Exploring the Bounds of Noise Tolerance within the Maximal Information Coefficient

Dongsheng Wang
Occidental College
California, USA
dwang3@oxy.edu

Yuhang Liu*
Institute of Computing Technology CAS
Beijing, China
liuyuhang@ict.ac.cn

Abstract

Our study delved into the intricate relationship between statistical correlation measures and noise contamination. By applying mathematical models and conducting experiments, we examined how noise fluctuations impair the credibility and interpretation of standard correlation coefficients. Extending the scope, we then further explored the interaction between noise levels and the Maximum Information Coefficient (MIC). Utilising a combination of theoretical derivations and empirical assessments, we clarified the effects of varying noise intensities on MIC's efficacy. A central aspect of this research is to obtain a relation for noise and MIC. We began with a naive approach in upper bounding noise providing us with a deeper understanding of the connection between noise and MIC. With this new nuanced understanding we derive a more effective method to lower bound the MIC with noise. Furthermore through mathematical proof and numerical data we show its consistency and validity.

CCS Concepts

• **Mathematics of computing → Nonparametric statistics**.

Keywords

Maximum Information Coefficient, Correlation Coefficient, Noise Level

ACM Reference Format:
Dongsheng Wang and Yuhang Liu. 2024. Exploring the Bounds of Noise Tolerance within the Maximal Information Coefficient. In *2024 7th International Conference on Computer Information Science and Artificial Intelligence (CISAI 2024), September 13–15, 2024, Shaoxing, China.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3703187.3703311

1 Introduction

Data is one of the most important resources in the modern world. Although it has always been present, only recently were we able to amass it in great quantities and derive insight from them. The emergence of a new resource demands faster, better, and more efficient ways of refining and processing. There is a growing need for the efficient processing of extremely large amounts of data.

CISAI 2024, September 13–15, 2024, Shaoxing, China

ACM ISBN 979-8-4007-0725-4/24/09
https://doi.org/10.1145/3703187.3703311

It is vital for data analysis to find relationships between variables. The importance of such a basic investigation is immense as it paves the way for potentially new and groundbreaking areas of investigation. Finding the exact relationship between variables allows us to manipulate one variable through another. It provides us with the insight hidden in the numerical immensity of data that is invisible through normal observation. When this insight is aligned with AI and machine learning, it produces staggeringly accurate predictions based on a multitude of variables.

The constraints of data science however, still apply despite the explosion of enthusiasm and demand in this field. The 4 V's of big data (i.e., volume, velocity, variety, and veracity) are the key framework in data science that helps us understand and manage datasets. In the modern world, volume, velocity, and variety are at an all-time high as the digitization of society has bolstered the production, collection, and variety of data. However, the veracity or the quality of data is still a realistic issue that has not been miraculously solved by digitization [3].

The obvious solution is to produce higher quality data, but an alternative could be to produce better ways of detection and analysis. The MIC is one such measure [1–4]. Maximum Information Coefficient (MIC) is a new statistical measure that has many promising properties [9]. It is a tool for detecting and quantifying the strength of any linear or nonlinear association between two variables in a dataset.

The MIC is a non-parametric measure that aims to measure the mutual information shared by two random variables. Mutual information is a fundamental concept in information theory that quantifies the amount of information one variable provides about another. Unlike Pearson's correlation coefficient, which is most informative for linear relationships, MIC is capable of detecting a wide range of associations, including nonlinear ones. The MIC is also relatively tolerant of noise and is equitable in that there is no bias towards specific types of functions [5–7].

Although the MIC is an extremely effective tool that may make up for noisy low-quality data, noise is still important to consider. Noise refers to unwanted variability or error that can distort signals, affecting the accuracy of measurements and results.

The noise level h in the context of this paper is defined as a uniform distribution with ranges $[-h, h]$ that is applied to the Y coordinate of each point. This definition simplifies testing and modelling without simplifying the effects of noise too much.

The goal of this paper is to examine the noise tolerances of MIC through both a lower and upper bound. We do this by examining the already present lower bound given by the source report and finding ways to potentially improve it through theory and numerical analysis. We also try to discuss the theory of an upper bound of noise. The goal is to verify the bounds of MIC and Noise and

examine the accuracy and tightness of each with respect to different functions.

The contributions of this paper are as follows.

- We investigated the relationship between correlation coefficients and noise levels, examining this connection from both theoretical and experimental perspectives. Through mathematical formulations and hands-on experimentation, our study aimed to illuminate how variations in noise affect the reliability and interpretation of correlation measures.
- We further examined the relationship between MIC and noise levels, approaching this investigation from both mathematical formulations and experimental analyses. This study aimed to elucidate how varying degrees of noise impact the estimation and interpretation of MIC, offering insights through theoretical derivation and practical validation.
- We identified the problems with a derived theoretical upper bound including accuracy and utility. We examined the source of its problem being a lower bound from the source text. After examining the proof and logic of this lower bound we derived a more effective and tighter lower bound. Through mathematical proof and numerical data we show this new improved lower bound to be consistently a tighter bound and a valid bound.
- With this new lower bound shown to be valid and tighter the application value of this result is promising. An existing more accurate lower bound on noise that is described in the text relies on the existing lower bound which we improved upon. A potential application of this research is an improvement of the lower bound mentioned before through using the more effective lower bound.

The rest of the paper is structured as follows. Section II introduces the framework of MIC, detailing its calculation process. Section III discusses lower bounding the correlation coefficient with respect to noise. Section IV investigates the boundary of noise level with respect to fixed thresholds of MIC, focusing on deriving an upper bound and lower bound for noise level. Section V explains the proof of a key lower bound equation, equation (21) to understand the logic and set the groundwork for improving this lower bound. Section VI explains the problems of equation 25 and by extension the theorised upper bound and proposes an improved lower bound. Section VII proves the proposition and section VIII contains experimental validations and discussion of results. Section IX surveys related work in the field, providing a broader context for our study. Lastly, Section X concludes the paper, summarizing key findings and suggesting avenues for future research.

2 Analysis of the process of MIC

Some preliminary concepts are needed to better understand the intuition of MIC. Entropy which is the measure of randomness of a variable is defined as:

$$H(X) = -\sum_{x} p(x)\log(p(x)) \tag{1}$$

Mutual information as the name implies measures how much knowing one variable will tell you about the other. In the MIC it is defined in terms of entropy:

$$I(X;Y) = H(X) - H(X|Y) \tag{2}$$

Intuitively it can be understood that mutual information measures dependencies between two variables by looking at the randomness of both. If X and Y are related, by knowing exactly what X is the entropy or randomness of Y can be reduced as there is a functional relationship. Mutual information measures how much entropy or uncertainty in one variable can be reduced by knowing the other [8].

The importance of identifying relations between variables cannot be understated, and the MIC is perhaps one of the most effective methods yet. Therefore, it is important to overview its methodology, definition, and key processes as done in the following.

The MIC process begins with a set of ordered pairs D which can be partitioned via a grid separating the x values into X bins and y values into Y bins; this is known as an x by y grid [1]. In Figure 1, the x bins would be going from left to right: $(2/8, 1/8, 2/8, 2/8, 1/8)$, and the y bins would be going from bottom to top: $(1/8, 2/8, 1/8, 2/8, 2/8)$.

