is related to In-Hospital Mortality (especially the worst Killip class) [95]. The other important variable is LDL, which is a good predictor of cardiac dysfunction [96].

Using a Directed Acyclic Graph (DAG), the NDA coefficient matrix can be visualized [97, 98]. In this example, the NDA coefficient is a high dimensional DAG [99]. Considering NDA values greater than 0.4, the DAG around In-Hospital Mortality node is shown in Fig. 20.

If variables with NDA coefficients of 0.029 or higher from Table 2 are selected as important features, we have Table 3 using a Neural Network (NN) model with Monte Carlo cross-validation for 20 times.

As shown in Table 3, the area under the curve (AUC) is 0.95. AUC shows the prediction performance of the model. This good AUC, compared to similar studies [100], shows the

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

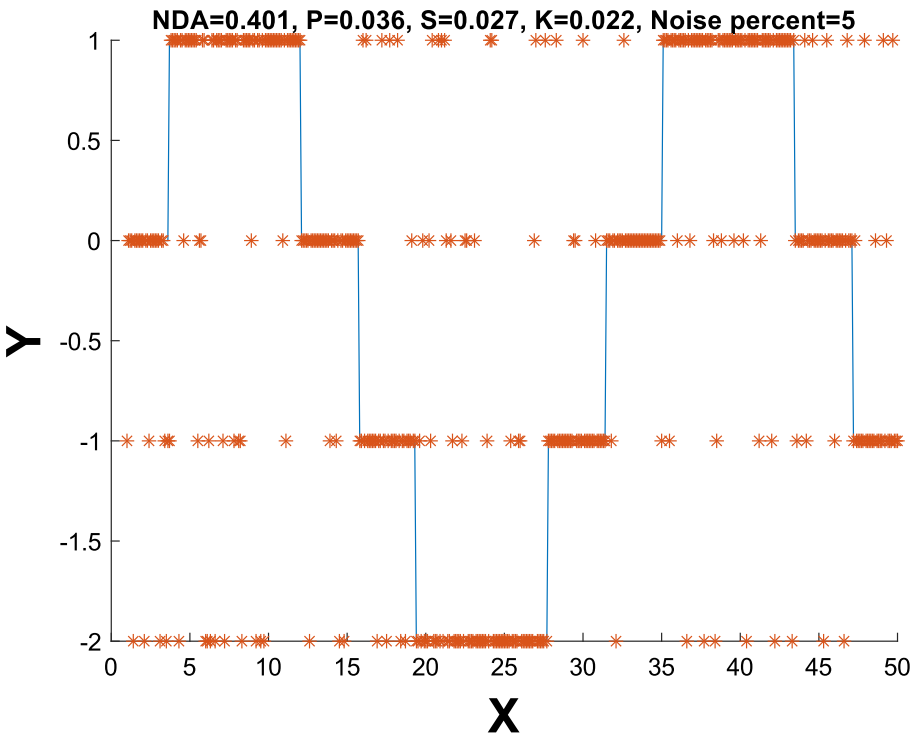

Fig. 16 Discrete (nominal) function with 4-levels

important features selected by NDA can properly predict the target variable (In-Hospital Mortality) together with neural networks.

5.7.2 Template matching example

NDA can have other applications like pattern template matching. Here is an experimental example of this application. In a study on rapid detection of clinical bacteria samples, we had to compare the pattern of the clinical samples with the pattern of standard bacteria. Figure 21 shows four standard bacteria patterns. Bacteria patterns show changes in the fluorescence intensity versus bacteria concentration. Patterns of clinical strains have similar profiles [101]. Methods such as mean square error (MSE) can be used to compare these patterns; but MSE is not a suitable measure, because the bacteria patterns are very similar to each other. In addition, there may be a bias between two patterns (as shown in Fig. 21), which leads to a bias in the MSE.

In this case, we need a good method to measure the matching or similarity of two patterns. If two patterns (clinical and standard pattern) are of the same bacteria, then the profile of one pattern is dependent on another profile. Simply speaking, one profile goes up and down by another profile. So, we plotted the intensity of one profile versus the intensity of another profile for each concentration. Ideally, the clinical pattern must obey the standard pattern, so the points of this scatter plot must be distributed around the $y = x + b$ regression line, where b is the bias between two patterns, which is related to the initial conditions of the experiment.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

