RESEARCH ARTICLE

A Bayesian measure of association that utilizes the underlying distributions of noise and information

Ishan Goel, Sukant Khurana*

CSIR-Central Drug Research Institute, Jankipuram Extension, Lucknow, Uttar Pradesh, India

* sukant.k@cdri.res.in, sukantkhurana@gmail.com

Abstract

We propose a new approach, Bayesian Probability of Association (BPA) which takes into account the probability distributions of information and noise in the variables and uses Bayesian statistics to predict associations better than existing approaches. Our approach overcomes the limitations of linearity of the relationship and normality of the data, assumed by the Pearson correlation coefficient. It is different from the current measures of association because considering information separately from noise helps identify the association in information more accurately, makes the approach less sensitive to noise and also helps identify causal directions. We tested the approach on 15 datasets with no underlying association and on 75 datasets with known causal relationships and compared the results with other measures of association. No false associations were detected and true associations were predicted in more than 90% cases whereas the Pearson correlation coefficient and mutual information content predicted associations for less than half of the datasets.

OPEN ACCESS

Citation: Goel I, Khurana S (2018) A Bayesian measure of association that utilizes the underlying distributions of noise and information. PLoS ONE 13(8): e0201185. <https://doi.org/10.1371/journal.pone.0201185>

Editor: Yun Li, University of North Carolina at Chapel Hill, UNITED STATES

Received: December 8, 2017

Accepted: July 10, 2018

Published: August 17, 2018

Copyright: © 2018 Goel, Khurana. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All the data used in this submission is available in the following public repositories: 1. <https://webdav.tuebingen.mpg.de/cause-effect/>. 2. <https://www.physionet.org/pn4/eegmmidb/>. The code used for this manuscript can be found at the following DOI: [10.5281/zenodo.1314680](https://doi.org/10.5281/zenodo.1314680).

Funding: The Central Drug Research Institute (CDRI) provided the salary of Dr. Sukant Khurana for the duration of this research. The CDRI manuscript number is 9718. No funds were available for this research. Ishan Goel received no

Introduction

Studying associations between variable is essential to almost all the branches of science and helps us understand the working and the design of a system. In neuroscience, studying associations between different brain regions helps us understand which brain regions are connected [1,2]. In financial markets, studying associations helps us understand which financial instruments are related [3–5]. In social sciences, studying associations helps us understand which factors in a society are associated to one another [6]. Associations are present in all branches of science and understanding such associations help us take decisions, create new designs and predict outcomes.

To study associations among variables, various quantitative approaches have been developed largely around the ideas of correlation analysis and information theory. Correlation analysis between two variables has been central to the study of associations and Pearson Correlation Coefficient has been an indispensable tool for it [7]. Pearson Correlation Coefficient quantifies the extent of linear association in two variables. In other words, Pearson correlation coefficient tests how tightly the scatter plot of two variables fits a line. The tightness of fit

salary for the duration of the work. The specific roles of this author is articulated in the 'author contributions' section.

Competing interests: The authors have declared that no competing interests exist.

is a good measure of association as it suggests that the variables are dependent on each other and when one is changed the other also changes. However, Pearson Correlation Coefficient suffers from the limitations of working only under the assumptions of normality of the underlying data and the linearity of the underlying association [8,9]. Further, attempts have also been made to compare the correlation coefficient with regression and identify the differences in the two [7,10]. Other types of measures have been developed to overcome the limitations of the same. Rank Correlation measures tackle the problem of normality by ranking the data and hence utilize only the order of the values instead of the actual values [9]. Spearman Rank Correlation Coefficient and Kendall Tau Rank Correlation Coefficient are two popular rank correlation coefficients. Other variations of the Pearson correlation coefficient like the circular correlation coefficient [11] and the scaled correlation coefficient [12] were also developed but on the same lines as the Pearson correlation coefficient. Circular correlation coefficient attempted to identify associations in angular values for instance in wind and ocean current directions. Scaled correlation coefficient looks at correlation over smaller time frames to identify the correlations in fast signal components. Further, polyserial and polychoric correlation coefficients also were developed to be robust to the non-normality of data, but they were developed to be used with variables that had a normally distributed underlying latent variable [13]. For instance, Likert scales are often used to measure normally distributed variables that cannot be measured otherwise [14].

Pearson Correlation Coefficient gives both false positives and false negatives because of its limitations of assumptions of linearity and normality. One type of the case where Pearson Correlation Coefficient gives a false negative is the case of testing a quadratic relationship. As the underlying data does not fit a straight line in this case, no matter how tight the relationship is, Pearson Correlation Coefficient is not very high. Further, as Pearson Correlation Coefficient assumes the underlying data to be normally distributed it gives more weight to values away from the mean. When the actual underlying distribution is not normal, Pearson Correlation Coefficient often ends up giving false positives if only the extreme values are correlated. Overcoming one part of the limitation, the distance correlation measures [15] guaranteed independence if the coefficient was 0 and hence theoretically ensured no false negatives.

Pearson Correlation Coefficient is calculated on data arranged into pairs and correlation was checked between two or more groups. However, a different measure was also required for looking at correlation within different groups. Intraclass correlation coefficient was developed which would test the similarity within different groups of data [16]. While these approaches were trying to improve Pearson Correlation Coefficient for specific cases, none of these approaches actually looked at the underlying probability distribution of data.