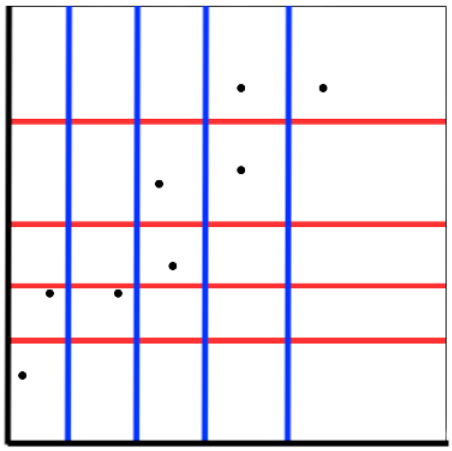

Figure 1: Example of graph partition.

Given this grid D where all x and y are positive integers, we can create two distributions. One distribution is obtained from the bins of x and the other from the bins of y where each bin's probability is the number of points in the bin divided by total points. So for a 2 by 2 grid where each grid has 1 point, the probability distribution for x would be $(1/2, 1/2)$ and y would be $(1/2, 1/2)$.

Define $I^*(D, x, y)$ to be the mutual information of the x by y grid whose partition produces the largest possible mutual information between the x bin distribution and y bin distribution.

$$I^*(D, x, y) = \max I(D \mid G) \tag{3}$$

Now define the term $I^*(D, x, y)/\log\min\{x, y\}$ in order to normalize the mutual information value. In the example diagram, the division value would be $\log(5)$ since both $|X|$ and $|Y|$ are 5. This division normalizes the values as different grid resolutions produce different maximum mutual information, so in order to normalize

them we divide each mutual information by the upper bound which is $\log \min\{x, y\}$.

$$M(D)_{x,y} = \frac{I^*(D, x, y)}{\log \min\{x, y\}} \tag{4}$$

The matrix containing all the normalized maximum mutual information values $I^*(D, x, y)/\log \min\{x, y\}$ that have a grid size less than a predefined $B(n)$ of sample size n is known as the characteristic matrix $M(D)$. Where $M_{xy}(D)$ is the $I^*(D, x, y)/\log \min\{x, y\}$ value of an x by y grid in the (x, y) cell of the matrix.

$$\text{MIC}(D) = \max_{xy<B(n)} \{M(D)_{x,y}\} \tag{5}$$

The MIC score is the largest entry in the characteristic matrix. The first two steps of fixing a grid size and finding the maximal mutual information is repeated until the entire matrix is full, the matrix size essentially limits how fine the grid can be and sets a maximal grid size.

The maximal grid size defined as $B(n)$ is usually set to $n^{0.6}$ [2], $B(n)$ is equivalent to the maximum number of cells of a grid or more simply the product of the maximum number of x bins and maximum number of y bins. Naturally a large $B(n)$ will produce a higher MIC and can be colloquially thought of as the sensitivity setting of the MIC's detection.

The immediate conclusion is that $B(n)$ should be set as high as possible so that detection is as thorough as possible. The problem however is the drastic runtime increase which makes high maximal grid sizes unfeasible to run. Setting $B(n)$ too high may also lead the MIC to mistake noise as minute relationships or focusing too much on irrelevant relationships. There is however still an incentive to have a high maximal grid size not only for more minute analysis but also to counteract the disrupting effects of noise.

Advantages of the MIC include: 1) Versatility: The MIC can detect various types of relationships (linear, non-linear, monotonic, and complex). 2) Universality: The MIC aims to be a universal dependence measure, suitable for a wide variety of data types. 3) Interpretability: The MIC score is easily interpreted, with higher values indicating stronger relationships.

Limitations include: 1) Computational Cost: Calculating MIC can be computationally intensive due to the need to test multiple grids. 2) False Discovery Rate: Without proper adjustment, MIC may lead to high false discovery rates in large datasets.

The symbols used in this paper and their meanings are presented in Table 1.

3 Correlation coefficient and noise

The correlation coefficient, a measure of linearity is a very traditional measure for finding linear relationships. It is defined as follows:

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{6}$$

where E is the expectation, μ_X and μ_Y are the means of X and Y and σ_X, σ_Y are the standard deviations of X and Y.

The correlation coefficient is given a direct equality as well as an upper bound from the source report. The following two inequalities are applied on the unit interval $f : [0, 1] \to [0, 1]$ and for any $k > 0$.

Table 1: Symbol Abbreviation.

Term	Definition
$H(X)$	Entropy, the measure of randomness
$I(X;Y)$	Mutual information, measure of how much entropy is reduced in X by knowing Y
$I^*(D, x, y)$	The maximum possible mutual information induced by a (x,y) grid on D
$M(D)_{x,y}$	The matrix where values of $I^*(D, x, y)$ normalized by dividing by $\log \min\{x, y\}$ are stored
$B(n)$	The maximal grid size which is equal to the product of the largest possible x bin and y bin, it can be thought of as the maximum resolution of the grid
$\rho_{X,Y}$	The correlation coefficient which measures how linear the relationship between X and Y is
$E(X)$	the expectation of X
μ_X	the mean of X
σ_X	the standard deviation of X
$R(k)^2$	the square of the correlation coefficient of $f(X)$ and $f(X) + E_k$
E_k	the uniform distribution on $[-k, k]$
σ^2	the standard deviation of $f(X)$
k	the noise level
$s_{\max}$	the maximum slope in absolute value of $f(X)$ on the unit interval
$s_{\min}$	the minimum slope in absolute value of $f(X)$ on the unit interval
C_0	The MIC score of noiseless data
C_h	A user set threshold MIC score with noise level h
C_k	A hypothetical MIC score value whose noise level when within the boundary defined will be guaranteed to be greater or equal to C_h
$MIC(F_k)$	The MIC score of F_k where F_k is the distribution $(X, f(X) + E_k)$
c	The number of columns which contain a point such that $\lvert f(x) - y_0 \rvert \leq k$ where f is a nowhere constant function and y_0 is the y coordinate that splits the points so that half are above this y value and half are below.
n	the number of points in the data set
α	a value less than $1/2$
$l(f, k)$	the fraction of the unit interval on which $\lvert f(x) - y_0 \rvert \leq k$
$S_{\text{avg},i}$	average slope of the ith column , defined as the gradient of the line between $(x_{\text{i}}, y_0 - k)$ and $(x_{\text{i+1}}, y_0 + k)$ where x_{i} and $x_{\text{i+1}}$ are the boundaries of the ith column.
S_{avg}	the smallest of $S_{\text{avg},i}$ across all columns

The equality is defined as follows:

$$R(k)^2 = \frac{\sigma^2}{\sigma^2 + \frac{k^2}{3}} = \frac{1}{1 + \frac{k^2}{3\sigma^2}} \tag{7}$$

where $R(k)^2$ is the square of the correlation coefficient of $f(X)$ and $f(X) + E_k$. E_k is the uniform distribution on $[-k, k]$. σ^2 is the standard deviation of $f(X)$. k is the noise level. This can be turned into the following upper bound:

$$R(k)^2 \leq \frac{1}{1 + \frac{4s_{\max}k^2}{3s_{\max}-2}} \tag{8}$$

where $s_{\max}$ is the maximum slope in absolute value of $f(X)$ on the unit interval.