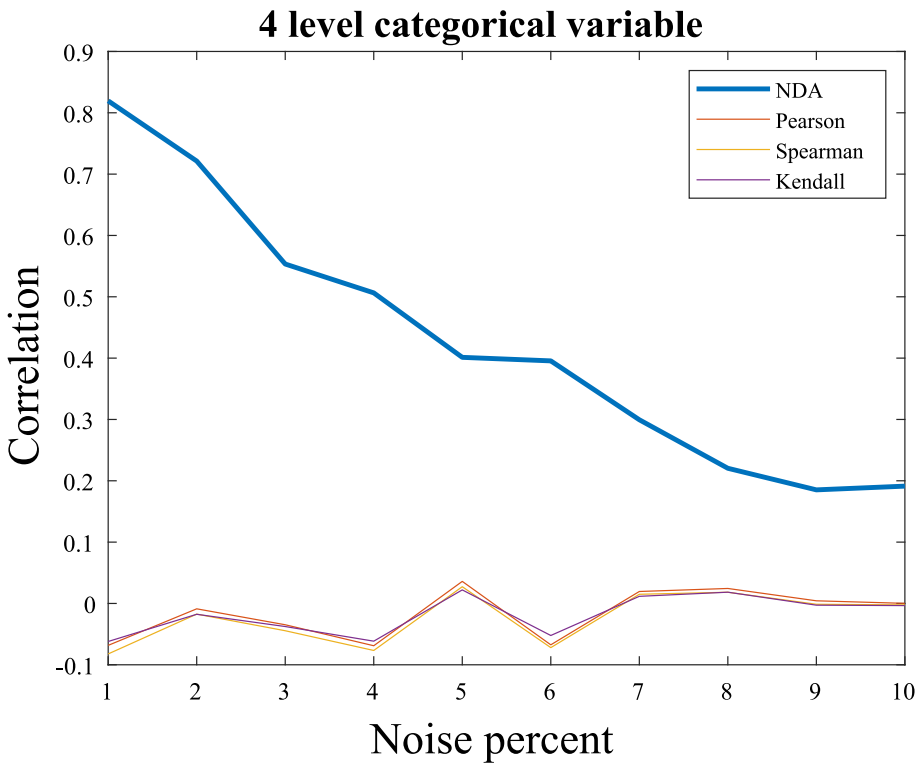

Fig. 17 Correlation changes versus noise of a discrete (nominal) 4-level function

Table 4 shows the NDA value between different clinical samples and standard patterns.

As shown in Table 4, the diagonal elements of Table 4 are greater than other elements in each row, so NDA can properly measure the similarity or matching of two patterns.

6 Discussions

Due to promising simulation and experimental results of NDA, it is strongly recommended to use NDA as a measure of dependence. Although there are similar attempts to measure dependence in the literature [50–52, 68, 102], the use of functionalizing technique and AST measure distinguishes NDA from other studies.

Although the NDA coefficient matrix reveals relations of a causal reasoning model, it would be more beneficial to plot the DAG graph of the NDA coefficient. In fact, DAG is a causal graph. This would help identify common-cause, common-effect, causal chains, or causal homeostasis relationships, using graph mining techniques or association rule mining methods [88, 103–109]. Subgraphs of DAG depict independent subsets of variables. DAG-related algorithms can be used for more studies [110].

The NDA method does not require prior knowledge about the data distribution, so it is distribution-free. As it was examined in the results section, NDA can be used for linear, nonlinear, numerical, and nominal data. Figures 14 and 15 show the common correlation

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

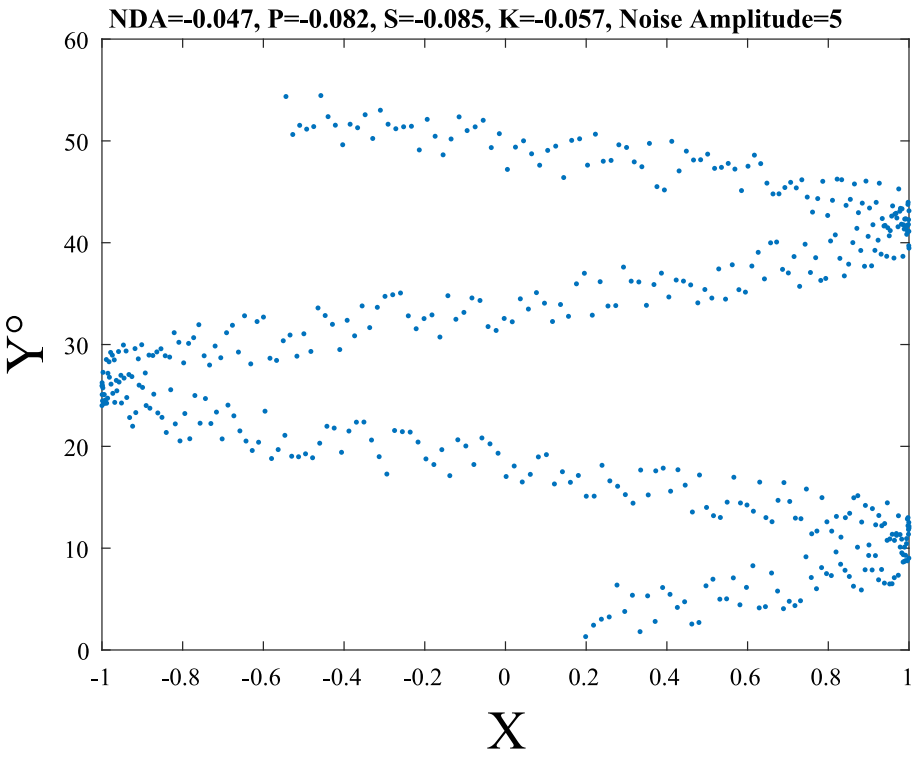

Fig. 18 A typical $Y = \sin^{-1}(X/5)$ + noise scatter plot. NDA and correlation coefficients are shown in the figure title. The unit of the Y-axis is in degrees

coefficients and NDA for sinusoidal function(s), whereas Figs. 18 and 19 show these values for the inverse sine function. Comparing these figures reveals that, in contrast to NDA, common correlation coefficients have similar results for a functional relation and its invers function. So, in contrast to NDA, Pearson, Spearman, and Kendall correlation coefficients can only measure correlation, especially for linear or monotonic functions. However, NDA can also be used to measure correlation using the formula:

$$\text{Correlation}(\boldsymbol{X}, \boldsymbol{Y}) = \max\{NDA(\boldsymbol{X}, \boldsymbol{Y}) + NDA(\boldsymbol{Y}, \boldsymbol{X})\}$$

Currently, NDA is a bivariate method, and the effect of other variables on the dependent variable appears as noise, but it can be further extended for multivariable processes by integrating neural networks and NDA. Neural networks and also regression methods can extract the optimum equation between each independent variable and dependent variable, and then NDA can measure the effect of each independent variable on the dependent variable to find the important feature or features.