With the development of Shannon's information theory [17], attempts were made to look at association by considering the distribution of the considered variables. Mutual Information Content compared the joint probability distribution of two variables with the product of their independent probability distribution to come up with a measure of association [18]. Mutual Information Content overcame the limitations of linearity and also normality but still did not consider information and noise separately. While mutual information content had a range from 0 to ∞ , it was adjusted to be in the range of 0 to 1 and the adjusted variant was called the normalized mutual information content so that the quantification of the association could be brought to a standard scale. Other measures like the correlation ratio [19] were also developed that could detect functional dependencies. While information theoretic approaches have given a new direction to associational analysis, they do not look at information and noise separately in the variable and have not provided a major improvement in the accuracy of results.

In our work to identify associations, we use the word noise to refer to the unassociated component in two variables. While this component may not always be undesirable in the data,

for the purpose of identifying association between the two variables, this component can be considered noise. Current works on association are highly sensitive to noise in the variables because they do not segregate information and noise. A fundamentally different approach to association is required that takes noise into account and is more robust. In some systems where different variables affect each other very strongly the amount of stochasticity in the system becomes very high. Because of this reason, it is not very easy to identify associations in these systems. One might argue that even if such associations are identified, such variables cannot be expected to behave in a strongly correlated manner due to the high noise. However, identifying associations is always beneficial for the study of the underlying system.

The underlying distribution of the associated component in two variables (which we refer to as information) holds important information for suggesting associations in the variables but has not been fully explored. Correlated values with lower probability are stronger indicators of association as compared to the values with higher probability. The low probability of independent occurrence of two rare events as compared to one being influenced with the other holds information that has not yet been utilized to study associations.

We first present the new approach to test associations and then suggest ways to make good approximations of the underlying distributions and the underlying relationship of the two variables. We then present the result with synthetic and real world data and show how the suggested approach works well even without any information of the underlying distribution and approximating it to be normal. We also suggest an extension of this approach which can potentially help identify causal directions in variables. Lastly, we show how the approach works well with even non-linear data.

Methods

We have developed a new approach to study associations in data and we present the derivation of the new approach and also explain how should the approach be used. We implemented the approach in Python 3.5 to test the results of the new approach. We have shared our implementation ([S1 File](#)) in pdf format and interested readers can copy the functions and use the code.

Calculating the Bayesian Probability of association

The data from two variables X and Y for n data points is given. As most real world variables are affected by multiple surrounding variables, they always have a noise component in them. Hence, both X and Y are assumed to be the sum of an information component (I) and a noise component (N). Considering the given variable to be a sum of information and a noise component leaves room for an independent component in the two variables and allows for the consideration of the dependent component to be perfectly associated. When the data from the two variables does not fit a curve perfectly, this noise component allows us to look at the possibility of association probabilistically rather than rejecting it straight away.

$$X = I_X + N_X \quad (1)$$

$$Y = I_Y + N_Y \quad (2)$$

We take two mutually exclusive hypotheses, first that the information components are related by a relationship f and second that the information components are independent. Hence, the following hypotheses are taken.

$$H_1 : I_X = f(I_Y) \quad (3)$$

$$H_2 : I_X \neq f(I_Y) \quad (4)$$

for some function f which we refer to as the relationship function. Please note that the first hypothesis assumes that one of the variables is generated stochastically while the other variable assumes its value as a direct consequence of the value of the first variable. For instance, consider the dataset of the sides and areas of squares that are independent and stochastically generated. Observe how, one of the variables can be taken to be stochastically distributed but the other is a direct consequence of the first. Also note that the given coefficient predicts or rejects association considering the specific relationship function and various choices of the same can be taken iteratively to test association. It must be noted here that although H_2 assumes I_X and I_Y to be independent for any relationship function, the suggested approach calculates the probability of association for a single relationship function and hence $P(H_2)$ is not the absolute probability of I_X and I_Y being independent. $P(H_2)$ is the conditional probability of independence given that the relationship function is f . The prior probabilities of 0.5 are assigned to both the hypotheses as there is no reason to believe the likelihood of one over the other until we start observing the data. If there is prior information available, these probabilities can be adjusted accordingly.

$$P(H_1) = 0.5 \quad (5)$$

$$P(H_2) = 0.5 \quad (6)$$

Also note that as the hypotheses are mutually exclusive and exhaustive given the relationship function, the sum of the probabilities of both the hypotheses should be 1 at all times.

$$P(H_1) + P(H_2) = 1 \quad (7)$$

Now, both the variables for one data point at a time are observed and the likelihood of observing that data point given H_1 and the likelihood of observing the data point given H_2 is calculated. The prior probabilities are updated using Bayesian updating. The posterior probability becomes the prior probability for the next data point and the algorithm is iterated over all data points.

Given both the variables for one data point, say x and y , the likelihood of observing the values given hypothesis 1 is calculated as

$$L(x, y|H_1) = \left[0.5 * \int_{i=-\infty}^{i=\infty} P(I_X = i) * P(N_X = x - i) * P(N_Y = y - f(i)) \right] + \left[0.5 * \int_{i=-\infty}^{i=\infty} P(I_Y = i) * P(N_X = x - f^{-1}(i)) * P(N_Y = y - i) \right] \quad (8)$$

where $P(e)$ refers to the probability of some event e and the probabilities will be given by the estimated underlying distribution of the variable. We explain in the next section on how to estimate these distributions. In the first bracket, we consider I_X to be the independent variable and I_Y to be the dependent variable. If the variable, I_X had assumed a value of i , N_X must have assumed a value of $x - i$ and N_Y must have assumed a value of $y - f(i)$. However, this value i could have been anything in the range of $-\infty$ to ∞ . Hence, we integrate over the range. Please note that $P(I_Y = f(i)|I_X = i) = 1$ as it is a direct consequence of the condition. Also note that all the other probabilities in the equation are independent of each other. There is an equal chance that I_Y was the independent variable and I_X was the dependent variable. Hence, we calculate the same considering the latter case which constitutes the second bracket and take an average of both the values by multiplying them by 0.5 and adding them.