4 Boundary of noise level tolerated by MIC

Let us define C_0 to be the MIC score of noiseless data for a relationship. Let C_h be a user-defined threshold value of MIC associated with noise level h. Naturally, $C_0 \geq C_h$, since higher noise levels typically result in lower MIC scores. Now, assume we have new data with noise level k and MIC score C_k. We will define upper and lower bounds for noise level k such that if it falls within this boundary, defined using C_h and other variables, C_k will be guaranteed to be greater than or equal to the threshold value C_h. We will also use an inequality from proposition 6.17[7]

$$C_k = \text{MIC}(D) \geq k - \frac{\log(l)}{\log(B(n)^\alpha)} \tag{9}$$

where k is the noise level, l is the maximum number of bins on the y axis or in other words the maximum number of unique y values, and $B(n)^\alpha$ is the maximal grid size raised to the power of alpha that is less than $\frac{1}{2}$. D is a set of N ordered pairs with distinct x values and whose y values are defined by functions $f_1, f_2, f_3, \ldots f_L$ such that for all (x, y) pairs in D, $y = f_i(x)$ for some i in the set [1,2,3 ... L]. Y values are $(B(n)^\alpha, k)$ partitionable. We can add an inequality to this equation to obtain:

$$C_k \geq k - \frac{\log(l)}{\log(B(n)^\alpha)} \geq C_h \tag{10}$$

Since we want the arbitrary MIC value with noise k to be above the threshold value. By rearranging we obtain:

$$k \geq C_h + \frac{\log(l)}{\log(B(n)^\alpha)} \tag{11}$$

Now using another equation from the source material, we can define an upper bound for noise level k[7]

$$\text{MIC}(F_k) \geq 1 - \frac{2c}{s_{\min}} \cdot k \tag{12}$$

where F_k is the distribution $(X, f(X) + E_k)$, E_k is the uniform distribution on $[-k, k]$ essentially noise, and $f(X)$ is the functional relationship without noise. k is the noise level. c is the number of columns which contain a point such that $|f(x) - y_0| \leq k$. f is a nowhere constant function and y_0 is the median of every point with respect to the y axis. $s_{\min}$ is the minimum slope on the interval of the data. Once again we add our threshold value as a lower bound giving:

$$C_k = \text{MIC}(F_k) \geq 1 - \frac{2c}{s_{\min}} \cdot k \geq C_h \tag{13}$$

Rearranging gives the following upper bound:

$$1 \geq C_h + \frac{2c}{s_{\min}} \cdot k \tag{14}$$

$$\frac{(1 - C_h)s_{\min}}{2c} \geq k \tag{15}$$

What we are left with is a boundary for the noise level k of C_k in terms of the boundary value and other factors such that it will guarantee C_k to be greater than the threshold value.

$$\frac{(1 - C_h)s_{\min}}{2c} \geq k \geq C_h + \frac{\log(l)}{\log(B(n)^\alpha)} \tag{16}$$

Please note that this is a sufficient but not necessary condition.

5 Detailed Proof Explanation

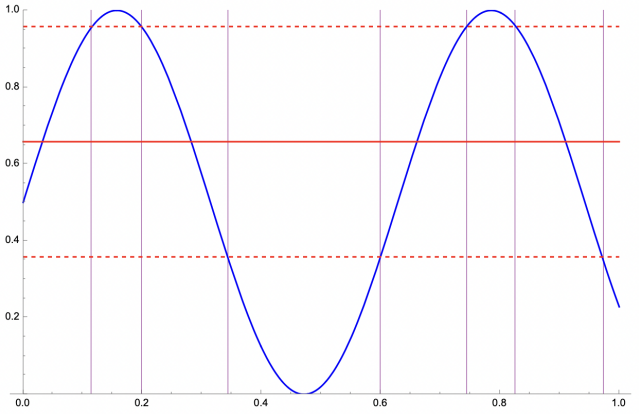

Figure 2: Example of partition for upper bound proof explanation

Corollary 6.25 [7] which is essential for the upper bound has a proof that is not well expanded on, therefore here we provide a more detailed explanation of the proof and reasoning. We start from the conclusion of 6.24 [7] which states the following:

Fix a nowhere-constant function f and a noise level k, and let y_0 be the y-value such that $1/2$ the probability mass of F_k is above y_0 and half is below it. Let $l(f, k)$ be the fraction of the unit interval on which $|f(x) - y_0| \leq k$. Then we have:

$$\text{MIC}(F_k) \geq 1 - l(f, k) \tag{17}$$

The proof of 6.24 which is relevant for 6.25 is directly copied from the source: *Draw a horizontal grid line at $y = y_0$. Every time f enters or exits the strip $[y_0 - k, y_0 + k]$, draw a vertical line. If the j-th column of our grid has $|f(x) - y_0| > k$, then $H_j^Y(F_k|G) = 0$. On the other hand, if it has $|f(x) - y_0| \leq k$, then we still have $H_j^Y(F_k|G) \leq 1$ because binary entropy never exceeds one. The result follows from Lemma 6.3.*[7]

Now from here, let c be the number of intervals on which $|f(x) - y_0| \leq k$, and let $s_{\min}$ be the minimum slope of f on those intervals. In Figure 2, c would be four and the intervals would be the first, third, forth and fifth column. To clarify $s_{\min}$ cannot be zero and the aforementioned intervals are all the columns resulting from the grid described in the proof of 6.24 who have a section of the function $f(x)$ inside the strip $[y_0 - k, y_0 + k]$. We know that the equation for slope is $\Delta y/\Delta x$, where Δy is the change in y and Δx is the change

in x. We can lower bound this fraction with the minimum possible slope and rearrange giving:

$$s_{\min} \leq \frac{\Delta y}{\Delta x}, \quad \Delta x \leq \frac{\Delta y}{s_{\min}} \tag{18}$$

Now consider that we will only apply the above to the columns who have sections of the function f within the strip $[y_0 - k, y_0 + k]$.

The column's boundaries are defined by when the function f enters and exits the strip $[y_0 - k, y_0 + k]$ therefore we can conclude that Δy will always be $2k$, giving the following:

$$\Delta x \leq \frac{2k}{s_{\min}} \tag{19}$$

An important clarification is that for the column to have non-zero slope, its boundaries must be different. The function can either enter from $y = y_0 - k$ and exit at $y = y_0 + k$ or enter from $y = y_0 + k$ and exit from $y = y_0 - k$. If the function enters and exits both from $y = y_0 - k$ (or $y = y_0 + k$) then at least a turning point must occur within the strip $[y_0 - k, y_0 + k]$ making the minimum slope zero.

An edge case to consider is the first column closest to the y axes. If the function, starts in the strip and has nonzero minimum slope then a problem occurs as we cannot guarantee Δy will be $2h$. This edge case can be easily resolved by just using an approximate slope of the strip where we approximate Δy to $2h$. The reason why this does not affect the validity of the bound will be explained further on. An example of this is shown in Figure 2.