In this study, the NDA value is dependent on the number of runs (step G in Method's overview). In future studies, NDA can be made independent of the number of runs, which makes NDA faster. Depending on the type of data (categorical or numerical), NDA produces an acceptable result with a few tens of samples.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

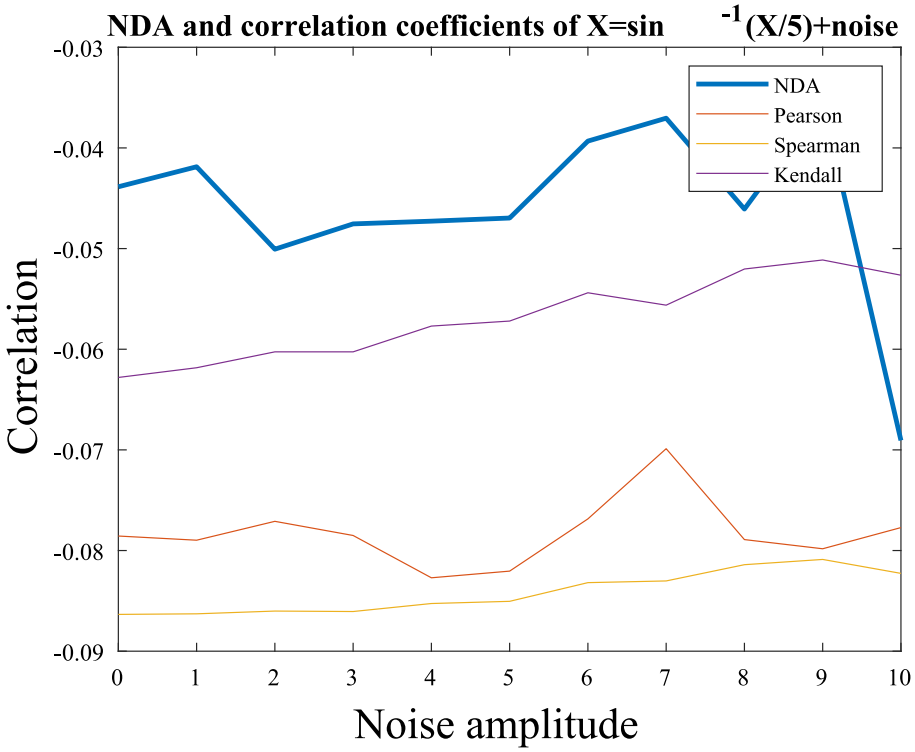

Fig. 19 Correlation of $Y = \sin^{-1}(X/5)$ + noise versus noise level

Using NDA, a good estimation of the minimal important change difference (MICD) [111, 112], the minimal important change (MIC) [113], the dependence between two signals [69, 114], and the entropy of a signal [102, 115, 116] can be obtained.

The simulation results and NDA results with real-world data in the results section show that NDA is robust to the noise. However, NDA is sensitive to the smoothness of the regression function, so there must be enough samples to get a good NDA result; otherwise, NDA will give misleading results.

While NDA provides a local measure of dependence of three successive points, it can be developed to measure global dependence by increasing the intervals between successive points of triangles [117, 118].

NDA still requires improvement; so, any modifications both to the method and MATLAB codes are welcomed.

7 Conclusions

In this research paper, we have introduced a groundbreaking method known as numerical dependency analysis (NDA), which opens up new possibilities in the field of statistical dependency. This innovative approach was designed to quantify the dependence of one variable on another. The novelty of this paper lies in the AST and functionalizing technique, which are

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 2 Comparison of NDA and correlation coefficients of In-Hospital Mortality. For more details about the variables in this table, refer to the related study [94]