We calculate the likelihood of observing x and y , given the second hypothesis as follows.

$$L(x, y|H_2) = \int_{i_x=-\infty}^{i_x=\infty} P(I_X = i_x) * P(N_X = x - i_x) * \int_{i_y=-\infty}^{i_y=\infty} P(I_Y = i_y) * P(N_Y = y - i_y) \quad (9)$$

Considering the second hypothesis, I_X and I_Y must have been generated independently as they have no underlying association. Hence, in the first integral we multiply the probability of I_X assuming a value of i_x and N_X assuming the value of $x - i_x$. We integrate this over all possible values of i_x . Similarly, we calculate the same for variable Y and multiply the two probabilities.

We then update the probabilities using the Bayesian approach as follows.

$$Posterior(H_1) = \frac{Prior(H_1) * L(x, y|H_1)}{Prior(H_1) * L(x, y|H_1) + Prior(H_2) * L(x, y|H_2)} \quad (10)$$

$$Posterior(H_2) = \frac{Prior(H_2) * L(x, y|H_2)}{Prior(H_1) * L(x, y|H_1) + Prior(H_2) * L(x, y|H_2)} \quad (11)$$

where Posterior refers to the posterior probability and Prior refers to the prior probabilities.

Observe how the given approach gives more weight to less probable outcomes only when they are encountered in both the variables. Assume that the relationship function f is the identity function and $I_X = I_Y$. When x and y , are encountered away from the mean but close to one another (thus leaving a small residue for the noise variable) $L(x, y|H_1)$ multiplies the low probability of encountering x and y only once while, $L(x, y|H_2)$ multiplies the low probability twice thus in effect making $L(x, y|H_1)$ higher than $L(x, y|H_2)$. Whereas when x and y , are encountered away from the mean but away from each other (away from the mean in the opposite directions, thus making the noise component large), the value of noise is also away from the mean in $L(x, y|H_1)$, thus multiplying a low probability component and reducing the overall value. Also observe, when x and y are not close enough $L(x, y|H_1)$ is reduced because the absolute sum of the two noise components is always equal to $|x - y|$ (given our relationship function f is the identity function, as $I_X = I_Y$, $|N_X + N_Y| = |x - y|$) where as it is not so for $L(x, y|H_2)$.

Note that in the rest of the paper, whenever we present and talk about results, we only present the probability of association which is the probability of the first hypothesis and refer to it as the Bayesian Probability of Association (BPA).

Estimation of underlying distributions and the relationship functions

Having the dataset, one cannot exactly know the underlying distribution of noise and information but one can make good approximations on the same. In specific studies where one has more domain knowledge of the underlying system, one can make more accurate estimations and have better results.

The central limit theorem states that the sum of different random variables tend to be normally distributed and hence, the normal distribution is a good estimate if nothing else is known about the data. It is worth noting that while Pearson Correlation Coefficient assumes its data to be normally distributed and quantifies association accordingly, BPA is not restricted to be used with the normal distribution and using more accurate distributions will only improve the accuracy of the result. We show with the real world data that this approximation works quite efficiently. When the variable is considered to be normally distributed and the variable X , is the sum of I_X and N_X , it is known that.

$$Mean(X) = Mean(I_X) + Mean(N_X) \quad (12)$$

$$Variance(X) = Variance(I_X) + Variance(N_X) \quad (13)$$

The mean of the distribution decides the location of the distribution and addition or subtraction of the mean does not change the shape of the distribution. To avoid estimating the actual mean and variance of all the four distributions, we suggest normalizing both the variables before calculating BPA.

Variance of I_X and N_X are indicative of the amount of information and noise in the variable respectively. As the total variance after normalization is 1, a proportion, α'_X can be estimated to be the variance of information and $(1 - \alpha'_X)$ can be estimated to be the variance to noise. Various combinations of α'_X and α'_Y can be tested to detect associations. Note that in this manuscript we will refer to the actual proportions of variance to be α_X and α_Y .

A good approximation to the relationship function can be found by curve fitting approaches and the domain knowledge of the variables being tested. For instance, consider the example consisting of various squares. While the sides have a dimension of length, area has a dimension of length square. This clearly suggests that the underlying relationship function is quadratic in nature.

Results

We tested the approach on 64 associated synthetic datasets, 64 unassociated synthetic datasets, 75 associated real world datasets, 15 unassociated real world datasets and five synthetic datasets with non-linear relationships. Interestingly, Bayesian Probability of Association was found to perform better than existing approaches in most datasets.

Fig 1 summarizes the need for associative measures, major current approaches and briefly introduces the new approach suggested in this paper.