Continuing with the explanation we now multiply both sides by c, the number of columns who have points within the strip $[y_0 - k, y_0 + k]$. Resulting in the following:

$$c \cdot \Delta x \leq \frac{2k}{s_{\min}} \cdot c \tag{20}$$

Another important clarification is needed, the columns we consider for $s_{\min}$ are less than or equal in number to c. This is because for c we consider columns who have pieces of $f(x)$ that enter and exit the strip on the same line and columns where $f(x)$ is within the strip and have a turning point inside the strip. We do not consider these columns when finding $s_{\min}$.

Continuing with the proof we can now say the following:

$$l(f,k) \leq c \cdot \Delta x \leq \frac{2k}{s_{\min}} \cdot c \tag{21}$$

To explain this, consider what $c \cdot \Delta x$ and $l(f,k)$ means. $l(f,k)$ is the percentage of x points who have y values inside the strip $[y_0 - k, y_0 + k]$. If you visually highlight all the segments on the x axis who have y value in the strip, the ratio of all the highlighted segments divided by the total length of the domain (which is one) is equivalent to $l(f,k)$, since we assume the x points are uniformly distributed in the domain $[0,1]$. Now consider what $c \cdot \Delta x$ represents. c is the total number of columns with function $f(x)$ within the strip, it can be essentially thought of as the number of visually highlighted segments mentioned before. Now consider what Δx is, this value is connected to $s_{\min}$, the minimum possible slope. If every Δy we consider is of equal value i.e., $2k$, then a minimum slope would mean the same thing as the largest Δx. Essentially, we are finding the largest Δx or in other words, the longest of the visually highlighted lengths mentioned before and multiplying that by the total number of lengths. This effectively upper bounds $l(f,k)$.

This provides us with corollary 6.25[7]:

$$\mathrm{MIC}(F_k) \geq 1 - \frac{2k}{s_{\min}} \cdot c \tag{22}$$

Now to explain why the approximation for Δy used for the edge cases mentioned above does not affect the validity, consider the following scenario. Let the edge column be the column that actually contains the minimum slope where the function f starts inside the $[y_0 - k, y_0 + k]$ strip. By using $2k$ instead of a smaller value and using the same minimum slope the value of the fraction of 2k and minimum slope is larger than the actual value leading to a slightly less tight lower bound. Although this makes the bound less accurate it is still a valid lower bound.

6 Problems of Eq. (22) and Potential Solutions

Although the method and reasoning is clear, in actual practice the utility of Eq. (22) and by extension the upper bound is limited. There are two reasons for this, the first and the less significant one is the relationship of $S_{\min}$, c and k. When rearranged into an upper bound $S_{\min}$ and c are needed in order to obtain an upper bound for k, however in practice to obtain $S_{\min}$ and c, k needs to be a known discrete value. Furthermore, because $S_{\min}$ and c depend on k, using different values of k will produce different bounds when we only want C_h (the user threshold value) to be influencing the bound. This problem is not overly problematic and is resolvable with a few more calculations. The inequality at equation (16) can still be used to find the maximum noise level tolerated for the MIC score to be above a certain user defined threshold. Simply keep increasing the noise level until an equality is obtained.This provides the upper bound as the equality indicates that if the noise level k increases anymore the minimum value of the MIC is no longer guaranteed.

The second problem however is far more troublesome as it is linked to the inherent accuracy of the source equation used for the upper bound. The accuracy of equation (22) is extremely low. The lower bound is often far too low to be of any use and in some cases is even negative. Naturally that would mean the upper bound derived from it is even less accurate. The reason for its inaccuracy is its usage of $S_{\min}$. As mentioned in section five $S_{\min}$ is used to lower bound the fraction $\Delta y/\Delta x$ and by extension the width or Δx of the column, however this lower bound is far too loose. For most functions the width of each column is overestimated by a significant amount when using $S_{\min}$ as a lower bound. This results in gross underestimates of the minimum MIC score.

It is also important to understand the cases when the lower bound in Eq. (22) gives a negative value. This happens when part of the lower or upper line of the strip $[y_0 - k, y_0 + k]$ is outside the y range of $[0,1]$, or when the median is too close to either 1 on 0. This results in the width of columns being overestimated even more as the Δy value in Eq. (22) which is normally $2k$ is actually less than that value, this is because the height of strip y is less than $2k$ due to part of the strip being outside the $[0,1]$ range of y. This results in the fraction $(2k/S_{\min}) \cdot c$ being larger than what it actually is resulting in a more loose lower bound.

An example of such a case is with the function $y = x^3$. The median is close to the x axis so adding even a small amount of

noise will lead to a strip where $y_0 - k$ is negative resulting in gross negative underestimations.

A potential solution for the issue of $S_{\min}$ is to use the average gradient of the column rather than the minimum gradient. This provides a significantly more accurate bound as the average gradient is the exact ratio of the width of the y strip divided by the width of the x column. This is because the average gradient in this case is defined as the gradient of the straight line that is the diagonal of the rectangle. Too elaborate, the rectangle mentioned is formed by the intersection of the two vertical lines forming the column's edges and the two horizontal lines that are the y strips boundary. By defining the average slope this way it becomes the exact ratio of the rectangles height to width which is $2k/\Delta x$. This is still valid as an upper bound as we still pick the smallest average gradient. When using the minimum slope we find the column with the largest x width and use a width larger than this value for all the columns. However, using the average slope, we find the column with the largest width and use that exact width for all the other columns, effectively upper bounding the total column width while being more accurate.

This change also better navigates the two edge cases mentioned previously. The first edge case was if the function starts in the strip so that using the value of $2k$ becomes an overestimate. The average slope deals with this perfectly as it only uses the gradient of the rectangles diagonal which already assumes the height of the rectangle is $2k$, so that ratio of $2k$ and average slope is exactly the columns width.

The second edge case was considering when the function had a turning point within the median y strip. The minimum slope method could only ignore these columns as they may generate a minimum slope of zero. The average slope method is able to utilize these columns in contrast. Since the slope is not dictated by the shape of the function inside the rectangle, only the rectangle itself it will never have a slope that is zero. This potentially allows the average slope method to have more accurate bounds.

A potentially even better method is to simply find the average gradient for each column and use that to bound instead. This results in each column using its own exact x width rather than an upper bound, giving an even tighter upper bound. The equation for such a bound would be as follows:

$$\mathrm{MIC}(F_k) \geq 1 - 2k \sum_{i=0}^{c} \frac{1}{S_{\mathrm{avg},i}} \tag{23}$$

where $S_{\mathrm{avg},i}$ is the average slope of the ith column.

In practice it may not be as effective. Functions which result in low numbers of columns like the exponential function or linear function will not be affected at all. This is because one method uses exact column widths while the other one uses a single column width for all columns, when there is only one column there is no difference. Furthermore functions like sin that will likely have identical columns will also be unaffected since the average slope for one column would be the average slope for all columns. This equation could potentially give a tighter result in cases where the function has many columns that are not uniform.