Number	Feature	NDA	Pearson	Spearman	Kendall
1	Worst.Killip.class	0.788	0.626	0.494	0.476
2	LDH	0.592	0.559	0.336	0.276
3	Highest.Cr (Creatinine)	0.286	0.315	0.254	0.221
4	First.EF	0.275	− 0.291	− 0.259	− 0.226
5	HF.in.hospital	0.248	0.298	0.298	0.298
6	Last.EF	0.244	− 0.276	− 0.248	− 0.217
7	Killip.class	0.207	0.378	0.331	0.326
8	CCS.class	0.175	0.235	0.206	0.204
9	CLCR (Creatinine Clearance)	0.171	− 0.197	− 0.214	− 0.175
10	SBP	0.157	− 0.195	− 0.179	− 0.152
11	Early.Cr	0.149	0.260	0.193	0.169
12	PIC	0.129	− 0.234	− 0.234	− 0.234
13	Highest.CKMB	0.123	0.018	− 0.039	− 0.032
14	Age	0.098	0.152	0.151	0.123
15	Staged.PCI	0.090	− 0.116	− 0.120	− 0.119
16	HR.on.qualifying.EKG	0.084	0.065	0.060	0.050
17	Early.Hb	0.063	− 0.129	− 0.112	− 0.093
18	BMI	0.060	− 0.022	− 0.030	− 0.025
19	AF.in.hospital	0.056	0.179	0.179	0.179
20	Patient delay	0.053	− 0.006	0.009	0.007
21	Current.smoking	0.051	− 0.023	− 0.039	− 0.039
22	Stent	0.048	− 0.135	− 0.135	− 0.135
23	TIMI.flow.after.dilatation	0.048	− 0.142	− 0.108	− 0.107
24	Highest.Troponin	0.043	0.031	− 0.004	− 0.003
25	Highest.CPK	0.042	0.158	0.129	0.105
26	HDL	0.041	− 0.019	− 0.037	− 0.030
27	No.territories	0.040	0.105	0.106	0.099
28	Lowest.Hb	0.037	− 0.117	− 0.103	− 0.085
29	PLT	0.032	− 0.033	− 0.037	− 0.030
30	LDL	0.029	− 0.031	− 0.038	− 0.032
31	Previous.stroke	0.025	0.106	0.098	0.098
32	AF.on.qualifying.EKG	0.020	0.105	0.105	0.105
33	Diabetes	0.018	0.058	0.065	0.065
34	TC	0.018	− 0.030	− 0.030	− 0.025
35	Gender	0.018	0.088	0.088	0.088
36	First.CPK	0.018	0.096	0.063	0.051
37	HLP	0.017	0.004	− 0.008	− 0.008

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 2 (continued)

Number	Feature	NDA	Pearson	Spearman	Kendall
38	TIMI.flow.before.dilatation	0.016	− 0.058	− 0.077	− 0.075
39	First.CKMB	0.014	0.089	0.082	0.069
40	Previous.AF	0.014	0.074	0.074	0.074
41	HTN	0.012	0.070	0.073	0.073
42	LMS.stenosis	0.012	0.114	0.114	0.114
43	Dialysis	0.011	0.079	0.084	0.084
44	System delay	0.010	− 0.008	0.058	0.052
45	Bleeding	0.009	0.008	0.024	0.024
46	Previous.PCI	0.007	0.057	0.044	0.044
47	CHF	0.006	0.065	0.057	0.056
48	Previous.PVD	0.005	0.091	0.081	0.081
49	Thrombectomy.duration.PCI	0.005	0.053	0.053	0.053
50	Old.MI	0.003	0.025	0.028	0.027
51	ESR	0.000	0.012	− 0.003	− 0.003
52	Transfusion	− 0.001	0.045	0.045	0.045
53	Previous.CABG	− 0.008	0.028	0.020	0.020
54	First.Troponin	− 0.012	− 0.029	− 0.025	− 0.025
55	TG	− 0.027	0.004	− 0.013	− 0.011

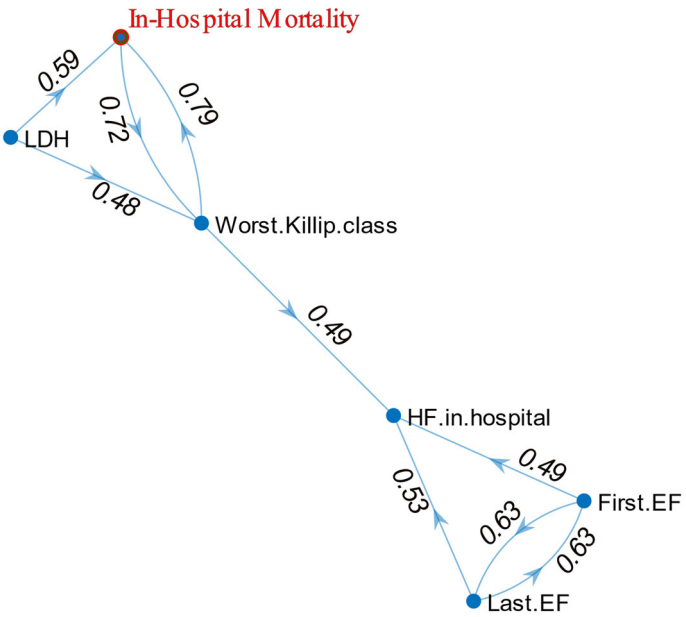

Fig. 20 DAG around In-Hospital Mortality. Red values are the top two NDA values in Table 2

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 3 Neural network results using Monte Carlo cross-validation. Values in each row are the average of 3 different runs

Test number	Specificity	Sensitivity	Accuracy	AUC	STD	MAE
1	0.971	0.941	0.956	0.966	0.116	0.075
2	0.941	0.941	0.941	0.947	0.122	0.079
3	1.000	0.882	0.941	0.928	0.110	0.067
4	0.882	0.971	0.926	0.943	0.142	0.106
5	0.941	1.000	0.971	0.968	0.186	0.150
6	0.971	1.000	0.985	0.971	0.148	0.102
7	0.971	0.882	0.926	0.926	0.164	0.110
8	0.912	0.971	0.941	0.931	0.126	0.072
9	0.941	0.941	0.941	0.924	0.144	0.039
10	0.971	0.882	0.926	0.954	0.126	0.068
11	0.941	0.971	0.956	0.943	0.080	0.041
12	1.000	0.941	0.971	0.962	0.134	0.103
13	0.941	0.971	0.956	0.965	0.168	0.125
14	1.000	1.000	1.000	0.971	0.087	0.058
15	0.971	0.912	0.941	0.966	0.168	0.111
16	0.941	0.971	0.956	0.959	0.157	0.107
17	0.941	1.000	0.971	0.969	0.119	0.066
18	0.941	1.000	0.971	0.964	0.144	0.108
19	1.000	0.912	0.956	0.965	0.113	0.080
20	0.912	0.941	0.926	0.933	0.104	0.053
Mean	0.954	0.952	0.953	0.953	0.133	0.086