Results on synthetic data

To test the approach, we generated 100 data points for two variables, X and Y using a random number generator (of the specifically chosen probability distribution) in the python programming language. In the first case which tested for associated data, I_X , N_X , I_Y , N_Y were generated by a normally distributed random variable of mean 0 and I_X was kept equal to I_Y (thus defining the relationship function as the identity function). X was calculated by the addition of I_X and N_X and similarly Y . The variance of the three variables were varied in a way that all combinations of α_X and α_Y in the range 0.01 to 0.99 at a gap of 0.15 were tested for all possible combinations of α'_X and α'_Y in the same range. It was observed that as long as α'_X and α'_Y (estimated

Fig 1. Summarization of the existing tools and their limitations and a brief introduction to our approach.

<https://doi.org/10.1371/journal.pone.0201185.g001>

Fig 2. Results given for various combinations of α_X and α_Y in associated datasets were better than Pearson correlation coefficient in the majority of cases. 9 heatmaps are shown in the figure, each representing one combination of α_X and α_Y . Each heatmap, calculates BPA for different combinations of α'_X and α'_Y . I_X was kept equal to I_Y for the dataset to be associated and X was calculated by the addition of I_X and N_X and similarly Y . The value of BPA obtained is encoded in color, the darker being higher. Further, below every heatmap the corresponding scatterplot for X and Y is shown. The straight black line shows the relationship function. On the top of every heatmap the normalized mutual information content (MI), the Bayesian Probability of Association (BPA, for the correct choice of α'_X and α'_Y which is equal to α_X and α_Y respectively) and the Pearson correlation value (PC) has been displayed. On the top of every scatter plot, the values of the Spearman Rank Correlation (SR) and the Kendall Tau Rank Correlation (KT) is also shown. It can be observed that when α'_X and α'_Y exceed α_X and α_Y respectively the results start to approach 0.

<https://doi.org/10.1371/journal.pone.0201185.g002>

proportions of information in both the variables) were less than α_X and α_Y (actual proportions of information in both the variables), the association was strongly predicted while the probability of association quickly dropped almost to 0 once the α'_X and α'_Y exceeded α_X and α_Y (Fig 2, S2 File). This can be explained as when the variance of information is estimated less than the actual value, probability of values far from the mean get multiplied twice in $L(x,y|H_2)$ thus reducing it. Also, in this scenario as the variance of the noise variable is high, large values of noise do not have a very low probability and thus do not reduce $L(x,y|H_1)$. Both these effects are reversed when the variance of information is estimated to be higher than it actually is and the results quickly fall towards 0. This property can be used to study causal directions as generally most of the noise of the cause is reflected in the effect, hence the α for cause would likely be higher than the α for effect [20].

For the sake of clarity, Fig 2 shows the results for 9 out of the 64 datasets that were tested. The detailed results are shown in S2 File. Further, for all the 64 datasets, the correct Bayesian Probability of Association (BPA for $\alpha'_X = \alpha_X$ and $\alpha'_Y = \alpha_Y$), the normalized mutual information content, the Pearson Correlation Coefficient, Spearman Rank Correlation and Kendall Tau Rank Correlation Coefficient were also calculated (S2 File).

In the second case, we tested the results of the approach for unassociated datasets. All the four variables were generated in the same way, except in this case I_X and I_Y were generated independently thus leaving no associated component in the two variables. It was also observed that in the second case when the data was not associated, the approach had a tendency to suggest false associations if α'_X and α'_Y are kept very low but not otherwise (Fig 3, S2 File). This can be explained by the same argument due to which true associations were strongly suggested even when the variance of information was estimated to be lower than its actual value. But false associations were suggested much less than true associations (Fig 3).

Similar to Fig 2, Fig 3 shows the results for 9 out of 64 unassociated datasets. S2 File shows the detailed results and lists the values of other correlation coefficients along with the correct BPA (BPA for $\alpha'_X = \alpha_X$ and $\alpha'_Y = \alpha_Y$).

The distribution of the various associational measures was tested for the 128 synthetic datasets (Fig 4). It was observed that while other measures had a more smoother distribution over 0 to 1, BPA (for $\alpha'_X = \alpha_X$ and $\alpha'_Y = \alpha_Y$) had a sharp peak around both 1 and 0. This suggested that BPA can be a better tool if our purpose is to observe the magnitude of correlation, but to classify two variables to be associated or not as it leaves a wider space (over 0 to 1) between the two classes and hence improving the confidence of our results. It was also observed that even though an equal number of datasets were associated and unassociated all the distributions were skewed towards 0, with the least skew in BPA (as even BPA fails for associated datasets with very low α_X and α_Y).

Results on real world data

The new approach was tested on 75 datasets from a cause-effect database [21] and 69 datasets were predicted to be associated. All 75 datasets had variables that had defined causal

Fig 3. Results given for various combinations of α_X and α_Y in unassociated datasets outperformed Pearson correlation coefficient in the majority of cases. The figure is arranged in the exact same way as Fig 2. I_X and I_Y were calculated independently for the dataset to be unassociated and X was calculated by the addition of I_X and N_X and similarly Y . The value of BPA obtained is encoded in color, the darker being higher. As in Fig 2, below every heatmap the corresponding scatterplot for X and Y is shown. The straight black line shows the relationship function. On the top of every heatmap the normalized mutual information content (MI), the Bayesian Probability of Association (BPA, for the correct choice of α'_X and α'_Y which is equal to α_X and α_Y respectively) and the Pearson correlation value (PC) has been displayed. On the top of every scatter plot, the values of the Spearman Rank Correlation (SR) and the Kendall Tau Rank Correlation (KT) is also shown.