7 Proposition proof

To prove both proposed equations, proving the more accurate one first, we start at Eq. (22):

$$\mathrm{MIC}(F_k) \geq 1 - \frac{2k}{s_{\min}} \cdot c \tag{24}$$

From Eq. (21), we know that $l(f,k)$ is less than $\frac{2k}{s_{\min}} \cdot c$, so we have the following:

$$\mathrm{MIC}(F_k) \geq 1 - l(f,k) \geq 1 - \frac{2k}{s_{\min}} \cdot c$$

$l(f,k)$ is equivalent to the exact percentage of points within the strip $[y_0 - k, y_0 + k]$ and therefore on a domain of $[0,1]$ with uniformly distributed x values, it can be approximated as the sum of the widths of the columns that have $|f(x) - y_0| \leq k$.

To represent this in terms of slope and k, we define an average slope $S_{\mathrm{avg},i}$. The way this is calculated is by getting the slope between the two opposite points of the rectangle formed by the lines $y = y_0 - k$, $y = y_0 + k$, $x = x_i$ and $x = x_{i+1}$, where $[x_i, x_{i+1}]$ are the x-boundaries of the column. The points in question are $(y_0 + k, x_{i+1})$ and $(y_0 - k, x_i)$. This creates a slope $S_{\mathrm{avg},i}$ such that:

$$S_{\mathrm{avg},i} = \frac{(y_0 + k) - (y_0 - k)}{x_{i+1} - x_i} = \frac{2k}{\text{width of the ith } x\text{-column}} \tag{25}$$

$$\text{width of ith } x\text{-column} = \frac{2k}{S_{\mathrm{avg},i}} \tag{26}$$

The summation of the widths of the columns, as mentioned before, is approximately $l(f,k)$. Therefore, the sum of the width of each column can be represented as follows:

$$l(f,k) = 2k \sum_{i=0}^{c} \frac{1}{S_{\mathrm{avg},i}} \tag{27}$$

Substituting $l(f,k)$ gives the following:

$$\mathrm{MIC}(F_k) \geq 1 - 2k \sum_{i=0}^{c} \frac{1}{S_{\mathrm{avg},i}} \tag{28}$$

This proves the more accurate of the two proposed equations. Now define a new slope s_{avg} where it is the smallest of all $S_{\mathrm{avg},i}$. To prove the more general equation where $s_{\min}$ is replaced with S_{avg}, we start with the following inequality:

$$s_{\min} \leq s_{\mathrm{avg}} \leq S_{\mathrm{avg},i} \tag{29}$$

This naturally holds for all i as $s_{\min}$ is the smallest possible slope across every column, whereas $S_{\mathrm{avg},i}$ is the average slope of the ith column.

$$1 - \frac{2k}{s_{\mathrm{avg}}} \cdot c \geq 1 - \frac{2k}{s_{\min}} \cdot c \tag{30}$$

As $s_{\mathrm{avg}} \leq S_{\mathrm{avg},i}$, we have the following:

$$1 - \frac{2kc}{s_{\mathrm{avg}}} \leq 1 - 2k \sum_{i=0}^{c} \frac{1}{S_{\mathrm{avg},i}} \tag{31}$$

And therefore:

$$1 - \frac{2kc}{s_{\min}} \leq 1 - \frac{2kc}{s_{\mathrm{avg}}} \leq 1 - 2k \sum_{i=0}^{c} \frac{1}{S_{\mathrm{avg},i}} = 1 - l(f,h) \leq \mathrm{MIC}(F_k) \tag{32}$$

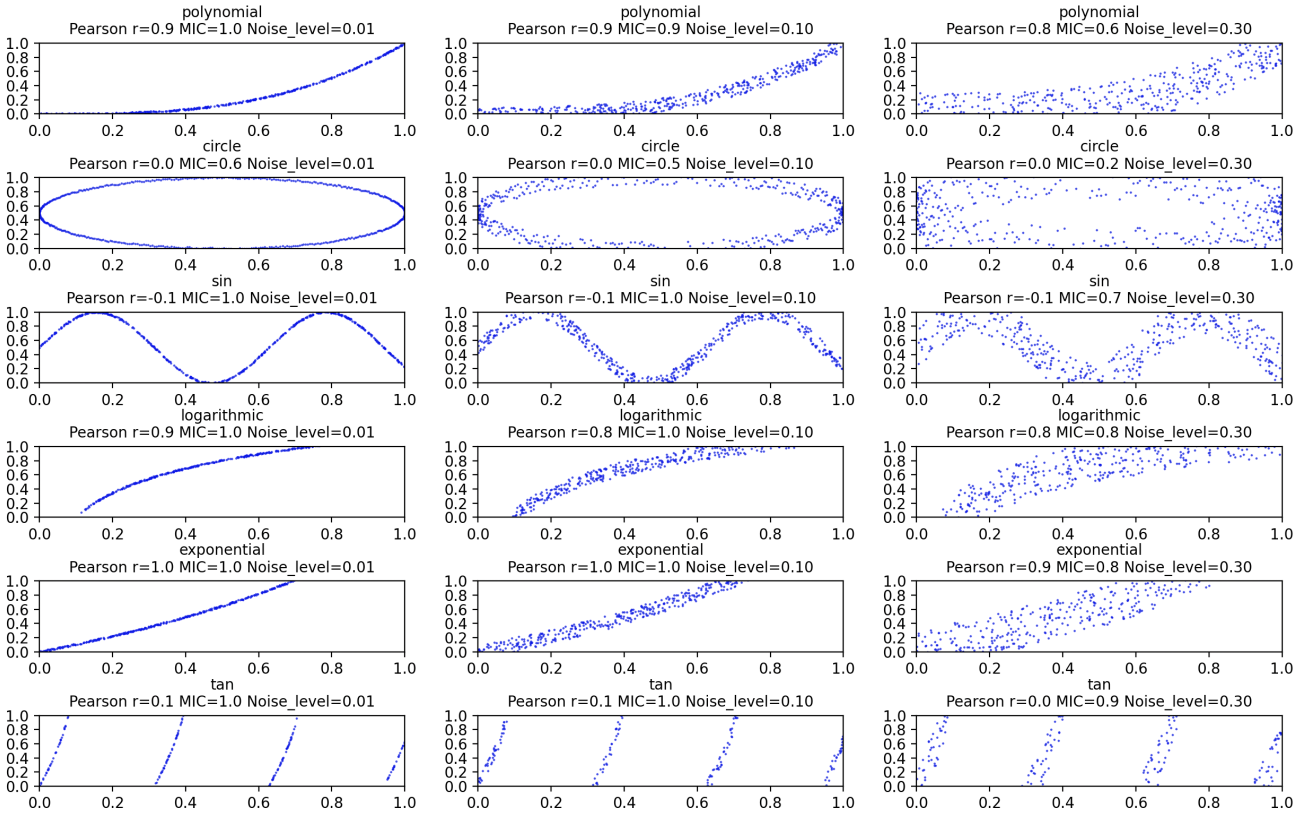

Figure 3: MIC, Correlation coefficient on different functions of various noise levels.

$$1 - \frac{2kc}{s_{\text{avg}}} \leq \text{MIC}(F_k) \tag{33}$$

Since $1 - 2kc/s_{\text{avg}}$ is between two valid lower bounds, it is also proved to be a valid lower bound. It is also a better lower bound than $1 - 2kc/s_{\text{min}}$, illustrating that using s_{avg} yields a better bound than s_{min}. Eq. (28) in theory is a more tight bound than Eq. (33), however in practice this only applies if the number of columns c is very high or the widths of these columns vary greatly. For most common functions this is not the case and so we choose to use Eq. (33) as it is cleaner and simpler to calculate.