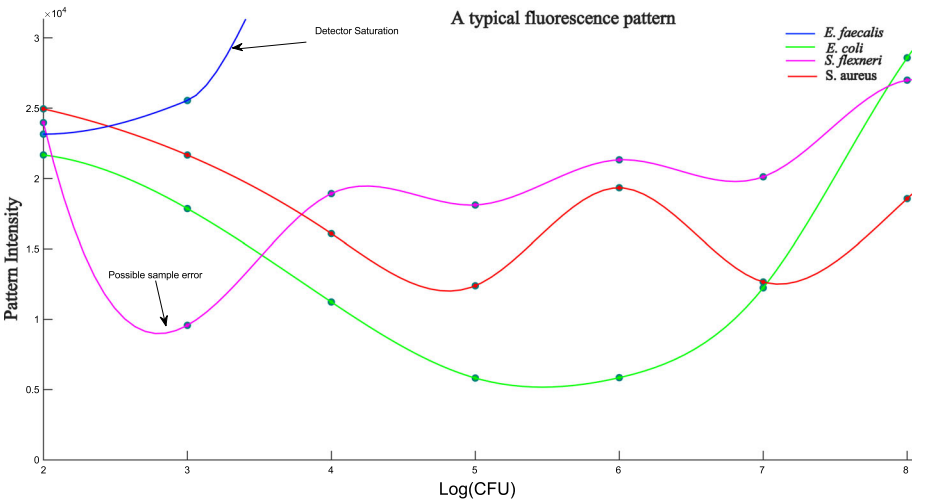

Fig. 21 Fluorescence pattern of standard bacteria. The vertical axis (dependent variable) shows the fluorescence intensity during the bacteria's activity, and the horizontal axis (independent variable) shows the concentration of bacteria. The unit of the horizontal axis is colony forming unit (CFU)

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 4 NDA of clinical sample pattern to the standard bacteria pattern

	Standard				
	Bacteria	*E. faecalis*	*E. coli*	*S. aureus*	*S. flexneri*
Clinical	*E. faecalis*	0.835	0.119	0.107	0.107
	E. coli	0.044	0.463	0.244	0.360
	S. aureus	− 0.238	0.137	0.220	− 0.119
	S. flexneri	0.186	0.054	0.380	0.396

integral components of the NDA method. The effectiveness and accuracy of NDA have been validated through both simulated and real-world data. NDA can be utilized in various other applications, such as measurement of correlation and template matching.

Acknowledgements The authors would like to express their special thanks to Mrs. Zohreh Keihan Jorshari for editing the manuscript, Dr. Meysam Olfatifar for epidemiological consults on STEMI patient data, Dr. Shahrokh Naseri for contributing in a part of the experimental study, and Sinus and Surgical Endoscopic Research Center, Mashhad University of Medical Sciences. The authors sincerely appreciate all the follow-ups on the manuscript and all the valuable comments and suggestions of the referees that have improved the quality of the paper.

Author contributions All authors contributed to the study's conception and design; A.A. and H.D. Supervised, directed and managed the study; A.Z. performed analysis and wrote the first draft of the manuscript. All authors read and approved the final manuscript.

Funding The present article has been extracted from the PhD thesis written by Abolfazl Zanghaei (Registration No. 22214) and financially supported (Grant No. 25332) by the School of Medicine, Shahid Beheshti University of Medical Sciences, and Mashhad University of Medical Sciences (Grant No. 4022644).

Data availability At the time of preparation of this article, just MATLAB codes for this method are available. All MATLAB codes are available at the following address: https://github.com/zanghaei/NDA. Although the MATLAB function that is available at the above address has a couple of options, in most cases, it is better to use the default options.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval and consent to participate The study was approved by the ethics board of Shahid Beheshti University of Medical Sciences (No. IR.SBMU.MSP.REC.1402.471) and Mashhad University of Medical Sciences (No. IR.MUMS.MEDICAL.REC.1403.074).

References

1. Mai J-F, Scherer M (2014) How to measure dependence? In: Mai J-F, Scherer M (eds) Financial engineering with copulas explained. Palgrave Macmillan UK, London, pp 35–48
2. Allen DE, McAleer M (2020) Generalized measures of correlation for asymmetry, nonlinearity, and beyond: some antecedents on causality. J Am Stat Assoc. 1–11
3. Kotz S, Drouet D (2001) Correlation and dependence. World Scientific
4. Rényi A (1959) On measures of dependence. Acta Math Hungar 10(3–4):441–451
5. Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960
6. Pearl J (2003) Statistics and causal inference: a review. TEST 12:281–345