<https://doi.org/10.1371/journal.pone.0201185.g003>

relationships (S3 File). It was also tested on 15 unassociated datasets and no false associations were found (S4 File). The datasets that were used for this study covered a wide range of disciplines. Some associated datasets included geographical variables like the relationship between altitude and temperature, and altitude and precipitation. Further, some associated datasets were taken from the abalone dataset that explored various biological associations. Also, physical associations were explored for instance the association between horsepower and mileage of a vehicle. Such a wide range of discipline shows that the approach is not limited to any branch and can be used for associational analysis in all branches of science.

Fig 4. Distribution of various associational measures over the range of 0 to 1 for all the synthetic datasets. The distribution of results of all the five measures over all the synthetic datasets is shown. For every dataset, only the correct BPA was chosen. Further, absolute values were taken for Pearson Correlation Coefficient, Spearman Rank Correlation Coefficient and Kendall Tau Rank Correlation to bring all the measures on the same scale. Observe how a more negative value also suggests a stronger correlation.

<https://doi.org/10.1371/journal.pone.0201185.g004>

Fig 5. Association was detected in 19 out of 25 datasets with underlying association in the low Pearson correlation coefficient (<0.35) group: BPA for combinations of various α'_X and α'_Y are shown as a heatmap for 9 datasets. Below every heatmap, a scatterplot of the actual data is shown. Also, a black line shows the relationship function that was calculated by curve fitting approaches. On the top left of the heatmap, the unique index of every dataset is presented. On the top of the heatmap Pearson correlation coefficient (PC) along with the p-value in brackets and the normalized mutual information content (MI) for the dataset is given. On the top of the scatterplot, the Spearman rank correlation coefficient (SR) and the Kendall tau rank correlation coefficient (KT) are given.

<https://doi.org/10.1371/journal.pone.0201185.g005>

The new approach was applied by estimating that all the four variables, I_X , N_X , I_Y , N_Y have a normal distribution. The mean of the variables was subtracted from the respective variable so that the mean can be estimated to be 0 for all the four variables. Further, variance of both the variables were calculated and all combinations of α'_X and α'_Y in the range 0.1 to 0.9 at a gap of 0.1 were tested. The underlying relationship was calculated by finding the linear regression function in the data. The fitted regression line has been shown in the figures (Figs 2–3, Figs 5–8). S3 and S4 Files should be referred for the slope and intercept values for all the datasets. Further, to predict the causal direction in the dataset we calculated two parameters,

$$S_X = \sum_{\alpha'_Y=0.1}^{0.9} \sum_{\alpha'_X=0.1}^{0.9} \alpha'_X * P(H_1)$$

$$S_Y = \sum_{\alpha'_Y=0.1}^{0.9} \sum_{\alpha'_X=0.1}^{0.9} \alpha'_Y * P(H_1)$$

As observed, when α'_X or α'_Y exceeds α_X and α_Y respectively, BPA tends to drop towards 0. Hence, the suggested parameters were used to identify which of the two variables had a higher information content, in other words which of the two was higher, α_X or α_Y . As explained, this information can help distinguish the cause from the effect as generally the noise content in the effect is higher than that in the cause [20].

The results on 27 selected datasets divided into three equal sized groups are presented (Figs 5, 6 and 7). Interested readers can refer to a detailed report of all the 75 datasets which can also be divided into three equal sized groups based on the absolute values of Pearson Correlation Coefficient (S3 File). BPA was assumed to be suggesting association if it suggested a probability of 1 for any combination of α'_X or α'_Y .

The first group had datasets with the absolute Pearson Correlation coefficient values of less than 0.35 (Fig 5) which is the low correlation group. Out of all the datasets that had a low correlation, it was observed that for 19 out of 25 datasets, BPA could identify the underlying association. The cases where BPA could not identify the association had Pearson Correlation coefficient values less than 0.18.

The second group had datasets with the Pearson correlation coefficient values from 0.35 to 0.63 (Fig 6) which was the moderate correlation group. Please note that we have taken the value of 0.63 for the sake equally dividing the three groups. Association was identified in all the 25 datasets that had a moderate Pearson Correlation value. It was also observed that in some cases where the scatterplot showed a tight relationship, Pearson Correlation Coefficient was low just because of the non-linearity of the association. BPA could identify association in all such scenarios.

The third group had datasets with high Pearson Correlation values higher than 0.63 (Fig 7). This group is the high correlation group. It was observed that association was identified for all the 25 datasets and for almost all combinations of α'_X or α'_Y . This suggests that for stronger associations, BPA will identify the association even with much more approximate estimations.

Fig 6. Association was detected in 25 out of 25 datasets with underlying association in the moderate Pearson correlation coefficient (0.35–0.63) group: The figure is arranged in exactly the same way as Figs 5 and 9 datasets have been shown. BPA for combinations of various α'_X and α'_Y are shown as a heatmap for 9 datasets. Below every heatmap, a scatterplot of the actual data is shown. Also, a black line shows the relationship function that was calculated by curve fitting approaches. On the top left of the heatmap, the unique index of every dataset is presented. On the top of the heatmap Pearson correlation coefficient (PC) along with the p-value in brackets and the normalized mutual information content (MI) for the dataset is given. On the top of the scatterplot, the Spearman rank correlation coefficient (SR) and the Kendall tau rank correlation coefficient (KT) are given.