8 Experimental Validations

Figure 3 depicts six different functions across six rows: a cubic polynomial, a circle centered at $(0.5, 0.5)$ with radius 0.5, a sine function, a logarithmic function, an exponential function shifted down one unit, and a tangent function.

The figure demonstrates that as the noise level increases from 0.01 to 0.1 to 0.3, the MIC score decreases uniformly for most functional relationships. The circle function however, consistently shows a lower MIC score of 0.6 even with minimal noise, contrasting with other functional relationships. This discrepancy agrees with the source text in that the MIC is only equitable for functional relationships, not necessarily for non functional relations.

Figure 4 offers a detailed analysis of noise level versus MIC score using the same functions as Figure 3, confirming the trend observed in Figure 3. MIC decreases uniformly with increasing noise levels across most relationships. Notably, the tangent function consistently displays a higher MIC score compared to other functional relationships, marking it as an outlier. Additionally, the graphs of exponential, sine, and logarithmic functions appear remarkably similar, with their MIC scores differing by at most 0.1 across the three noise levels examined. The polynomial graph is also relatively similar although that is seen only in shape and less in the data of Figure 3.

Figure 5 presents the correlation coefficient scores plotted against noise levels. Similar to MIC, the correlation coefficient uniformly decreases with increasing noise. However, it has a consistently low score for non linear relationships such as sin, tangent and circular. This reaffirms its inability to detect nonlinear relationships. This figure reinforces the findings from MIC analysis. It indicates that MIC and correlation coefficient metrics align in their observations of functional linear relationships amidst varying noise levels but differ on non linear or non functional relationships.

Figure 6 and 7 depict comparisons of three different MIC lower bounds to the actual MIC noise relationship. As stated in the legend, the blue line is the original lower bound provided by the source work which utilizes the minimum slopes of the column as a lower

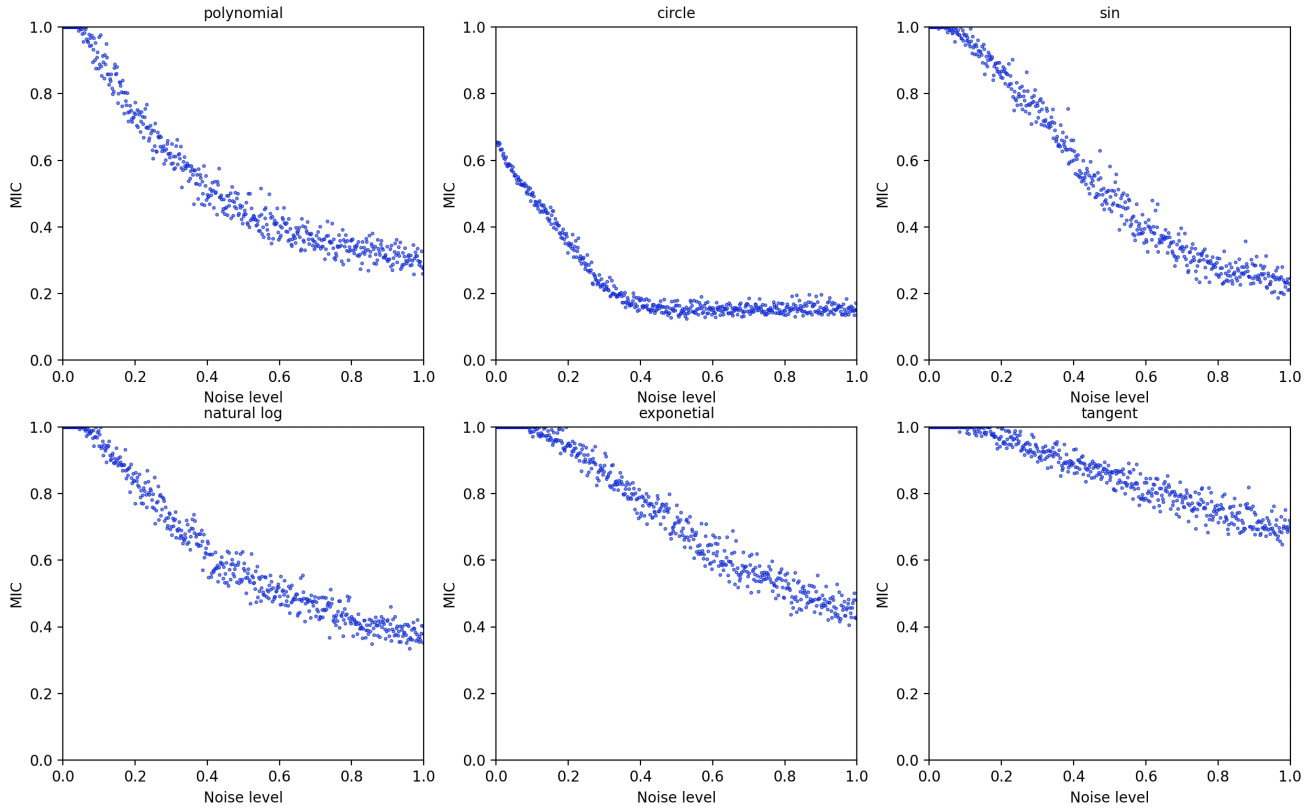

Figure 4: MIC score against noise level across different functions.

bound. The yellow line is the proposed improvement. Instead of using the smallest slope the average slope is used. The green line is a similar method that is slightly more accurate. Instead of using the smallest of all the average slopes, the exact average slope for each column is used.

Observing Figure 6 it can be seen that the green and yellow line are consistently above the blue line. This validates the claim that using the average slope instead of the minimum slope will consistently produce more accurate lower bounds for the MIC. Looking closer at Figure 6, the sin function graph illustrate a small detail. The green line is slightly above the yellow line. What this means is that the green line is a slightly tighter upper bound than the yellow line. This agrees with predictions as the green line is the lower bound using the more accurate summation of all average slopes corresponding to Eq. (28) whereas the yellow line utilizes the minimum average slope corresponding to Eq. (33). The fact that this discrepancy between these two measures appeared in the sin function also agrees with predictions. Eq. (28) will only distinguish itself with Eq. (33) when the function results in multiple columns. The sin function will intercept the median strip multiple times creating multiple columns validating the prediction.

As shown in Figure 7, the tangent graph further validates this claim. The green line in this graph is also slightly above the yellow line illustrating how Eq. (28) is slightly more accurate than Eq. (33). The tangent functions shape also intersects the median strip multiple times creating multiple columns. This confirms the prediction that Eq. (28) is more accurate when multiple columns are formed.

Every other graph in Figure 6 and Figure 7 show the green and yellow line being identical. All of these functions also happen to only intersect the median strip once forming a single column. This validates that the previous predictions contrapositive is also true. If multiple columns are not formed, then Eq. (28) is not anymore accurate than Eq. (33).