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

7. Olivetti E, Greiner S, Avesani P (2015) Statistical independence for the evaluation of classifier-based diagnosis. Brain informatics 2:13–19
8. Zanghaei A et al (2023) Prediction of renal transplantation outcome using artificial neural networks and investigating important risk factors. Urologiia 4:82–89
9. Austin PC (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res 46(3):399–424
10. Raza M, Qamar U (2018) A heuristic based dependency calculation technique for rough set theory. Pattern Recogn 81:309–325
11. Bagirov AM, Rubinov A, Yearwood J (2002) A heuristic algorithm for feature selection based on optimization techniques. Heuristic and optimization for knowledge discovery. IGI Global, University of New South Wales, Australia, pp 13–27
12. Dokeroglu T, Deniz A, Kiziloz HE (2022) A comprehensive survey on recent metaheuristics for feature selection. Neurocomputing
13. Rebelo de Sá C (2019) Variance-Based Feature Importance in Neural Networks. 306–315
14. Leray P, Gallinari P (1999) Feature selection with neural networks. Behaviormetrika 26(1):145–166
15. Olden JD, Joy MK, Death RG (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol Model 178(3–4):389–397
16. Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of relieff and rrelieff. Mach Learn 53(1):23–69
17. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. Journal of machine learning research. 3: 1157–1182
18. Liu H, Motoda H (2007) Computational methods of feature selection. CRC Press
19. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28
20. Li J et al (2017) Feature selection: a data perspective. ACM Comput surveys (CSUR) 50(6):1–45
21. Kumar V, Minz S (2014) Feature selection: a literature review. SmartCR 4(3):211–229
22. Kira K, Rendell LA (1992) A practical approach to feature selection. Machine learning proceedings 1992. Elsevier, pp 249–256
23. Yang W, Wang K, Zuo W (2012) Neighborhood component feature selection for high-dimensional data. J Comput 7(1):161–168
24. Strobl C et al (2008) Conditional variable importance for random forests. BMC Bioinform 9(1):1–11
25. Greenwell BM, Boehmke BC, McCarthy AJ (2018) A simple and effective model-based variable importance measure. arXiv:1805.04755
26. Wu X et al (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
27. Janssen J, Guan V, Robeva E (2023) Ultra-marginal feature importance: Learning from data with causal guarantees. In: International conference on artificial intelligence and statistics
28. Ezekiel M, Fox KA (1959) Methods of correlation and regression analysis: linear and curvilinear
29. Ryan TP (2008) Modern regression methods. Wiley
30. Dette H, Siburg K, Stoimenov P (2013) A copula-based non-parametric measure of regression dependence. Scand J Stat 40(1):21–41
31. Tjur T (2009) Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. Am Stat 63(4):366–372
32. Schober P, Boer C, Schwarte LA (2018) Correlation coefficients: appropriate use and interpretation. Anesth Analg 126(5):1763–1768
33. Wang Q, Shen Y, Zhang JQ (2005) A nonlinear correlation measure for multivariable data set. Physica D 200(3):287–295
34. Mukaka MM (2012) Statistics corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Med J J Med Assoc Malawi 24(3):69–71
35. Wang Y et al (2015) Efficient test for nonlinear dependence of two continuous variables. BMC Bioinform 16(1):260
36. Benesty J et al (2009) Pearson correlation coefficient. Noise reduction in speech processing. Springer, Berlin, pp 1–4
37. Yang L, Frees EW, Zhang Z (2020) Nonparametric estimation of copula regression models with discrete outcomes. J Am Stat Assoc 115(530):707–720
38. Giannerini S, Maasoumi E, Dagum EB (2015) Entropy testing for nonlinear serial dependence in time series. Biometrika 102(3):661–675
39. Chernyshov KR (2021) System identifiability and structure identification: input and output variables selection based on consistent measures of dependence. IFAC-PapersOnLine 54(14):132–137
40. Siburg KF, Stoimenov PA (2010) A measure of mutual complete dependence. Metrika 71:239–251
41. Micheas AC, Zografos K (2006) Measuring stochastic dependence using ϕ-divergence. J Multivar Anal 97(3):765–784