<https://doi.org/10.1371/journal.pone.0201185.g006>

Overall, it was observed that association were predicted for 69 out of 75 datasets, whereas the absolute value of Pearson correlation coefficient was more than 0.5 for only 35 datasets (Fig 4). Further, Normalized Mutual Information Content was usually lower than Pearson Correlation Coefficient. Spearman Rank Correlation Coefficient and Kendall Tau Rank Correlation performed similar to the Pearson Correlation Coefficient. It was also observed that for datasets which had high Pearson correlation coefficient (more than 0.5), most of the combinations of α'_X or α'_Y in the tested range predicted the probability of association to be close to 1. Further, for 31 of the datasets, the causal direction could not be predicted because of high association between the variables because of which highest tested values of α'_X or α'_Y probably remained less than α_X and α_Y . This led to $BPA \approx 1$ for all combinations of α'_X or α'_Y . Hence, for these cases $S_X \approx S_Y$. For such cases, higher resolution of the combinations of α'_X or α'_Y are expected to reveal the correct causal direction. In the remaining cases 35 times the causal direction was predicted correctly, and 9 times incorrectly.

To show that the suggested approach identifies cases of independence correctly, we took 15 examples from an EEG dataset [22–24]. We took combinations of data from different recordings and calculated the results in exactly the same way as for the cause-effect dataset. The results were quite low in all the fifteen cases for all the combinations of α'_X or α'_Y (Fig 8, S4 File) and no false associations were detected.

Results with non-linear relationship

We tested BPA for various underlying functions and checked if the new approach could identify non-linear associations any better than other measures (Fig 9). It was observed that BPA performed better than all the Pearson Correlation Coefficient, the Normalized Mutual Information Content, Spearman Rank Correlation Coefficient and Kendall Tau Rank Correlation Coefficient in identifying non-linear associations. We took five relationship functions and checked if the new approach could identify associations with different estimations of the relationship function. It was observed in all the cases the approach could identify association when the correct relationship function was estimated (Fig 9A). The variance of noise was kept as half the variance of information so that other measures could work well. It was also observed that the Normalized Mutual Information Content was higher than the three correlation coefficients and was more robust to different functions but was still lesser than BPA (Fig 9B).

However, it should be noted that although the correlation coefficients do not work very good in identifying non-linear relationships, the same way that we estimate the relationship function for BPA, we can estimate a relationship function and transform the data with the estimated relationship function before calculating the coefficients.

Please note that to apply the approach to different relationship functions, one needs to change the integral limits in the likelihood calculation according to the domain and the range of the relationship function.

Conclusion and discussion

The new approach calculates the probability of two given datasets being associated by utilizing the underlying probability distributions of information and noise in the variables. Further, to

Fig 7. Association was detected in 25 out of 25 datasets with underlying association in the high Pearson correlation coefficient (>0.63) group: The figure is arranged in exactly the same way as Fig 5. BPA for combinations of various α'_x and α'_y are shown as a heatmap for 9 datasets. Below every heatmap, a scatterplot of the actual data is shown. Also, a black line shows the relationship function that was calculated by curve fitting approaches. On the top left of the heatmap, the unique index of every dataset is presented. On the top of the heatmap Pearson correlation coefficient (PC) along with the p-value in brackets and the normalized mutual information content (MI) for the dataset is given. On the top of the scatterplot, the Spearman rank correlation coefficient (SR) and the Kendall tau rank correlation coefficient (KT) are given.

<https://doi.org/10.1371/journal.pone.0201185.g007>

take into account all the data points, it uses Bayesian ideas to update the beliefs on association in the dataset. Results show that the suggested approach makes good predictions in the absence of any knowledge about the underlying system and further knowledge of the underlying system will only further improve the results. The suggested approach was found to be performing better than both Pearson Correlation Coefficient and Normalized Mutual Information Content. Also, the suggested approach provides a method where externally known information about the system can be incorporated in the calculation of the probability of association for better results.

External information about the system can be incorporated at basically three points in the given system. Firstly, in the underlying distributions. As explained, probability distributions of the underlying variables provide valuable information in the prediction of association. By leaving the option of choosing an underlying distribution, this approach also overcomes the limitation of the assumption of normality faced by Pearson correlation coefficient. Secondly, external information can also be incorporated into the underlying relationship function and hence BPA can identify associations that are non-linear in nature. However, it must be noted that the same can also be done for Pearson Correlation Coefficient by transforming the variables based on the known function. Lastly, in the initial probabilities assigned to H_1 and H_2 . If there is a reason to believe one over the other, the prior belief can be incorporated in the system by changing the initial probabilities of H_1 and H_2 .

While associations have been studied for a long time, they have never been studied by considering noise and information separately in a variable. Our approach, by differentiating between noise and information leads to a better estimation of association as the sensitivity to noise in the coefficient can be controlled by setting the distribution of noise. We expect this approach to be developed on the same lines as other correlation coefficients. Other correlation approaches are used in canonical correlation analysis [25], cross-correlation analysis and time lagged correlation analysis [26]. These approaches apply correlation coefficients in different ways to identify associations in a system. Our approach is valid for all of the given methods. Further, we expect this approach to become the primary tool in the analysis of variables with a large amount of noise and where external domain specific inputs can be given to improve results.

To understand the properties of BPA, we analyzed all combinations of α'_x and α'_y but to just study associations, faster algorithmic approaches like binary search [27] and randomized searches [28] can be used to find the optimal values of α_x and α_y . However, there cannot be a fixed approach to choose values of α_x and α_y that reveal true association among variables. The noise to information ratio varies in variables in different branches of science and hence the results of BPA depend on domain specific knowledge. We expect such methods to be used when the suggested approach gains popularity. Further, such methods will support higher resolutions of the values α'_x and α'_y and in turn also improve the prediction of causal directions.