Observing both Figure 6 and 7 all the lower bounds except for the circle function are valid, in that they are not above the red line. Both the original lower bound and the improved method are above the red line indicating the lower bound is above the actual value. This implies this method of lower bounding does not work for non functional relationships. This agrees with the source report as it is defined only to work for a nowhere constant function.

Despite the fact the accuracy of the improved method is consistently higher than the original method using the minimum slope, the accuracy of the lower bound is still disappointingly inconsistent. This is partly due to the fact the original lower bound was already widely inconsistent. Through theory and data observation, a key observations about factors that influence the effectiveness of the lower bound can be made.

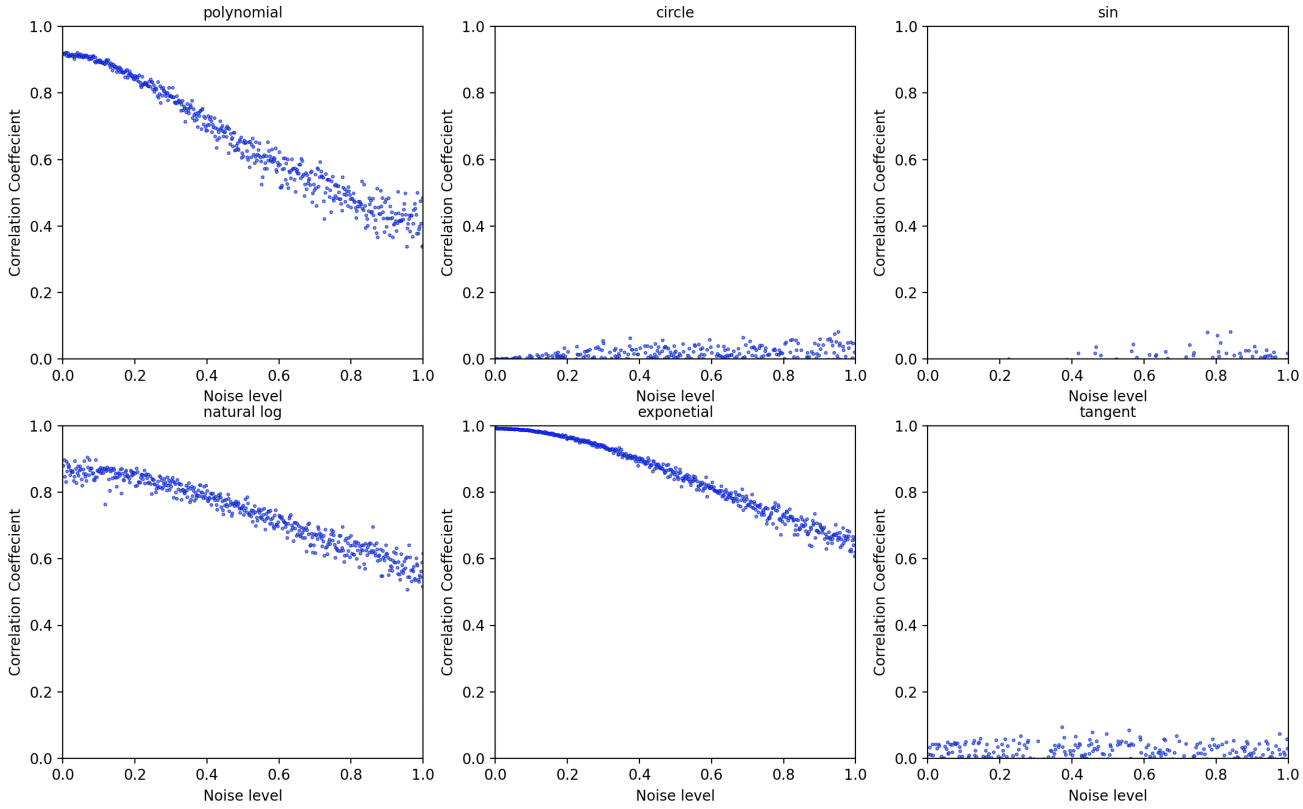

Figure 5: Correlation coefficient against noise level across different functions.

The observation comes from comparing the exponential and logarithmic function. The logarithmic function has a worse lower bound in that it becomes negative relatively quickly. The exponential functions lower bound is able to follow the actual value with higher degrees of accuracy for longer. This discrepancy is due to the distribution of points. The median value for the exponential function is much closer to 0.5 (the middle) than the median value for the logarithmic function. The consequence of this is the exponential function can withstand more noise level before the median strips range is outside the y range of [0,1]. The logarithmic functions median strip in contrast, will be outside the y range of [0,1] with lower noise levels. Once the median strips range is outside the [0,1] y range, the values of column width begin to be grossly overestimated. The height of the median strip is no longer 2k but less, resulting in the lower bound being less accurate, as the fraction to be subtracted is larger than it should be. So to summarise, functions whose median is close to 0.5 will likely have better lower bounds.

More critical readers may notice that the sin function directly contradicts this statement as they may assume the median is close to 0.5 however Figure 6 shows it does not perform well. The function used however had a period of pi/5 so the median is closer to 0.7. Furthermore the tangent function can be seen as the strongest evidence backing this claim. The tan function has the best lower bound out of all six function. Due to its unique shape, its median is extremely close to 0.5 when the period is set to pi/10.

9 Related Work

MIC has garnered significant attention within the realm of data analysis due to its effectiveness in detecting and quantifying a diverse range of relationships. MIC has been employed in various domains, ranging from bioinformatics to financial data analysis, where its ability to uncover complex dependencies has proven invaluable.

Despite the extensive application and theoretical developments surrounding MIC, there exists a relative dearth of research that focuses explicitly on the impact of noise on its performance and reliability. While noise is a ubiquitous component in real-world datasets and known to affect the accuracy and interpretability of statistical measures, studies investigating how different levels and types of noise influence the MIC remain limited. This gap in knowledge hinders a comprehensive understanding of MIC's resilience under noisy conditions and its optimal usage in practical scenarios where noise is inevitable.

There lies a clear need for further exploration into the relationship between MIC and noise. There is need to ascertain the best practices for mitigating noise-induced biases, refining MIC-based