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

42. Studený M, Vejnarová J (1998) The multiinformation function as a tool for measuring stochastic dependence. In: Jordan MI (ed) Learning in graphical models. Springer, Dordrecht, pp 261–297
43. van Dorp JR, Duffey M (1999) Statistical dependence in risk analysis for project networks using Monte Carlo methods. Int J Prod Econ 58(1):17–29
44. Lancaster HO (1972) Development of the notion of statistical dependence. Mathe Chronicle 2:1–16
45. Zvárová J (1974) On measures of statistical dependence. Časopis pro Pěstování Matematiky 99(1):15–29
46. Tjøstheim D, Otneim H, Støve B (2022) Statistical dependence: beyond pearson's ρ. Stat Sci 37(1):90–109
47. Chatterjee S (2021) A new coefficient of correlation. J Am Stat Assoc 116(536):2009–2022
48. Deb N., P. Ghosal, and B. Sen (2020) Measuring association on topological spaces using kernels and geometric graphs. arXiv preprint arXiv:2010.01768
49. Móri TF, Székely GJ (2019) Four simple axioms of dependence measures. Metrika 82(1):1–16
50. Strothmann C, Dette H, Siburg KF (2024) Rearranged dependence measures. Bernoulli 30(2):1055–1078
51. Azadkia M, Chatterjee S (2021) A simple measure of conditional dependence. Ann Stat 49(6):3070–3102
52. Fuchs S (2024) Quantifying directed dependence via dimension reduction. J Multivar Anal 201:105266
53. Junker RR, Griessenberger F, Trutschnig W (2021) Estimating scale-invariant directed dependence of bivariate distributions. Comput Stat Data Anal 153:107058
54. Trutschnig W (2011) On a strong metric on the space of copulas and its induced dependence measure. J Math Anal Appl 384(2):690–705
55. Chatterjee S (2024) A survey of some recent developments in measures of association. Probability and stochastic processes: a volume in Honour of Rajeeva L. Karandikar. 109–128
56. Sadeghi B (2022) Chatterjee correlation coefficient: a robust alternative for classic correlation methods in geochemical studies-(including "TripleCpy" Python package). Ore Geol Rev 146:104954
57. Wiesel JC (2022) Measuring association with Wasserstein distances. Bernoulli 28(4):2816–2832
58. Griessenberger F, Junker RR, Trutschnig W (2022) On a multivariate copula-based dependence measure and its estimation. Electron J Stat 16(1):2206–2251
59. Auddy A, Deb N, Nandy S (2024) Exact detection thresholds and minimax optimality of Chatterjee's correlation coefficient. Bernoulli 30(2):1640–1668
60. Wei Z, Kim D (2018) On multivariate asymmetric dependence using multivariate skew-normal copula-based regression. Int J Approximate Reasoning 92:376–391
61. Boonmee T, Tasena S (2016) Measure of complete dependence of random vectors. J Math Anal Appl 443(1):585–595
62. Steuer R et al (2002) The mutual information: detecting and evaluating dependencies between variables. Bioinformatics 18:S231–S240
63. Qu G, Hariri S, Yousif M (2005) A new dependency and correlation analysis for features. IEEE Trans Knowl Data Eng 17(9):1199–1207
64. Reshef YA et al (2016) Measuring dependence powerfully and equitably. J Mach Learn Res 17(211):1–63
65. Póczos B, Ghahramani Z, Schneider J (2012) Copula-based kernel dependency measures. arXiv preprint arXiv:1206.4682
66. Mohammadi M, Emadi M, Amini M (2022) Testing bivariate independence based on α-divergence by improved probit transformation method for copula density estimation. Commun Stat Simul Comput. p. 1–19
67. Bahraoui T, Kolev N (2021) New measure of the bivariate asymmetry. Sankhya A 83(1):421–448
68. Tasena S, Dhompongsa S (2016) Measures of the functional dependence of random vectors. Int J Approxim Reason 68:15–26
69. Gretton A et al. (2005) Measuring statistical dependence with Hilbert-Schmidt norms. In: International conference on algorithmic learning theory. Springer
70. Tasena S, Dhompongsa S (2013) A measure of multivariate mutual complete dependence. Int J Approximate Reason 54(6):748–761
71. Schmid F et al (2010) Copula-based measures of multivariate association. Springer, Berlin
72. Azadkia M, Chatterjee S (2019) A simple measure of conditional dependence. arXiv preprint arXiv:1910.12327
73. Kinon A (2022) Empirical measure of multivariate complete dependence. Thai J Math 20(4):1491–1504
74. Delicado P, Smrekar M (2009) Measuring non-linear dependence for two random variables distributed along a curve. Stat Comput 19(3):255
75. Zheng S, Shi N-Z, Zhang Z (2012) Generalized measures of correlation for asymmetry, nonlinearity, and beyond. J Am Stat Assoc 107(499):1239–1252
76. Engle R (2002) Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. J Bus Econ Stat 20(3):339–350
77. Ritter M et al. (2018) Efficient test for nonlinear dependence of two continuous variables. pp 107–120