To predict causal directions more efficiently some information in this approach remains unexploited. As α'_x and α'_y exceed α_x and α_y , we observe that BPA starts to tend to 0, a property which can be exploited to get an idea of the value of α_x and α_y . While in many cases, the final

Fig 8. Results for 15 datasets with no underlying association: Data from different EEG recordings was taken and the channel (electrode position) was picked at random. As two different recordings must not have any underlying association, BPA was checked for different combinations of α'_x and α'_y and it was found that the results were pretty low. The figure is arranged in exactly the same way as Figs 5 and 9 datasets have been shown. Below every heatmap, a scatterplot of the actual data is shown. Also, a black line shows the relationship function that was calculated by curve fitting approaches. On the top left of the heatmap, the unique index of every dataset is presented. On the top of the heatmap Pearson correlation coefficient (PC) along with the p-value in brackets and the normalized mutual information content (MI) for the dataset is given. On the top of the scatterplot, the Spearman rank correlation coefficient (SR) and the Kendall tau rank correlation coefficient (KT) are given.

<https://doi.org/10.1371/journal.pone.0201185.g008>

value of BPA might be almost 1 for all tested values of α_X and α_Y , hence failing to predict causal directions. In such scenarios, the rate of increase of BPA with respect to the number of data points consumed can help us predict which of the two, α_X or α_Y are higher.

However, there are some limitations of BPA that we have identified. Firstly, it is more computationally intensive than the Pearson Correlation Coefficient. With the advances in computing technologies, we strongly believe that this will hardly be a hindrance in the development of BPA. Secondly, the approach inherently requires some brute-force searching for an appropriate estimation of the underlying distributions and the relationship function. But we have also shown that the approach gives accurate results even in cases where approximate estimations are made. Lastly, just like the Pearson Correlation Coefficient, BPA is not devoid of false positives or false negatives and some false results are inevitable.

We expect BPA to have applications in all branches of science. Further, we expect that BPA will help advance the study of associations by extending them beyond correlation analysis. It is worth noting that even though association and correlation are related concepts, they do not

Fig 9. The new approach could identify associations in cases of non-linearity: A), Five different relationship functions were taken and the results of the new approach were tested for various combinations of actual and estimated relationship functions. It was observed that the new approach could identify associations when the correct relationship function was estimated. B), Results were compared with the Pearson Correlation Coefficient (PC), the Spearman Rank Correlation Coefficient (SR), the Kendall Tau Rank Correlation Coefficient (KT) and the normalized mutual information content (MI). It was observed that while the three correlation coefficients could not identify associations in all the cases, Bayesian Probability of Association (BPA) could. Normalized Mutual Information Content was lower than the suggested approach but as seen it was higher than Pearson Correlation Coefficient.

<https://doi.org/10.1371/journal.pone.0201185.g009>

imply each other. Two variables which are associated may not have a high correlation between them due to a large amount of noise. Hence, we expect associational analysis to have another layer of analysis beyond correlations. It should also be noted that while correlation by definition is a continuous variable, association is binary. Two variables either will be associated or not. We believe BPA is a good measure of association as it classifies results into 0 or 1 for most cases (Fig 4). We expect BPA to be developed further to quantify the amount of noise in the variables and to study causal directions in all cases. Further, with the advancement in computing power we expect BPA to become fast and be widely used.

Supporting information

S1 File. Code for calculating Bayesian Probability of association. The file shows the code for calculating Bayesian Probability of Association written in python 3.5.
(PDF)

S2 File. Detailed results for synthetic datasets. The detailed results for synthetic datasets are presented as heatmaps and also as a table.
(PDF)

S3 File. Detailed results for associated real world datasets. The detailed results for the real world associated datasets are presented including a heatmap of BPA results, a scatter plot and the numerical values of all the measures compared in the manuscript.
(PDF)

S4 File. Detailed results for unassociated real world datasets. The detailed results for the real world unassociated datasets are presented in the same manner as S3 File.
(PDF)

Acknowledgments

Central Drug Research Institute (CDRI) provided the salary of Dr. Sukant Khurana for the duration of this research. The CDRI manuscript number is 9718.

Author Contributions

Conceptualization: Ishan Goel.

Data curation: Ishan Goel.

Formal analysis: Ishan Goel.

Methodology: Ishan Goel.

Project administration: Sukant Khurana.

Resources: Sukant Khurana.

Software: Ishan Goel.

Supervision: Sukant Khurana.

Validation: Ishan Goel, Sukant Khurana.

Visualization: Ishan Goel.

Writing – original draft: Ishan Goel.

Writing – review & editing: Ishan Goel, Sukant Khurana.