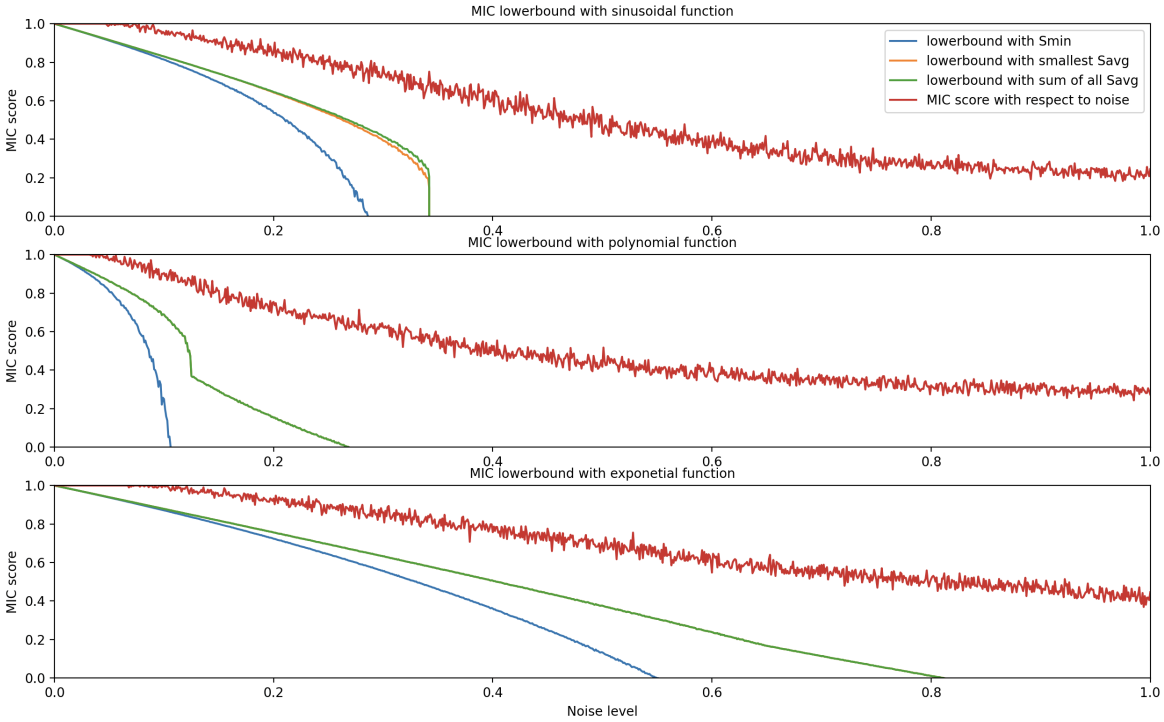

Figure 6: Comparisons of different MIC score lowerbounds to actual MIC score.

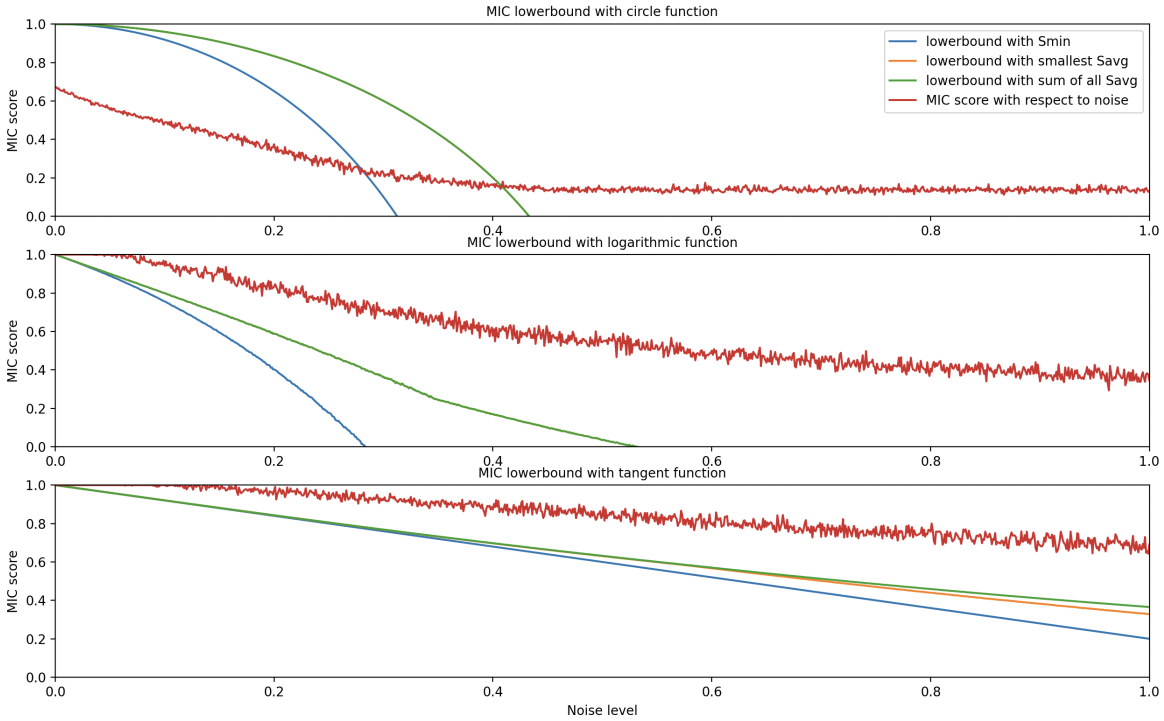

Figure 7: Comparisons of different MIC score lowerbounds to actual MIC score.

methodologies, and guiding researchers and practitioners in leveraging MIC more effectively in noisy data environments.

10 Conclusions

This study embarked on a comprehensive exploration of the intricate relationship between statistical measures of association and the presence of noise in datasets, focusing on correlation coefficients and the Maximum Information Coefficient (MIC). By integrating theoretical frameworks with empirical assessments, this paper aimed to decipher the impact of noise variability on the trustworthiness and interpretation of these measures.

Through a parallel methodology entailing mathematical modeling and experimental evaluations, the study elucidated the manner in which different noise intensities influence the estimation and understanding of MIC. The combination of theoretical deduction and practical substantiation provided a holistic view of MIC's resilience under noisy conditions.

The culmination of these efforts allowed for a significant improvement on a previous lower bound. This improvement is shown to be consistently more tight and able to manage edge cases better. Although not shown in this report this improvement could potentially significantly increase the accuracy of other results that utilize this inequality as the form of the inequality is relatively unchanged.

References

[1] Lazarsfeld, J., Johnson, A., Adeniran, E. (2022). Differentially Private Maximal Information Coefficients. In *Proceedings of the 162nd International Conference on Machine Learning*, July 17-23, 2022, PMLR (Vol. 162, pp. 2-3). Retrieved from https://proceedings.mlr.press/v162/.

[2] Albanese, D. (2013). MinePy: A Python library for Maximal Information-based Nonparametric Exploration. Retrieved from https://minepy.readthedocs.io/en/latest/python.html.

[3] No formal author name given. (2023, March 23). Difference between Data Science and Artificial Intelligence. GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/difference-between-data-science-and-artificial-intelligence/.

[4] NumPy Developers. (2008). `numpy.random.multivariate_normal`. NumPy Documentation. Retrieved from https://numpy.org/doc/stable/reference/random/generated/numpy.random.multivariate_normal.html.

[5] Wikipedia contributors. (2024). Covariance matrix. Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Covariance_matrix.

[6] NumPy Developers. (2023, March 21). NumPy `empty()` function - `numpy.empty()` in Python. w3resource. Retrieved from https://www.w3resource.com/numpy/array-creation/empty.php.

[7] 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. DOI: 10.1126/science.1205438

[8] Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). John Wiley & Sons, Inc.

[9] Reshef, Y. A., Reshef, D. N., Finucane, H. K., Sabeti, P. C., Mitzenmacher, M. (2016). Measuring Dependence Powerfully and Equitably. *Journal of Machine Learning Research*, 17, 5-16.