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

78. Wang Y et al (2017) Bagging nearest-neighbor prediction independence test: an efficient method for nonlinear dependence of two continuous variables. Sci Rep 7(1):12736
79. Tucker LR (1958) Determination of parameters of a functional relation by factor analysis. Psychometrika 23(1):19–23
80. Jiang L et al (2018) A test of general asymmetric dependence. J Appl Economet 33(7):1026–1043
81. Livan G, Rebecchi L (2012) Asymmetric correlation matrices: an analysis of financial data. Eur Phys J B 85(6):213
82. Manner H (2010) Testing for asymmetric dependence. Stud Nonlinear Dyn Econom. https://doi.org/10.2202/1558-3708.1658/html
83. Støve B, Tjøstheim D (2014) Asymmetric dependence patterns in financial returns: An empirical investigation using local Gaussian correlation. Essays Nonlinear Time Series Econom. 307
84. Madsen L, Birkes D (2013) Simulating dependent discrete data. J Stat Comput Simul 83(4):677–691
85. Genest C, Nešlehová JG (2014) Modeling dependence beyond correlation. Statistics in action: a canadian outlook. 59–78
86. Que X et al (2021) A similarity measurement with entropy-based weighting for clustering mixed numerical and categorical datasets. Algorithms 14(6):184
87. Raghu VK, Poon A, Benos PV (2018) Evaluation of causal structure learning methods on mixed data types. Proc Mach Learn Res 92:48–65
88. James RG, Crutchfield JP (2017) Multivariate dependence beyond Shannon information. Entropy 19(10):531
89. Pearl J (2000) Models, reasoning and inference. CambridgeUniversityPress, Cambridge
90. Imbens GW, Rubin DB (2015) Causal inference in statistics, social, and biomedical sciences. Cambridge University Press
91. Pearl J (2009) Causal inference in statistics: an overview
92. Retherford RD, Choe MK (2011) Statistical models for causal analysis. Wiley
93. Asuero AG, Sayago A, González AG (2006) The Correlation Coefficient: An Overview. Crit Rev Anal Chem 36(1):41–59
94. Janjani P et al (2023) Creatinine clearance is key to solving the enigma of sex difference in in-hospital mortality after STEMI: propensity score matching and mediation analysis. PLoS ONE 18(5):e0284668
95. Hashmi KA et al (2020) Risk assessment of patients after ST-segment elevation myocardial infarction by killip classification: an institutional experience. Cureus 12(12):e12209
96. Zhang H et al (2024) High serum lactate dehydrogenase as a predictor of cardiac insufficiency at follow-up in elderly patients with acute myocardial infarction. Arch Gerontol Geriatr 117:105253
97. Digitale JC, Martin JN, Glymour MM (2022) Tutorial on directed acyclic graphs. J Clin Epidemiol 142:264–267
98. Kalisch M, Bühlman P (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J Mach Learn Res 8(3):613
99. Andrews B, Ramsey J, Cooper GF (2019) Learning high-dimensional directed acyclic graphs with mixed data-types. Proc Mach Learn Res 104:4–21
100. Goriki Y et al (2020) A novel predictive model for in-hospital mortality based on a combination of multiple blood variables in patients with st-segment-elevation myocardial infarction. J Clin Med 9(3):852
101. Zanghaei A et al (2023) Rapid identification of bacteria by the pattern of redox reactions rate using 2′, 7′-dichlorodihydrofluorescein diacetate. Biochem Biophys Res Commun 678:78–83
102. Huang X, Shang HL, Pitt D (2022) Permutation entropy and its variants for measuring temporal dependence. Aust N Z J Stat 64(4):442–477
103. Mudumba B, Kabir MF (2024) Mine-first association rule mining: An integration of independent frequent patterns in distributed environments. Decision Anal J 10:100434
104. Shrier I, Platt RW (2008) Reducing bias through directed acyclic graphs. BMC Med Res Methodol 8:1–15
105. Williams TC et al (2018) Directed acyclic graphs: a tool for causal studies in paediatrics. Pediatr Res 84(4):487–493
106. Suttorp MM et al (2015) Graphical presentation of confounding in directed acyclic graphs. Nephrol Dial Transplant 30(9):1418–1423
107. Van Cauwenberg J, De Paepe A, Poppe L (2023) Lost without a cause: time to embrace causal thinking using Directed Acyclic Graphs (DAGs). Int J Behav Nutr Phys Act 20(1):145
108. Lee K-Y, Li L, Li B (2024) Functional directed acyclic graphs. J Mach Learn Res 25(78):1–48
109. Washio T, Motoda H (2003) State of the art of graph-based data mining. ACM SIGKDD Explor Newsl 5(1):59–68
110. Kalisch M, Bühlmann P (2008) Robustification of the PC-algorithm for directed acyclic graphs. J Comput Graph Stat 17(4):773–789

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

111. Dekker J, de Boer M, Ostelo R (2023) Minimal important change and difference in health outcome: an overview of approaches, concepts, and methods. Osteoarthritis and cartilage
112. Copay AG et al (2007) Understanding the minimum clinically important difference: a review of concepts and methods. The Spine Journal 7(5):541–546
113. Terwee CB et al (2021) Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res 30(10):2729–2754
114. Schmitz A (2000) Measuring statistical dependence and coupling of subsystems. Phys Rev E 62(5):7508
115. Holland PW, Wainer H (2012) Differential item functioning. Routledge
116. Kallenberg O (2017) Random measures, theory and applications. Springer
117. Berentsen GD, Kleppe TS, Tjøstheim DB (2014) Introducing localgauss, an R package for estimating and visualizing local Gaussian correlation. J Stat Softw 56:1–18
118. Lacal V, TjØstheim D (2017) Local Gaussian autocorrelation and tests for serial independence. J Time Ser Anal 38(1):51–71

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Abolfazl Zanghaei has received BSc in electrical engineering, MSc in biomedical engineering from Tehran University of Medical Science, PhD in biomedical engineering from Shahid Beheshti University of Medical Sciences. He has the experience of a postdoctoral fellowship in Mashhad University of Medical Sciences. His research areas are artificial intelligence, image and signal processing, and medical instrumentation. His research focuses on machine learning methods and mathematical models in medicine, especially causal inference methods in medical studies.

Hassan Doosti is a senior lecturer and program director of the Master of Data Science at Macquarie University. He specializes in statistics, focusing on flexible modeling techniques, especially when data is not balanced or fully representative. He has developed and applied models that adapt to complex data patterns, providing strong insights and predictions. He is passionate about using statistics to find actionable insights and support decision-making, whether in medical studies or across various business applications.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Ali Ameri received his BSc and MSc in Electrical Engineering from Sharif University of Technology and Shahid Beheshti University, respectively. In 2010, he obtained his PhD in biomedical engineering from the University of New Brunswick, Fredericton, Canada. He joined the Department of Biomedical Engineering at Shahid Beheshti University of Medical Sciences, Tehran, Iran, in 2016, where he is currently an associate professor. His research interests include the application of artificial intelligence in biomedical engineering.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com