References

1. Bin G, Gao X, Yan Z, Hong B, Gao S. An online multi-channel SSVEP-based brain-computer interface using a canonical correlation analysis method. *J Neural Eng*. IOP Publishing; 2009; 6: 46002. <https://doi.org/10.1088/1741-2560/6/4/046002> PMID: 19494422
2. Lin Z, Zhang C, Wu W, Gao X. Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs. *IEEE Trans Biomed Eng*. 2007; 54: 1172–1176. <https://doi.org/10.1109/TBME.2006.886577> PMID: 17549911
3. Chiang TC, Jeon BN, Li H. Dynamic correlation analysis of financial contagion: Evidence from Asian markets. *J Int Money Financ*. 2007; 26: 1206–1228. <https://doi.org/10.1016/j.jimonfin.2007.06.005>
4. Barndorff-Nielsen OE, Shephard N. *Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics*. Econometrica. Blackwell Publishing Ltd; 2004; 72: 885–925. <https://doi.org/10.1111/j.1468-0262.2004.00515.x>
5. Min H-G, Hwang Y-S. Dynamic correlation analysis of US financial crisis and contagion: evidence from four OECD countries. *Appl Financ Econ*. Routledge; 2012; 22: 2063–2074. <https://doi.org/10.1080/09603107.2012.698161>
6. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. *Proc Natl Acad Sci. National Academy of Sciences*; 2013; 110: 5802–5805. <https://doi.org/10.1073/pnas.1218772110> PMID: 23479631
7. Zou KH, Tuncali K, Silverman SG. Correlation and Simple Linear Regression. *Radiology*. 2003; 227: 617–628. <https://doi.org/10.1148/radiol.2273011499> PMID: 12773666
8. Speed T. A Correlation for the 21st Century. *Science* (80-). 2011; 334: 1502–1503. Available: <http://science.sciencemag.org/content/334/6062/1502.full>
9. Hauke J, Kossowski T. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaest Geogr*. 2011; 30. <https://doi.org/10.2478/v10117-011-0021-1>
10. Asuero AG, Sayago A, González AG. The Correlation Coefficient: An Overview. *Crit Rev Anal Chem*. Taylor & Francis Group; 2006; 36: 41–59. <https://doi.org/10.1080/10408340500526766>
11. Jammalamadaka SR, Sengupta A. Topics in Circular Statistics. *Statistics (Ber)*. 2001; 5: e20505. <https://doi.org/10.1111/j.1442-2050.2010.01169.x>
12. Nikolić D, Mureşan RC, Feng W, Singer W. Scaled correlation analysis: A better way to compute a cross-correlogram. *Eur J Neurosci*. Blackwell Publishing Ltd; 2012; 35: 742–762. <https://doi.org/10.1111/j.1460-9568.2011.07987.x> PMID: 22324876
13. Drasgow F. Polychoric and Polyserial Correlations. *Encyclopedia of Statistical Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2006. <https://doi.org/10.1002/0471667196.ess2014.pub2>
14. Allen IE, Seaman CA. Likert Scales and data analyses. *Qual Prog*. 2007; 40: 64.
15. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Stat. Institute of Mathematical Statistics*; 2007; 35: 2769–2794. <https://doi.org/10.1214/009053607000000505>
16. Koch GG, Koch G. G. Intraclass Correlation Coefficient. *Encyclopedia of Statistical Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2006. <https://doi.org/10.1002/0471667196.ess1275.pub2>
17. Shannon CE. A mathematical theory of communication. *ACM SIGMOBILE Mob Comput Commun Rev*. ACM; 2001; 5: 3. <https://doi.org/10.1145/584091.584093>
18. Schreiber T. Measuring Information Transfer. *Phys Rev Lett. American Physical Society*; 2000; 85: 461–464. <https://doi.org/10.1103/PhysRevLett.85.461> PMID: 10991308
19. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. *Med Image Anal*. Elsevier; 2001; 5: 143–156. [https://doi.org/10.1016/S1361-8415\(01\)00036-6](https://doi.org/10.1016/S1361-8415(01)00036-6) PMID: 11516708
20. Mooij J, Peters J, Janzing D, Zscheischler J. Distinguishing cause from effect using observational data: methods and benchmarks. *J Mach*. 2016; Available: <http://www.jmlr.org/papers/volume17/14-518/14-518.pdf>
21. Lichman M. UCI Machine Learning Repository. In: University of California, Irvine, School of Information and Computer Sciences [Internet]. 2013. Available: <http://archive.ics.uci.edu/ml>
22. Gerwin Schalk. BCI2000 [Internet]. Available: <http://www.schalklab.org/research/bci2000>
23. Schalk G., McFarland D.J., Hinterberger T., Birbaumer N., Wolpaw JR. BCI2000: A General-Purpose Brain-Computer Interface (BCI) System. *IEEE Trans Biomed Eng*. 2004; 51: 1034–1043. <https://doi.org/10.1109/TBME.2004.827072> PMID: 15188875
24. HE. GAALGLHJPCMRMJMGPCKS. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. In: *Circulation*. 2000 pp. e215–e220.

25. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* MIT Press 238 Main St., Suite 500, Cambridge, MA 02142–1046 USA journals-info@mit.edu; 2004; 16: 2639–2664. <https://doi.org/10.1162/0899766042321814> PMID: 15516276
26. Härdle W, Simar L. Applied Multivariate Statistical Analysis. *Technometrics*. 2003; 47: 486. <https://doi.org/10.1198/tech.2005.s319>
27. Nowak R. Generalized binary search. 2008 46th Annual Allerton Conference on Communication, Control, and Computing. IEEE; 2008. pp. 568–574. <https://doi.org/10.1109/ALLERTON.2008.4797609>
28. Droste S, Jansen T, Wegener I. Upper and lower bounds for randomized search heuristics in black-box optimization. *Theory Comput Syst.* Springer-Verlag; 2006; 39: 525–544. <https://doi.org/10.1007/s00224-004-1177-z>